Análisis y Modelación de Datos

Aplicaciones para la Clasificación de Granos Oryza Sativa L con Inteligencia Artificial

1era Edición

SciELa

Área: U - Computación y tecnología de la información

Análisis y Modelación de Datos:

Aplicaciones para la Clasificación de Granos Oryza Sativa L con Inteligencia Artificial ©

Edición: Primera

Autores:

José Julián Coronel Reyes
Héctor Ramiro Carvajal Romero
Jessica Maribel Quezada Campoverde
Carlota Rosario Delgado Vera
Ramiro Fernando Carrión Durán

SciELa

Area: U - Computer and information technology

Data Analysis and Modeling:

Applications for the Classification of Oryza Sativa L with Artificial Intelligence ©

Edition: First

Authors:

José Julián Coronel Reyes
Héctor Ramiro Carvajal Romero
Jessica Maribel Quezada Campoverde
Carlota Rosario Delgado Vera
Ramiro Fernando Carrión Durán

SciELa

ISBN: 978-9942-7264-6-9

Primera Edición, Febrero 2025 ©

Análisis y Modelación de Datos: Aplicaciones para la Clasificación de Granos Oryza Sativa L con Inteligencia Artificial

ISBN digital: 978-9942-7264-6-9
DOI: https://doi.org/10.62131/978-9942-7264-6-9

Editado por: © Editorial Investigativa Latinoamericana (SciELa)

Quevedo, Los Ríos, Ecuador

→ **E-mail:** admin@editorial-sciela.org
→ **Código Postal:** 120303
→ **WEB:** https://editorial-sciela.org

Este libro se sometió a arbitraje bajo el sistema de doble ciego (peer review) y antiplágio. Este producto investigativo cumple con la Declaración de Principios de Budapest, San Francisco, México, Helsinki y Firma del Marco del MIT.

Dirección editorial: Lic. Alexander Fernando Haro, MSI.

→ **Revisor (1):** Ing. Fabian Alberto Gallardo Gonzaga, Mg.
→ **Revisor (2):** Ing. Miguel Ángel Toaza Garces, Mg.
→ **Revisor (3):** Ing. Cynthia Lisseth Medina Maldonado, Mg.

Sistema de clasificación decimal DEWEY

→ 338.1 - Agricultura

Clasificación comercial internacional - THEMA

→ U - Computación y tecnologías de la información
→ UX - Informática aplicada
→ UXT - Aplicaciones informáticas para la industria y la tecnología

AUTORES

José Julián Coronel Reyes

- **ORCID:** https://orcid.org/0000-0002-7883-5388
- **Correo:** coroneljulian@live.com
- **Ciudad/País:** Machala - Ecuador

- **Filiación:** Universidad Técnica de Machala

- Economista Agropecuario por la Universidad Técnica de Machala, Diplomado en Minería de Datos por la Universidad Central de Venezuela, Máster en Sistema de Información con mención en Inteligencia de Negocios y Analítica de Datos Masivos por la UNEMI de Ecuador. Ponente en congresos locales e internacionales, consultor Agropecuario y facilitador en el manejo de Software estadísticos y gestores bibliográficos. Sus intereses de investigación son Business Intelligence (BI), Machine Learning (ML), Deep Learning (DP), además del uso de tecnologías para el mejoramiento de la productividad agropecuaria. Cuenta con publicaciones indexadas, varias de ellas en revistas de alto impacto en los índices de JCR y SJR. Actualmente es Coordinador de Planificación de la Red Municipal de Salud Machala EP y también es Docente del Instituto Superior Tecnológico España en programas de maestrías.

Héctor Ramiro Carvajal Romero

- **ORCID:** https://orcid.org/0000-0001-6303-6295
- **Correo:** hcarvajal@utmachala.edu.ec
- **Ciudad/País:** Machala - Ecuador

- **Filiación:** Universidad Técnica de Machala

- Doctor en Análisis Económico y Estrategia Empresarial (Universidade A Coruña). Magíster en Administración de Empresas (UPS). Especialista en gerencia Educativa (UASB). Graduado de Ingeniero Comercial especialidad Administración de Empresas (UTMACH). Docente a nivel Pregrado en la Facultad de Ciencias Agropecuarias (UTMACH) desde el año 2016.

Jessica Maribel Quezada Campoverde

- **ORCID:** https://orcid.org/0000-0003-2760-4827
- **Correo:** jquezada@utmachala.edu.ec
- **Ciudad/País:** Machala - Ecuador

- **Filiación:** Universidad Técnica de Machala

- Mi trayectoria profesional se ha centrado en la docencia universitaria y la vinculación con la sociedad en el campo de la agronomía. Como docente en la Universidad Técnica de Machala, he impartido conocimientos y formado a futuros ingenieros agrónomos en la Facultad de Ciencias Agropecuarias.Mi compromiso con la agricultura sostenible y la seguridad alimentaria me ha llevado a participar en diversos proyectos de vinculación con la sociedad. Estos proyectos han buscado soluciones innovadoras y prácticas para mejorar la producción agrícola, promover el acceso a alimentos nutritivos y fomentar el desarrollo de comunidades agrícolas sostenibles.

Carlota Rosario Delgado Vera

- **ORCID:** https://orcid.org/0000-0002-8527-6078
- **Correo:** cdelgado@uagraria.edu.ec
- **Ciudad/País:** Guayaquil - Ecuador

- **Filiación:** Universidad Agraria del Ecuador

- Profesional con amplia trayectoria en el ámbito de la ingeniería en computación y la docencia universitaria. Cuento con una formación académica sólida, que incluye una Maestría en Sistemas de Información Gerencial, una Maestría en Educación Informática, una Ingeniería en Computación e Informática y una Tecnología en el mismo campo. Mi experiencia investigativa se ha centrado en el desarrollo de algoritmos de inteligencia artificial, contribuyendo al avance de esta disciplina. Adicionalmente, he participado activamente en proyectos de vinculación con la comunidad y he colaborado como revisora de libros especializados. Mi compromiso es integrar conocimiento, innovación y docencia para formar profesionales capaces de enfrentar los retos tecnológicos del futuro.

Ramiro Fernando Carrión Durán

- **ORCID:** https://orcid.org/0009-0009-1417-5534
- **Correo:** ramiro.carrion@educacion.gob.ec
- **Ciudad/País:** Loja - Ecuador

- **Filiación:** Colegio de Bachillerato PCEI "Teniente Hugo Ortiz"

- Economista por la Universidad Internacional del Ecuador, con un Máster Universitario en Investigación de Mercados por la Universidad Internacional de La Rioja y un Máster en Sistemas de Información con mención en Inteligencia de Negocios y Analítica de Datos Masivos por la UNEMI. Actualmente, se desempeña como docente y coordinador en el Colegio de Bachillerato PCEI "Teniente Hugo Ortiz".

Índice

CAPÍTULO I.

Definiendo el Problema

CAPÍTULO II.

Marco Teórico Referencial

Prefacio

«La agricultura es la profesión propia del sabio, la más adecuada al sencillo y la ocupación más digna para todo hombre libre». **Marco Tulio Cicerón (106 - 43 a.C.)**

El presente libro plantea como objetivo general de investigación, aplicar las diferentes técnicas de aprendizaje automático en la clasificación de variedades de granos en arroz. La calidad del grano de arroz se determina habitualmente mediante inspecciones visuales y mediciones manuales, método que es lento, subjetivo y propenso a errores humanos. Por lo tanto, la industria demanda una técnica rápida y precisa que pueda clasificar el grano de arroz a bajo costo y estandarizarlo. Por medio del aprendizaje automático se puede desarrollar modelos de decisión en entornos eminentemente complejos y no lineales.

Este estudio se enmarca en un diseño de investigación cuantitativa, consiste en la recolección de datos numéricos para su posterior análisis e interpretación utilizando herramientas de análisis matemático y estadístico para describir y explicar fenómenos mediante datos numéricos. El tipo de la investigación es causal comparativa dado que se comparan diferentes técnicas de aprendizaje automático en el proceso de clasificación de granos de arroz.

La clasificación fue binaria, se consideró como población de estudio a los 3.260 granos de dos variedades del Instituto Nacional de Investigación Agropecuaria 14 (1.630) y 17 (1.630). Las técnicas que fueron empleadas están basadas en cinco modelos utilizando los siguientes algoritmos: regresión logística, Perceptrón Multicapa, Máquina

de Vectores de Apoyo, Bosque Aleatorio y K Vecino Más Cercano. Las mediciones estadísticas de la matriz de confusión como resultado de la clasificación se utilizaron como métricas de rendimiento. Los resultados permiten concluir que Máquina de Vectores de Soportes el mejor método de clasificación, debido a su mejor predicción de los valores obtenidos de verdaderos positivos, y verdaderos negativos, además, la media de la precisión del modelo fue de 93,33%, superior a los otros modelos. El desarrollo de este trabajo permite llegar a la conclusión de que el uso del aprendizaje automático en la industria arrocera, es aplicable para el apoyo en la toma las decisiones basadas en el desarrollo del proceso de clasificación del grano.

Capítulo I.

Definiendo el Problema

Capítulo I.

Definiendo el Problema

1.1. Introducción

El arroz *(Oryza sativa L)* es una fuente alta de nutrientes en la dieta humana. El cereal más importante y más cultivado después del trigo y el maíz en todo el mundo es el arroz (Krishna et al., 2022; Sultana et al., 2022). Del mismo modo, también desempeña un papel relevante en la agricultura del Ecuador.

La presentación de los granos se determina por las características físicas, el tamaño, el color, la salud interna y la diversidad aumenta el valor de mercado.

La clasificación manual de la variedad de arroz es un proceso difícil. Además, este método requiere tiempo y es ineficiente, sobre todo cuando se trabaja con grandes volúmenes de producción (Ndikuryayo et al., 2022). Por lo tanto, la industria demanda con urgencia

"La revolución del arroz comienza cuando el ojo humano cede el paso a la precisión de la inteligencia artificial"

una técnica rápida y precisa que pueda clasificar el grano de arroz a bajo costo y ampliamente estandarizado.

En este sentido, la IA es una herramienta informática muy valiosa cuyas técnicas son cada vez más utilizadas en diversos campos y además permiten crear sistemas que imiten un comportamiento inteligente.

Por esta razón, con la ayuda de la IA permite desarrollar modelos de decisión en entornos eminentemente complejos y no lineales. En la actualidad, estas técnicas son tendencia en países con tecnología desarrollada y alta inversión en I+D+i (Mishra & Tyagi, 2022; Xu et al., 2022).

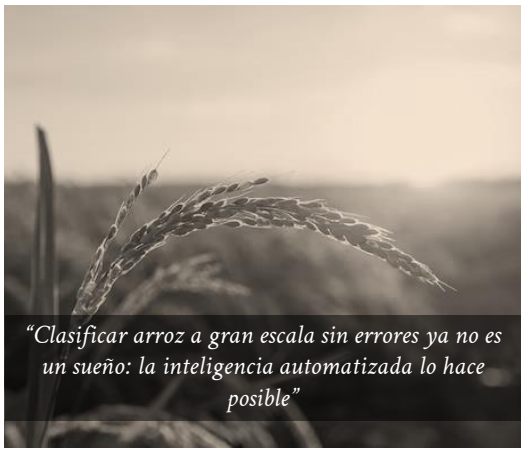
"En la era de la tecnología, cada grano de arroz cuenta su propia historia: la IA se encarga de contarla con máxima exactitud"

Los algoritmos de ML pueden ser entrenados de manera supervisada para el análisis de los datos agrícolas. En tal sentido, algunas técnicas de ML se están popularizando como buenas alternativas a las técnicas clásicas, ya que se basan en el reconocimiento de patrones (Greener et al., 2022; Rolnick et al., 2022).

Además, se han aplicado algoritmos para la evaluación de la calidad de los alimentos, incluida la carne de atún, cerdo y salmón (Chen et al., 2022; Medeiros et al., 2021), huevos de pato y carne de camello (Dong et al., 2021; Molaei et al., 2021), frutas y verduras (Ali & Dildar, 2021), en salud (Moya et al., 2017). La combinación de técnicas estadísticas multivariantes se ha convertido en una poderosa herramienta para hacer frente a diversos problemas en el sector alimentario.

Recientemente se han publicado trabajos que utilizan algoritmos para clasificar variedades de granos en arroz utilizando diferentes métodos analíticos (Aggarwal et al., 2022; Khatri et al., 2022; Komal et al., 2022; Tosawadi et al., 2022)., entre otros trabajos.

Por esta razón, siguiendo los argumentos de trabajos anteriores, se planteó la hipótesis de que el uso de técnicas de ML se puede aplicar como un medio para respaldar la toma de decisiones basada en datos en los sistemas de producción agrícola.



"Clasificar arroz a gran escala sin errores ya no es un sueño: la inteligencia automatizada lo hace posible"

En la actualidad las industrias arroceras han logrado incrementos en sus producciones debido a su implementación de nuevas maquinarias agrícolas que permiten procesar 1 hasta 10 tn/hora.

Sin embargo, la mayoría de las empresas carecen de tecnologías al momento de clasificar granos de arroz, mientras otro método es la clasificación manual, pero es un proceso difícil. Además, este método requiere tiempo y es ineficiente, sobre todo cuando se trabaja con grandes volúmenes de producción. Por lo tanto, la industria demanda con urgencia una técnica rápida y precisa que pueda clasificar el grano de arroz a bajo costo y ampliamente estandarizado (Ndikuryayo et al., 2022).

1.2. Planteamiento del problema

La Organización de las Naciones Unidas para la Alimentación y la Agricultura indica que para el año 2030 la producción mundial de

arroz tendrá que aumentar un 40% para satisfacer la necesidad alimentaria de millones de habitantes. En este contexto, el rol del sector agrícola arrocero adquiere un protagonismo mayúsculo, en la provisión de alimentos y materias primas (Amos et al., 2022).

Ecuador cuenta con unas buenas zonas agroecológicas, en donde puede desarrollarse el cultivo de arroz, así mismo, las provincias que tiene la mayor producción de la gramínea son Guayas, Los Ríos, Manabí y El Oro.

Sin embargo, la calidad del grano arroz se juzga en función de los atributos, que podrían clasificarse de forma intrínseca, entre las importante tenemos su textura, olor y sabor, mientras que en otro escenario es extrínsecas que tiene que ver mucho con la marca, empaquetado y etiquetado (Zahra et al., 2022).

Además, las características visuales de los granos de arroz son atributos de búsqueda importantes que afectan las decisiones de compra de los consumidores y, por lo tanto, se utilizan como algunos de los primeros criterios de selección como el tamaño, longitud, grosor, color del grano en los diferentes países y mercados (Ali et al., 2022). Es precisamente en este punto, donde se evidencia una oportunidad de aplicar técnicas ML, que permitan reducir el tiempo de clasificación y minimizar los costos generados por esta actividad.

Durante esta etapa, una persona se encarga del proceso de clasificación del arroz, sin embargo, este método usualmente no es eficaz y el tiempo empleado para esta actividad es extenso, como consecuencia se generan demoras en el proceso.

Al tratarse de un método que no es totalmente certero, también existe la probabilidad de errores y falta de homogenización en los granos de arroz.


"Innovar en la agricultura es asegurar el alimento del mañana; la IA es la llave que abre esa puerta"

Tomando en cuenta esta premisa, es necesario mejorar el proceso de clasificación de los granos de arroz, considerando para esto sus características morfológicas, tamaño, entre otras. Para cumplir con esta actividad se puede recurrir al uso de la tecnología, particularmente al aprendizaje automático con la finalidad de mejorar los tiempos de clasificación, al mismo tiempo que se optimiza los costos de producción.

En la actualidad existen diferentes algoritmos de aprendizaje automático, sin embargo, no todos ellos tienen la misma precisión, particularmente para la clasificación del arroz, actividad que requiere bastante precisión, por lo cual surge la necesidad de investigar su funcionamiento para este tipo de producto específicamente.


"En la cadena productiva, el verdadero secreto está en los datos: el ML convierte simples números en decisiones inteligentes"

Según Vecchio et al. (2022), manifiesta que las universidades, centros de investigación públicos y privados, cumple un rol muy fundamental en el proceso de capacitación al agricultor o empresario que está vinculado en esta actividad.

En este sentido, es necesario implementar programas que permitan a los arroceros aplicar nuevas tecnologías para incrementar la calidad del grano y su productividad. Esta investigación busca utilizar técnicas ML con la finalidad de generar un modelo que permita clasificar el grano de arroz por sus características morfológicas. Tomando en cuenta esto, se plantea como pregunta de investigación:

¿Cuál será el algoritmo de aprendizaje automático más eficaz para realizar el proceso de clasificación de los granos de arroz?

1.3. Objetivo general

→ Aplicar las diferentes técnicas de aprendizaje automático en la clasificación de variedades de granos en arroz.

1.4. Objetivos específicos

→ Investigar de manera bibliográfica las diferentes teorías sobre la clasificación de arroz.

→ Establecer los modelos de aprendizaje automático más apropiados para dar solución al problema del proceso de clasificación de granos de arroz.

→ Evaluar el rendimiento de los modelos de aprendizaje automático propuesto para la clasificación de variedades de granos en arroz.

1.5. Justificación

El arroz es un producto muy rico en hidratos de carbono y almidón. Además, tiene una gran participación en la alimentación, por ser nutritivo y económico, y también es muy utilizado en el ámbito industrial.

La calidad del grano de arroz se determina habitualmente mediante inspecciones visuales y mediciones manuales, sin embargo, este método es lento, subjetivo y propenso a errores humanos (Krishna et al., 2022; Sultana et al., 2022).

Los instrumentos de laboratorio son costosos y pueden dificultar su aplicación entre las pequeñas empresas, especialmente en los países en desarrollo. Por lo tanto, es importante desarrollar un método alternativo que sea rápido, preciso, menos complicado y a bajo costo (Ndikuryayo et al., 2022).

Por esta razón, dentro de la revisión del marco teórico, se puede denotar la importancia de aplicar algoritmos de aprendizaje automático para la clasificación del grano de arroz de acuerdo sus características morfológicas e incluso durante la etapa del proceso productivo en el que se encuentra.

El aprendizaje automático (ML) es una de las principales áreas de investigación de la Inteligencia Artificial (IA), es una técnica de análisis auxiliar más prometedora para entender este complejo sistema, abriendo nuevos retos y oportunidades al mundo de la investigación. La principal aportación está relacionada con el análisis orientado a apoyar las decisiones (Greener et al., 2022; Rolnick et al., 2022).

El objetivo del ML es construir modelos matemáticos directamente a partir de muestras de datos sin instrucciones explícitas, ya que algunas tareas pueden ser fácilmente resueltas por los humanos, pero es difícil explicar explícitamente cómo las resuelven.

El ML llena este vacío dejando que los ordenadores aprendan automáticamente modelos de mapeo a partir de las muestras de datos, que pueden proyectar las muestras de datos a su salida deseada.

En los últimos años, en el mundo se vienen desarrollando algunos trabajos de investigación como las citadas en el marco teórico que aborda casos similares. Debido a los problemas antes mencionados, este documento pretende desarrollar una nueva solución para identificar los tipos de granos de arroz utilizando el enfoque basado en algoritmos ML. Esto permitirá un avance importante en materia de soporte inteligente a la toma de decisiones.

"Del laboratorio al campo: la inteligencia automatizada siembra progreso y cosecha precisión"

Capítulo II.

Marco Teórico Referencial

Capítulo II.

Marco Teórico Referencial

2.1. Conceptos generales

2.1.1. El cultivo de arroz

Uno de los cereales de mayor consumo y relevancia nacional e internacional es el arroz, este producto es base de la alimentación de más de la mitad de la población a nivel mundial (Chipana Valero et al. 2022). De acuerdo con Álvarez Hernández et al. (2018), este cereal suministra aproximadamente el 20% de la energía alimentaria. El arroz es fundamental para la seguridad mundial, su producción se lleva a cabo utilizando métodos de labranza convencionales y sistemas de riego por inundación (Fernández Rodríguez et al. 2021).

Esta planta tiene su origen en el sur de China, ya que se cultivaba a orillas del río Azul en el quinto milenio antes de cristo. Se clasifican en dos especies de arroz: *Oryza sativa L. y Oryza glaberrima S.,* esta última es poco cultivada debido a

la restricción que existe de su cultivo en la zona oeste de África (Álvarez et al. 2021). Corresponde a la familia de las Poaceae, es un cereal que sirve de alimento para el ser humano (López et al.).

De acuerdo con la FAO (2022), en la actualidad está previsto un incremento de la producción mundial de arroz para el año 2022-2023, la cual se espera que ascienda a 512,6 millones de toneladas de arroz elaborado, mientras que su utilización mundial será de 518,3 millones de toneladas, el descenso interanual será de 0,7% como consecuencia de las conmociones pronosticadas debido a su uso comercial y como pienso.

2.1.2. Producción de arroz en Ecuador

Dentro de la agricultura ecuatoriana el arroz constituye un rubro de alto valor, es la principal fuente de carbohidratos desde inicios del siglo XX, incluso en algunas zonas del Ecuador donde se produce la patata; la superficie total cultivada es de aproximadamente 340.000 ha, de estas, el 41% utiliza sistema de riego, y el 45% de explotaciones no superan las 5 hectáreas (Murillo et al. 2022).

De acuerdo con el Instituto Nacional de Estadísticas y Censos (2021), para el año 2021 en Ecuador existían 342.967 hectáreas (Ha) de arroz sembradas, de las cuales se obtuvo una producción de 1.504.214 toneladas métricas. El mayor volumen de producción se encuentra en la región costa 323.230 Ha, las cuales se distribuyen

principalmente en las provincias de Guayas con 204.874 Ha y Los Ríos con 104.165 ha.

En Ecuador se siembran principalmente las variedades INIAP 14 (33,7%), INIAP 11 (10,4%) e INIAP 15 (4,7%), producidas por el Instituto Nacional de Investigaciones Agropecuarias (INIAP), además de otras distribuidas por la empresa Procesadora Nacional de Alimentos (PRONACA), entre estas se encuentran las variedades SFL 09 (29,6%) y SFL011 (7%). Otras variedades presentes en el mercado ecuatoriano son: Yuma, Conquistador y San Juan, esta última proviene de la empresa colombiana INTEROC S.A (Zambrano et al. 2019).

2.1.3. Etapas del proceso de producción de arroz

La producción de arroz se inicia con la siembra, cuidado, irrigación y control de plagas, el grano está listo para ser cosechado cuando presenta una humedad interna de aproximadamente el 25%, una vez cosechado es transportado a los molinos donde se realiza su procesamiento, el mismo que consta de las siguientes etapas: secado, limpieza, descascarado, blanqueo y clasificación, de esta manera es posible conseguir un grano listo para su comercialización (Hernández-Cuello et al., 2021).

"La mayor parte de la producción de arroz en Ecuador proviene de Guayas y Los Ríos"

Dos de las etapas más importantes del proceso son la limpieza y la clasificación. La limpieza consiste en separar los granos de arroz de las sustancias extrañas, mientras que durante la clasificación se

separa los quebrados de los resistentes, también se los puede clasificar por el color, en manchados y rayados (Cinar and Koklu 2019).

2.1.4. Clasificación de granos de arroz

Parte importante del proceso es el control y clasificación de granos, esta es realizada generalmente por un perito, es una actividad tediosa y repetitiva, por lo que existen grandes posibilidades de cometer errores. Usualmente se mide la longitud de los granos por medio de un calibre, esta medición se realiza tomando en cuenta una muestra representativa, los granos que resulten defectuosos serán extraídos de la muestra control (Acosta et al. 2017).

"El proceso de producción incluye preparación del terreno, siembra y cosecha."

La clasificación manual de granos de arroz, resulta un proceso costoso y demanda de mucho tiempo, está limitada por la experiencia que tengan los evaluadores, es por esto que durante los últimos años se ha recurrido a diferentes alternativas que permitan valorar la clasificación y calidad del arroz. Estos incluyen algunos parámetros geométricos como la longitud o el perímetro, la tasa de fractura, la blancura, o las grietas que pueda presentar el grano (Koklu et al. 2021).

El constante incremento de la demanda de arroz crea la necesidad de producir y clasificar los granos con mayor rapidez, por lo que en la actualidad se están creando alternativas que permitan realizar el proceso de manera más rápida y eficiente, que permitan la obtención de

resultados precisos, una de las alternativas utilizadas actualmente es el aprendizaje automático (Ibrahim et al. 2019).

2.1.5. Escalas de clasificación del arroz

De acuerdo con el Servicio Ecuatoriano de Normalización (INEN), se utiliza la siguiente escala para el tamaño del arroz:

Tabla 1.

Clasificación del arroz de acuerdo al tamaño

Clase 1.	**Extra largo**	Granos con longitud mínima de 7,0 mm. Se tolera máximo el 20% de mezcla de otros granos largos
Clase 2.	**Largo**	Granos con longitud entre 6,0 mm y 6,99 mm. Se tolera máximo el 20% de mezcla de otros granos medios
Clase 3.	**Medio**	Granos con longitud entre 5,0 mm y 5,99 mm. Se tolera máximo el 10% de otros granos cortos
Clase 4.	**Corto**	Granos con longitud menor de 5,0 mm
Clase 5.	**Mezcla**	Granos mezclados de dos clases o más de las clases mencionadas

2.1.6. Determinación del grado

Con la finalidad de determinar el grado se puede tomar en cuenta las recomendaciones del Instituto Internacional de Investigación del Arroz (IIIA):

→ ***Granos rojos:*** consiste en granos de arroz enteros o quebrados, que presentan un color rojo nítido o estrías de color rojizo

en la cutícula (Internacional de Investigación del Arroz (2021).

→ ***Granos tizosos:*** estos son granos enteros o quebrados en los cuales se evidencia el proceso de entizamiento, características tizosas o harinosas, total o parcialmente sobre la extensión del grano (Internacional de Investigación del Arroz (2021).

"El grado del grano se evalúa por su estado físico y proporción de granos partidos."

→ ***Granos tizosos totales:*** son aquellos granos enteros o quebrados en los que es posible ver en más de la extensión de un grano entero el proceso de entizamiento. Esta denominación también incluye a los granos inmaduros (Internacional de Investigación del Arroz (2021).

→ ***Granos tizosos parciales:*** son granos enteros o quebrados en los que se puede observar el proceso de entizamiento, solamente sobre algunos sectores del grano, no alcanzando la mitad del grano. En esta denominación se incluyen aquellos granos conocidos como panza blanca (Internacional de Investigación del Arroz (2021).

→ ***Granos dañados:*** son todos aquellos granos de arroz, enteros o quebrados, que tienen alteraciones procedentes de hongos, fermentaciones, heladas, calentamiento u otros orígenes (Internacional de Investigación del Arroz (2021).

→ ***Materia extraña:*** consiste en todo aquel material que no sea arroz pilado, incluso el arroz no descascarado o paddy (Internacional de Investigación del Arroz (2021).

→ ***Granos quebrados:*** son granos que no pasan de entre ¼ y ¾ del tamaño total del grano entero (Internacional de Investigación del Arroz (2021).

2.2. Aprendizaje Automático

El concepto de aprendizaje automático fue propuesto por primera vez por Arthur Samuel en 1959 (Zhou, 2021). Por aquel entonces, la investigación en aprendizaje automático se centraba principalmente en modelos estadísticos simples en el campo de los juegos de ordenador (Alpaydin, 2021). Tras varias décadas de desarrollo, Tom M. Mitchell propuso en 1997 una definición más formal del aprendizaje automático, que lo define desde la perspectiva de la mejora automática del rendimiento de los modelos matemáticos a partir de experiencias (Bonaccorso, 2017).

A lo largo de la historia se han desarrollado muchas técnicas de aprendizaje automático. Algunos hitos dignos de mención son la retropropagación de redes neuronales desarrollada en la década de 1970, la máquina de vectores de soporte desarrollada en la década de 1990, y el aprendizaje profundo desarrollado en la década de 2000 (Molnar, 2020; Wang et al., 2016).

El ML puede clasificarse en tres paradigmas principales (Alpaydin, 2021): aprendizaje supervisado, aprendizaje no supervisado y aprendizaje por refuerzo. En el aprendizaje supervisado, cada muestra de datos se empareja sistemáticamente con una etiqueta determinada.

Un modelo toma las muestras de datos como entrada y aprende a realizar predicciones lo más cercanas posible a sus etiquetas correspondientes, lo que significa que el proceso de aprendizaje está "supervisado" por las etiquetas reales (Zhou, 2021). Por el contrario, las muestras de datos no tienen ninguna etiqueta en el aprendizaje no supervisado. El modelo de mapeo intenta desvelar los puntos comunes subyacentes en las muestras de datos.

El aprendizaje por refuerzo se sitúa entre el aprendizaje supervisado y el no supervisado. En el aprendizaje por refuerzo, las muestras de datos no tienen etiquetas directas. En su lugar, se da una medida para cada acción, conocida como recompensa (Bonaccorso, 2017). El aprendizaje automático (ML) se ha convertido en una valiosa herramienta para extraer información versátil de datos complejos mediante tareas de regresión o clasificación en la agricultura (Jagtap et al., 2022). El ML puede describirse como un área de la informática que estudia algoritmos y técnicas para automatizar soluciones a problemas desafiantes con métodos de programación tradicionales (Elbadawi et al., 2021).

Un algoritmo de ML tiene como objetivo aprender un modelo o un conjunto de reglas a partir de un conjunto de datos etiquetados para que los datos indiquen etiquetas en el otro conjunto de datos que puedan predecir correctamente (Chen et al., 2021).

2.2.1. Aprendizaje automático en la agricultura

Kiratiratanapruk et al. (2020), en su investigación utilizaron cuatro técnicas de aprendizaje automático estadístico (LR, LDA, k-NN y SVM) y cinco modelos preentrenados (VGG16, VGG19, Xception,

InceptionV3 e InceptionResNetV2), con la finalidad de comparar el rendimiento de la clasificación de granos de arroz, para esto se clasificó las muestras en grupos y subgrupos colectivos. El mejor nivel de precisión fue obtenido por medio del método SVM al 90,61%, 82,71% y 83,9% en los subgrupos 1 y 2 y el grupo colectivo; en lo que respecta a las técnicas de aprendizaje profundo, la mejor precisión se obtuvo con los modelos InceptionResNetV2, esta fue del 95,15%.

Un asunto importante para el almacenamiento, empaque y transporte de granos es el contenido de humedad, es por esto que Liu et al. (2022), en su investigación desarrollaron un dispositivo portátil conformado por tres partes: un módulo de circuito de microondas, un módulo de cálculo en tiempo real y un software para exponer los resultados, con la finalidad de medir la humedad del arroz con cáscara por medio de sensores microstrip de microondas asistido por estrategias de aprendizaje automático. Entre los modelos de predicción utilizados, el que presentó mejor rendimiento, precisión y estabilidad fue el modelo de bosque aleatorio (R2 = 0,99, RMSE = 0,28, MAE = 0,26), su rendimiento fue relativamente estable, con un error absoluto medio máximo del 0,55%.

Un indicador clave para tener una cosecha de arroz apropiada es el contenido de humedad de los granos, tomando en cuenta esto, Yang et al. (2021), evaluaron el nivel de humedad de la cosecha de granos por medio del aprendizaje automático en teléfonos inteligentes para establecer el momento adecuado para la cosecha, las imágenes de panículas individuales fueron tomadas mediante el uso de teléfonos

inteligentes y fueron corregidos con una placa de corrección espectral-geométrica. Se construyeron cuatro modelos de aprendizaje automático, estos incluyeron bosque aleatorio, perceptrón multicapa, regresión de vector de soporte (SVR) y regresión lineal multivariada, de estos, el más adecuado para medir el grado de humedad fue el modelo SVR con un error absoluto medio de 1.23%.

Dheer et al. (2019), en su investigación recurren al uso de modelos de aprendizaje automático con la finalidad de diseñar un sistema de inspección de variedades de arroz, para esto se tomó en cuenta modelos clasificadores simples como Análisis Discriminante Lineal, Regresión Logística, K-Nearest Neighbor's (KNN) y método Naïve-Bayes. El método que presentó un mayor nivel de exactitud fue KNN, con un 99,16% de exactitud, 9,12% de precisión, y 99,12% de recuperación.

La clasificación de los granos de arroz es una dificultad que debe superarse, para esto, en muchos casos se está recurriendo a los sistemas inteligentes y la automatización, Arora et al. (2020), en su investigación utilizaron técnicas de procesamiento de imágenes y aprendizaje automático para este fin; como resultado se obtuvo que el sistema puede capturar varios parámetros de manera exitosa, tomando en cuenta para esto las imágenes de muestra de los diferentes tipos de arroz, los cuales fueron almacenados en un archivo CSV para finalmente ser procesados, el sistema también evalúa el volumen de granos de arroz e implementa varios algoritmos para el procesamiento de imágenes y de aprendizaje automático.

2.2.2. Clases de aprendizaje automático supervisado

El método de aprendizaje automático supervisado se divide principalmente en dos categorías:

- → Regresión o capacidad de predecir valores continuos.
- → Clasificación que es la categorización de valores categóricos.

El proceso de clasificación comienza con el entrenamiento del modelo utilizando datos de entrenamiento para identificar el tipo o clase de la característica introducida (Ndikuryayo et al., 2022). Un clasificador binario etiqueta los datos como pertenecientes a uno de los dos grupos de salida. Esta tesis de maestría se centrará en utilizar algoritmos de aprendizaje supervisado puede explicarse cómo sigue.

2.2.3. Algoritmos de aprendizaje supervisado

Los modelos de clasificación son un método de gran importancia utilizado en diversos campos. En la determinación de clases, los modelos de clasificación se utilizan para determinar a qué clase pertenecen los datos. El modelo de clasificación es un modelo que funciona haciendo predicciones (Waleed et al., 2021). El propósito de la clasificación es hacer uso de las características comunes de los datos para analizar los datos en cuestión. En nuestro estudio, se crearon modelos utilizando sistemas LR (Regresión Logística), MLP (Perceptrón Multicapa), SVM (Máquina de Vectores de Apoyo), RF (Bosque Aleatorio) y k- NN (K

Vecino Más Cercano) para clasificar los granos de arroz según sus características.

2.2.4. K-Nearest Neighbor (k-NN)

El método k-NN es un algoritmo de aprendizaje no paramétrico. k-NN utiliza la distancia euclidiana como parámetro en nombre de la clasificación del conjunto de datos, donde K representa el número de vecinos, para calcular la distancia entre los datos (Pavani & Augusta Sophy Beulet, 2022).

La k-NN está pensada para clasificar datos de muestra cuya clase es desconocida. Por esta razón, la distancia a los datos de muestra se calcula con el conjunto de datos preclasificados en el conjunto de entrenamiento. Dado que hay una cierta cantidad de datos a probar, los datos de prueba se procesan con todos los datos existentes individualmente (Rekha Sundari et al., 2021). Los datos de prueba tendrán muchos vecinos que están cerca de ellos en términos de todas las medidas.

El algoritmo tiene algunas ventajas y desventajas. Una de las ventajas más importantes del algoritmo k-NN es la eficiencia y la otra es la flexibilidad. El algoritmo es eficiente en términos de simplicidad, velocidad y escalabilidad (Jagtap et al., 2022).

La flexibilidad del algoritmo facilita la gestión de conjuntos de datos que no pueden explicarse mediante relaciones lineales o no lineales por la complejidad de las relaciones. La desventaja más importante

del algoritmo k-NN es que, a diferencia de otros algoritmos de minería de datos, no da una idea de qué variables son importantes en las nuevas predicciones (Malik et al., 2021).

2.2.5. Máquina Soporte Vectorial (SVM)

En los modelos SVM se utilizan diferentes funciones de núcleo. En este estudio, la clasificación se realizó utilizando la función de núcleo polinómico.

El objetivo del algoritmo SVM es encontrar un hiperplano que separe el conjunto de datos en un número predefinido y discreto de clases que sean las más consistentes con los ejemplos de entrenamiento (Kok et al., 2021). El término separación óptima del hiperplano se refiere a la frontera de decisión que minimiza las clasificaciones erróneas durante el paso de entrenamiento.


"La máquina de soporte vectorial encuentra límites óptimos entre clases"

La SVM tiene la capacidad de clasificar datos en forma de lineal en el espacio bidimensional, planar en el espacio tridimensional e hiperplano en el espacio multidimensional con mecanismos de separación (Abdullah & Abdulazeez, 2021). La SVM realiza el proceso de clasificación encontrando el mejor hiperplano que separa los datos pertenecientes a las clases.

Las SVM tienen características similares a otros algoritmos de clasificación. Es especialmente similar a las redes neuronales, pero más parecido al algoritmo K-NN. Al igual que el algoritmo K- NN, la SVM determina sus vecinos basándose en los datos de muestra que

se le presentan al algoritmo y asume que las estimaciones se realizan para los nuevos datos (Banerjee & Madhumathy, 2022).

2.2.6. Bosque aleatorio (RF)

RF es un clasificador compuesto por múltiples DT. Para realizar una nueva clasificación, cada DT proporciona una clasificación para las entradas. Después, RF evalúa las clasificaciones y selecciona la estimación que tiene más votos. La RF tiene la capacidad de gestionar un gran número de variables en un conjunto de datos. También tiene bastante éxito en la predicción de datos incompletos (Rakhra et al., 2021). El mayor inconveniente de la RF es su falta de repetibilidad. Además, el modelo final y los resultados posteriores son difíciles de interpretar.

Los árboles se generan extrayendo un subconjunto de muestras de ejercicio (un proceso de embolsado). Aproximadamente dos tercios de las muestras se utilizan para desarrollar un modelo. En cambio, el tercio restante de las muestras se utiliza para estimar el rendimiento del modelo mediante la votación por mayoría (Suresh et al., 2021). El algoritmo produce árboles con altas varianzas y bajas tendencias, aumentando el bosque hasta un número de árboles definido por el usuario. El promedio de las probabilidades de asignación de clase determinadas por todos los árboles generados es la decisión final de clasificación.

2.2.7. Regresión logística (LR)

El LR es uno de los modelos estadísticos más utilizados. En el LR, la variable dependiente se estima a partir de una o más variables. El LR aclara la relación entre las variables dependientes y las independientes con el menor número de variables (Arumugam et al., 2022).

En el LR no es necesario crear una distribución normal de las variables. Dado que los valores previstos en la LR son probabilidades, la LR se limita a 0 y 1. Esto se debe a que la LR predice su probabilidad, no a sí misma, en los resultados (Arumugam et al., 2022). En este estudio, se utilizó el método Newton para la optimización durante la clasificación con la ayuda de LR.

2.2.8. Perceptrón multicapa (MLP)

En la actualidad, se han desarrollado muchos modelos de redes neuronales artificiales para su uso con fines específicos, y el MLP es uno de los más utilizados de estos modelos (Bakthavatchalam et al., 2022). En MLP, la secuencia de neuronas está

en capas, y hay una capa oculta entre ellas, junto con dos capas principales.

El MLP puede contener más de una capa oculta. La capa de entrada, que es la primera de las capas principales, es la capa donde se leen los datos y contiene información sobre el problema que hay que resolver. La capa de salida, que es la segunda capa principal, es la capa donde se definen las clases y se reciben las salidas de la información procesada en la red. La capa oculta es la capa donde se realizan las operaciones intermedias sobre los datos entre las capas principales (P. Xu et al., 2021).

El MLP tiene tantas neuronas como el número de características, y los datos son proporcionados por un flujo de datos en una dirección desde la capa de entrada hasta la capa de salida. Además, es posible controlar y modificar la estructura de la red durante el periodo de entrenamiento (Goyal et al., 2022). En este estudio, hay 4 capas ocultas y también se utilizó la función de activación sigmoidea.

2.3. Validación cruzada

La validación cruzada es un método de predicción de errores desarrollado con el objetivo de mejorar la seguridad de la clasificación.

La validación cruzada divide el conjunto de datos de forma aleatoria en un número determinado de subconjuntos para el entrenamiento y la prueba. Acepta uno de los subconjuntos como conjunto de prueba y el sistema se entrena con los conjuntos restantes (Mourtzinis et

al., 2021). Este proceso se repite hasta el número de conjuntos de datos y se prueba el sistema. Debido a estas características del conjunto de datos, se consideró apropiado utilizar métodos de clasificación. Los algoritmos de clasificación entrenaron el modelo observando los patrones de los datos en el conjunto de entrenamiento (Kasinathan et al., 2021).

De este modo, clasificó los datos de forma que no se pudieran ver antes de una manera muy precisa. En este estudio, los núcleos de granos de arroz se modelaron utilizando clasificadores LR, MLP, SVM, RF y k-NN con la ayuda del lenguaje de programación R y Rstudio. Son métodos de aprendizaje automático que se utilizan frecuentemente en problemas de clasificación. Además de estos métodos, también se han probado otros métodos de aprendizaje automático y los métodos utilizados en el estudio han obtenido resultados de clasificación más significativos que otros.

2.4. Medidas de desempeño

La creación de un nuevo modelo necesario para los problemas de clasificación o el uso de modelos existentes y el éxito de este modelo se calculó por el número de estimaciones precisas (Nosratabadi et al., 2021). Esto es efectivo en la precisión de la clasificación más que en la estimación de si el modelo es bueno o no. Por eso se utiliza la matriz de confusión para explicar las evaluaciones predictivas de la clasificación.

Por lo tanto, la matriz de confusión se utiliza para explicar las evaluaciones predictivas de la clasificación. La matriz que proporciona información sobre las clases reales con las clases estimadas realizadas a través de un modelo de clasificación basado en los datos de la prueba es la matriz de confusión (Koklu et al., 2021). La Matriz de confusión tiene cuatro parámetros, VP (verdaderos positivos) y VN (verdaderos negativos) representan respectivamente el número de ejemplos positivos y negativos clasificados correctamente, mientras que FP (falsos positivos) y FN (falsos negativos) representan respectivamente el número de ejemplos positivos y negativos clasificados incorrectamente (Mena, 2008).

Capítulo III.

Delimitación Técnica Metodológica

Capítulo III.

Delimitación Técnica Metodológica

3.1. Tipo y diseño de investigación

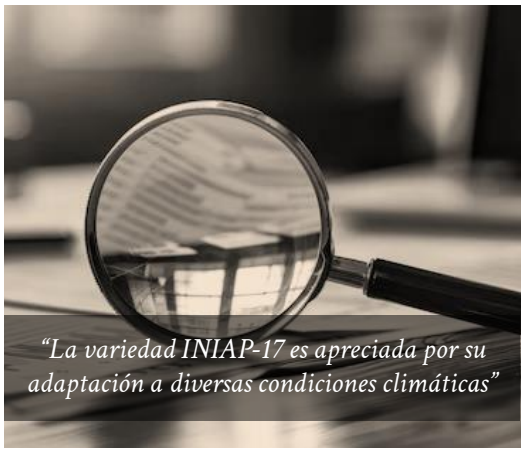
"La comparación de algoritmos descubre diferencias clave en la clasificación de datos"

Este estudio se enmarca en un diseño de investigación cuantitativa. Acorde a Ochoa et al. (2020) este diseño de investigación consiste en la recolección de datos numéricos para su posterior análisis e interpretación utilizando herramientas de análisis matemático y estadístico para describir y explicar fenómenos a través de datos numéricos. El tipo de la investigación es causal comparativa dado que se comparan diferentes técnicas de aprendizaje automático en el proceso de clasificación de granos de arroz. Según Sánchez et al. (2021) este tipo de investigación destaca porque depende de factores de comparación y deja al investigador la decisión de la selección de características que se desean comparar. En el caso de este estudio se comparan diferentes medidas de

rendimiento en cinco tipos de algoritmos para la correcta clasificación de dos diferentes variedades de arroz.

3.2. La población y muestra

3.2.1. Características de la población



Entre los arroces certificados que se cultivan en el Ecuador, se han seleccionado para el presente estudio, la especie INIAP-14, que cuenta con una gran superficie de plantación desde 1.999, y la especie INIAP-17, cultivada desde 2.007. Si se observan las características generales de la variedad INIAP14, tienen un aspecto ancho, largo, vidrioso y opaco. El peso del grano de mil es de 23-26 gr. Al observar las características generales de la especie INIAP-17, tienen un aspecto ancho y largo, vidrioso y opaco. El peso de mil de grano es de 22-28 gr.

En este estudio, la distribución de 3.260 granos de arroz obtenida como resultado del procesamiento de las imágenes de ambas especies.

3.2.2. Delimitación de la población

En el estudio se utilizó un conjunto de datos pertenecientes a las dos variedades de arroz (INIAP 14 e INIAP 17), que suelen cultivarse en la costa ecuatoriana. El conjunto de datos constó de 3.260 de granos

de arroz (1.630 para INIAP-14 y 1.630 para INIAP 17) que fueron recolectados en campo, los mismo fueron transportados a laboratorio. Donde la cámara utilizada para el estudio tiene 2,2 megapíxeles, una resolución de 2048×1088 y resolución completa a una frecuencia de imagen máxima de 53,7 fps. Dispone de funciones como el balance de blancos y la corrección de contraluz.

La cámara utilizada en el estudio se colocó en una caja cerrada con un dispositivo de iluminación en su interior y una estructura para evitar que la luz del entorno exterior. El color de fondo de la caja se seleccionó como negro para facilitar el procesamiento de la imagen. Los tamaños se diseñaron de forma que las imágenes pudieran captarse desde un área de 14 cm de ancho y 18 cm de largo. La altura de la cámara se fijó a 15 cm. Las imágenes resultantes se grabaron transfiriéndolas al ordenador.

En este estudio no fue necesario calcular la muestra, ya que se trabajó con toda la población para determinar que método de aprendizaje automático da mejores resultados en cuanto a la clasificación de arroz basado en las medidas de rendimiento (Dominguez-Lara y Merino-Soto, 2018). Se evaluaron en total siete variables morfológicas descritas en la Tabla 2.

Tabla 2.

Variables morfológicas para evaluar

No	*Nombre*	*Explicación*
1	*Área (A)*	Devuelve el número de píxeles dentro de los límites del grano de arroz.
2	*Perímetro (P)*	Calcula la circunferencia mediante el cálculo de la distancia entre píxeles alrededor de los límites del grano de arroz.

3	*Longitud del eje mayor (L)*	La línea más larga que se puede dibujar en el grano de arroz, es decir, la distancia del eje principal.
4	*Longitud del eje menor (I)*	La línea más larga de un grano de arroz que se puede trazar perpendicularmente al eje mayor.
5	*Excentricidad*	Mide la redondez de la elipse, que tiene los mismos momentos que el grano de arroz.
6	*ConvexArea (CA)*	Devuelve el recuento de píxeles de la cáscara convexa más pequeña de la región formada por el grano de arroz.
7	*Extensión (EX)*	Devuelve la proporción de la región formada por el grano de arroz a los píxeles de la caja delimitadora.

Fuente: tomado del trabajo de Ozkan & Koklu (2017).


"La cámara de alta resolución captura detalles esenciales para el procesamiento de imágenes"

Dado que el conjunto de datos es irregular, es decir, no hay el mismo número de datos pertenecientes a cada variedad de arroz, y el éxito de la clasificación no es alto, por lo que se vuelve necesario utilizar métricas adicionales como la exhaustividad, y la exactitud, respecto a otros estudios en los que los datos han sido regulares y han utilizado otro tipo de métricas (sensibilidad) (Koklu et al., 2021).

Las métricas utilizadas van de 0 a 1. Debido a que el éxito de clasificación de los modelos utilizados podría no ser alto, no se puede hacer una comparación cuando se redondean estos valores. Por ello, los valores de estas medidas se muestran como porcentajes.

Posteriormente, se realizó la validación cruzada que es un método de predicción de errores desarrollado con el objetivo de mejorar la seguridad de la clasificación. La validación cruzada es un método utilizado para medir objetivamente la precisión de los modelos de clasificación. En este método, el conjunto de datos se divide en el mismo número de partes según el valor numérico especificado.

El valor numérico especificado se denomina k. 1 / k; parte del conjunto de datos se reserva para la prueba, k-1 y parte se reserva para el entrenamiento del modelo (training). (Koklu et al., 2021). Este proceso continúa hasta que cada parte del conjunto de datos se utiliza como parte de prueba. Este proceso se repite k veces. El éxito general de clasificación del modelo en el conjunto de pruebas se obtiene tomando la media aritmética de los éxitos de clasificación obtenidos como resultado de estas operaciones.

La validación cruzada dividió el conjunto de datos de forma aleatoria en un número determinado de subconjuntos para realizar el entrenamiento y prueba de los datos. Este proceso se repitió hasta terminar con la población de conjuntos de datos y poner a prueba los modelos.

Finalmente, para la evaluación de los modelos, se aplicó el estadístico Kappa, para determinar qué modelo es el mejor a través de la

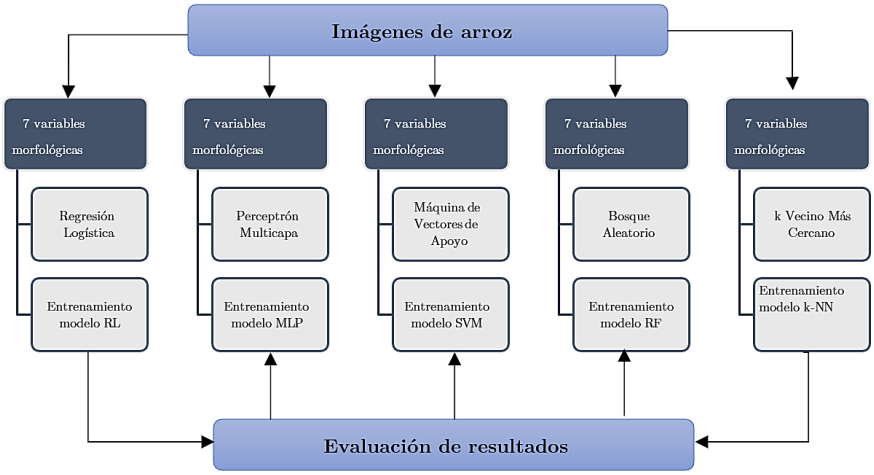
interpretación de los rangos de valores que van desde -1 a 1, interpretándose valores cercanos a -1 como el peor modelo, y valores cercanos a +1 como el mejor modelo.

3.3. Los métodos y las técnicas

Las técnicas que fueron empleadas están basadas en cinco métodos o modelos de aprendizaje automático llamados también algoritmos. Estos modelos son Regresión Logística (LR), Perceptrón Multicapa (MLP), Máquina de Vectores de Apoyo (SVM), Bosque Aleatorio (RF) y k Vecino Más Cercano (k- NN). Estos modelos se compararon a través de las medidas de rendimiento para determinar el mejor modelo para clasificar los granos de arroz. En la Figura 1, se muestra el diagrama del modelo propuesto para la clasificación de las dos variedades de arroz.

Figura 1.

Diagrama del modelo propuesto



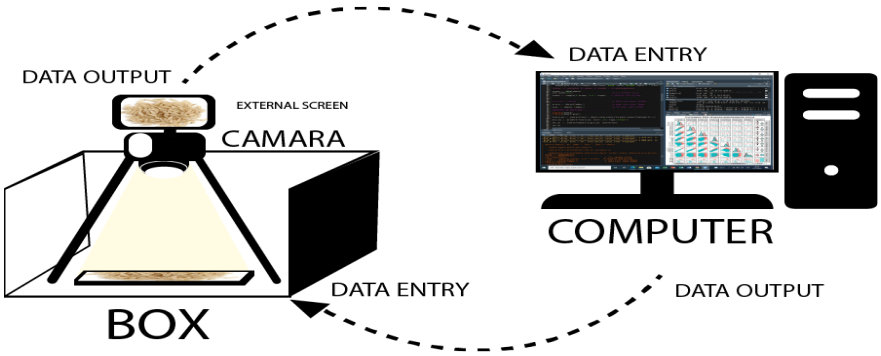
3.4. Procesamiento estadístico de la información

Una vez tomadas en laboratorio las fotografías de los 3.260 granos de arroz de las dos variedades, se procedió a calcular las características morfológicas de forma del grano. En esta sección se explican las operaciones preliminares sobre las imágenes para obtener procesos de extracción y clasificación de características de la forma más precisa. El procesamiento de imágenes es fundamental, ya que afecta directamente al resultado de la inferencia y clasificación de características. Por este motivo, se tuvieron en cuenta a la hora de diseñar la fase de procesamiento de la imagen. El procesamiento de la imagen se ha llevado a cabo con ayuda del lenguaje de programación R y Rstudio, y sus librerías. Las imágenes tomadas de la cámara se han convertido a escala de grises e imágenes binarias para prepararlas para la inferencia de características morfológicas. En total se calcularon siete variables morfológicas que fueron el área, perímetro, longitud del eje mayor y menor, excentricidad, área convexa, y extensión.

Para explorar estas características, se realizó un análisis estadístico descriptivo exhaustivo (Figura 2). Se calcularon medidas de tendencia central y dispersión, como la media aritmética, desviación estándar, varianza, error estándar, coeficiente de variación, asimetría y curtosis, proporcionando una comprensión profunda de la distribución y comportamiento de las variables. Posteriormente, los modelos se entrenaron y evaluaron utilizando una matriz de confusión como base para calcular métricas predictivas, como la especificidad, precisión, exactitud y el coeficiente Kappa. Este enfoque permitió cuantificar la eficacia de cada modelo en un escenario de datos desbalanceados, ya que las dos variedades de arroz presentaban tamaños de muestra desiguales. La integración de estas métricas y la robustez de los modelos permitió abordar este desafío y garantizar resultados confiables y aplicables a futuras investigaciones y aplicaciones agrícolas.

Figura 2.

El esquema de adquisición de imágenes



Esta matriz proporcionó información sobre las clases reales con las clases estimadas realizadas a través de un modelo de clasificación basado en los datos de la prueba es la matriz de confusión. La Matriz de confusión tiene cuatro parámetros; vp: verdadero positivo, fp: falso positivo, fn: falso negativo, y vn: verdadero negativo. Para las mediciones del rendimiento de la clasificación de las dos variedades de arroz, se utilizaron criterios de éxito como la exactitud, precisión, exhaustividad, especificidad y puntuación.

La matriz de confusión se utiliza para medir el rendimiento de clasificación de los métodos de aprendizaje automático. Esta matriz facilita la búsqueda de conexiones entre el rendimiento del clasificador y los resultados de las pruebas. La matriz de confusión proporciona información sobre la clasificación correcta e incorrecta de las muestras positivas y la clasificación correcta e incorrecta de las muestras negativas. En la Tabla 3 se muestra una matriz de confusión de dos clases.

Tabla 3.

Matriz de confusión para dos clases

		Clase estimada	
		Positivo	*Negativo*
Clase actual	*Positivo*	Verdadero positivo (VP)	Falso negativo (FN)
	Negativo	Falso positivo (FP)	Verdadero negativo (VN)

Fuente: tomado del trabajo de Martínez (2018).

El conjunto de datos sobre el arroz de este estudio consta de dos clases. Por este motivo, se utilizó la matriz de confusión de dos clases en los procesos de clasificación. A partir de los valores VP, FP, FN y VN de la matriz de confusión, se realizan cálculos estadísticos y se puede analizar en detalle el rendimiento de los clasificadores (Martínez et al. 2018). Las métricas obtenidas de los cálculos estadísticos para la matriz de confusión de dos clases, las fórmulas utilizadas para calcular estas métricas e información sobre el propósito para el que se utilizan las métricas se muestran en la Tabla 4.

Tabla 4.

Medidas de rendimiento y fórmulas de cálculo para la clasificación de dos clases

N°	Nombre	Formula	Explicación
1	Exactitud	$\frac{tp+tn}{tp+fp+tn+fn}$ X 100	Se utiliza para medir la proporción de predicción verdadera en todas las muestras incluidas en la evaluación.
2	Precisión	$\frac{tp}{tp+fp}$ X 100	Se utiliza para medir la relación entre las muestras positivas clasificadas con precisión y muestras positivas clasificadas con precisión y las muestras positivas totales estimadas.
3	Exhaustividad	$\frac{tp}{tp+fn}$ X 100	Se utiliza para medir la proporción de valores positivos correctamente clasificados.
4	Especificidad	$\frac{tp}{tp+fn}$ X 100	Se utiliza para medir la proporción de valores negativos correctamente clasificados.
5	Puntuación F1	$\frac{2tp}{2tp+fp+fn}$ X 100	Representa la media armónica entre los valores de Exhaustividad y Precisión.

Fuente: tomado del trabajo de Ozkan & Koklu (2017).

Capítulo IV.

Desarrollo Descriptivo e Inferencial

Capítulo IV.

Desarrollo Descriptivo e Inferencial

4.1. Estadística descriptiva

De un total de 3.260 de granos de arroz pertenecientes a las dos variedades de arroz, se evaluaron 7 características morfológicas para cada grano. En la tabla 5, se muestra los resultados de la estadística descriptiva para las dos variedades (INIAP- 14 y INIAP 17) en el cual se calculó para cada característica (área, perímetro, eje mayor y menor, entre otras) la media aritmética, desviación estándar, varianza, error estándar, coeficiente de variación, asimetría y kurtosis.

Tabla 5.

Estadística descriptiva de las características morfológicas (INIAP-14:n=1.630; INIAP-17: n=1.630)

	Área		Perímetro		Eje mayor		Eje menor	
	INIAP-14	INIAP-17	INIAP-14	INIAP-17	INIAP-14	INIAP-17	INIAP-14	INIAP-17
Media	14164,76	11550,88	487,48	429,43	205,49	176,29	88,78	84,48
D.E.	1288,59	1041,69	22,23	20,14	10,36	9,36	5,35	5,3
Var(n-1)	1660462,98	1085125,3	494,07	405,81	107,36	87,57	28,59	28,11
E.E.	31,85	22,3	0,55	0,43	0,26	0,2	0,13	0,11
CV	9,1	9,02	4,56	4,69	5,04	5,31	6,02	6,28
Asimetría	0,01	-0,06	-0,16	0,04	-0,07	0,28	-0,01	-0,35
Kurtosis	-0,04	0,35	-0,1	0,23	0,05	0,15	0,45	0,59

	Excentricidad		Área convexa		Extensión	
	INIAP- 14	INIAP- 17	INIAP- 14	INIAP- 17	INIAP- 14	INIAP- 17
n	2180	1630				
Media	0,9	0,88	14496,37	11800,56	0,65	0,67
D.E.	0,01	0,02	1311,04	1062,48	0,08	0,07
Var(n-1)	1,80E-04	3,60E-04	1718837,17	1128859,99	0,01	0,01
E.E.	3,30E-04	4,10E-04	32,4	22,74	2,00E-03	1,50E-03
CV	1,48	2,17	9,04	9	12,61	10,8
Asimetría	-0,51	-0,19	-0,02	-0,05	0,42	0,38
Kurtosis	1,14	0,52	-0,05	0,38	-1,05	-1,05

Se puede observar en la Tabla 5, que la variedad INIAP-17 tiene una menor área promedio por grano respecto a la variedad INIAP-14, es decir que su grano es más grande, además, posee un mayor perímetro, lo que significa que compensa teniendo un mayor borde del grano. La variedad INIAP-14 también posee una longitud del eje mayor y menor más grandes que INIAP-17 lo que representa que los granos son más alargados y anchos. Esto a su vez significa que INIAP-14 tendría mayor cantidad de almidón en su composición, lo que requeriría una mayor cantidad de agua para su cocción. En cuanto a la variabilidad de los datos se agrega que tomando como referencia el coeficiente de variabilidad relativa (Var. coef.) se puede decir que hay una variación inherente a cada una de las características que se considera como igual estadísticamente, por lo tanto, los dos tipos de grano de arroz son comparables entre si a través de sus promedios en cada una de las características. Además, agregando el componente de la simetría en sus distribuciones cercanos a cero, se habla de una relativa normalidad. Con este último resultado se pueden utilizar pruebas paramétricas para evaluar formalmente la igualdad o desigualdad de estas características entre ambas especies de arroz.

"El coeficiente de variabilidad confirma la comparabilidad entre las dos variedades"

Los estadísticos descriptivos proporcionaron una base sólida para la comprensión inicial de los datos antes de aplicar técnicas de aprendizaje automático. Permitieron comprender la distribución de los datos, identificación de valores atípicos, evaluación de tendencias centrales, y caracterización de la variabilidad que son esenciales para el éxito de la clasificación cualitativa del grano de arroz mediante métodos de aprendizaje automático. Al realizar la prueba de T de Student para demostrar la igualdad de cada característica en cada

tipo de grano, se pudo constatar que si existen diferencias significativas entre las distintas características morfológicas de ambas variedades (p<0,001), para todas las características de ambas variedades de arroz, tal como lo muestra la Tabla 5.

De acuerdo con la problemática abordada, los métodos de análisis y predicción utilizados en el presente trabajo, este resultado de desigualdad en las características entre las variedades de arroz resulta beneficioso. Dado que en un problema de clasificación resulta fundamental tener una separación clara entre características que van a permitir la realización de la clasificación de las especies de grano, un resultado demostrado formalmente como estadísticamente significativo aporta un punto de partida firme y sustancial sobre una clasificación adecuada. Al tener una separación clara entre las características se pueden mitigar de mejor forma posibles fallos durante el proceso de entrenamiento y clasificación ya que existe un umbral evidente entre las variedades de grano de arroz.

Tabla 6.

Prueba de T de Student a dos colas

Variable	Variedad 1	Variedad 2	n1	n2	Media 1	Media 2	T	p-valor
Área	INIAP-14	INIAP-17	1.630	1.630	14.164	11.550	67,2	<0,0001
Perímetro	INIAP-14	INIAP-17	1.630	1.630	487,4	429,43	83,1	<0,0001
Longitud eje mayor	INIAP-14	INIAP-17	1.630	1.630	205,4	176,29	89,8	<0,0001
Longitud eje menor	INIAP-14	INIAP-17	1.630	1.630	88,78	84,48	24,6	<0,0001
Excentricidad	INIAP-14	INIAP-17	1.630	1.630	0,90	0,88	47,2	<0,0001
Área convexa	INIAP-14	INIAP-17	1.630	1.630	14496,3	11800,5	68,1	<0,0001
Extensión	INIAP-14	INIAP-17	1.630	1.630	0,65	0,67	-7,24	<0,0001

Una vez realizada la estadística descriptiva, se aplicaron los diferentes modelos de aprendizaje automático partiendo de la matriz de

confusión para explicar las evaluaciones predictivas de la clasificación del grano de arroz. Esta matriz proporcionó información sobre las clases reales con las clases predichas realizadas a través de un modelo de clasificación basado en los datos de la partición reservada para prueba. La capacidad de cada uno de los modelos para realizar predicciones correctas se analizó a partir de los casos clasificados correctamente frente a los errores y el total de observaciones que el modelo pretende predecir.

Asimismo, dado que el proceso de creación y entrenamiento de estos modelos de clasificación corresponden a un proceso de experimentación que mejore sus resultados, es necesario realizar la configuración de sus hiperparámetros tomando en cuenta el mejor resultado durante la fase de los entrenamientos de los modelos. De esta forma, dentro de la Tabla 7 se muestran los hiperparámetros utilizados en cada uno de los modelos de clasificación utilizados.

Tabla 7.

Hiperparámetros utilizados en cada uno de los modelos

Modelo	Parámetros
LR	Estimado a través del método de máxima verosimilitud
MLP	**Épocas:** 10 **Tamaño de batch:** 32 **Función de salida:** Sigmoide **Funciones de activación:** ReLU **Optimizador:** Adam **Ratio de aprendizaje:** 0.05 **Función a optimizar**: Entropía Binaria **Métrica de evaluación y ajuste:** Precisión

SVM	**Kernel:** RBF **Grado:** 3 **Tolerancia:** 0.001
RF	**Criterio:** Gini **Profundidad máxima:** Sin limite **Numero de hojas máximo:** Sin limite **Características máximas:** Sin limite **Separador:** Best
KNN	**Distancia:** Minkowski **Número de vecinos:** 5

4.2. Matriz de confusión

Una vez definidos los hiperparámetros utilizados para cada uno de los modelos se puede realizar en entrenamiento y evaluación a partir de la matriz de confusión (Tabla 8) obtenida del cruce entre la variedad de arroz observada y la variedad de arroz predicha por el modelo de clasificación, este cruce de valores observados y predichos son los que se utilizan para el cálculo de las métricas de precisión, sensibilidad, especificidad y F1-score del modelo para cada uno de los modelos entrenados, en función de los Verdaderos positivos (Vp), Falsos positivos (Fp), Verdaderos negativos (Vn) y Falsos negativos (Fn). Cabe resaltar la importancia de una transformación adicional a los datos para poder reducir el costo computacional del entrenamiento de los modelos. Esta transformación realizada corresponde a una normalización de los datos a través del máximo y

mínimo valor de cada una de las características de la siguiente manera x* =(x-min{x})/(max{x} – min{x}).

Tabla 8.

Matriz de confusión de los algoritmos utilizados (a) LR, (b) MLP, (c) SVM, (d) RF y (e) k-NN

Matriz de LR		Variedad Predicha	
Clase real	INIAP-14	1.274 (Vp)	121 (Fp)
	INIAP-17	107 (Fn)	1.758 (Vn)

(a)

Matriz de MLP		Clase estimada	
Clase real	INIAP-14	1.270	125
	INIAP-17	108	1.757

(b)

Matriz de SVM		Clase estimada	
Clase real	INIAP-14	1.285	110
	INIAP-17	108	1.765

(c)

Matriz de RF		Clase estimada	
Clase real	INIAP-14	1.260	128
	INIAP-17	120	1.746

(d)

Matriz de k-NN		Clase estimada	
Clase real	INIAP-14	1.214	181
	INIAP-17	192	1.674

(e)

En función de los resultados, el modelo con la mejor precisión y sensibilidad fue el SVM. Si se prioriza los valores predichos correctamente, el algoritmo SVM lideró con 3.050, seguido de LR con 3.032, MLP con 3.027, RF con 3.006 y k-NN con 2.888 aciertos en cada modelo respectivamente.

Capítulo V.

Rendimiento y Conclusiones

Capítulo V.

Rendimiento y Conclusiones

5.1. Medidas de rendimiento

En el contexto de la clasificación de granos de arroz mediante métodos de aprendizaje automático, las métricas de rendimiento son herramientas esenciales para evaluar la eficacia y la precisión de los modelos. Para las métricas de rendimiento de la clasificación, se calcularon criterios de éxito como la Accuracy, Sensivity, Specificity, Precisión, F1-Score y, Puntuación-F1 y Kappa coefficient, en base a la matriz de confusión para cada modelo. Los resultados de la medición del rendimiento de la clasificación se indican en la Tabla 9.

Tabla 9.

Medidas de rendimiento promedio de los modelos utilizados

Métricas	SVM	MLP	LR	RF	k-NN
Accuracy	93.33	92.85	93.01	92.38	88.56
Sensitivity	92.11	92.16	92.25	91.30	86.34
Specificity	94.23	93.36	93.56	93.17	90.24
Accuracy	92.25	91.04	91.33	90.78	87.03
F1-score	92.18	91.60	91.79	91.04	86.68
Kappa Coefficient	86.36	85.38	85.70	84.41	76.66

El modelo SVM presentó el mejor desempeño general, destacándose con una exactitud del 93.33%, una sensibilidad del 92.11%, una especificidad del 94.23%, y un coeficiente Kappa de 86.36, lo que indica un ajuste superior al azar en la clasificación de las clases reales de granos de arroz respecto a las estimadas. El modelo MLP obtuvo una exactitud del 92.85%, una sensibilidad de 92.16%, y un Kappa de 85.38, resultados muy cercanos a los del modelo SVM, mostrando un buen equilibrio entre las métricas de precisión y especificidad. Por su parte, el modelo LR alcanzó una exactitud del 93.01%, una sensibilidad de 92.25%, y un coeficiente Kappa de 85.70, lo que lo sitúa en un desempeño competitivo, aunque ligeramente inferior al SVM en ajuste global.

El modelo RF logró una exactitud del 92.38% y un coeficiente Kappa de 84.41, manteniendo un buen nivel de desempeño, aunque algo menor en comparación con los modelos anteriores. El modelo k-NN mostró un desempeño más bajo en todas las métricas, con una exactitud del 88.56%, una sensibilidad del 86.34%, y un Kappa de 76.66, indicando que su capacidad para clasificar correctamente es más limitada frente a los otros algoritmos.

Para la obtención de los promedios al conjunto de datos se lo dividió en cinco particiones (k-folds). Se realizó un bucle a través de las cinco particiones. En cada iteración, una de las particiones se utilizó como conjunto de prueba, mientras que las otras cuatro se utilizaron como conjunto de entrenamiento. Para cada fold se entrenó los cincos modelos utilizando el conjunto de entrenamiento asociado. Se evaluó el rendimiento de cada modelo en el conjunto de prueba, calculando las métricas de interés para cada fold. Esto generó un conjunto de valores de métricas para cada modelo. Después de completar las cinco iteraciones, finalmente se calculó el promedio de cada métrica de rendimiento para cada modelo.

El uso de 5 folds se determinó dado que el número de datos es extenso, una partición en demasiadas particiones resultaría en un proceso y tiempo de entrenamiento demasiado largo. Por convención, el número de folds se fija de 5 a 10 cuando la data es extensa, por lo que, en este caso al tener más de 3000 observaciones, se considera una data amplia, sugiriendo y fijando el número de particiones en 20% para cada una de las particiones realizadas. Los hallazgos destacan que el modelo de SVM demostró la mejor precisión y sensibilidad en comparación con otros algoritmos evaluados. Esto sugiere que SVM es eficaz en identificar correctamente instancias de la variedad de arroz, superando a otros modelos en esta categoría específica.

5.2. Discusión de resultados


"El modelo LR destaca como el más efectivo para clasificar las variedades de arroz"

Los hallazgos destacan que el modelo de SVM demostró la mejor precisión y sensibilidad en comparación con otros algoritmos evaluados. Esto sugiere que SVM es eficaz en identificar correctamente instancias de la variedad de arroz, superando a otros modelos en esta categoría específica. Al examinar estudios previos en la clasificación de variedades de arroz como el (Nga et al., 2021)) también identificó al SVM como un modelo eficaz para la clasificación de variedades de arroz, con resultados comparables en términos de precisión y sensibilidad. Sin embargo, es importante señalar que las métricas numéricas específicas pueden variar en comparación con otros estudios debido a diferencias en conjuntos de datos, preprocesamiento y configuración

experimental. Por ejemplo, (Song et al., 2024) informó resultados similares con los Vp pero obtuvo variaciones en los resultados al priorizar Vn, lo cual destaca la sensibilidad de los resultados a diferentes condiciones experimentales.

Los resultados del presente estudio confirman la robustez de SVM, especialmente en la clasificación precisa de instancias positivas (Vp) y negativas (Vn). La literatura también señala que, a comparación de los otros modelos, SVM es un modelo lineal que ha demostrado también ser eficaz en problemas de clasificación binaria. Sin embargo, es importante tener en cuenta que los algoritmos pueden presentar limitaciones, por ejemplo, en el estudio de (Anami et al., 2018), aunque se destaca la utilidad del algoritmo de SVM en la clasificación de cultivos y granos, éste puede ser limitado en la captura de relaciones no lineales. Por otro lado, si bien, estudios han resaltado la capacidad de los MLP para aprender representaciones jerárquicas, pueden requerir grandes cantidades de datos para entrenar de manera efectiva, y pueden ser propensos al sobreajuste en conjuntos de datos pequeños (Ahad et al., 2023).

Por otro lado, aunque el algoritmo del RF es eficaz, la interpretación de cada árbol individual dentro del bosque puede ser desafiante ya que puede tener un rendimiento limitado en conjuntos de datos pequeños o altamente desequilibrados, lo que podría llevar al sobreajuste. Finalmente, el rendimiento de k-NN puede depender significativamente de la elección del parámetro k. Valores de k demasiado pequeños pueden hacer que el modelo sea sensible a ruido, mientras que valores demasiado grandes pueden reducir la capacidad del modelo para capturar patrones locales (Montesinos López et al., 2022). Los resultados del presente trabajo respaldan la eficacia de SVM en la clasificación de variedades de arroz. La comparación con estudios anteriores destaca consistencias y variabilidades, subrayando la

importancia de considerar el contexto específico de la aplicación al seleccionar el modelo óptimo.

Para futuras investigaciones, se sugiere explorar estrategias adicionales de preprocesamiento de datos o la aplicación de técnicas avanzadas de aprendizaje automático para mejorar aún más el rendimiento del modelo en este desafiante problema de clasificación

5.3. Conclusiones

Durante la etapa de clasificación del arroz se pueden evaluar parámetros geométricos como la longitud o el perímetro, la tasa de fractura, la blancura, o las grietas que pueda presentar el grano; o también se puede recurrir a las teorías de clasificación internacionales, una de estas la del Instituto Internacional de Investigación del Arroz, en la actualidad se está recurriendo con mayor frecuencia al uso del aprendizaje automático, por medio del cual es posible clasificar al arroz de manera más apropiada.

Para la clasificación se realizaron diferentes modelos utilizando los siguientes algoritmos LR, MLP, SVM, RF y k- NN, que son las técnicas de aprendizaje automático más utilizadas. Las mediciones estadísticas de la matriz de confusión como resultado de la clasificación se utilizaron como métricas de rendimiento. Las líneas de arroz utilizadas para esta investigación fueron INIAP-14 e INIAP-17.

El presente trabajo sugiere que el modelo SVM exhibió el mejor rendimiento global en términos de precisión y sensibilidad en la clasificación de granos de arroz. Esto lo convierte en una opción sólida y eficiente para la tarea específica. Además, SVM demostró ser particularmente eficaz en la identificación precisa de las variedades de arroz, liderando en la categoría de Vp. Esto es crucial en la aplicación práctica de clasificación correcta de las variedades de arroz INIAP 14 e INIAP 17. Asimismo, el modelo SVM mostró robustez al clasificar correctamente instancias de la otra variedad de arroz (o grano diferente) con la mayor cantidad de Vn. Esto indica su capacidad para minimizar los Falsos Positivos y proporciona una buena especificidad. SVM es conocido por su capacidad para manejar conjuntos de datos complejos y no lineales. En el contexto de la clasificación de granos de arroz, donde las relaciones pueden ser no lineales, la capacidad de SVM para encontrar límites de decisión óptimos es una fortaleza práctica.

Dada su alta precisión y sensibilidad, SVM se presenta como la elección principal cuando la clasificación precisa de variedades de arroz es la prioridad. Esto puede ser crucial en aplicaciones agrícolas o industriales donde se requiere una distinción clara entre diferentes tipos de granos. Al implementar SVM en un entorno práctico, es esencial considerar el contexto específico de la aplicación. Evaluar factores como la interpretabilidad del modelo, la disponibilidad de recursos computacionales y la facilidad de implementación es fundamental para tomar decisiones informadas. Dado que el rendimiento de los modelos puede depender de las características específicas del conjunto de datos, se sugiere realizar evaluaciones continuas del rendimiento de SVM en entornos de producción y, si es necesario, realizar ajustes o considerar otras técnicas de aprendizaje automático.

5.4. Recomendaciones

Se recomienda la revisión de nuevas teorías de clasificación del arroz a nivel nacional e internacional, por ejemplo, la clasificación del Instituto Ecuatoriano de Normalización (INEN), y otras clasificaciones que puedan existir con la finalidad de contar con más argumentos de clasificación que permitan a su vez probar los diferentes modelos de aprendizaje automático.

"Revisar estándares de clasificación nacionales e internacionales para validación"

Realizar nuevas investigaciones revisando otros modelos aparte de los analizados en esta investigación que permitan tener nuevos argumentos, y que a su vez hagan posible tener una industria arrocera más eficiente, esto permitirá reducir el tiempo de clasificación, evitar pérdidas previsibles y minimizar los costos generados por esta actividad.

Se recomienda que, dentro de la industria arrocera se haga uso del modelo ML ya que como se puede apreciar en los resultados obtenidos, es más eficiente al momento de realizar la clasificación del arroz, por lo tanto, es una alternativa importante para este sector productivo.

Anexos

Referencias, Listas y Soportes

Referencias bibliográficas

Abdullah, D. M., & Abdulazeez, A. M. (2021). Machine Learning Applications based on SVM Classification A Review. Qubahan Academic Journal, 1(2), 81–90. https://doi.org/10.48161/qaj.v1n2a50

Acosta, C., Sampallo, G. M., González Thomas, A., Cleva, M., & Liska, D. (2017, October 10). Detección e identificación de defectos en granos de arroz empleando visión artificial. IX Congreso Argentino de AgroInformática (CAI 2017) – JAIIO 46 - CLEI 43 (Córdoba, 2017). http://sedici.unlp.edu.ar/handle/10915/62834

Aggarwal, S., Suchithra, M., Chandramouli, N., Sarada, M., Verma, A., Vetrithangam, D., Pant, B., & Ambachew Adugna, B. (2022). Rice Disease Detection Using Artificial Intelligence and Machine Learning Techniques to Improvise Agro- Business. Scientific Programming, 2022. https://doi.org/10.1155/2022/1757888

Ali, I., Iqbal, A., Ullah, S., Muhammad, I., Yuan, P., Zhao, Q., Yang, M., Zhang, H., Huang, M., Liang, H., Gu, M., & Jiang, L. (2022). Effects of Biochar Amendment and Nitrogen Fertilizer on RVA Profile and Rice Grain Quality Attributes. Foods (Basel, Switzerland), 11(5). https://doi.org/10.3390/foods11050625

Álvarez Hernández, J. C., Tapia Vargas, L. M., Hernández Pérez, A., Barrios Gomez, E. J., & Pardo Melgarejo, S. (2018). Estabilidad productiva de variedades avanzadas de arroz grano largo

delgado en Michoacán México. Revista Mexicana de Ciencias Agricolas, 9(3), 629–637.

Álvarez, R., Reyes, E., Ramos, N., & Valera, E. (2021). "LIBERTAD FL": Nuevo cultivar de arroz de riego para Venezuela. Revista. http://revistas.uns.edu.pe/index.php/PUNKURI/article/view/6

Alpaydin, E. (2021). Machine Learning, revised and updated edition. MIT Press. https://play.google.com/store/books/details?id=2nQJEAAAQBAJ

Amos, B., Aidoo, R., Osei Mensah, J., Adzawla, W., Appiah-Twumasi, M., Akey, E. A., & Bannor, R. K. (2022). Rice Marketing Outlets, Commercialization, and Welfare: Insights From Rural Ghana. Journal of International Food & Agribusiness Marketing, 1–27. https://doi.org/10.1080/08974438.2021.2022556

Arora, B., Bhagat, N., Saritha, L. R., & Arcot, S. (2020). Rice Grain Classification using Image Processing & Machine Learning Techniques. 2020 International Conference on Inventive Computation Technologies (ICICT), 205–208.

Arumugam, K., Swathi, Y., Sanchez, D. T., Mustafa, M., Phoemchalard, C., Phasinam, K., & Okoronkwo, E. (2022). Towards applicability of machine learning techniques in agriculture and energy sector. Materials Today: Proceedings, 51, 2260–2263. https://doi.org/10.1016/j.matpr.2021.11.394

Bakthavatchalam, K., Karthik, B., Thiruvengadam, V., Muthal, S., Jose, D., Kotecha, K., & Varadarajan, V. (2022). IoT Framework for Measurement and Precision Agriculture: Predicting the Crop Using Machine Learning Algorithms. Technologies, 10, 13. https://doi.org/10.3390/technologies10010013

Banerjee, I., & Madhumathy, P. (2022). IoT Based Agricultural Business Model for Estimating Crop Health Management to Reduce Farmer Distress Using SVM and Machine Learning. In P. K. Pattnaik, R. Kumar, & S. Pal (Eds.), Internet of Things and Analytics for Agriculture, Volume 3 (pp. 165–183). Springer Singapore. https://doi.org/10.1007/978-981-16-6210-2_8

Bhujel, & Shakya. (2022). Rice Leaf Diseases Classification Using Discriminative Fine Tuning and CLR on EfficientNet. Journal of Soft Computing Paradigm. https://irojournals.com/jscp/V4/I3/06.pdf

Bonaccorso, G. (2017). Machine Learning Algorithms. Packt Publishing Ltd. https://play.google.com/store/books/details?id=_-ZDDwAAQBAJ

Ceavichay, K., Valenzuela, D., y Cornejo, F. (2013). Caracterización Física, Tecnológica y Reológica de Tres Variedades de Arroz Pilado Ecuatoriano, Cosecha Invierno. ESPOL, 1(1), 1-8.

Cinar, I., & Koklu, M. (2019). Classification of Rice Varieties Using Artificial Intelligence Methods. International Journal of Intelligent Systems and Applications in Engineering, 7(3), Art. 3. https://doi.org/10.18201/ijisae.2019355381

Coello, V. C., y Garces, C. C. (2013). Análisis de propiedades térmicas durante gelatinización en tres variedades de arroz iniap aplicando el calorímetro diferencial de barrido (dsc) [Tesis de Pregrado, ESPOL]. http://www.dspace.espol.edu.ec/handle/123456789/21612

Chen, Z., Goh, H. S., Sin, K. L., Lim, K., Chung, N. K. H., & Liew, X. Y. (2021). Automated Agriculture Commodity Price

Prediction System with Machine Learning Techniques. In arXiv [cs.LG]. arXiv. http://arxiv.org/abs/2106.12747

Chipana Valero, C., Manzaneda Delgado, F. F., & Choque Tarqui, C. E. (2022). Valuación de dos variedades de arroz (Oryza sativa L.), en dos sistemas de manejo de suelos en Sapecho Alto- Beni. Revista de Investigación e Innovación Agropecuaria y de Recursos Naturales, 9(1), 3–9.

Cinar, I., & Koklu, M. (2022). Identification of Rice Varieties Using Machine Learning Algorithms. Jiangsu Nong Ye Xue Bao = Journal of Agricultural Sciences. https://dergipark.org.tr/en/pub/ankutbd/issue/68522/862482

Dheer, P., Singh, R. K., & Others. (2019). Identification of indian rice varieties using machine learning classifiers. Plant Archives, 19(1), 155–158.

Dominguez-Lara, S., y Merino-Soto, C. (2018). Evaluación de las malas especificaciones en modelos de ecuaciones estructurales. Revista Argentina de Ciencias del Comportamiento, 10(2), 19-24.

Elbadawi, M., Gaisford, S., & Basit, A. W. (2021). Advanced machine-learning techniques in drug discovery. Drug Discovery Today, 26(3), 769–777. https://doi.org/10.1016/j.drudis.2020.12.003

Fernández, D., Liso, Á. A., Abades, D. P., Piñeiro, A. L., Gómez, S., Sánchez, J., Martín, C., & Vicente, L. (2021). Estrategias de riego y laboreo con aplicación de enmienda orgánica para mejorar las propiedades del suelo y las producciones de arroz (Oryza sativa L.). Phytoma España: La revista profesional de sanidad vegetal, 328, 50–53.

Goyal, N., Kumar, S., & Saraswat, M. (2022). Detection of Unhealthy citrus leaves using Machine Learning Technique. 2022 12th International Conference on Cloud Computing, Data Science & Engineering (Confluence), 591–595. https://doi.org/10.1109/Confluence52989.2022.9734162

Greener, J. G., Kandathil, S. M., Moffat, L., & Jones, D. T. (2022). A guide to machine learning for biologists. Nature Reviews. Molecular Cell Biology, 23(1), 40–55. https://doi.org/10.1038/s41580-021-00407-0

Herath, H.M.K.K.M.B., Karunasena, G.M.K.B., & Prematilake, R.D.D. (2022). Computer Vision for Agro-Foods: Investigating a Method for Grading Rice Grain Quality in Sri Lanka. In N. Kumar, C. Shahnaz, K. Kumar, M. Abed Mohammed, & R. S. Raw (Eds.), Advance Concepts of Image Processing and Pattern Recognition: Effective Solution for Global Challenges (pp. 21–34). Springer Singapore. https://doi.org/10.1007/978-981-16-9324-3_2

Hernández-Cuello, G., Morejón-Mesa, Y., Monzón-Monrabal, L. L., Díaz-Ruiz, D., & Domínguez-Calvo, G. (2021). Manejo durante la cosecha del arroz y su influencia en la calidad del secado industrial. Revista, 11(2), 39–43.

Ibrahim, S., Zulkifli, N. A., & Sabri, N. (2019). Rice grain classification using multi-class support vector machine (SVM). IAES. http://download.garuda.kemdikbud.go.id/article.php?article=1494270&val=151&title=Rice%20grain%20classification%20using%20multi-class%20support%20vector%20machine%20SVM

International Rice Research Institute (2021). Clasificación del arroz. https://www.irri.org/our-solutions/irri-education

Jagtap, S. T., Phasinam, K., Kassanuk, T., Jha, S. S., Ghosh, T., & Thakar, C. M. (2022). Towards application of various machine learning techniques in agriculture. Materials Today: Proceedings, 51, 793–797. https://doi.org/10.1016/j.matpr.2021.06.236

Jin, B., Zhang, C., Jia, L., Tang, Q., Gao, L., Zhao, G., & Qi, H. (2022). Identification of Rice Seed Varieties Based on Near-Infrared Hyperspectral Imaging Technology Combined with Deep Learning. ACS Omega, 7(6), 4735-4749. https://doi.org/10.1021/acsomega.1c04102

Kasinathan, T., Singaraju, D., & Uyyala, S. R. (2021). Insect classification and detection in field crops using modern machine learning techniques. Information Processing in Agriculture, 8(3), 446–457. https://doi.org/10.1016/j.inpa.2020.09.006

Khatri, A., Agrawal, S., & Chatterjee, J. M. (2022). Wheat Seed Classification: Utilizing Ensemble Machine Learning Approach. Scientific Programming, 2022. https://doi.org/10.1155/2022/2626868

Kiratiratanapruk, K., Temniranrat, P., Sinthupinyo, W., Prempree, P., Chaitavon, K., Porntheeraphat, S., & Prasertsak, A. (2020). Development of Paddy Rice Seed Classification Process using Machine Learning Techniques for Automatic Grading Machine. Journal of Sensors, 2020. https://doi.org/10.1155/2020/7041310

Koklu, M., Cinar, I., & Taspinar, Y. S. (2021). Classification of rice varieties with deep learning methods. Computers and Electronics in Agriculture, 187, 106285. https://doi.org/10.1016/j.compag.2021.106285

Kok, Z. H., Mohamed, A. R., Alfatni, M. S. M., & Khairunniza-Bejo, S. (2021). Support Vector Machine in Precision Agriculture: A review. Computers and Electronics in Agriculture, 191, 106546. https://doi.org/10.1016/j.compag.2021.106546

Komal, Sethi, G. K., & Bawa, R. K. (2022). A prototype of automatic rice variety identification system using artificial intelligence techniques. AIP Conference Proceedings, 2455(1), 040004. https://doi.org/10.1063/5.0100827

Krishna, C. V., Suchitra, B., Sujihelen, L., Roobini, M. S., Cherukullapurath, S., & Jesudoss, A. (2022). Quality Analysis of Rice Grains Using Morphological Techniques. 2022 International Conference on Communication, Computing and Internet of Things (IC3IoT), 1–5. https://doi.org/10.1109/IC3IOT53935.2022.9767925

Liu, J., Qiu, S., & Wei, Z. (2022). Real-Time Measurement of Moisture Content of Paddy Rice Based on Microstrip Microwave Sensor Assisted by Machine Learning Strategies. Chemosensors, 10(10), 376.

Liu, W., Zeng, S., Wu, G., Li, H., & Chen, F. (2021). Rice Seed Purity Identification Technology Using Hyperspectral Image with LASSO Logistic Regression Model. Sensors, 21(13), Art. 13. https://doi.org/10.3390/s21134384

López, G. M., Miranda, R. P., Hernández, A. G., & Sánchez, E. U. R. (2021). Rice production potential technology (Oryza sativa L.) in the state of Tabasco, Mexico and it contribution to food

sovereignty. https://chapingo-cori.mx/rchsat/rchsat/article/download/9/8

Malik, P., Sengupta, S., & Jadon, J. S. (2021). Comparative Analysis of Soil Properties to Predict Fertility and Crop Yield using Machine Learning Algorithms. 2021 11th International Conference on Cloud Computing, Data Science & Engineering (Confluence), 1004–1007. https://doi.org/10.1109/Confluence51648.2021.9377147

Mena, Luis (2008). Aprendizaje automático a partir de conjuntos de datos no balanceados y su aplicación en el diagnóstico y pronóstico médico [tesis doctoral]. Instituto Nacional De Astrofísica, Óptica y Electrónica. https://inaoe.repositorioinstitucional.mx/jspui/bitstream/1009/533/1/MenaCaLJ.p df

Mishra, S., & Tyagi, A. K. (2022). The Role of Machine Learning Techniques in Internet of Things-Based Cloud Applications. In S. Pal, D. De, & R. Buyya (Eds.), Artificial Intelligence-based Internet of Things Systems (pp. 105–135). Springer International Publishing. https://doi.org/10.1007/978-3-030-87059-1_4

Molnar, C. (2020). Interpretable Machine Learning. Lulu.com. https://play.google.com/store/books/details?id=jBm3DwAAQBAJ

Mourtzinis, S., Esker, P. D., Specht, J. E., & Conley, S. P. (2021). Advancing agricultural research using machine learning algorithms. Scientific Reports, 11(1), 17879. https://doi.org/10.1038/s41598-021-97380-7

Murillo, W. J. J., Pérez, E. M. G., & Murillo, C. A. J. (2022). Aplicación de modelo matemático para la variación de ingresos de

productores de arroz. La Troncal – Cañar. Ciencia Latina Revista Científica Multidisciplinar, 6(1), 4824–4838.

Ndikuryayo, C., Ndayiragije, A., Kilasi, N., & Kusolwa, P. (2022). Breeding for Rice Aroma and Drought Tolerance: A Review. Agronomy, 12(7), 1726. https://doi.org/10.3390/agronomy12071726

Nosratabadi, S., Ardabili, S., Lakner, Z., Mako, C., & Mosavi, A. (2021). Prediction of Food Production Using Machine Learning Algorithms of Multilayer Perceptron and ANFIS. Collection FAO: Agriculture, 11(5), 408. https://doi.org/10.3390/agriculture11050408

Ochoa, R., Nava, N., & Fusil, D. (2020). Comprensión epistemológica del tesista sobre investigaciones cuantitativas, cualitativas y mixtas. Orbis: revista de Ciencias Humanas, 15(45), 13-22.

Pavani, S., & Augusta Sophy Beulet, P. (2022). Prediction of Jowar Crop Yield Using K-Nearest Neighbor and Support Vector Machine Algorithms. Futuristic Communication and Network Technologies, 497–503. https://doi.org/10.1007/978-981-16-4625-6_49

Rakhra, M., Soniya, P., Tanwar, D., Singh, P., Bordoloi, D., Agarwal, P., Takkar, S., Jairath, K., & Verma, N. (2021). Crop Price Prediction Using Random Forest and Decision Tree Regression:-A Review. Materials Today: Proceedings. https://doi.org/10.1016/j.matpr.2021.03.261

Ramadhani, F. (2022). Mapping of multitemporal rice (Oryza sativa L.) growth stages using remote sensing with multi-sensor and machine learning: a thesis dissertation presented in partial fulfilment of the requirements for the degree of Doctor of Philosophy in Earth Science at Massey University, Manawatū,

New Zealand [Massey University]. https://mro.massey.ac.nz/handle/10179/17409

Ramírez Morales, I. (2018). Estudio de aplicabilidad de técnicas de inteligencia artificial en el sector agropecuario. https://ruc.udc.es/dspace/handle/2183/20325

Rekha Sundari, M., Siva Rama Krishna, G., Sai Naveen, V., & Bharathi, G. (2021). Crop Recommendation System Using K-Nearest Neighbors Algorithm. Proceedings of 6th International Conference on Recent Trends in Computing, 581–589. https://doi.org/10.1007/978-981-33-4501-0_54

Rolnick, D., Donti, P. L., Kaack, L. H., Kochanski, K., Lacoste, A., Sankaran, K., Ross, A. S., Milojevic-Dupont, N., Jaques, N., Waldman-Brown, A., Luccioni, A. S., Maharaj, T., Sherwin, E. D., Mukkavilli, S. K., Kording, K. P., Gomes, C. P., Ng, A. Y., Hassabis, D., Platt, J. C., … Bengio, Y. (2022). Tackling Climate Change with Machine Learning. ACM Comput. Surv., 55(2), 1–96. https://doi.org/10.1145/3485128

Sánchez Molina, A. A., Murillo Garza, A., Sánchez Molina, A. A., & Murillo Garza, A. (2021). Enfoques metodológicos en la investigación histórica: Cuantitativa, cualitativa y comparativa. Debates por la historia, 9(2), 147-181. https://doi.org/10.54167/debates-por-la-historia.v9i2.792

Sultana, S., Faruque, M., & Islam, M. R. (2022). Rice grain quality parameters and determination tools: a review on the current developments and future prospects. International Journal of Food Properties, 25(1), 1063–1078. https://doi.org/10.1080/10942912.2022.2071295

Suresh, N., Ramesh, N. V. K., Inthiyaz, S., Priya, P. P., Nagasowmika, K., Kumar, K. V. N. H., Shaik, M., & Reddy, B. N.

K. (2021). Crop Yield Prediction Using Random Forest Algorithm. 2021 7th International Conference on Advanced Computing and Communication Systems (ICACCS), 1, 279–282. https://doi.org/10.1109/ICACCS51430.2021.9441871

Tosawadi, T., Kasetkasem, T., & Others. (2022). Automatic rice plant disease evaluation method based on anomaly detection and deep learning [Kasetsart University]. https://ethesis.lib.ku.ac.th/dspace/handle/123456789/921

Vecchio, Y., Di Pasquale, J., Del Giudice, T., Pauselli, G., Masi, M., & Adinolfi, F. (2022). Precision farming: what do Italian farmers really think? An application of the Q methodology. Agricultural Systems, 201, 103466. https://doi.org/10.1016/j.agsy.2022.103466

Waleed, M., Um, T.-W., Kamal, T., & Usman, S. M. (2021). Classification of Agriculture Farm Machinery Using Machine Learning and Internet of Things. Symmetry, 13(3), 403. https://doi.org/10.3390/sym13030403

Wang, H., Lei, Z., Zhang, X., Zhou, B., & Peng, J. (2016). Machine learning basics. Deep Learning, 98–164. http://whdeng.cn/Teaching/PPT_01_Machine%20learning%20Basics.pdf

Xu, P., Yang, R., Zeng, T., Zhang, J., Zhang, Y., & Tan, Q. (2021). Varietal classification of maize seeds using computer vision and machine learning techniques. Journal of Food Process Engineering, 44(11). https://doi.org/10.1111/jfpe.13846

Yang, M.-D., Hsu, Y.-C., Tseng, W.-C., Lu, C.-Y., Yang, C.-Y., Lai, M.-H., & Wu, D.- H. (2021). Assessment of Grain Harvest Moisture Content Using Machine Learning on Smartphone

Images for Optimal Harvest Timing. Sensors, 21(17). https://doi.org/10.3390/s21175875

Zahra, N., Hafeez, M. B., Nawaz, A., & Farooq, M. (2022). Rice production systems and grain quality. Journal of Cereal Science, 105, 103463. https://doi.org/10.1016/j.jcs.2022.103463

Zambrano, C. E., & Andrade Arias, M. S. (2019). Factores que inciden en la productividad del cultivo de arroz en la provincia Los Ríos. Revista Universidad Y. http://scielo.sld.cu/scielo.php?script=sci_arttext&pid=S2218- 36202019000500270

Zhou, Z.-H. (2021). Machine Learning. Springer Nature. https://play.google.com/store/books/details?id=ctM-EAAAQBAJ

Reactivos

Capitulo I.

¿Cuál es el cereal más cultivado en el mundo después del trigo y el maíz?

A. La cebada

B. El arroz

C. La avena

D. El sorgo

ANSWER: B

¿Qué factor no influye en la presentación de los granos de arroz?

A. Tamaño

B. Color

C. Salud interna

D. Precio

ANSWER: D

¿Cuál es la principal desventaja de la clasificación manual del arroz?

A. Es un proceso económico

B. Es un método eficiente

C. Requiere mucho tiempo y es ineficiente

D. Es ampliamente estandarizado

ANSWER: C

¿Qué herramienta informática se menciona como una alternativa eficiente para la clasificación del arroz?

A. Robótica

B. Nanotecnología

C. Inteligencia Artificial

D. Biotecnología

ANSWER: C

¿Cuál es una ventaja del aprendizaje automático (ML) en la clasificación del arroz?

A. Depende exclusivamente de la intervención humana

B. Se basa en el reconocimiento de patrones

C. No requiere modelos matemáticos

D. Es más lento que la clasificación manual

ANSWER: B

¿En qué sector se han aplicado los algoritmos de aprendizaje automático además del arroz?

A. Minería

B. Evaluación de la calidad de alimentos

C. Turismo

D. Educación

ANSWER: B

¿Cuál es el principal reto que enfrenta la industria arrocera en la clasificación del grano?

A. Falta de demanda del producto

B. Escasez de tierras cultivables

C. Métodos ineficientes de clasificación

D. Exceso de producción mundial

ANSWER: C

¿Qué característica del arroz influye en la decisión de compra de los consumidores?

A. Cantidad de agua utilizada en su cultivo

B. Características visuales del grano

C. Marca del fertilizante utilizado

D. Nombre del productor

ANSWER: B

¿Qué se busca con el uso de aprendizaje automático en la clasificación del arroz?

A. Optimizar los tiempos de clasificación y reducir costos

B. Aumentar la producción de arroz sin mejorar su calidad

C. Eliminar completamente la intervención humana en la agricultura

D. Reducir la demanda de arroz en el mercado

ANSWER: A

¿Por qué es importante desarrollar métodos alternativos para la clasificación del arroz?

A. Porque los métodos actuales son rápidos y eficientes

B. Porque las inspecciones visuales son subjetivas y propensas a errores

C. Porque la clasificación manual es más precisa
D. Porque no existen técnicas automatizadas disponibles
ANSWER: B

Capítulo II.

¿Cuál es el porcentaje aproximado de la energía alimentaria que suministra el arroz a nivel mundial?
A. 10%
B. 20%
C. 30%
D. 40%
ANSWER: B

¿Cuál de las siguientes especies de arroz es poco cultivada debido a restricciones en la zona oeste de África?
A. Oryza sativa L.
B. Oryza glaberrima S.
C. Oryza indica
D. Oryza japonica
ANSWER: B

¿Qué provincia ecuatoriana tiene la mayor superficie cultivada de arroz?
A. Manabí
B. El Oro
C. Guayas
D. Los Ríos
ANSWER: C

¿Cuál es una de las etapas más importantes en el procesamiento del arroz?

A. Transporte
B. Cosecha
C. Limpieza
D. Comercialización
ANSWER: C

Según el INEN, ¿qué longitud mínima deben tener los granos de arroz extra largo?
A. 5,0 mm
B. 6,0 mm
C. 6,99 mm
D. 7,0 mm
ANSWER: D

¿Qué tipo de grano de arroz presenta un color rojo nítido o estrías de color rojizo en la cutícula?
A. Granos tizosos
B. Granos rojos
C. Granos dañados
D. Granos quebrados
ANSWER: B

¿Qué paradigma de aprendizaje automático utiliza etiquetas en los datos de entrenamiento?
A. Aprendizaje supervisado
B. Aprendizaje no supervisado
C. Aprendizaje por refuerzo
D. Aprendizaje profundo
ANSWER: A

¿Cuál de los siguientes algoritmos se considera un modelo de aprendizaje supervisado utilizado en la clasificación del arroz?

A. Redes neuronales convolucionales
B. K-Nearest Neighbors (k-NN)
C. Algoritmos genéticos
D. Clustering jerárquico

ANSWER: B

¿Cuál es la principal ventaja del algoritmo de Bosque Aleatorio (RF)?

A. Es fácil de interpretar
B. Maneja un gran número de variables
C. No requiere datos de entrenamiento
D. Es un modelo determinista

ANSWER: B

¿Qué parámetro de la matriz de confusión representa la cantidad de ejemplos negativos clasificados correctamente?

A. VP (verdaderos positivos)
B. FP (falsos positivos)
C. FN (falsos negativos)
D. VN (verdaderos negativos)

ANSWER: D

Capítulo III.

¿Qué tipo de diseño de investigación se utilizó en este estudio?

A. Cualitativo
B. Experimental

C. Cuantitativo
D. Mixto
ANSWER: C

¿Por qué se considera que la investigación es de tipo causal comparativa?
A. Porque analiza solo una técnica de clasificación
B. Porque compara diferentes técnicas de aprendizaje automático
C. Porque estudia únicamente las características del arroz
D. Porque no incluye análisis estadístico
ANSWER: B

¿Cuántos granos de arroz se analizaron en total en el estudio?
A. 2.500
B. 3.000
C. 3.260
D. 3.500
ANSWER: C

¿Qué variedades de arroz fueron utilizadas en este estudio?
A. INIAP-10 e INIAP-12
B. INIAP-11 e INIAP-15
C. INIAP-14 e INIAP-17
D. INIAP-16 e INIAP-18
ANSWER: C

¿Cuál es la finalidad principal de la validación cruzada en el estudio?
A. Mejorar la precisión de la clasificación
B. Acelerar el proceso de clasificación
C. Eliminar el uso de métricas estadísticas

D. Evitar el uso de algoritmos de aprendizaje automático

ANSWER: A

¿Cuál de las siguientes no es una variable morfológica evaluada en el estudio?

A. Área

B. Perímetro

C. Color

D. Excentricidad

ANSWER: C

¿Cuál es el propósito del estadístico Kappa en la evaluación del modelo?

A. Determinar qué modelo es más eficiente en la clasificación

B. Calcular el peso de los granos de arroz

C. Definir nuevas variedades de arroz

D. Evaluar el contenido nutricional del arroz

ANSWER: A

¿Qué algoritmo de aprendizaje automático no fue utilizado en este estudio?

A. Regresión Logística (LR)

B. Máquina de Vectores de Apoyo (SVM)

C. Redes Neuronales Convolucionales (CNN)

D. Perceptrón Multicapa (MLP)

ANSWER: C

¿Qué proceso se llevó a cabo con las imágenes antes de la clasificación del arroz?

A. Se imprimieron para su análisis manual

B. Se convirtieron a escala de grises e imágenes binarias

C. Se descartaron las imágenes de menor calidad

D. Se almacenaron sin procesamiento previo

ANSWER: B

¿Qué parámetro de la matriz de confusión mide la proporción de valores negativos correctamente clasificados?

A. Exactitud

B. Precisión

C. Exhaustividad

D. Especificidad

ANSWER: D

Capítulo IV.

¿Cuántas características morfológicas se evaluaron en los granos de arroz en este estudio?

A. 5

B. 6

C. 7

D. 8

ANSWER: C

¿Qué variedad de arroz presentó mayor área promedio por grano?

A. INIAP-14

B. INIAP-17

C. Ambas tienen el mismo tamaño

D. No se evaluó el área promedio

ANSWER: A

¿Qué prueba estadística se utilizó para comparar las características morfológicas de ambas variedades de arroz?

A. ANOVA
B. Prueba de T de Student
C. Chi-cuadrado
D. Regresión lineal
ANSWER: B

¿Cuál de los siguientes es un hallazgo clave del análisis de las características del arroz?
A. No existen diferencias significativas entre las variedades
B. INIAP-17 tiene mayor longitud de eje mayor que INIAP-14
C. INIAP-14 es más alargado y ancho que INIAP-17
D. INIAP-14 requiere menos agua para su cocción
ANSWER: C

¿Qué métrica estadística se utilizó para evaluar la simetría en la distribución de los datos?
A. Coeficiente de variación
B. Media aritmética
C. Kurtosis
D. Asimetría
ANSWER: D

¿Qué modelo de aprendizaje automático obtuvo la mejor precisión en la clasificación de los granos de arroz?
A. Regresión Logística (LR)
B. Perceptrón Multicapa (MLP)
C. Máquina de Vectores de Soporte (SVM)
D. k-Vecinos más Cercanos (k-NN)
ANSWER: C

¿Qué transformación se realizó en los datos antes de entrenar los modelos de clasificación?

A. Conversión a escala logarítmica
B. Normalización mediante el rango (mínimo y máximo)
C. Eliminación de valores atípicos
D. Conversión a datos categóricos
ANSWER: B

¿Cuál fue el modelo de aprendizaje automático con menor cantidad de aciertos en la clasificación?
A. Regresión Logística (LR)
B. Bosque Aleatorio (RF)
C. Perceptrón Multicapa (MLP)
D. k-Vecinos más Cercanos (k-NN)
ANSWER: D

¿Qué hiperparámetro NO fue utilizado en el modelo de Máquina de Vectores de Soporte (SVM)?
A. Kernel: RBF
B. Grado: 3
C. Función de activación: ReLU
D. Tolerancia: 0.001
ANSWER: C

¿Qué valor de p se obtuvo en la prueba de T de Student para todas las características evaluadas entre ambas variedades de arroz?
A. > 0.05
B. < 0.01
C. < 0.001
D. 0.5
ANSWER: C

Capítulo V.

¿Qué modelo de aprendizaje automático obtuvo el mejor desempeño en la clasificación del arroz?

A. Regresión Logística (LR)

B. Perceptrón Multicapa (MLP)

C. Máquina de Vectores de Soporte (SVM)

D. k-Vecinos más Cercanos (k-NN)

ANSWER: C

¿Cuál fue la métrica de exactitud obtenida por el modelo SVM?

A. 92.38%

B. 93.33%

C. 88.56%

D. 91.79%

ANSWER: B

¿Qué métrica se utiliza para evaluar la capacidad de un modelo para clasificar correctamente instancias negativas?

A. Precisión

B. Sensibilidad

C. Especificidad

D. Puntuación-F1

ANSWER: C

¿Cuántas particiones (k-folds) se utilizaron en la validación cruzada del estudio?

A. 3

B. 5

C. 7

D. 10

ANSWER: B

¿Qué métrica refleja la proporción de valores positivos correctamente clasificados?

A. Exactitud

B. Precisión

C. Sensibilidad

D. Coeficiente Kappa

ANSWER: C

¿Cuál fue el modelo con el menor coeficiente Kappa en la clasificación del arroz?

A. SVM

B. MLP

C. RF

D. k-NN

ANSWER: D

¿Cuál de los siguientes algoritmos puede ser sensible al ruido si se elige un valor de k demasiado pequeño?

A. SVM

B. k-Vecinos más Cercanos (k-NN)

C. Bosque Aleatorio (RF)

D. Regresión Logística (LR)

ANSWER: B

¿Qué factor puede limitar el desempeño del modelo Bosque Aleatorio (RF)?

A. Su facilidad de interpretación

B. Su baja capacidad de generalización

C. Su tendencia al sobreajuste en datos pequeños

D. Su dependencia exclusiva de datos numéricos

ANSWER: C

¿Qué se recomienda para futuras investigaciones en la clasificación de arroz?

A. Utilizar únicamente el modelo SVM en todas las aplicaciones
B. Explorar nuevos modelos de aprendizaje automático
C. Evitar el uso de métricas estadísticas en la evaluación de modelos
D. Eliminar la validación cruzada en los experimentos

ANSWER: B

¿Cuál es una de las principales ventajas de SVM en la clasificación del arroz?

A. Su facilidad para manejar datos con relaciones no lineales
B. Su alto costo computacional
C. Su dependencia de datos categóricos
D. Su baja precisión en clasificación

ANSWER: A

Glosario

Capítulo I.

→ **Arroz (Oryza sativa L.)**: Cereal ampliamente cultivado en el mundo y una fuente principal de carbohidratos en la dieta humana.

→ **Clasificación manual de arroz**: Proceso de separación de granos por sus características físicas realizado de forma artesanal.

→ **Industria arrocera**: Sector productivo dedicado al cultivo, procesamiento y distribución de arroz.

→ **Inteligencia Artificial (IA)**: Rama de la informática que permite la automatización de tareas a través de algoritmos y modelos computacionales.

→ **Machine Learning (ML)**: Subcampo de la IA que permite a los sistemas aprender y mejorar su desempeño a partir de datos sin ser programados explícitamente.

→ **Patrones de clasificación**: Características o atributos utilizados para agrupar datos de manera automática mediante algoritmos de IA.

→ **Agricultura de Precisión**: Uso de tecnologías avanzadas para optimizar la producción agrícola y mejorar la eficiencia.

→ **I+D+i**: Investigación, Desarrollo e Innovación, proceso de mejora tecnológica aplicada a diferentes sectores.

→ **FAO (Organización de las Naciones Unidas para la Alimentación y la Agricultura)**: Organismo internacional que promueve la seguridad alimentaria.

→ **Proyección de producción de arroz**: Estimación del aumento en la producción de arroz necesaria para satisfacer la demanda mundial.

→ **Atributos intrínsecos del arroz**: Características del grano como textura, olor y sabor que influyen en su aceptación en el mercado.

→ **Atributos extrínsecos del arroz**: Características relacionadas con el empaquetado, etiquetado y marca del producto.

→ **Automatización de procesos agrícolas**: Implementación de tecnologías para mejorar la eficiencia en la producción agrícola.

→ **Centros de investigación agrícola**: Instituciones encargadas de desarrollar nuevas tecnologías y mejorar los procesos agrícolas.

→ **Homogeneización de granos**: Proceso de estandarización del tamaño y forma de los granos de arroz.

→ **Investigación bibliográfica**: Revisión de estudios previos sobre clasificación de arroz.

→ **Modelos de ML para clasificación**: Identificación de los algoritmos más adecuados para el problema.

→ **Calidad del grano de arroz**: Conjunto de características físicas y nutricionales que determinan su valor comercial.

→ **Mediciones manuales de calidad**: Evaluaciones visuales y manuales para determinar la calidad del arroz.

→ **Costo de inspecciones de calidad**: Gastos asociados a la evaluación manual de la calidad del arroz.

→ **Errores humanos en clasificación**: Fallos en la separación de granos debido a la variabilidad en la percepción humana.

→ **Aprendizaje automático como alternativa**: Uso de ML para mejorar la clasificación del arroz y reducir costos.

→ **Modelos de apoyo a la toma de decisiones**: Algoritmos que facilitan la clasificación y estandarización en la industria arrocera.

→ **Proyectos de investigación en clasificación de arroz**: Estudios previos sobre el uso de tecnología en la industria arrocera.

Capitulo II.

→ **Producción de arroz en Ecuador**: Actividad económica centrada en el cultivo y procesamiento del arroz en el país.

→ **Variedades de arroz**: Tipos de arroz diferenciados por su morfología, origen y características de cultivo.

→ **Cultivo de arroz**: Proceso agrícola que abarca desde la siembra hasta la cosecha del arroz.

→ **Sistemas de riego**: Métodos utilizados para suministrar agua a los cultivos de arroz, incluyendo riego por inundación y riego tecnificado.

→ **Proceso de producción de arroz**: Etapas que incluyen la siembra, cosecha, procesamiento y distribución del arroz.

→ **Escalas de clasificación del arroz**: Estándares utilizados para categorizar los granos según su tamaño y calidad.

→ **Determinación del grado de arroz**: Evaluación de parámetros como color, textura y presencia de defectos.

→ **Materia extraña en el arroz**: Elementos ajenos al grano como cáscaras, polvo y residuos orgánicos.

→ **Aprendizaje automático en la agricultura**: Aplicación de modelos de ML para optimizar procesos agrícolas y mejorar la calidad del producto.

→ **Redes neuronales artificiales**: Algoritmos inspirados en la estructura del cerebro humano utilizados para el reconocimiento de patrones.

→ **Matriz de confusión**: Tabla de evaluación del desempeño de un modelo de clasificación.

→ **Precisión de modelos**: Medida de la exactitud de un algoritmo en la clasificación de datos.

→ **Técnicas de validación**: Métodos estadísticos para evaluar la efectividad de los modelos de clasificación.

Capitulo III.

→ **Investigación cuantitativa**: Enfoque metodológico basado en la recopilación y análisis de datos numéricos para identificar patrones y relaciones.

→ **Diseño de investigación causal comparativo**: Estrategia que permite analizar diferencias entre grupos o variables para identificar relaciones de causa y efecto.

→ **Población de estudio**: Conjunto total de elementos sobre los cuales se quiere obtener información en una investigación.

→ **Muestra representativa**: Subconjunto seleccionado de la población que conserva sus características esenciales para el estudio.

→ **Parámetros morfológicos del arroz**: Características físicas del grano, como tamaño, forma y estructura.

→ **Captura de datos por imagen**: Proceso de adquisición de imágenes de los granos de arroz mediante cámaras especializadas para su posterior análisis.

→ **Preprocesamiento de imágenes**: Conjunto de técnicas utilizadas para mejorar la calidad de las imágenes antes de su análisis automatizado.

→ **Normalización de datos**: Transformación de variables para ajustarlas a una escala estándar y facilitar su procesamiento en modelos matemáticos.

→ **Validación cruzada**: Método de evaluación de modelos de clasificación mediante la división de datos en subconjuntos de entrenamiento y prueba.

→ **Segmentación de imágenes**: División de una imagen en regiones homogéneas para facilitar el análisis de las características de los granos.

→ **Modelo de regresión logística**: Algoritmo estadístico utilizado para predecir la probabilidad de pertenencia a una clase.

→ **Optimización de hiperparámetros**: Proceso de ajuste de los parámetros de un modelo de aprendizaje automático para mejorar su desempeño.

→ **Evaluación del sesgo del modelo**: Análisis de posibles desviaciones en los resultados del modelo debido a limitaciones en los datos de entrenamiento.

→ **División en k-folds**: Estrategia de validación cruzada en la que los datos se dividen en k subconjuntos para mejorar la robustez del modelo.

→ **Errores de clasificación**: Diferencias entre las predicciones del modelo y las categorías reales en los datos de prueba.

→ **Software de análisis estadístico**: Herramientas computacionales utilizadas para realizar pruebas y evaluar modelos de clasificación.

Capítulo IV.

→ **Estadística descriptiva**: Rama de la estadística que se encarga de recopilar, organizar, resumir y presentar datos de manera informativa.

→ **Medidas de tendencia central**: Indicadores estadísticos que resumen un conjunto de datos en un valor representativo, como la media, mediana y moda.

→ **Medidas de dispersión**: Parámetros estadísticos que describen la variabilidad de los datos, como la desviación estándar, varianza y coeficiente de variación.

→ **Distribución de frecuencias**: Representación de cómo se distribuyen los valores de una variable dentro de un conjunto de datos.

→ **Asimetría**: Medida estadística que indica el grado de simetría de la distribución de los datos en torno a su media.

→ **Curtosis**: Indicador que describe la forma de la distribución de los datos en relación con la normalidad, determinando si tiene colas más pesadas o ligeras.

→ **Inferencia estadística**: Conjunto de métodos que permiten realizar conclusiones sobre una población a partir de una muestra representativa.

→ **Prueba de hipótesis**: Procedimiento estadístico para determinar si existe suficiente evidencia en una muestra para inferir una afirmación sobre la población.

→ **Nivel de significancia**: Probabilidad de cometer un error tipo I en una prueba estadística, generalmente expresado como un valor alfa (α).

→ **Intervalo de confianza**: Rango de valores dentro del cual se espera que se encuentre un parámetro poblacional con un cierto nivel de confianza.

→ **Prueba t de Student**: Método estadístico utilizado para comparar las medias de dos grupos y determinar si las diferencias observadas son significativas.

→ **Análisis de varianza (ANOVA)**: Técnica estadística que permite comparar las medias de tres o más grupos para detectar diferencias significativas.

→ **Coeficiente de correlación**: Medida que evalúa la relación y la fuerza de asociación entre dos variables cuantitativas.

→ **Regresión lineal**: Modelo estadístico utilizado para analizar la relación entre una variable dependiente y una o más variables independientes.

→ **P-valor**: Probabilidad de obtener un resultado tan extremo como el observado, asumiendo que la hipótesis nula es verdadera.

→ **Residuales**: Diferencias entre los valores observados y los valores predichos por un modelo estadístico.

→ **Análisis de componentes principales (PCA)**: Técnica de reducción de dimensionalidad que transforma variables correlacionadas en un conjunto de variables no correlacionadas.

→ **Tamaño del efecto**: Medida de la magnitud de una relación o diferencia estadística entre grupos.

→ **Distribución normal**: Patrón de distribución de datos en forma de campana donde la mayor parte de los valores se agrupan alrededor de la media.

→ **Homocedasticidad**: Suposición estadística de que la varianza de los errores en un modelo es constante en todos los niveles de una variable independiente.

→ **Multicolinealidad**: Situación en la que dos o más variables independientes en un modelo de regresión están altamente correlacionadas entre sí.

Capítulo V.

→ **Medidas de rendimiento**: Indicadores utilizados para evaluar la eficacia de un modelo de aprendizaje automático o estadístico.

→ **Exactitud (Accuracy)**: Porcentaje de predicciones correctas realizadas por un modelo en relación con el total de predicciones.

→ **Precisión (Precision)**: Proporción de predicciones positivas correctas en relación con todas las predicciones positivas realizadas.

→ **Sensibilidad (Recall o Tasa de Verdaderos Positivos)**: Capacidad de un modelo para identificar correctamente los casos positivos dentro de la muestra.

→ **Especificidad**: Capacidad de un modelo para identificar correctamente los casos negativos dentro de la muestra.

→ **Puntuación F1 (F1-score)**: Media armónica entre la precisión y la sensibilidad, útil para evaluar modelos con datos desbalanceados.

→ **Coeficiente Kappa de Cohen**: Medida de concordancia entre observaciones reales y predichas, ajustando por la probabilidad de coincidencia aleatoria.

→ **Curva ROC (Receiver Operating Characteristic)**: Gráfico que evalúa el desempeño de un modelo de clasificación mediante la tasa de verdaderos positivos y la tasa de falsos positivos.

→ **Área Bajo la Curva (AUC-ROC)**: Medida de la capacidad de discriminación de un modelo de clasificación.

→ **Overfitting (Sobreajuste)**: Situación en la que un modelo se ajusta demasiado a los datos de entrenamiento y pierde capacidad de generalización.

→ **Underfitting (Subajuste)**: Ocurre cuando un modelo es demasiado simple y no capta patrones significativos en los datos.

→ **Regularización**: Técnica utilizada para reducir el sobreajuste en los modelos de aprendizaje automático.

→ **Evaluación cruzada (Cross-validation)**: Método de validación que divide los datos en subconjuntos de entrenamiento y prueba para mejorar la generalización del modelo.

→ **Errores tipo I y tipo II**: Tipo I ocurre cuando se rechaza incorrectamente una hipótesis nula verdadera (falso positivo). Tipo II ocurre cuando no se rechaza una hipótesis nula falsa (falso negativo).

→ **Rendimiento computacional**: Evaluación de la eficiencia de un modelo en términos de velocidad de ejecución y consumo de recursos.

→ **Generalización del modelo**: Capacidad de un modelo de aprendizaje automático para realizar predicciones precisas sobre datos no vistos.

→ **Comparación de modelos**: Proceso de análisis para seleccionar el modelo más adecuado en función de métricas de rendimiento.

→ **Benchmarking en modelos de ML**: Evaluación comparativa de diferentes modelos de aprendizaje automático en términos de rendimiento y eficiencia.

ISBN: 978-9942-7264-6-9

1era Edición