PONTIFICIA UNIVERSIDAD CATÓLICA DEL ECUADOR

FACULTAD DE INGENIERÍA

MAESTRÍA EN SISTEMA DE INFORMACIÓN,MENCIÓN DATA SCIENCE

TESIS

Técnicas de fusión de datos para la predicción de plagas en cultivos de pitahaya en el cantón joya de los sachas

Autor: Santiago Israel Logroño Naranjo, MSc.

Director: Ing. Wilson Gustavo Chango PhD.

Quito 2024

DEDICATORIA

A mi familia, por su apoyo incondicional y por ser siempre mi fortaleza en los momentos más difíciles. Su amor, paciencia y palabras de aliento han sido el motor que me ha impulsado a alcanzar esta meta. A mi novia, cuyo amor y comprensión han sido mi refugio y motivación constante. Gracias por creer, por cada sonrisa, por cada gesto de cariño y por estar a mi lado en este camino.

A mi fiel compañera de cuatro patas, cuya presencia y lealtad han llenado de alegría y consuelo mis días. Gracias por el amor incondicional que solo tú sabes dar.

AGRADECIMIENTO

En primer lugar, quiero expresar mi más sincero agradecimiento a mi familia. Su amor incondicional, apoyo constante y palabras de aliento fueron la base que me sostuvo a lo largo de este camino. Sin su comprensión y paciencia, este logro no habría sido posible.

Quiero también agradecer al PhD. Wilson Chango, mi Director de Tesis, por su invaluable guía, paciencia y dedicación a lo largo de este proceso. Su conocimiento y orientación fueron fundamentales para la realización de este trabajo. Gracias por brindarme la confianza y el apoyo necesarios para superar cada obstáculo y por compartir conmigo su vasto conocimiento y experiencia.

A la Ing. Yadira Vargas, Directora de la Estación Experimental Central de la Amazonía del INIAP, le extiendo mi más profundo agradecimiento por su colaboración y apoyo en el desarrollo de esta investigación. Su liderazgo y compromiso con el avance científico en la región amazónica fueron inspiradores, y su aporte fue esencial para la realización de este estudio.

También deseo agradecer a todos los profesores de la maestría, quienes con su dedicación y enseñanzas me brindaron las herramientas necesarias para alcanzar este logro. Cada una de sus clases no solo enriqueció mi formación académica, sino que también me proporcionó valiosas lecciones que aplicaré en mi vida profesional.

RESUMEN

En la región amazónica de Ecuador, especialmente en el cantón Joya de los Sachas, la producción de pitahaya ha crecido considerablemente desde 2015 debido a su alta demanda en mercados nacionales e internacionales. Este crecimiento ha traído consigo desafíos relacionados con la gestión de plagas y enfermedades que afectan la calidad y cantidad de la producción. A pesar de los avances en la producción de pitahaya, los métodos actuales de manejo de plagas no han sido completamente efectivos, resultando en pérdidas económicas significativas. La literatura evidencia una brecha en la implementación de tecnologías avanzadas que integren datos meteorológicos y de campo para mejorar la predicción y gestión de plagas. Específicamente, en la región de Joya de los Sachas, existe una necesidad urgente de desarrollar métodos más eficaces para abordar esta problemática. Este estudio propone abordar la brecha identificada mediante la implementación de técnicas de fusión de datos que integren información climática y de campo, optimizando así la predicción y gestión de plagas en cultivos de pitahaya. Se aplican técnicas de aprendizaje no supervisado y reducción de dimensionalidad para clasificar datos relacionados con la clorofila en plantas, utilizando algoritmos como MeanShift y MiniBatchKMeans. Se emplea PCA, IPCA y KernelPCA para reducir la dimensionalidad y mejorar la precisión de los modelos de clasificación. Los resultados indican que PCA y KernelPCA con kernel lineal son los métodos más efectivos para la reducción de dimensionalidad en este contexto, con una precisión de hasta 0.9699 en la clasificación de plantas con y sin plaga. Sin la fusión de ciertos datos, se observa una ligera disminución en la precisión, lo que sugiere que la fusión de datos es beneficiosa en este contexto. Este estudio subraya la importancia de la fusión de datos y la implementación de técnicas avanzadas de reducción de dimensionalidad para mejorar la predicción de plagas en cultivos de pitahaya. La investigación proporciona un modelo que puede ser replicado en otras regiones, promoviendo prácticas agrícolas más sostenibles y resilientes en la Amazonía ecuatoriana y más allá.

Palabras clave: Pitahaya, Gestión de plagas, Fusión de datos, Aprendizaje no supervisado, Reducción de dimensionalidad

INDICE

INDICE DE TABLAS

INDICE DE FIGURAS

Capítulo 1

Introducción.

La implementación de gemelos digitales en la agricultura ha transformado el control de plagas, mejorando la precisión en la predicción y aumentando la eficiencia económica, lo que ha dado lugar a una nueva era en la producción agrícola sostenible. En la región amazónica de Ecuador, específicamente en el cantón Joya de los Sachas, la producción de pitahaya ha experimentado un crecimiento significativo en los últimos años, impulsada por la alta demanda en mercados nacionales e internacionales. Sin embargo, este crecimiento ha traído consigo desafíos relacionados con la gestión de plagas y enfermedades, que afectan tanto la calidad como la cantidad de la cosecha. A pesar de que las prácticas agrícolas actuales son avanzadas, no siempre resultan suficientes para prevenir y controlar las infestaciones, lo que genera pérdidas significativas para los productores locales.

El objetivo de la investigación es desarrollar e implementar un sistema de fusión de datos que integre información meteorológica y de campo para mejorar la predicción y gestión de plagas en los cultivos de pitahaya en el cantón Joya de los Sachas, promoviendo prácticas agrícolas sostenibles y resilientes. La pregunta de investigación plantea si la fusión de datos meteorológicos, agronómicos y de sensores, mediante técnicas de aprendizaje profundo, puede mejorar significativamente la precisión en la predicción de plagas en cultivos de pitahaya en el cantón Joya de los Sachas, permitiendo una gestión más eficiente y sostenible de este cultivo.

La necesidad de este estudio se fundamenta en la brecha identificada entre la literatura existente y las prácticas agrícolas actuales en la región amazónica ecuatoriana. La implementación de técnicas avanzadas de fusión de datos permitirá mejorar la predicción y gestión de plagas en los cultivos de pitahaya, reduciendo las pérdidas económicas y promoviendo prácticas agrícolas más sostenibles. Este estudio no solo contribuirá al desarrollo de la agricultura en la Amazonía ecuatoriana, sino que también proporcionará un modelo replicable en otras regiones con condiciones similares, potenciando la resiliencia agrícola a nivel global.

1.1. Antecedentes.

A nivel mundial, la implementación de gemelos digitales en la agricultura ha revolucionado el control de plagas, optimizando la precisión en las predicciones y aumentando la eficiencia económica. Un estudio pionero de Dai et al. (2024) demostró cómo un sistema de gemelos digitales puede gestionar eficazmente las plagas en cultivos de pimiento, alcanzando una precisión de predicción del 88.01% mediante el uso de algoritmos de bosques aleatorios. Este enfoque no solo mejora la precisión en la detección de plagas, sino que también incrementa significativamente la rentabilidad de los cultivos.

La integración de datos ambientales secuenciales y modelos de aprendizaje profundo ha demostrado ser una herramienta poderosa para predecir riesgos de plagas y enfermedades. Lee y Yun. (2023) desarrollaron un modelo que permite controlar el entorno para prevenir de manera proactiva la aparición de plagas y enfermedades, logrando una alta precisión en la predicción de enfermedades en fresas.

Por otro lado, la combinación de datos de sensores instalados en vehículos aéreos no tripulados (UAVs) y satélites, como lo demostró Maimaitijiang et al. (2020)**,** ha mejorado significativamente el monitoreo de cultivos y la predicción de plagas. Esta combinación permite obtener información detallada sobre el estado de los cultivos y predecir parámetros críticos como el índice de área foliar y la biomasa aérea.

Dhanaraj et al. (2024) propusieron un enfoque innovador que combina redes residuales de fusión multirrama y sistemas de detección de plagas basados en IoT para mejorar la capacidad de detección y clasificación de plagas en grandes campos agrícolas. Al analizar el sonido ambiental, este método permite identificar y clasificar diferentes tipos de plagas con alta precisión.

En Ecuador, específicamente en la Amazonía, el cultivo de pitahaya ha experimentado un crecimiento exponencial. Sin embargo, este cultivo enfrenta desafíos significativos relacionados con la gestión de plagas y enfermedades. Estudios como el realizado por el INIAP Tinoco et al. (2020) han destacado la importancia de implementar prácticas de manejo integrado de plagas para garantizar la sostenibilidad de la producción de pitahaya. Asimismo, Diéguez-Santana et al. (2022) subrayaron la necesidad de integrar tecnologías avanzadas en la cadena de valor de la pitahaya para

mejorar la eficiencia y la sostenibilidad.

La aplicación de estas tecnologías en el cultivo de pitahaya en la Amazonía ecuatoriana presenta una oportunidad única para mejorar la productividad y la sostenibilidad de este cultivo. Mediante la integración de datos meteorológicos, de campo y de sensores, y el empleo de técnicas de aprendizaje profundo, se puede desarrollar un sistema de alerta temprana que permita detectar y controlar las plagas de manera más efectiva.

La implementación de tecnologías avanzadas en la agricultura, como los gemelos digitales y el aprendizaje profundo, ofrece un gran potencial para mejorar la gestión de plagas y aumentar la sostenibilidad de los cultivos. La aplicación de estas tecnologías en el cultivo de pitahaya en la Amazonía ecuatoriana puede servir como un modelo para otras regiones y cultivos, contribuyendo al desarrollo de una agricultura más eficiente y respetuosa con el medio ambiente

1.2. Justificación

En la región amazónica de Ecuador, particularmente en el cantón Joya de los Sachas, la producción de pitahaya ha experimentado un crecimiento significativo desde el año 2015. Este crecimiento ha sido impulsado por la alta demanda en los mercados nacionales e internacionales, convirtiéndose en una fuente crucial de ingresos para los agricultores locales. Sin embargo, este auge ha venido acompañado de desafíos importantes, especialmente en la gestión de plagas y enfermedades que afectan tanto la calidad como la cantidad de la producción de pitahaya. A pesar de los avances en las prácticas agrícolas actuales, estas han demostrado ser insuficientes para controlar de manera efectiva las infestaciones de plagas, lo que subraya la necesidad urgente de adoptar nuevas tecnologías y enfoques integrados para mejorar la sostenibilidad y rentabilidad de este cultivo en la región.

La lucha contra las plagas en los cultivos de pitahaya no solo representa un desafío local, sino que también tiene implicaciones significativas para la economía agrícola de la región amazónica del Ecuador. Las pérdidas económicas derivadas de la mala gestión de plagas pueden ser devastadoras para los pequeños y medianos agricultores, quienes dependen en gran medida de este cultivo para su sustento. Este estudio busca abordar

esta problemática mediante la implementación de técnicas de fusión de datos, que no solo permitirán una mejor predicción de la aparición de plagas, sino que también proporcionarán herramientas más precisas para la gestión integrada de plagas, contribuyendo así a la sostenibilidad y competitividad de la producción de pitahaya en la región.

La implementación exitosa de tecnologías de fusión de datos en la predicción y gestión de plagas podría marcar un punto de inflexión en la agricultura de la región amazónica. Este enfoque innovador tiene el potencial de reducir significativamente las pérdidas económicas asociadas con las plagas, al mejorar la precisión en la detección temprana y permitir la aplicación oportuna de medidas preventivas. Además, al fomentar prácticas agrícolas más sostenibles y resilientes, este estudio contribuirá a la conservación del medio ambiente y al fortalecimiento de la economía local, generando un impacto positivo en la calidad de vida de los agricultores y en la competitividad del cultivo de pitahaya en los mercados globales.

Este estudio no solo se centra en la aplicación práctica de tecnologías de fusión de datos, sino que también aporta un valor teórico significativo al campo de la agricultura de precisión. Al integrar información meteorológica y de campo en modelos predictivos avanzados, la investigación ampliará el conocimiento sobre cómo estos factores interactúan para influir en la aparición de plagas. Además, la metodología propuesta, que incluye el uso de sensores montados en UAVs y satélites, proporciona una base sólida para futuras investigaciones en el manejo fitosanitario de cultivos. Este enfoque podría ser replicado o adaptado a otros cultivos en diferentes regiones, ampliando su impacto y utilidad.

Los resultados de esta investigación tendrán una aplicabilidad directa en la mejora de las prácticas agrícolas en la producción de pitahaya. Los modelos predictivos desarrollados proporcionarán a los agricultores herramientas efectivas para anticipar y gestionar las plagas, reduciendo la necesidad de intervención química y mejorando la sostenibilidad del cultivo. Además, la adopción de estas tecnologías podría incentivar una mayor inversión en el sector agrícola de la región, promoviendo la innovación y el desarrollo económico. A largo plazo, los beneficios incluyen una mayor resiliencia frente a las plagas, una producción más eficiente y un mejor posicionamiento de la pitahaya ecuatoriana en el mercado global.

La investigación propuesta es pionera en la integración de técnicas avanzadas de fusión de datos en la gestión de plagas en cultivos de pitahaya. A diferencia de estudios previos, que se han centrado en prácticas agrícolas tradicionales o en el uso de tecnologías de manera aislada, este estudio combina múltiples fuentes de datos y tecnologías para ofrecer una solución más completa y efectiva. La originalidad radica en la aplicación de estas técnicas en un contexto específico, como el cantón Joya de los Sachas, lo que permite abordar los desafíos únicos de la región y proponer soluciones adaptadas a sus necesidades.

Esta investigación tiene el potencial de transformar la gestión de plagas en la producción de pitahaya en la región amazónica del Ecuador. Al combinar avances tecnológicos con un enfoque integrado de fusión de datos, el estudio no solo aborda un problema crítico para los agricultores locales, sino que también contribuye al desarrollo de prácticas agrícolas más sostenibles y eficientes. La importancia de esta investigación radica en su capacidad para mejorar la competitividad de la pitahaya en el mercado global, al tiempo que promueve la sostenibilidad económica y ambiental de la región.

1.3. Planteamiento del Problema

En la región amazónica de Ecuador, particularmente en el cantón Joya de los Sachas, la producción de pitahaya ha experimentado un notable crecimiento en los últimos años, impulsada por su alta demanda tanto en mercados nacionales como internacionales. Este aumento en la producción se ha intensificado desde el año 2015, con un enfoque en mejorar las prácticas agrícolas y en aumentar la productividad del cultivo. No obstante, esta expansión ha conllevado desafíos importantes relacionados con la gestión de plagas y enfermedades que afectan tanto la calidad como la cantidad de la producción de pitahaya.

Los estudios existentes han resaltado la importancia de un manejo adecuado del suelo y de la implementación de medidas preventivas para el control de plagas en los cultivos de pitahaya. Según un estudio del INIAP, las prácticas agrícolas actuales no siempre son suficientes para prevenir y controlar las infestaciones de plagas, lo que resulta en pérdidas significativas para los productores (Tinoco et al., 2020). Además, investigaciones sobre la economía circular en la cadena agroalimentaria de pitahaya han

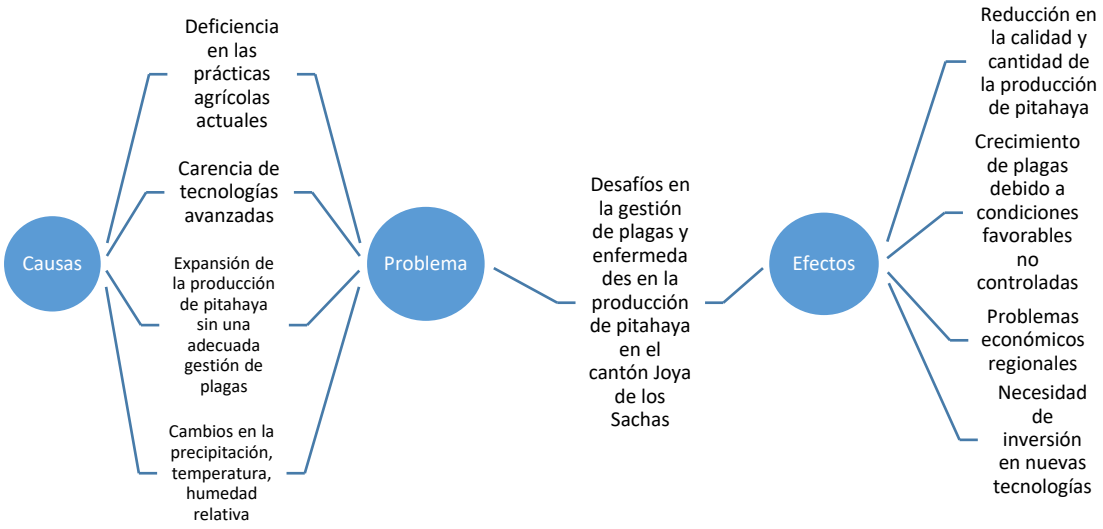
indicado que una gestión inadecuada de residuos puede agravar los problemas de plagas y enfermedades, creando condiciones favorables para su proliferación (Diéguez-Santana et al., 2022). Estos estudios subrayan una brecha en la implementación de tecnologías avanzadas y prácticas sostenibles que integren de manera efectiva la gestión de plagas.

La necesidad de este estudio radica en abordar la brecha identificada tanto en la literatura como en la práctica agrícola actual. Es crucial desarrollar e implementar técnicas de fusión de datos que integren información meteorológica y de campo para mejorar la predicción y gestión de plagas en los cultivos de pitahaya. Este enfoque no solo contribuirá a reducir las pérdidas económicas causadas por plagas, sino que también promoverá prácticas agrícolas más sostenibles y resilientes. El estudio se llevará a cabo mediante la recopilación y análisis de datos meteorológicos y de campo, la implementación de técnicas de fusión de datos y la evaluación de su efectividad en la predicción de plagas. Los objetivos de este estudio incluyen mejorar la sostenibilidad de la producción de pitahaya en la Amazonía ecuatoriana y proporcionar un modelo replicable en otras regiones con condiciones similares.

La pregunta de investigación es: ¿Qué técnicas de fusión de datos son más efectivas para mejorar la predicción de plagas en cultivos de pitahaya en el cantón Joya de los Sachas, considerando la integración de datos automáticos provenientes de sensores y datos manuales de pitahaya injertada y sin injertar?

En esta investigación se consideran variables independientes que integran factores climáticos y agronómicos, tales como precipitación, temperatura, humedad relativa, velocidad del viento, cobertura de sombra y tratamiento aplicado a los cultivos. Se anticipa que estas variables tendrán un impacto considerable en la aparición de plagas. Por otro lado, la variable dependiente seleccionada es el índice de clorofila, el cual actúa como un indicador de la salud de las plantas y su susceptibilidad a las plagas.

Figura 1. Árbol de Problemas, Desafío en gestión de plagas en la pitahaya



Fuente: Logroño S, 2024

1.4. Objetivos de la Investigación

1.4.1. Objetivo general

Implementar un modelo de predicción de plagas en cultivos de pitahaya utilizando técnicas avanzadas de fusión de datos, con el fin de optimizar el manejo integrado de plagas y mejorar la eficiencia productiva en la región amazónica de Ecuador.

1.4.2. Objetivos específicos.

- Identificar los factores climáticos y agronómicos que influyen en la aparición de plagas en cultivos de pitahaya.
- Recopilar y analizar datos meteorológicos y de campo provenientes de sensores y observaciones manuales en cultivos de pitahaya injertada y sin injertar.
- Desarrollar modelos predictivos utilizando técnicas de fusión de datos para anticipar la aparición de plagas.
- Evaluar la efectividad de los modelos predictivos en la gestión de plagas,

comparándolos con los métodos tradicionales.

1.5. Alcance

El alcance de este estudio se centra en desarrollar y evaluar un sistema de predicción de plagas en cultivos de pitahaya, utilizando técnicas avanzadas de fusión de datos. Específicamente, la investigación se enfoca en:

Área geográfica: Cantón Joya de los Sachas, región amazónica de Ecuador.

Cultivo: Pitahaya.

Plagas: Se identificarán y priorizarán las plagas más relevantes para la región y el cultivo.

Datos: Se utilizarán datos meteorológicos, agronómicos y de sensores para desarrollar los modelos predictivos.

Modelos: Se emplearán técnicas de fusión de datos como el aprendizaje automático y las redes neuronales para crear modelos predictivos.

Evaluación: Se evaluará la precisión y eficacia de los modelos desarrollados en comparación con métodos tradicionales.

Implementación: Se propondrá un modelo de manejo integrado de plagas basado en los resultados de la investigación.

Consideraciones Adicionales para el Alcance:

Colaboración con productores: Es fundamental establecer una estrecha colaboración con los productores de pitahaya para obtener información relevante y garantizar la adopción de las recomendaciones del estudio.

Monitoreo a largo plazo: Se recomienda realizar un monitoreo a largo plazo de los resultados obtenidos para evaluar la sostenibilidad y la eficacia de las recomendaciones a largo plazo.

Sensibilización: Es importante sensibilizar a los productores sobre la importancia de la tecnología y la adopción de prácticas agrícolas sostenibles.

Capítulo 2

Marco Teórico

2.1. Agricultura de precisión

La agricultura de precisión es una de las tecnologías avanzadas para mejorar la eficiencia y productividad agrícola mediante el uso de datos y análisis en tiempo real. Este enfoque combina tecnologías avanzadas como el Internet de las Cosas (IoT), sensores inteligentes, drones, análisis de grandes volúmenes de datos y sistemas de información geográfica (GIS). La agricultura inteligente permite una gestión precisa de los recursos agrícolas, optimizando el uso de agua, fertilizantes y pesticidas, y mejorando la sostenibilidad y rentabilidad de las explotaciones agrícolas (Friha et al., 2021; Karunathilake et al., 2023).

La agricultura de precisión consiste en utilizar técnicas y tecnologías avanzadas para gestionar de manera efectiva las variaciones en tiempo y espacio que ocurren dentro de los campos agrícolas. Esto se logra mediante la recopilación de datos detallados sobre las condiciones del suelo, el clima, el crecimiento de las plantas y otros factores relevantes, lo que permite a los agricultores tomar decisiones informadas y oportunas. La incorporación de sistemas GPS, drones y sensores en los campos ayuda a monitorear y controlar las operaciones agrícolas con una precisión sin precedentes (Karunathilake et al., 2023).

2.1.1.1. Tecnologías Aplicadas

Las tecnologías aplicadas en la agricultura inteligente incluyen una variedad de herramientas y sistemas que permiten la recolección y análisis de datos en tiempo real:

2.1.1.1.1. Internet de las Cosas (IoT)

El Internet de las Cosas (IoT) es una red de dispositivos interconectados que se comunican entre sí y con su entorno para recopilar y compartir datos. En la agricultura inteligente, los dispositivos IoT son fundamentales, ya que permiten un monitoreo constante de las condiciones del suelo, el clima y la salud de las plantas, proporcionando información clave para optimizar las prácticas agrícolas. Estos dispositivos incluyen

sensores de humedad del suelo, sensores de temperatura, sensores de luz, estaciones meteorológicas y más. Los datos recopilados por estos dispositivos se envían a una plataforma centralizada donde se analizan para proporcionar información en tiempo real a los agricultores, permitiéndoles tomar decisiones informadas sobre riego, fertilización y protección de cultivo (Friha et al., 2021; Karunathilake et al., 2023)

El uso de IoT en la agricultura de precisión permite la creación de un entorno interconectado donde todos los aspectos del cultivo se pueden monitorear y gestionar de manera eficiente. Por ejemplo, los sensores de humedad del suelo pueden detectar cuándo una planta necesita agua, lo que activa un sistema de riego automático para suministrar la cantidad exacta de agua necesaria. Las estaciones meteorológicas equipadas con sensores IoT pueden proporcionar datos climáticos precisos, como la temperatura, la humedad, la velocidad del viento y la precipitación, lo que ayuda a los agricultores a prever y prepararse para condiciones climáticas adversas (Padhiary et al., 2024)

2.1.1.1.2. Drones

Los drones, también conocidos como vehículos aéreos no tripulados (UAVs), son una de las tecnologías más innovadoras y útiles en la agricultura de precisión. Equipados con cámaras de alta resolución y una variedad de sensores, los drones pueden sobrevolar los campos de cultivo y capturar imágenes detalladas y datos en tiempo real sobre el estado de las plantas, las condiciones del suelo y otros factores ambientales. Esta capacidad de monitoreo aéreo permite a los agricultores identificar problemas rápidamente y tomar medidas correctivas antes de que se conviertan en serios (Karunathilake et al., 2023).

Los drones proporcionan una vista aérea de los campos, permitiendo a los agricultores evaluar el estado general de los cultivos, detectar signos de estrés hídrico, enfermedades, plagas y deficiencias de nutrientes. Las imágenes de alta resolución capturadas por los drones pueden revelar detalles que no son visibles desde el nivel del suelo, facilitando una gestión más precisa de los cultivos (Karunathilake et al., 2023).

2.1.1.1.3. Análisis de Big Data

El análisis de big data ha dado grandes resultados en la agricultura de precisión. Esta tecnología permite recopilar, almacenar y analizar enormes cantidades de datos que los sistemas tradicionales no pueden manejar. En el ámbito agrícola, el big data brinda a los agricultores la capacidad de tomar decisiones más acertadas, proporcionando información detallada y en tiempo real sobre diversos factores que influyen en el crecimiento y salud de los cultivos.(Karunathilake et al., 2023)

El análisis de big data en la agricultura abarca diversas etapas esenciales. En primer lugar, se lleva a cabo la recolección de datos, donde la información proviene de múltiples fuentes como sensores IoT, drones, satélites, maquinaria agrícola y sistemas de gestión agrícola. Estos datos abarcan desde las condiciones del suelo y el clima, hasta la salud de las plantas y el rendimiento de los cultivos. Posteriormente, los datos recopilados se almacenan en bases de datos avanzadas, que tienen la capacidad de manejar grandes volúmenes de información. Frecuentemente, se emplea la tecnología de la nube para asegurar que estos datos se almacenen de manera segura y puedan ser accesibles desde cualquier lugar (San Emeterio de la Parte et al., 2023).

A continuación, los datos que son almacenados son sometidos a un análisis ocupando algoritmos avanzados y técnicas de ML para poder identificar patrones y tendencias. Este proceso permite anticipar problemas futuros, como brotes de plagas o enfermedades, y optimizar el uso de recursos como agua y fertilizantes. Finalmente, los resultados del análisis de big data se presentan de forma visual mediante gráficos, mapas y otros medios interactivos. Esto proporciona a los agricultores la interpretación de los datos y toma de decisiones informadas y precisas, fundadas en información detallada y en tiempo real (Osinga et al., 2022).

2.1.1.1.4. Sistemas de Información Geográfica (GIS)

Los Sistemas de Información Geográfica permiten almacenar, analizar, manipular y mapear información georreferenciada, lo cual es esencial para gestionar datos espaciales y temporales sobre las condiciones del suelo, el clima, la salud de las plantas

y el rendimiento de los cultivos. En el ámbito agrícola, los GIS facilitan la toma de decisiones basadas en datos precisos y detallados, optimizando así las operaciones y mejorando la eficiencia en el uso de recursos. Además, proporcionan a los agricultores una manera eficaz de visualizar y analizar datos espaciales, lo que se traduce en una mejor planificación y gestión de los recursos agrícolas (San Emeterio de la Parte et al., 2023).

Los GIS integran datos provenientes de múltiples fuentes, como sensores IoT, imágenes satelitales y drones, proporcionando una visión integral y detallada del entorno agrícola. Esta integración de datos permite a los agricultores tener una comprensión más completa y precisa de sus campos, mejorando así la eficiencia y sostenibilidad de sus prácticas agrícolas.

2.1.2. Tipos de plagas en la pitahaya

La pitahaya, conocida como fruta del dragón, es un cultivo valioso tanto por su valor nutricional como medicinal. Sin embargo, su producción enfrenta numerosos desafíos, especialmente relacionados con plagas y enfermedades. Entre las plagas más comunes destacan los insectos chupadores como áfidos, trips y cochinillas, que dañan las plantas al succionar su savia y debilitar su estructura, además de ser vectores de enfermedades virales. El control de estas plagas requiere un manejo integrado que combine métodos biológicos, químicos y culturales, según el Instituto Nacional de Investigaciones Agropecuarias (INIAP) (Tinoco et al., 2020).

En cuanto a las enfermedades, la pitahaya es susceptible a varias, incluyendo la antracnosis, la pudrición del fruto y del tallo, y el chancro del tallo. La antracnosis, causada por especies del género Colletotrichum, es especialmente destructiva, manifestándose como manchas negras en frutos y tallos, lo que reduce significativamente la calidad del producto. Para gestionar estas enfermedades, se recomiendan prácticas culturales como el manejo de la temperatura y el uso de biopesticidas y agentes de control biológico como Bacillus subtilis, que han demostrado ser efectivos contra varios patógenos (Balendres & Bengoa, 2019).

Por otro lado, es fundamental el monitoreo constante de datos meteorológicos y de

los índices de clorofila de las plantas. Conocer las condiciones climáticas permite anticipar y mitigar los efectos negativos del clima, mientras que los índices de clorofila ofrecen una medida precisa de la salud de las plantas, ayudando a detectar deficiencias nutricionales y enfermedades en etapas tempranas. Esta combinación de estrategias no solo mejora la precisión en la aplicación de tratamientos, sino que también optimiza el uso de recursos y minimiza el impacto ambiental, contribuyendo así a una producción de pitahaya más sostenible y rentable.

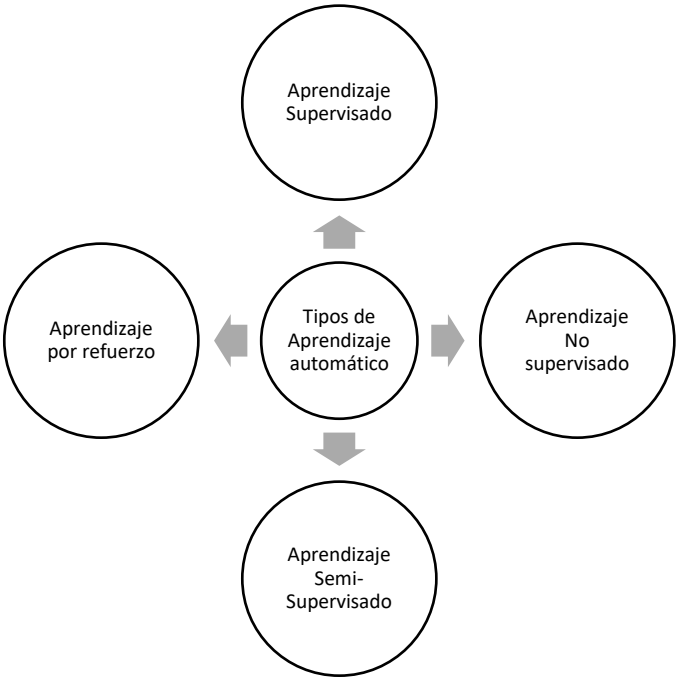
2.1.3. Machine Learning

El aprendizaje automático (ML) es una rama de la inteligencia artificial que se centra en desarrollar algoritmos que permiten a las computadoras aprender y mejorar a partir de los datos, sin requerir una programación explícita para cada tarea en particular. Este método facilita la detección de patrones y relaciones en grandes cantidades de datos, lo que resulta en predicciones y decisiones más precisas y efectivas. En el ámbito agrícola, ML se emplea para optimizar la gestión de los cultivos y el uso eficiente de los recursos, entre otras aplicaciones (Benos et al., 2021).

2.1.3.1. Tipos de aprendizaje en Machine Learning

El aprendizaje automático se puede categorizar en varios tipos según cómo se utilicen y gestionen los datos durante el proceso de entrenamiento. En la Fig 2 sedetalla los diferentes tipos de ML.

Figura 2. Principales tipos de aprendizaje automático



Fuente: Logroño S, 2024

2.1.3.1.1. Aprendizaje supervisado

El aprendizaje supervisado es una técnica clave en el campo del aprendizaje automático, donde los modelos se entrenan utilizando un conjunto de datos etiquetados, lo que significa que cada entrada en los datos de entrenamiento está asociada con una salida conocida. Este enfoque permite al modelo aprender la relación entre las entradas y salidas, y luego utilizar ese conocimiento para hacer predicciones precisas sobre nuevos datos. Mientras el modelo se entrena, va ajustando sus parámetros internos con el objetivo de minimizar la discrepancia entre las predicciones que realiza y los resultados reales obtenidos durante el entrenamiento. Esta capacidad de ajuste continuo es lo que permite al aprendizaje supervisado aplicarse exitosamente en una amplia variedad de dominios, incluyendo la clasificación de imágenes, el reconocimiento de voz y la predicción de tendencias en series temporales, tal como se discute en el artículo de (Sarker, 2021).

2.1.3.1.2. Aprendizaje no supervisado

El aprendizaje no supervisado es una técnica donde los modelos se entrenan utilizando datos no etiquetados, lo que permite descubrir patrones ocultos y estructuras subyacentes en grandes volúmenes de información. A diferencia del aprendizaje supervisado, en el aprendizaje no supervisado no se proporciona al modelo ninguna salida esperada durante el entrenamiento; en su lugar, el modelo explora los datos en busca de similitudes o agrupaciones naturales. Esta capacidad de identificar y categorizar datos de manera autónoma hace que el aprendizaje no supervisado sea fundamental en áreas como la visión por computadora, el reconocimiento de voz y la creación de sistemas de inteligencia artificial autónomos. Entre las técnicas más comunes se incluyen el clustering, la detección de anomalías, y los autoencoders, cada una de las cuales ofrece soluciones específicas para problemas complejos sin la necesidad de intervención humana en el proceso de etiquetado de datos (Naeem et al., 2023).

2.1.3.1.3. Aprendizaje semi supervisado

El aprendizaje semi-supervisado es una técnica que combina elementos del aprendizaje supervisado y no supervisado, aprovechando la combinación de datos con etiquetas y sin etiquetas para mejorar la precisión y efectividad de los modelos. Esta técnica es especialmente valiosa en situaciones donde etiquetar datos es un proceso costoso o que consume mucho tiempo, mientras que los datos sin etiquetar son abundantes y fáciles de obtener. Al aprovechar ambos tipos de datos, el aprendizaje semi-supervisado permite desarrollar modelos que no solo son más precisos, sino que también se adaptan mejor a nuevos escenarios, superando a los métodos tradicionales que solo usan un tipo de datos (C A Padmanabha Reddy et al., 2018).

2.1.3.1.4. Aprendizaje por refuerzo

El aprendizaje por refuerzo permite a los agentes de software y a las máquinas identificar de forma autónoma el comportamiento óptimo en un entorno o contexto determinado, con el fin de maximizar su eficiencia y rendimiento. Este enfoque, centrado en la interacción con el entorno, se basa en un sistema de recompensas y penalizaciones, donde el objetivo es utilizar el conocimiento adquirido para tomar decisiones que maximicen las recompensas o minimicen los riesgos. El aprendizaje por refuerzo es una herramienta poderosa para entrenar modelos de inteligencia artificial, especialmente en áreas que requieren alta automatización y optimización de la eficiencia

operativa, como la robótica, la conducción autónoma, la fabricación y la logística de la cadena de suministro. Sin embargo, no es la opción más adecuada para resolver problemas simples o básicos (Sarker, 2021).

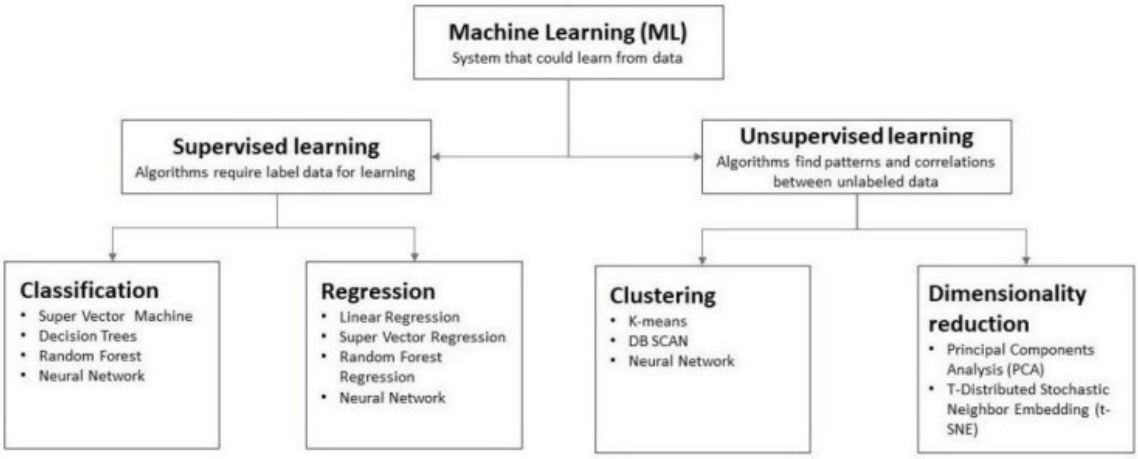
2.1.3.2. Algoritmos de Machine Learning

El uso de algoritmos de Macchine Learning (ML) en la agricultura ha permitido avances significativos en la gestión y optimización de cultivos. Estos algoritmos son esenciales para analizar grandes volúmenes de datos y proporcionar insights que mejoran la toma de decisiones agrícolas. En la Tabla 1, se presenta una descripción de los principales algoritmos de machine learning, organizados por su tipo de aprendizaje y categoría, para ofrecer una visión clara de las herramientas disponibles en este campo. A continuación, se presentan algunos algoritmos de ML más utilizados en la agricultura.

Tabla 1. Algoritmos de Machine Learning

Tipo de Aprendizaje	***Categoría***	***Algoritmo***	***Descripción***
Supervisado	Clasificación	Máquinas de Soporte Vectorial (SVM)	Algoritmo de clasificación que encuentra el hiperplano óptimo para separar las clases en un espacio de características.
Supervisado	Clasificación	Árboles de Decisión	Algoritmo que clasifica datos organizándolos en un árbol, donde cada nodo representa una prueba de una característica.
Supervisado	Clasificación y Ensamble	Bosques Aleatorios (Random Forest)	Método de ensamble que combina múltiples árboles de decisión para mejorar la precisión y reducir el riesgo de sobreajuste.
Supervisado y no Supervisado	Clasificación y Clustering	Redes Neuronales (Neural Network)	Modelos inspirados en el cerebro humano, capaces de aprender patrones complejos y manejar grandes volúmenes de datos.
Supervisado	Regresión	Regresión Lineal	Modelo estadístico que predice un valor continuo basado en la relación lineal entre las variables independientes y dependiente.
No Supervisado	Clustering	K-means	Algoritmo de clustering que agrupa datos en K clusters basados en la similitud de las características.
No Supervisado	Clustering	DBSCAN	Algoritmo de clustering que identifica clusters de alta densidad, separándolos de las regiones de baja densidad.
No Supervisado	Reducción de Dimensionalidad	Análisis de Componentes Principales (PCA)	Técnica de reducción de dimensionalidad que transforma variables originales en componentes principales, preservando la variabilidad máxima.
No Supervisado	Reducción de Dimensionalidad	T-Distributed Stochastic Neighbor Embedding (t-SNE)	Técnica de reducción de dimensionalidad que visualiza datos de alta dimensión en un espacio de menor dimensión, preservando la estructura de los datos.

Figura 3. Algoritmos de Machine Learning



Fuente: (Chango et al., 2022)

2.1.4. Tipos de fusión de datos

La fusión de datos es un elemento clave en el análisis de datos multimodales, ya que permite combinar información de distintas fuentes para mejorar la comprensión y facilitar la toma de decisiones. (Chango et al., 2022)destacan que las técnicas de fusión de datos se categorizan principalmente según el momento en que se realiza la fusión. Entre estas técnicas se encuentran la fusión temprana, que consiste en concatenar las características de diferentes fuentes en un único vector; la fusión tardía, que integra las predicciones de clasificadores independientes construidos a partir de diversas fuentes; y la fusión híbrida, que combina aspectos de ambos enfoques.

2.1.4.1. Fusión temprana

La fusión temprana es un enfoque que se implementa en las etapas iniciales del procesamiento de datos. En este método, las características derivadas de diversas fuentes de datos se combinan en un solo conjunto antes de ser introducidas en un modelo de aprendizaje. Este enfoque permite al modelo aprender directamente de la integración de todas las características

disponibles, capturando las relaciones entre diferentes tipos de datos desde el inicio del proceso como se muestra en la Figura 2.

2.1.4.2. Fusión tardía

La fusión tardía ocurre después de que cada fuente de datos ha sido procesada por separado. En este enfoque, se entrena un modelo individual para cada conjunto de datos, y luego se combinan las predicciones de estos modelos para llegar a una decisión final (Figura 2b). Este método es especialmente beneficioso cuando cada tipo de datos ofrece información complementaria, ya que mantener la independencia en el procesamiento puede mejorar la precisión de las predicciones finales (Chango et al., 2022).

2.1.4.3. Fusión híbrida

La fusión híbrida compone aspectos de la fusión temprana y fusión tardía. Con este enfoque, algunas características se combinan desde el principio, mientras que otras se procesan de forma independiente y se unen más adelante en la fase de toma de decisiones (Figura 2c). Esta flexibilidad permite aprovechar lo mejor de ambos métodos, maximizando el uso de la información disponible y ajustándose mejor a las particularidades de cada problema (Chango et al., 2022).

Capítulo 3

Metodología

Para el proceso de Machine Learning, se utilizará scikit-learn, una biblioteca de Python que proporciona herramientas completas para el desarrollo de modelos predictivos y de clustering. Los componentes principales incluyen:

Preprocesamiento de Datos: Incluye escalado y normalización con herramientas como StandardScaler y MinMaxScaler, así como la división de datos mediante train_test_split.

Selección y Entrenamiento de Modelos: Ofrece modelos supervisados como regresión lineal y SVM, así como modelos no supervisados como K-Means y PCA.

Evaluación de Modelos: Permite el uso de métricas de rendimiento (precisión, recall, F1-score, MSE) y la validación cruzada con cross_val_score.

Optimización y Ajuste de Hiperparámetros: Facilita el ajuste de hiperparámetros mediante GridSearchCV y RandomizedSearchCV.

Implementación y Predicción: Permite realizar predicciones sobre nuevos datos utilizando el método predict.

Visualización y Análisis: Se integra con bibliotecas como matplotlib y seaborn para la visualización de resultados.

En resumen, scikit-learn proporciona una solución integral para el ciclo completo de Machine Learning, abarcando desde la preparación de datos hasta la evaluación y predicción de modelos.

3.1. Diseño del estudio

Este estudio se ha diseñado como una investigación experimental, enfocada en evaluar y aplicar técnicas avanzadas de fusión de datos para mejorar la predicción de plagas en los cultivos de pitahaya. Dado que era esencial validar la efectividad de diferentes algoritmos de aprendizaje automático en un entorno controlado, se optó por un enfoque experimental. Para ello, se recopilaron datos de diversas fuentes, como sensores IoT, registros meteorológicos y datos en campo. Estos datos fueron integrados y analizados para detectar patrones y correlaciones que pudieran anticipar la aparición de plagas. Este enfoque no solo permite observar los resultados de

los algoritmos en acción, sino también comparar su rendimiento en términos de precisión y eficacia, lo que proporciona una base sólida para tomar decisiones informadas en la gestión agrícola.

La investigación experimental llevada a cabo en este estudio se diseñó para evaluar el impacto de la integración de múltiples fuentes de datos en la mejora de la predicción de plagas en los cultivos. A través de la experimentación controlada, se implementaron y probaron varios modelos de aprendizaje automático, con especial énfasis en los enfoques supervisados y no supervisados.

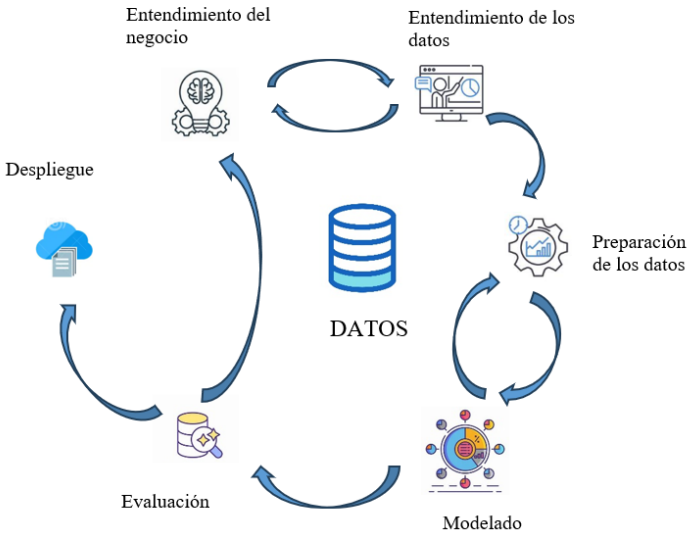
3.2. Variables y medidas

Las variable dependiente es la predicción y las variables independientes son las técnicas de fusión de datos que son utilizadas en los diferentes modelos de machine Learning para el análisis en el campo agrícola, específicamente en la pitahaya.

3.3. Procedimiento

CRISP-DM (Proceso Estándar Intersectorial para la Minería de Datos) es una metodología ampliamente utilizada en la industria para llevar a cabo proyectos de minería de datos de manera ordenada y eficaz. Esta metodología se estructura en seis fases principales: comprensión del negocio, comprensión de los datos, preparación de los datos, modelado, evaluación y el despliegue.. Cada una de estas etapas está diseñada para asegurar una transición fluida entre las diferentes fases del proyecto, garantizando que los objetivos del negocio estén alineados con los resultados técnicos. CRISP-DM es altamente apreciado por su flexibilidad y su capacidad para adaptarse a una variedad de contextos y sectores, ofreciendo un marco sólido para gestionar proyectos complejos de análisis de datos, desde la fase inicial de conceptualización hasta la implementación final de los modelos (Schröer et al., 2021). En esta investigación, se emplea la metodología CRISP-DM para orientar la recolección, procesamiento e integración de datos, así como para evaluar la eficacia de los modelos predictivos en la detección y prevención de plagas en los cultivos de pitahaya.

Figura 4. Ciclo de la Metodología CRISP-DM



Fuente: Logroño S, 2024

3.3.1. Entendimiento del negocio

En esta primera fase, se identifican los objetivos clave del proyecto relacionados con la agricultura de precisión, específicamente en la predicción de plagas en cultivos de pitahaya. Esto incluye comprender las necesidades de los agricultores y los desafíos que enfrentan, como la prevención de plagas y la optimización de los rendimientos. El objetivo es alinear el proceso de minería de datos con estos objetivos empresariales, definiendo claramente los resultados esperados y cómo estos contribuirán a mejorar la gestión agrícola.

3.3.2. Entendimiento de los datos

En esta fase, se recopilan y exploran los datos relevantes, que pueden incluir variables como la humedad del suelo, la temperatura y datos históricos de plagas. Se utiliza scikit-learn para realizar un análisis exploratorio inicial, que incluye la verificación de la calidad de los datos, la

identificación de patrones preliminares y la preparación de los datos para su posterior procesamiento.

3.3.3. Preparación de los datos

En esta fase, el objetivo principal es convertir los datos crudos en un formato óptimo para el modelado. Esto incluye limpiar los datos, seleccionando las características más relevantes y combinando información de diferentes fuentes. Con scikit-learn, se aplican técnicas como la imputación para manejar valores faltantes y la creación de variables derivadas que puedan mejorar la calidad del conjunto de datos. Además, se utilizan métodos de reducción de dimensionalidad, como el Análisis de Componentes Principales (PCA), para simplificar el conjunto de datos, manteniendo al mismo tiempo la mayor parte de la información valiosa.

3.3.4. Modelado

En la fase de modelado, se eligen y entrenan los modelos de Machine Learning que mejor se adaptan a la tarea en cuestión. Usando scikit-learn, se pueden aplicar algoritmos supervisados como la regresión lineal y las máquinas de soporte vectorial (SVM), así como algoritmos no supervisados como K-Means para realizar clustering. Los modelos se entrenan utilizando los datos preparados en la fase anterior, y se ajustan los hiperparámetros con técnicas como GridSearchCV y RandomizedSearchCV para asegurar que el rendimiento del modelo sea óptimo.

3.3.5. Evaluación

En esta fase, se evalúa la eficacia de los modelos desarrollados mediante el uso de métricas de rendimiento como precisión, recall, F1-score y error cuadrático medio (MSE). Scikit-learn proporciona herramientas como cross_val_score para realizar una validación cruzada y asegurar que el modelo generaliza bien a datos no vistos. Esta evaluación rigurosa permite determinar si el modelo cumple con los objetivos definidos en la fase de entendimiento del negocio.

3.3.6. Despliegue

En la fase final, el modelo se implementa en un entorno de producción, donde puede empezar a hacer predicciones con nuevos datos. Con scikit-learn, esto se logra mediante el método predict, que aplica el modelo entrenado para generar predicciones sobre datos recién ingresados. Además, se pueden crear visualizaciones de los resultados utilizando herramientas como matplotlib y

seaborn, lo que facilita que los usuarios finales, como los agricultores o gestores de cultivos, interpreten y apliquen los resultados de manera efectiva en su trabajo diario.

3.4. Análisis de datos

En el presente estudio, se han aplicado técnicas avanzadas de Machine Learning para el análisis de datos, enfocadas en la predicción y detección de plagas en cultivos agrícolas. Se emplearon algoritmos supervisados, como las máquinas de soporte vectorial (SVM) y regresión logística, así como algoritmos no supervisados, como el clustering con K-means. Además, se implementaron métodos de preprocesamiento de datos, como la normalización y la selección de características, utilizando herramientas proporcionadas por la biblioteca Scikit-learn. Estos enfoques han demostrado ser eficaces en estudios previos para mejorar la precisión en la predicción de fenómenos agronómicos y en la optimización de modelos predictivos.

3.5. Consideraciones éticas

Para asegurar la integridad y privacidad de los datos empleados en este estudio, se procedió a su anonimización antes de cualquier proceso de análisis. Este paso es fundamental para cumplir con las normativas de protección de datos, garantizando que la identidad de los individuos o las fuentes de información no pueda ser rastreada, lo que protege la confidencialidad de las personas involucradas. Este enfoque es especialmente relevante en investigaciones que manejan grandes volúmenes de datos, donde el riesgo de reidentificación puede ser considerable. La anonimización, por tanto, no solo es una medida de cumplimiento legal, sino también un compromiso ético para preservar la privacidad en contextos de investigación.

3.6. Limitaciones del estudio

Una de las principales limitaciones de este estudio es su dependencia de datos históricos, que pueden no representar con exactitud las condiciones actuales o futuras en el ámbito agrícola. Aunque los datos históricos ofrecen una valiosa visión a lo largo del tiempo, su utilidad en escenarios dinámicos puede verse reducida debido a cambios en las condiciones ambientales, las prácticas agrícolas o la aparición de nuevas plagas.

3.7. Reproducibilidad

Para garantizar la reproducibilidad de los resultados de este estudio, todo el código fuente y los modelos desarrollados se pondrán a disposición del público a través de la plataforma GitHub. Este enfoque no solo facilita que otros investigadores puedan validar y replicar los resultados, sino que también promueve la colaboración y el progreso en el campo de la agricultura de precisión. La transparencia en la metodología y el acceso abierto a los recursos son pilares esenciales de la ciencia reproducible, ya que permiten que los hallazgos sean verificados y aplicados en una variedad de contextos.

Capítulo 4

Resultados

Cuando efectuamos la fusión de datos, las técnicas se pueden caracterizar de distintas maneras de acuerdo al ámbito de aplicación. Estas técnicas se pueden clasificar en función del momento en que se lleva a cabo la fusión, dando lugar a tres tipos principales, según w. Chango et al. (2023):

Fusión a nivel de características o temprana: Este método de fusión implica combinar las diferentes características de datos provenientes de diversas fuentes en un solo vector compuesto por elementos heterogéneos.

Fusión a nivel de decisión o tardía: En este enfoque, se crea primero un clasificador para cada fuente de datos por separado, para luego fusionar las predicciones ofrecidas por los distintos clasificadores.

Fusión híbrida: Este enfoque combina los métodos anteriores en un único proceso de fusión, utilizando tanto la fusión a nivel de características como la fusión a nivel de decisión.

La Tabla 2 categoriza los artículos seleccionados según el punto de fusión o el momento en que se realiza la fusión (temprana, tardía e híbrida). Es significativo recalcar que algunos trabajos pueden asomar en varias categorías, ya que han utilizado distintos puntos temporales para la fusión de datos.

Tabla 2. Categorización de papers por punto de fusión

Punto de Fusión	***Explicación***	***Estudios***
Temprana	Concatenación de las características de las diferentes fuentes de datos	Andrade et al., 2016; Bahreini et al., 2016; Chango, Cerezo, & Romero, 2021; Chango, Cerezo, Sanchez-Santillan, et al., 2021; Gadaley et al., 2020; Giannakos et al., 2019; N. L. Henderson, Rowe, Mott, & Lester, 2019; N. L. Henderson, Rowe, Mott, Brawner, et al., 2019; N. Henderson et al., 2020; Liu et al., 2019; Mao et al., 2019; Nan Liao et al., 2019; Olsen et al., 2020; Peng & Nagao, 2021; Prieto et al., 2018; Shankar et al., 2019; Wu et al., 2020; Xu et al., 2019; Yue et al., 2019; Di Mitri et al., 2017; Sharma et al., 2019; Hussain et al., 2011
Tardía	Fusión de las predicciones de cada clasificador (cada uno creado a partir de una fuente de datos)	. Chango, Cerezo, & Romero, 2021; Chango, Cerezo, Sanchez-Santillan, et al., 2021; J. Chen et al., 2014; Daoudi et al., 2021; Peng & Nagao, 2021; Wu et al., 2020; N. L. Henderson, Rowe, Mott, & Lester, 2019; N. L. Henderson, Rowe, Mott, Brawner, et al., 2019; Monkaresi et al., 2017
Híbrida	Una combinación de los dos enfoques anteriores	Brodny, 2017; Chango, Cerezo, & Romero, 2021; Chango, Cerezo, Sanchez-Santillan, et al., 2021; Luo et al., 2020
Otros	Enfoques que no encajan en ninguno de los tres descritos anteriormente	Li et al., 2020; Qu et al., 2021; Worsley, 2014; Mao et al., 2015

Fuente: (Chango et al., 2022)

4.1. Experimento 1 Fusión temprana

Se lleva a cabo la fusión de datos utilizando dos conjuntos de información distintos. El primer conjunto de datos proviene de sensores instalados en el entorno de cultivo, los cuales capturan variables ambientales y de cultivo en tiempo real, tales como temperatura, humedad y otros factores climáticos relevantes. Estos datos proporcionan una visión detallada de las condiciones que afectan el crecimiento y la salud de las plantas de pitahaya.

El segundo conjunto de datos se origina directamente de los cultivos de pitahaya e incluye información sobre las características físicas y el estado de las plantas, como el índice de clorofila, el tamaño de los frutos y cualquier signo visible de plagas o enfermedades. Estos datos son fundamentales para evaluar la calidad de la producción y detectar problemas específicos en los cultivos.

El objetivo de fusionar estos dos tipos de datos es integrar la información ambiental capturada por los sensores con los datos específicos de las plantas de pitahaya. Esta integración permite una evaluación más completa y precisa del estado de los cultivos, facilitando la identificación de patrones y relaciones entre las condiciones ambientales

y la salud de las plantas. Al combinar datos de múltiples fuentes, el experimento pretende mejorar la capacidad predictiva y la gestión de los cultivos, optimizando el manejo de plagas y enfermedades y promoviendo prácticas agrícolas más efectivas y sostenibles.

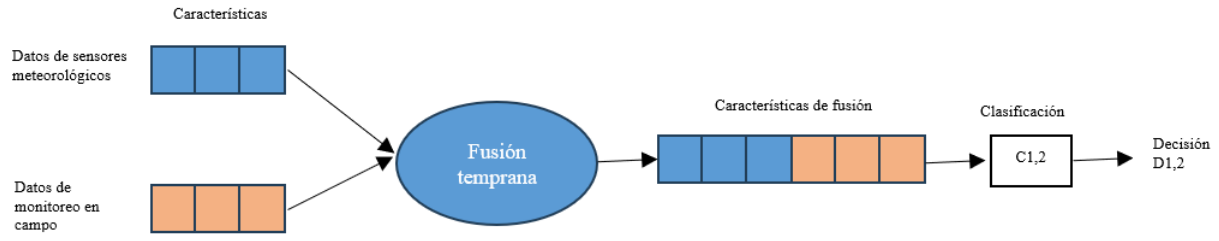
Tabla 3. Fuente de datos de estación meteorológica.

rain	temperature	rh	dew_point	wind_speed	gust_speed	wind_ direction	planta	fruto	severidad	incidencia
0	22.47	92.958	21.352	0	0	145	PL001	1	0	0
0	22.47	92.958	21.352	0	0	145	PL001	2	0	0
0	22.47	92.958	21.352	0	0	145	PL001	3	10	1
0	22.47	92.958	21.352	0	0	145	PL001	4	0	0
0	22.47	92.958	21.352	0	0	145	PL001	5	0	0

Tabla 4. Fuente de datos de monitoreo en campo

meses	Descripción	Tratamiento	Repetición	Ubicación	Index	Bringht
febrero	pitahaya sin injertar	1	1	S	219	2
febrero	pitahaya sin injertar	1	1	O	224	2
febrero	pitahaya sin injertar	1	1	S	191	2
febrero	pitahaya sin injertar	1	1	E	215	1
febrero	pitahaya sin injertar	1	1	O	329	3
febrero	pitahaya sin injertar	1	1	S	260	2
febrero	pitahaya sin injertar	1	1	E	194	1
febrero	pitahaya sin injertar	1	1	N	300	1
febrero	pitahaya sin injertar	1	1	S	261	2

Figura 5. Esquema de Fusión de datos temprana



Fuente: Logroño S, 2024

Utilizamos el algoritmo de clustering K-means para realizar una primera segmentación de los datos, agrupando las plantas en dos categorías: susceptibles a plagas y no susceptibles. Esta elección se basa en la premisa de que variables como el índice de clorofila y el índice de vigor son indicadores de la salud de la planta y, por lo tanto, de su susceptibilidad a enfermedades y plagas.

Al asignar un valor de 0 o 1 a cada muestra según su grupo, podemos establecer una prioridad de intervención para cada planta. Esta clasificación inicial nos permitirá focalizar nuestros esfuerzos en aquellas plantas que presentan un mayor riesgo de infestación.

Tabla 5. Agrupación de datos

Agrupamos en dos categorías
X = dataset.drop('descripcion', axis=1) kmeans = MiniBatchKMeans(n_clusters=2, batch_size=8).fit(X) print("Total de centros: " , len(kmeans.cluster_centers_)) print(kmeans.predict(X)) dataset['group'] = kmeans.predict(X) print(dataset)

Tabla 6. Resultado datos fusionado

	rain	temperature	rh	dew_point	wind_speed	gust_speed	wind_direction	meses	año	descripción	tratamiento	repeticion	ubicacion	indexcl	bright	sombra	group
0	0	29.826	65.107	22.777	0.5	1.5	156	2	2022	0	1	1	2	219	2	74.3	0
0	0	29.826	65.107	22.777	0.5	1.5	156	2	2022	0	1	1	4	224	2	76.7	0
0	0	29.826	65.107	22.777	0.5	1.5	156	2	2022	0	1	1	2	191	2	65	0
0	0	29.826	65.107	22.777	0.5	1.5	156	2	2022	0	1	1	3	215	1	73	0
0	0	29.826	65.107	22.777	0.5	1.5	156	2	2022	0	1	1	4	329	3	112	0
0	0	29.826	65.107	22.777	0.5	1.5	156	2	2022	0	1	1	2	260	2	88	0
0	0	29.826	65.107	22.777	0.5	1.5	156	2	2022	0	1	1	3	194	1	66	0
0	0	29.826	65.107	22.777	0.5	1.5	156	2	2022	0	1	1	1	300	1	100.7	0
0	0	29.826	65.107	22.777	0.5	1.5	156	2	2022	0	1	1	2	261	2	88.3	0
0	0	29.826	65.107	22.777	0.5	1.5	156	2	2022	0	1	1	4	335	2	113.7	0
0	0	29.826	65.107	22.777	0.5	1.5	156	2	2022	0	1	1	2	200	3	68.3	0
0	0	29.826	65.107	22.777	0.5	1.5	156	2	2022	0	1	1	2	188	1	63.7	0
0	0	29.826	65.107	22.777	0.5	1.5	156	2	2022	0	1	1	4	177	2	61	0
0	0	29.826	65.107	22.777	0.5	1.5	156	2	2022	0	1	1	2	251	1	84.7	0

El valor de SCORE representa una métrica utilizada para evaluar la calidad de la

reducción de dimensionalidad. Un valor de SCORE más alto indica que el método ha sido más efectivo en preservar la información original de los datos después de la transformación. En otras palabras, cuanto más cercano a 1 sea el valor de SCORE, mejor será la representación de los datos en el espacio de menor dimensión.

Tabla 7. Evaluación de Reducción de dimensionalidad

Reducción de dimensionalidad	**Valor**
SCORE KPCA linear	0.9624
SCORE KPCA poly	0.9248
SCORE KPCA rbf	0.8496
SCORE PCA:	0.9699
SCORE IPCA:	0.9398

Análisis de los Resultados de Reducción de Dimensionalidad

- KPCA con Kernel Linear: 0.9624

El rendimiento con el kernel lineal es bastante alto, con una precisión de 0.9624. Esto indica que el kernel lineal ha sido efectivo para capturar las relaciones lineales en los datos y ha proporcionado una buena capacidad predictiva.

- KPCA con Kernel Polinómico: 0.9248

El rendimiento con el kernel polinómico es el más alto de todos los métodos de KPCA, con una precisión de 0.9248. Esto sugiere que el kernel polinómico ha sido particularmente eficaz para capturar las relaciones no lineales en los datos, proporcionando una mejor representación para la clasificación.

- KPCA con Kernel RBF: 0.8496

El rendimiento con el kernel RBF es notablemente más bajo, con una precisión de 0.8496. Aunque el kernel RBF es conocido por su capacidad para manejar complejidades y no linealidades en los datos, en este caso, su rendimiento no ha sido tan bueno como los kernels lineal y polinómico.

- PCA (Análisis de Componentes Principales): 0.9699

El rendimiento de PCA es el más alto de todos los métodos, con una precisión de 0.9699. Esto indica que la reducción de dimensionalidad mediante PCA ha sido muy efectiva, proporcionando una representación que captura la mayor parte de la varianza

en los datos, lo que resulta en una alta capacidad predictiva.

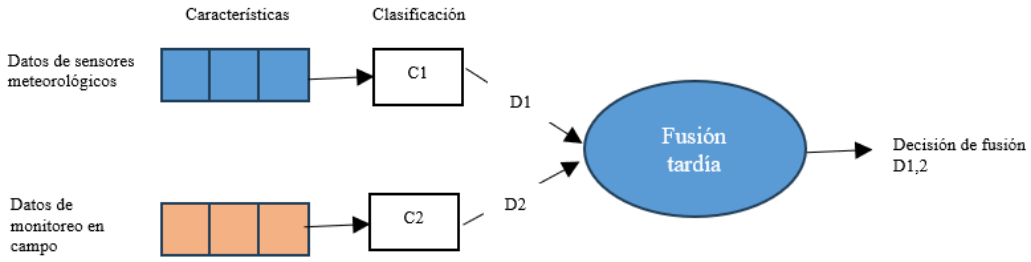
- IPCA: 0.9398

El rendimiento de IPCA es considerablemente bueno, con una precisión de 0.9398. Aunque es menos efectivo en comparación con PCA, sigue siendo una técnica valiosa, especialmente cuando se trabaja con grandes conjuntos de datos que requieren procesamiento en lotes.

Experimento

4.2. Experimento 1 Fusión tardía

Figura 6. Esquema de Fusión de datos tardía



Fuente: Logroño S, 2024

Tabla 8. Fuente de datos de estación meteorológica – data1

rain	temperature	rh	dew_point	wind_speed	gust_speed	wind_ direction	meses	año	sombra	group
0	29.826	65.107	22.777	0.5	1.5	156	2	2022	74.3	0
0	29.826	65.107	22.777	0.5	1.5	156	2	2022	76.7	0
0	29.826	65.107	22.777	0.5	1.5	156	2	2022	65	0
0	29.826	65.107	22.777	0.5	1.5	156	2	2022	73	0
0	29.826	65.107	22.777	0.5	1.5	156	2	2022	112	0
0	29.826	65.107	22.777	0.5	1.5	156	2	2022	88	0
0	29.826	65.107	22.777	0.5	1.5	156	2	2022	66	0
0	29.826	65.107	22.777	0.5	1.5	156	2	2022	100.7	0
0	29.826	65.107	22.777	0.5	1.5	156	2	2022	88.3	0
0	29.826	65.107	22.777	0.5	1.5	156	2	2022	113.7	0

- KPCA con Kernel Linear: 0.9548

El rendimiento del kernel lineal es alto, con una precisión de 0.9548. Esto indica que el kernel lineal ha sido muy efectivo en capturar las relaciones lineales en los datos, proporcionando una capacidad predictiva robusta.

- KPCA con Kernel Polinómico: 0.9022

El rendimiento con el kernel polinómico es bueno pero menor que el del kernel lineal, con una precisión de 0.9022. Aunque el kernel polinómico captura relaciones no lineales, en este caso no ha superado al kernel lineal.

- KPCA con Kernel RBF: 0.8496

El rendimiento con el kernel RBF es el más bajo de los métodos de KPCA, con una precisión de 0.8496. Aunque el kernel RBF es conocido por su capacidad para manejar complejidades en los datos, su rendimiento ha sido inferior en comparación con los otros kernels en este conjunto de datos.

- PCA: 0.9624

El rendimiento de PCA es alto, con una precisión de 0.9624. Esto sugiere que PCA ha sido muy efectivo en la reducción de dimensionalidad mientras mantiene la mayoría de la información relevante para la clasificación, lo que resulta en una alta capacidad predictiva.

- IPCA: 0.9548

El rendimiento de IPCA coincide con el de KPCA con kernel lineal, con una precisión de 0.9548. Esto indica que IPCA ha sido igualmente efectivo en la reducción de dimensionalidad y en la preservación de la capacidad predictiva, especialmente en conjuntos de datos grandes o cuando el procesamiento por lotes es necesario.

Tabla 9. Evaluación de reducción de dimensionalidad data 1

data 1	
Reducción de dimensionalidad	Valor
SCORE KPCA linear	0.9548
SCORE KPCA poly	0.9022
SCORE KPCA rbf	0.8496
SCORE PCA:	0.9624
SCORE IPCA:	0.9548

En general, PCA y KPCA con kernel lineal han demostrado ser las técnicas más

efectivas en este conjunto de datos, mientras que KPCA con kernel RBF ha sido menos efectivo. IPCA también ha mostrado un rendimiento comparable al de PCA y KPCA con kernel lineal, lo que lo hace adecuado para conjuntos de datos grandes.

Tabla 10. Fuente de datos de monitoreo en campo - data 2

meses	año	descripcion	tratamiento	repeticion	ubicacion	indexcl	bringht	group
2	2022	0	1	1	2	219	2	0
2	2022	0	1	1	4	224	2	0
2	2022	0	1	1	2	191	2	0
2	2022	0	1	1	3	215	1	0
2	2022	0	1	1	4	329	3	0
2	2022	0	1	1	2	260	2	0
2	2022	0	1	1	3	194	1	0
2	2022	0	1	1	1	300	1	0
2	2022	0	1	1	2	261	2	0

- KPCA con Kernel Lineal: 0.9323

El rendimiento del kernel lineal es alto, con una precisión de 0.9323. Esto indica que el kernel lineal es efectivo para capturar relaciones lineales en los datos, proporcionando un modelo predictivo robusto.

- KPCA con Kernel Polinómico: 0.9398

El rendimiento con el kernel polinómico es ligeramente superior al del kernel lineal, con una precisión de 0.9398. Esto sugiere que las relaciones no lineales capturadas por el kernel polinómico han mejorado la capacidad predictiva del modelo en este conjunto de datos.

- KPCA con Kernel RBF: 0.8045

El rendimiento con el kernel RBF es considerablemente más bajo, con una precisión de 0.8045. Aunque el kernel RBF es potente para manejar datos no lineales y complejos, en este caso no ha proporcionado un rendimiento tan bueno como los otros kernels. Esto podría indicar que el patrón no lineal que intenta capturar el kernel RBF no es tan relevante o está sobreajustado a los datos.

- PCA: 0.9699

El rendimiento de PCA es el más alto entre todas las técnicas, con una precisión de 0.9699. Este resultado indica que PCA ha logrado reducir la dimensionalidad del conjunto de datos mientras preserva la mayoría de la información relevante para la

clasificación, lo que resulta en un modelo con una excelente capacidad predictiva.

- IPCA: 0.8796

El rendimiento de IPCA es menor que el de PCA, con una precisión de 0.8796. Aunque IPCA es útil en escenarios donde los datos son grandes y necesitan ser procesados en lotes, su capacidad para preservar la información relevante no ha sido tan eficaz como la de PCA en este caso.

Tabla 11. Evaluación reducción de dimensionalidad data 2

data 2	
Reducción de dimensionalidad	Valor
SCORE KPCA linear	0.9323
SCORE KPCA poly	0.9398
SCORE KPCA rbf	0.8045
SCORE PCA:	0.9699
SCORE IPCA:	0.8796

En general, PCA es la técnica más recomendada para este conjunto de datos, seguida de cerca por KPCA con Kernel Polinómico.

Tabla 12. Evaluación de fusión de la data 1 y data 2

Fusion data 1 y data 2	
Reducción de dimensionalidad	**Valor**
SCORE KPCA linear	0,94355
SCORE KPCA poly	0,921
SCORE KPCA rbf	0,82705
SCORE PCA:	0,96615
SCORE IPCA:	0,9172

- KPCA con Kernel Lineal: 0.94355

El kernel lineal ha demostrado un rendimiento sólido con una precisión de 0.94355. Este resultado es consistente con la capacidad del kernel lineal para manejar relaciones lineales en los datos, lo que sugiere que las relaciones lineales tienen un peso significativo en este conjunto de datos fusionado.

- KPCA con Kernel Polinómico: 0.921

El rendimiento del kernel polinómico es ligeramente inferior al del kernel lineal,

con una precisión de 0.921. Esto indica que las relaciones no lineales capturadas por el kernel polinómico son menos predominantes o relevantes en este conjunto de datos fusionado.

- KPCA con Kernel RBF: 0.82705

El kernel RBF ha obtenido la precisión más baja entre las técnicas de KPCA, con un valor de 0.82705. Aunque el kernel RBF es eficaz para modelar relaciones complejas no lineales, en este caso, parece no haber capturado las características más relevantes de los datos fusionados, lo que se traduce en un rendimiento inferior.

- PCA: 0.96615

PCA ha logrado el mejor rendimiento en este conjunto de datos fusionado, con una precisión de 0.96615. Este resultado indica que PCA es muy efectivo para capturar la estructura subyacente de los datos fusionados, preservando la mayor parte de la información relevante para la clasificación.

- IPCA: 0.9172

IPCA ha mostrado un rendimiento moderado con una precisión de 0.9172. Aunque IPCA es útil para procesar grandes volúmenes de datos de manera incremental, en este caso, no ha alcanzado la misma precisión que PCA. Sin embargo, sigue siendo una opción viable para escenarios donde se requiere procesamiento por lotes.

Tabla 13. Comparación de los resultados de técnicas de fusión

Reducción de dimensionalidad	Fusión tardía	Fusión temprana
SCORE KPCA linear	0,94355	0.9624
SCORE KPCA poly	0,921	0.9248
SCORE KPCA rbf	0,82705	0.8496
SCORE PCA:	0,96615	0.9699
SCORE IPCA:	0,9172	0.9398

- KPCA Linear

Fusión Temprana: 0.9624

Fusión Tardía: 0.94355

La fusión temprana ha resultado en una mayor precisión (0.9624) en comparación con la fusión tardía (0.94355). Esto sugiere que, al aplicar el kernel lineal, la integración temprana de los datos permitió capturar mejor las relaciones lineales presentes en el

conjunto de datos.

- KPCA Poly

  Fusión Temprana: 0.9248

  Fusión Tardía: 0.921

  Similar a KPCA Linear, la fusión temprana ha tenido un rendimiento ligeramente superior (0.9248) frente a la fusión tardía (0.921), indicando que las relaciones polinómicas fueron modeladas con mayor efectividad cuando los datos fueron fusionados de manera temprana.

- KPCA RBF

  Fusión Temprana: 0.8496

  Fusión Tardía: 0.82705

  La fusión temprana ha logrado una mejor precisión (0.8496) que la fusión tardía (0.82705) al usar el kernel RBF. Esto sugiere que la complejidad no lineal de los datos es mejor capturada cuando se combinan tempranamente.

- PCA (Análisis de Componentes Principales)

  Fusión Temprana: 0.9699

  Fusión Tardía: 0.96615

  PCA ha mostrado un rendimiento ligeramente superior en la fusión temprana (0.9699) en comparación con la tardía (0.96615). Esto indica que la reducción de dimensionalidad mediante PCA es más efectiva cuando se aplica después de una integración temprana de los datos.

- IPCA (Incremental PCA)

  Fusión Temprana: 0.9398

  Fusión Tardía: 0.9172

  IPCA también muestra un rendimiento mejor en la fusión temprana (0.9398) frente a la fusión tardía (0.9172). Dado que IPCA maneja datos en bloques, la fusión temprana parece haber facilitado una mayor retención de la variabilidad explicada en cada bloque.

Capítulo 5

Discusión

En esta sección se discutirá el desempeño de diversas técnicas de reducción de dimensionalidad empleando datos de dos experimentos que utilizan métodos de fusión temprana y tardía. El objetivo es interpretar los resultados obtenidos y evaluar las implicaciones teóricas y prácticas de los diferentes métodos, identificando sus fortalezas y limitaciones. También se ofrecerán sugerencias para futuras investigaciones y se concluirá con una reflexión sobre los hallazgos.

La fusión temprana demostró un rendimiento superior con un valor SCORE de 0.9624 en comparación con 0.94355 en la fusión tardía. Esto sugiere que la integración de datos en una etapa temprana permite una mejor captura de las relaciones lineales presentes en los datos, posiblemente debido a una mejor preservación de la información relevante antes de aplicar la reducción de dimensionalidad.

A pesar de que la diferencia es mínima, la fusión temprana resultó ligeramente superior. Esto podría indicar que las relaciones no lineales se capturan de manera más efectiva cuando los datos se integran antes de la reducción dimensional.

La fusión temprana también ofreció un mejor rendimiento con el kernel RBF, lo que podría reflejar una mayor capacidad para modelar complejidades no lineales en los datos cuando se realiza una integración temprana.

PCA mostró un rendimiento muy cercano entre ambas técnicas de fusión, pero la fusión temprana resultó en una ligera mejora. Esto indica que PCA, al capturar la mayor parte de la varianza en los datos, se beneficia mínimamente de la integración temprana.

IPCA también mostró un mejor rendimiento con la fusión temprana, lo que puede estar relacionado con una mejor preservación de la variabilidad en datos procesados en bloques.

Los resultados sugieren que la fusión temprana de datos tiende a proporcionar una mejor preservación de la información relevante en la reducción de dimensionalidad,

especialmente para métodos que capturan relaciones lineales y no lineales. Esto tiene implicaciones para la teoría de reducción dimensional, indicando que la estrategia de fusión puede influir significativamente en la calidad de la representación de los datos.

En aplicaciones prácticas, la elección entre fusión temprana y tardía puede depender de la naturaleza de los datos y de los objetivos del análisis. La fusión temprana podría ser preferible en escenarios donde la preservación de las relaciones originales es crucial, mientras que la fusión tardía podría ser útil en contextos donde el procesamiento por lotes es necesario o más eficiente.

Los resultados pueden estar influenciados por las características específicas de los conjuntos de datos utilizados, como el tamaño, la complejidad y la estructura de los datos. El valor SCORE es una métrica útil, pero puede no capturar todos los aspectos de la efectividad de la reducción dimensional. Otras métricas o métodos de evaluación podrían proporcionar una visión más completa.

Investigar técnicas adicionales o combinaciones de métodos podría proporcionar una visión más amplia sobre la efectividad de la reducción dimensional en diferentes contextos. Aplicar estos métodos a conjuntos de datos con diversas características y estructuras podría validar y extender los hallazgos actuales. Además, experimentar con la optimización de parámetros de los métodos de reducción dimensional podría mejorar la capacidad predictiva y la preservación de información.

Los resultados de los experimentos indican que la fusión temprana de datos tiende a ofrecer mejores resultados en la reducción de dimensionalidad en comparación con la fusión tardía, especialmente para métodos que capturan relaciones lineales y no lineales. PCA y KPCA con kernel lineal han demostrado ser las técnicas más efectivas en la preservación de la información y capacidad predictiva. Las limitaciones del estudio sugieren la necesidad de investigación adicional en diferentes contextos y con otros métodos. Las sugerencias para futuras investigaciones proporcionan una dirección para ampliar y profundizar el análisis de la reducción dimensional en diversos escenarios.

Conclusiones

Los resultados obtenidos evidencian la importancia de la elección de la técnica de reducción de dimensionalidad y el momento de la fusión de datos. El PCA se consolidó como una herramienta robusta y versátil, capaz de capturar la mayor parte de la

variabilidad en los datos. El KPCA complementó a PCA, demostrando ser eficaz en la captura de relaciones no lineales, especialmente cuando se utiliza el kernel lineal.

La fusión temprana de los datos resultó ser una estrategia beneficiosa, ya que permitió una mejor preservación de la información relevante y facilitó la aplicación de las técnicas de reducción de dimensionalidad. Esto tiene implicaciones importantes para futuras investigaciones, ya que sugiere que el orden en el que se combinan los datos puede influir significativamente en los resultados obtenidos.

En términos prácticos, estos hallazgos ofrecen directrices para la selección de la técnica de reducción de dimensionalidad más adecuada en diferentes escenarios. PCA es una excelente opción para una amplia gama de aplicaciones, mientras que KPCA puede ser más adecuado cuando se sospecha la existencia de relaciones no lineales complejas. La fusión temprana se recomienda como estrategia general, aunque es importante evaluar cada caso de forma individual

Recomendaciones

Los resultados obtenidos en este estudio abren nuevas vías para futuras investigaciones. Se recomienda continuar experimentando con la optimización de los parámetros de los métodos de reducción de dimensionalidad, ya que ajustes finos pueden mejorar significativamente su desempeño. Además, es crucial explorar combinaciones de técnicas y métodos híbridos para abordar problemas más complejos.

La diversidad de los conjuntos de datos utilizados en futuras investigaciones es fundamental para validar la generalizabilidad de los hallazgos. Evaluar el desempeño de las técnicas en datos con diferentes características y tamaños permitirá determinar su robustez y limitaciones.

En términos de aplicaciones prácticas, la elección entre fusión temprana y tardía dependerá de las características específicas del problema y de los recursos computacionales disponibles. La fusión temprana, aunque ofrece ventajas en términos de preservación de la información, puede ser computacionalmente más costosa.

Para evaluar de manera más exhaustiva la calidad de la reducción dimensional, se sugiere utilizar un conjunto de métricas complementarias al valor SCORE. Estas métricas pueden incluir medidas de error de reconstrucción, capacidad de generalización y interpretabilidad de los resultados.

IV. Referencias

Balendres, M. A., & Bengoa, J. C. (2019). Diseases of dragon fruit (Hylocereus species): Etiology and current management options. In *Crop Protection* (Vol. 126). Elsevier Ltd. https://doi.org/10.1016/j.cropro.2019.104920

Benos, L., Tagarakis, A. C., Dolias, G., Berruto, R., Kateris, D., & Bochtis, D. (2021). Machine learning in agriculture: A comprehensive updated review. In *Sensors* (Vol. 21, Issue 11). MDPI AG. https://doi.org/10.3390/s21113758

C A Padmanabha Reddy, Y., Viswanath, P., & Eswara Reddy, B. (2018). Semi-supervised learning: a brief review. *International Journal of Engineering & Technology*, *7*(1.8), 81. https://doi.org/10.14419/ijet.v7i1.8.9977

Chango, W., Lara, J. A., Cerezo, R., & Romero, C. (2022). A review on data fusion in multimodal learning analytics and educational data mining. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, *12*(4). https://doi.org/10.1002/widm.1458

Dai, M., Shen, Y., Li, X., Liu, J., Zhang, S., & Miao, H. (2024). Digital Twin System of Pest Management Driven by Data and Model Fusion. *Agriculture*, *14*(7), 1099. https://doi.org/10.3390/agriculture14071099

Dhanaraj, R. K., Ali, Md. A., Sharma, A. K., & Nayyar, A. (2024). Deep Multibranch Fusion Residual Network and IoT-based pest detection system using sound analytics in large agricultural field. *Multimedia Tools and Applications*, *83*(13), 40215–40252. https://doi.org/10.1007/s11042-023-16897-3

Diéguez-Santana, K., Sarduy-Pereira, L. B., Sablón-Cossío, N., Bautista-Santos, H., Sánchez-Galván, F., & Ruíz Cedeño, S. D. M. (2022). Evaluation of the Circular Economy in a Pitahaya Agri-Food Chain. *Sustainability (Switzerland)*, *14*(5). https://doi.org/10.3390/su14052950

Friha, O., Ferrag, M. A., Shu, L., Maglaras, L., & Wang, X. (2021). Internet of Things for the Future of Smart Agriculture: A Comprehensive Survey of Emerging Technologies. In *IEEE/CAA Journal of Automatica Sinica* (Vol. 8, Issue 4, pp. 718–752). Institute of Electrical and Electronics Engineers Inc. https://doi.org/10.1109/JAS.2021.1003925

Karunathilake, E. M. B. M., Le, A. T., Heo, S., Chung, Y. S., & Mansoor, S. (2023). The Path to Smart Farming: Innovations and Opportunities in Precision Agriculture. In *Agriculture (Switzerland)* (Vol. 13, Issue 8). Multidisciplinary Digital Publishing Institute (MDPI). https://doi.org/10.3390/agriculture13081593

Lee, S., & Yun, C. M. (2023). A deep learning model for predicting risks of crop pests and diseases from sequential environmental data. *Plant Methods*, *19*(1). https://doi.org/10.1186/s13007-023-01122-x

Maimaitijiang, M., Sagan, V., Sidike, P., Daloye, A. M., Erkbol, H., & Fritschi, F. B. (2020). Crop monitoring using satellite/UAV data fusion and machine learning. *Remote Sensing*, *12*(9). https://doi.org/10.3390/RS12091357

Naeem, S., Ali, A., Anam, S., & Ahmed, M. M. (2023). An Unsupervised Machine Learning Algorithms: Comprehensive Review. *International Journal of Computing and Digital Systems*, *13*(1), 911–921. https://doi.org/10.12785/ijcds/130172

Osinga, S. A., Paudel, D., Mouzakitis, S. A., & Athanasiadis, I. N. (2022). Big data in agriculture: Between opportunity and solution. *Agricultural Systems*, *195*. https://doi.org/10.1016/j.agsy.2021.103298

Padhiary, M., Saha, D., Kumar, R., Sethi, L. N., & Kumar, A. (2024). Enhancing precision agriculture: A comprehensive review of machine learning and AI vision applications in all-terrain vehicle for farm automation. In *Smart Agricultural Technology* (Vol. 8). Elsevier B.V. https://doi.org/10.1016/j.atech.2024.100483

San Emeterio de la Parte, M., Martínez-Ortega, J. F., Hernández Díaz, V., & Martínez, N. L. (2023). Big Data and precision agriculture: a novel spatio-temporal semantic IoT data management framework for improved interoperability. *Journal of Big Data*, *10*(1). https://doi.org/10.1186/s40537-023-00729-0

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

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

Schröer, C., Kruse, F., & Gómez, J. M. (2021). A systematic literature review on applying CRISP-DM process model. *Procedia Computer Science*, *181*, 526–534. https://doi.org/10.1016/j.procs.2021.01.199

Tinoco, L., Bastidas, S., Chuquimarca, J., Macas, J., & Viera, W. (2020). *Manual del Cultivo de Pitahaya para la Amazonía Ecuatoriana Instituto Nacional de Investigaciones Agropecuarias Estación Experimental Central de la Amazonía*.