

UNIVERSIDAD NACIONAL PEDRO RUIZ GALLO

FACULTAD DE AGRONOMÍA

ESCUELA PROFESIONAL DE AGRONOMÍA



TESIS PARA OPTAR EL TÍTULO PROFESIONAL DE:

INGENIERO AGRÓNOMO

**“Modelos de machine learning para el pronóstico del peso de bayas de
arándano *Vaccinium corymbosum* L.”**

INVESTIGADORES:

Cancino Chavez, Gustavo Franco

Quispe Solano, Nilson

ASESORA:

M. Sc. Peralta Inga, Maruja

Lambayeque, 2026

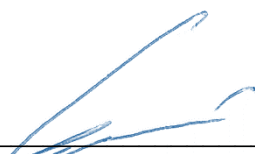
Modelos de machine learning para el pronóstico del peso de bayas de arándano
Vaccinium corymbosum L.

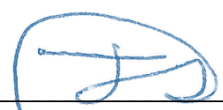
POR:

Cancino Chavez, Gustavo Franco
Quispe Solano, Nilson

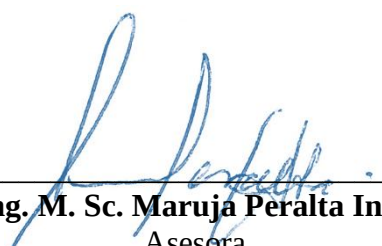
Presentada a la Facultad de Agronomía de la Universidad Nacional Pedro Ruiz gallo, para
optar el Título Profesional de:
INGENIERO AGRÓNOMO

APROBADO POR:



Ing. M. Sc. Eduardo Exequiel Deza León
Presidente

Ing. Juan Elías Villegas Cubas
Secretario

Ing. Mg. Víctor Gustavo Hernández Jiménez
Vocal

Ing. M. Sc. Maruja Peralta Inga
Asesora

LAMBAYEQUE, 2026



UNIVERSIDAD NACIONAL PEDRO RUIZ GALLO
FACULTAD DE AGRONOMÍA
UNIDAD DE INVESTIGACIÓN



ACTA DE SUSTENTACIÓN DE TESIS N°027-2026-D-FAG

En la ciudad de Lambayeque a los siete días del mes de mayo del año dos mil veintiséis, siendo las ocho y media de la mañana, se reunieron en el Auditorio de la Facultad de Agronomía los Miembros del Jurado evaluador de la tesis titulada: **"MODELOS DE MACHINE LEARNING PARA EL PRONÓSTICO DEL PESO DE BAYAS DE ARÁNDANO *Vaccinium corymbosum* L."**, mediante Resolución N°141-2025-D-FAG, de fecha 16 de junio del 2025, se nombra el jurado de tesis; con Resolución N°261-2025-D-FAG, de fecha 24 de octubre del 2025, se aprueba el proyecto de Tesis; con la finalidad de evaluar y calificar la Sustentación de la Tesis antes mencionada; el jurado está conformado por los siguientes docentes:

M.Sc. Eduardo Exequiel Deza León
 Ing. Juan Elías Villegas Cubas
 M.Sc. Víctor Gustavo Hernández Jiménez
 M.Sc. Maruja Peralta Inga

Presidente
 Secretario
 Vocal
 Patrocinadora

El acto de Programación de Sustentación fue autorizado por **RESOLUCIÓN N°091-2026-D-FAG**, de fecha 04 de mayo del 2026.

La Tesis fue presentada y sustentada por el Bachiller **CANCINO CHAVEZ GUSTAVO FRANCO**, tuvo una duración de 70 minutos. Después de la sustentación y absueltas las preguntas de los Miembros de Jurado, se procedió a la calificación respectiva otorgándole el calificativo de 1.8 en la escala vigesimal, con mención MUY BUENO.

Por lo que queda **APTO (A)** para obtener el Título Profesional de Ingeniero (a) Agrónomo (a) de acuerdo con la Ley Universitaria N° 30220 y el Art. 46° del Reglamento de Grados y Títulos de la Universidad Nacional Pedro Ruiz Gallo.

Siendo las 9:40 AM, se dio por concluido el presente acto académico, dándose conformidad el presente acto con las firmas de los Miembros de Jurado.

M.Sc. Eduardo Exequiel Deza León
 Presidente

Ing. Juan Elías Villegas Cubas
 Secretario

M.Sc. Víctor Gustavo Hernández Jiménez
 Vocal

M.Sc. Maruja Peralta Inga
 Patrocinadora

Observación:



UNIVERSIDAD NACIONAL PEDRO RUIZ GALLO

FACULTAD DE AGRONOMÍA

UNIDAD DE INVESTIGACIÓN



ACTA DE SUSTENTACIÓN DE TESIS N°026-2026-D-FAG

En la ciudad de Lambayeque a los siete días del mes de mayo del año dos mil veintiséis, siendo las ocho y media de la mañana, se reunieron en el Auditorio de la Facultad de Agronomía los Miembros del Jurado evaluador de la tesis titulada: **"MODELOS DE MACHINE LEARNING PARA EL PRONÓSTICO DEL PESO DE BAYAS DE ARÁNDANO *Vaccinium corymbosum* L."**, mediante Resolución N°141-2025-D-FAG, de fecha 16 de junio del 2025, se nombra el jurado de tesis; con Resolución N°261-2025-D-FAG, de fecha 24 de octubre del 2025, se aprueba el proyecto de Tesis; con la finalidad de evaluar y calificar la Sustentación de la Tesis antes mencionada; el jurado está conformado por los siguientes docentes:

M.Sc. Eduardo Exequiel Deza León

Ing. Juan Elías Villegas Cubas

M.Sc. Víctor Gustavo Hernández Jiménez

M.Sc. Maruja Peralta Inga

Presidente

Secretario

Vocal

Patrocinadora

El acto de Programación de Sustentación fue autorizado por RESOLUCIÓN N°091-2026-D-FAG, de fecha 04 de mayo del 2026.

La Tesis fue presentada y sustentada por el Bachiller **QUISPE SOLANO NILSON**, tuvo una duración de 70 minutos. Después de la sustentación y absueltas las preguntas de los Miembros de Jurado, se procedió a la calificación respectiva otorgándole el calificativo de 12 en la escala vigesimal, con mención MUY BUENO.

Por lo que queda **APTO (A)** para obtener el Título Profesional de Ingeniero (a) Agrónomo (a) de acuerdo con la Ley Universitaria N° 30220 y el Art. 46° del Reglamento de Grados y Títulos de la Universidad Nacional Pedro Ruiz Gallo.

Siendo las 9:40 AM, se dio por concluido el presente acto académico, dándose conformidad el presente acto con las firmas de los Miembros de Jurado.

M.Sc. Eduardo Exequiel Deza León
Presidente

Ing. Juan Elías Villegas Cubas
Secretario

M.Sc. Víctor Gustavo Hernández Jiménez
Vocal

M.Sc. Maruja Peralta Inga
Patrocinadora

Observación:

.....

.....

.....

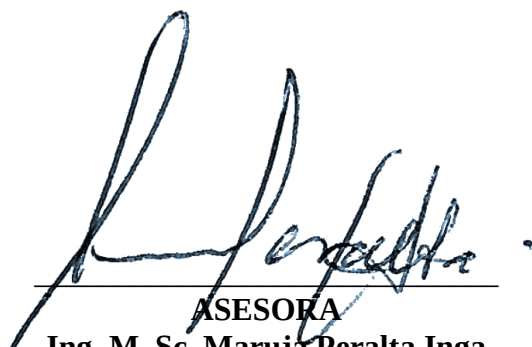
CONSTANCIA DE VERIFICACIÓN DE ORIGINALIDAD

Yo Maruja Peralta Inga usuario revisor de Tesis Titulado: Modelos de machine learning para el pronóstico del peso de bayas de arándano *Vaccinium corymbosum* L. Cuyos autores son; Gustavo Franco, Cancino Chavez con DNI N° 47116114 y Nilson, Quispe Solano con DNI N° 47822890; declaro que la evaluación realizada por el programa informático, ha arrojado un porcentaje de similitud 14%, verificables en el Resumen del Reporte automatizado de similitudes que se acompaña.

El suscrito analizó reporte y concluyó que cada una de las coincidencias detectadas dentro del porcentaje de similitud permitido no constituye plagio y que el documento cumple con la integridad científica y con las normas para el uso de citas y referencias establecidas en los protocolos respectivos.

Se cumple con adjuntar el Recibo Digital a efectos de la trazabilidad respectiva del proceso.

Lambayeque, 4 de mayo del 2026



ASESORA
Ing. M. Sc. Maruja Peralta Inga
DNI: 16725113

REPORTE DE INFORME DE SIMILITUD

Modelos de machine learning para el pronóstico del peso de bayas de arándano *Vaccinium corymbosum* L.

INFORME DE ORIGINALIDAD

14%	12%	2%	4%
INDICE DE SIMILITUD	FUENTES DE INTERNET	PUBLICACIONES	TRABAJOS DEL ESTUDIANTE

FUENTES PRIMARIAS

1	repositorio.lamolina.edu.pe Fuente de Internet	2%
2	repositorio.uct.edu.pe Fuente de Internet	2%
3	Submitted to Universidad Nacional Agraria La Molina Trabajo del estudiante	1%
4	hdl.handle.net Fuente de Internet	1%
5	apirepositorio.unu.edu.pe Fuente de Internet	<1%
6	Submitted to Universidad Nacional de Trujillo Trabajo del estudiante	<1%
7	Submitted to Universidad Nacional Pedro Ruiz Gallo Trabajo del estudiante	<1%
8	repositorio.unjfsc.edu.pe Fuente de Internet	<1%
9	repositorio.esan.edu.pe Fuente de Internet	<1%
10	repositorio.unprg.edu.pe:8080 Fuente de Internet	<1%
11	revistas.pucp.edu.pe Fuente de Internet	<1%

ASESORA

Ing. M. Sc. Maruja Peralta Inga
DNI: 16725113

RECIBO DIGITAL



Recibo digital

Este recibo confirma que su trabajo ha sido recibido por Turnitin. A continuación podrá ver la información del recibo con respecto a su entrega.

La primera página de tus entregas se muestra abajo.

Autor de la entrega: GUSTAVO FRANCO CANCINO CHAVEZ NILSON QUISPE SOLANO
 Título del ejercicio: Pregrado
 Título de la entrega: Modelos de machine learning para el pronóstico del peso de b...
 Nombre del archivo: Tesis_Proyeccion_de_Peso_BB_29-03-26.docx
 Tamaño del archivo: 2.01M
 Total páginas: 83
 Total de palabras: 18,663
 Total de caracteres: 107,123
 Fecha de entrega: 04-may-2026 10:19a.m. (UTC-0500)
 Identificador de la entrega: 2951237229



UNIVERSIDAD NACIONAL
 PEDRO RUIZ GALLO
 FACULTAD DE AGRONOMÍA
 ESCUELA PROFESIONAL DE AGRONOMÍA



TESIS

Modelos de machine learning para el pronóstico del peso de bovinos
 (Tesis de grado)

PARA OBTENER EL TÍTULO DE:
 Ingeniero agrónomo

Asesor por:

Ing. M. Sc. Maruja Peralta Inga
 DNI: 16725113

Asesor:

M. Sc. Francisco Márquez

Carla Rodríguez Pita

2024

ASESORA

Ing. M. Sc. Maruja Peralta Inga
 DNI: 16725113

Derechos de autor 2026 Turnitin. Todos los derechos reservados.

Dedicatoria

DEDICATORIA

Dedico este trabajo en primer lugar a Dios por haberme dado la vida y permitirme llegar hasta este momento tan importante de mi formación humana y profesional.

A mis padres Leonel Cancino Sánchez y Celia Chavez Chavez, por su amor, apoyo incondicional y sacrificio durante este tiempo que me permitió llegar a esta meta académica, gracias a ustedes por enseñarme el valor de la perseverancia y el rigor.

Cancino Chávez, Gustavo Franco

Dedicado a mis padres Toribio Quispe Herrera y Teresa Amelia Solano Alejandría; por su constante apoyo en mi formación profesional, fueron un pilar fundamental para poder lograr este objetivo.

Quispe Solano, Nilson

Agradecimientos

Agradezco a Dios, a mi familia y hermanos por brindarme su apoyo, motivación y paciencia durante todo este proceso de formación profesional necesario para alcanzar esta meta, también expreso mi más profundo agradecimiento a mi asesora por su invaluable guía, apoyo, paciencia y rigor académico durante este camino largo de mi proyecto de tesis.

Cancino Chávez, Gustavo Franco

Expreso mi más sincero agradecimiento a mis padres, mi hermano Juan Gabriel Quispe Solano, quienes con su esfuerzo, sacrificio y fortaleza necesaria me guiaron ahi alcanzar mis objetivos académicos.

A si mismo agradezco a mi asesora de tesis, por su orientación, paciencia y aporte académico durante el proceso de investigación.

Quispe Solano, Nilson

Resumen

Considerando la relevancia económica del calibre como indicador de calidad y precio, en este estudio se tuvo como objetivo desarrollar modelos de machine learning capaces de estimar con precisión el peso de las bayas de arándano a partir de su calibre. Se construyó un preprocesamiento reproducible mediante el paquete `recipes` de `tidymodels` para garantizar la calidad e integridad de los registros de calibre y peso de bayas de arándano obtenidos durante la campaña 2023-2024 en Lambayeque, Perú. El análisis exploratorio evidenció una correlación casi perfecta entre calibre y peso ($r=0,98$) y una tendencia decreciente del peso a lo largo del periodo ($r=-0,70$). Tras aplicar transformaciones logarítmicas y selección de variables mediante regresión Lasso, se evaluaron cuatro enfoques de predicción: regresión lineal simple, regresión polinómica de segundo grado, árbol de decisión y bosque aleatorio. La calibración de hiperparámetros de los modelos de árbol se realizó mediante optimización bayesiana basada en RMSE y su desempeño se comparó usando RMSE, MAE, MAPE y R^2 en validación cruzada repetida ($5 \text{ folds} \times 20 \text{ repeticiones}$) y en prueba externa. El modelo polinómico mostró el mejor equilibrio entre precisión y robustez, con un MAE de 0,05 g, un MAPE cercano al 2,5 % y $R^2=0,97$ en ambos escenarios. A pesar de que el bosque aleatorio alcanzó resultados competitivos, la simplicidad y transparencia del modelo polinómico facilitaron su implementación en entornos productivos para optimizar la planificación de cosecha y la clasificación de frutos. Estos hallazgos confirmaron que, ante relaciones casi lineales, los métodos paramétricos bien calibrados pueden igualar o superar el rendimiento de algoritmos más complejos.

Palabras clave: arándano, calibre de baya, peso de fruto, aprendizaje automático, regresión polinómica, bosque aleatorio.

Abstract

Machine learning models for forecasting the berry weight of *Vaccinium corymbosum* L. blueberry, Olmos District, Lambayeque Province and Region, 2025.

Considering the economic relevance of size as an indicator of quality and price, this study aimed to develop machine learning models capable of accurately estimating blueberry berry weight from its size. A reproducible preprocessing was constructed using the recipes package of tidymodels to ensure the quality and integrity of the size and weight records of blueberry berries obtained during the 2023–2024 season in Lambayeque, Peru. Exploratory analysis revealed an almost perfect correlation between size and weight ($r = 0.98$) and a decreasing trend in weight over the period ($r = -0.70$). After applying logarithmic transformations and variable selection via Lasso regression, four prediction approaches were evaluated: simple linear regression, second-degree polynomial regression, decision tree, and random forest. Hyperparameter calibration of the tree models was performed by Bayesian optimization based on RMSE, and their performance was compared using RMSE, MAE, MAPE, and R^2 in repeated cross-validation (5 folds $\times$ 20 repetitions) and external testing. The polynomial model demonstrated the best balance between accuracy and robustness, achieving an MAE of 0.05 g, a MAPE close to 2.5 %, and $R^2 = 0.97$ in both scenarios. Although the random forest produced competitive results, the simplicity and transparency of the polynomial model facilitated its implementation in production environments to optimize harvest planning and fruit classification. These findings confirmed that, given nearly linear relationships, well-calibrated parametric methods can match or outperform more complex algorithms.

Key words: blueberry, berry size, fruit weight, machine learning, polynomial regression, random forest.

Índice

Dedicatoria	3
Agradecimientos.....	9
Resumen	10
Abstract	11
Índice	12
Índice de tablas	13
Índice de figuras.....	14
I. INTRODUCCIÓN	1
II. DISEÑO TEÓRICO	3
2.1. Antecedentes de la investigación	3
2.2. Bases teóricas	8
2.3. Bases conceptuales	28
III. DISEÑO METODOLÓGICO	30
3.1. Tipo y nivel de investigación.....	30
3.2. Diseño de investigación.....	30
3.3. Población y muestra.....	31
3.4. Técnicas e instrumentos de recolección de datos.	31
3.5. Procedimientos.....	32
3.6. Plan de procesamiento y análisis de datos.....	32
3.7. Matriz de consistencia.....	34
IV. RESULTADOS	35
4.1. Análisis exploratorio	36
4.2. Pre-procesamiento de los datos	41
4.3. Validación de modelos	44
V. DISCUSIÓN DE RESULTADOS.....	56
VI. CONCLUSIONES	61
VII. RECOMENDACIONES	62
VIII. REFERENCIAS.....	63

Índice de tablas

Tabla 1. <i>Operacionalización de variables</i>	29
Tabla 2. <i>Matriz de consistencia</i>	34
Tabla 3. <i>Diccionario de variables</i>	37

Índice de figuras

Figura 1. <i>Relación entre el peso de baya y las variables cuantitativas.</i>	37
Figura 2. <i>Relación entre el peso de baya y la variable fecha.</i>	38
Figura 3. <i>Relación entre el peso de baya y la variable mes.</i>	39
Figura 4. <i>Métricas de poder predictivo de las variables exógenas.</i>	40
Figura 5. <i>Distribución y función de probabilidad acumulada del peso de bayas en los conjuntos de datos original, de entrenamiento y de prueba.</i>	41
Figura 6. <i>Comparación de valores predichos y observados en validación cruzada para los modelos.</i>	45
Figura 7. <i>Comparación de valores predichos y observados en el conjunto de prueba para los modelos.</i>	47
Figura 8. <i>Distribución de errores frente al peso real de la baya en el conjunto de prueba.</i>	49
Figura 9. <i>Distribución de errores porcentuales frente al peso real de la baya en el conjunto de prueba.</i>	51
Figura 10. <i>Distribución de métricas de desempeño en validación cruzada y prueba para los modelos evaluados.</i>	53

I. INTRODUCCIÓN

El cultivo del arándano representa una actividad agrícola de alta importancia económica debido a la creciente demanda global de esta fruta, reconocida por sus propiedades nutricionales y su valor en mercados internacionales (Tinoco et al., 2023), siendo la calidad y el peso de las bayas factores clave que influyen directamente en los precios y en la competitividad de los productores (Zapata et al., 2013). Dentro de las características físicas de las bayas, el calibre es un indicador importante que no solo determina su aceptación comercial, sino que también está relacionado con el rendimiento y la eficiencia de la cosecha.

Actualmente, el pronóstico del peso de las bayas basado en el calibre enfrenta desafíos significativos debido a la variabilidad intrínseca de los cultivos, influenciada por factores genéticos, ambientales y manejo agronómico. Los métodos tradicionales de estimación, basados en promedios generales o modelos lineales simples, no logran capturar adecuadamente la complejidad del comportamiento del peso de la baya, generando limitaciones en la planificación de la cosecha, el manejo poscosecha y la toma de decisiones comerciales (Defilippi et al., 2023).

En este contexto, el uso de técnicas de machine learning ha cobrado relevancia al permitir modelar relaciones no lineales propias de los sistemas agrícolas. Sin embargo, su aplicación específica para predecir el peso de la baya de arándano a partir del calibre aún es limitada, lo que evidencia una brecha en el conocimiento (Muñoz, 2024).

Frente a ello, la presente investigación se orientó al desarrollo y evaluación de modelos de machine learning para estimar el peso de la baya (g) en función del calibre, con el propósito de mejorar la precisión del pronóstico y aportar herramientas útiles para la toma de decisiones en la producción de arándano. A partir de ello, se planteó la siguiente pregunta

de investigación: ¿Cuál modelo de machine learning permite predecir el peso de baya (en g) del arándano en función de su calibre?

Se estableció como objetivo general:

- Desarrollar modelos predictivos basados en técnicas de machine learning para pronosticar con precisión el peso de bayas de arándano en función de su calibre.

Y los objetivos específicos fueron:

- Preprocesar los datos históricos de calibre y peso de bayas de arándano para mejorar la calidad de los datos y el desempeño de los modelos predictivos.
- Evaluar el desempeño de los modelos predictivos basados en técnicas de machine learning con los modelos estadísticos tradicionales para identificar la técnica más precisa y robusta.
- Seleccionar el modelo predictivo con mejor desempeño mediante la evaluación de diferentes algoritmos de machine learning, optimizando sus hiperparámetros.

II. DISEÑO TEÓRICO

2.1. Antecedentes de la investigación

Dávila y Vásquez (2025) desarrollaron una investigación orientada a predecir los calibres semanales de bayas de arándano (*Vaccinium corymbosum* L.) empleando variables fenológicas y técnicas de aprendizaje automático. Se llevó a cabo un riguroso procesamiento de datos que incluyó la limpieza, transformación y codificación de variables mediante técnicas como *One-Hot Encoding*, la generación de variables temporales y la transformación con *natural splines*. El conjunto de datos incluyó observaciones fenológicas clave como el número de frutos cuajados por planta, las cuales fueron seleccionadas con base en su poder predictivo. La validación del modelo se estructuró en dos niveles: una validación interna y otra externa, ambas bajo esquemas de partición tipo Hold-Out para series temporales y validación cruzada temporal (*Time Series Cross Validation*), garantizando una adecuada evaluación de la capacidad de generalización. Se probaron tres enfoques principales de modelado: regresión lineal múltiple, *Random Forest* (RF) y *Modelos Mixtos Aditivos Generalizados* (GAMM). Entre estos, el algoritmo *Random Forest* mostró el mejor desempeño predictivo, alcanzando un error porcentual absoluto medio (MAPE) de $2.44 \pm 0.13\%$, superando consistentemente a los otros modelos tanto en la fase de validación cruzada como en los conjuntos de prueba internos y externos. Este resultado confirmó su capacidad para mantener un bajo sesgo y alta precisión dentro del rango de calibres históricos observados. Se concluyó que el éxito en la predicción de calibres depende en gran medida de una adecuada ingeniería de variables y del uso de datos fenológicos relevantes.

Li et al. (2025) desarrollaron en su investigación una herramienta de fenotipado automatizado para estimar el número, tamaño y peso promedio de bayas de arándano mediante visión computacional, con el propósito de superar las limitaciones del conteo

manual en la investigación y mejoramiento genético de esta fruta. El estudio propuso dos enfoques de procesamiento de imágenes, uno tradicional basado en transformadas de Hough y segmentación Watershed, y otro basado en redes neuronales profundas, específicamente el modelo YOLOv5s modificado con los módulos GhostNet y biFPN para mejorar la eficiencia en dispositivos móviles. Se recolectaron 198 imágenes provenientes de tres conjuntos de datos que incluían cultivares comerciales y genotipos avanzados. Las imágenes fueron tomadas en condiciones controladas, y cada baya fue contada y pesada manualmente para obtener datos de referencia. El modelo tradicional usó detección de círculos para identificar las bayas, seguido de filtros de color y forma para excluir falsos positivos. En paralelo, se entrenaron modelos YOLOv5 y sus variantes modificadas, utilizando técnicas de aumento de datos y normalización de tamaño para mejorar la generalización. Para las pruebas en móviles, los modelos fueron convertidos a formato TensorFlow Lite con cuantización de 16 bits. Los resultados mostraron que el modelo basado en YOLOv5 modificado logró una precisión promedio (mAP) de hasta 93.5%, y una correlación superior al 96% entre el tamaño estimado de las bayas y su peso medido, con un error absoluto medio inferior a 0.10 gramos. El modelo basado en YOLOv5s, con modificaciones GhostNet y biFPN (YOLOv5s-C), ofreció un balance óptimo entre velocidad de inferencia (143.5 ms por imagen), tamaño del modelo (5.84 millones de parámetros) y precisión (mAP₅₀₋₉₅ de 91.9%). En conclusión, el estudio demostró que la visión computacional combinada con redes neuronales profundas puede reemplazar efectivamente las prácticas manuales de fenotipado de bayas de arándano.

Alegría (2024), con el objetivo de determinar un modelo de regresión lineal para estimar el peso del arándano en función de su calibre en los arándanos maduros en la Parcela 71 del Fundo Agricultor en Chao, realizó una investigación explicativa y experimental, con un modelo de regresión para cuantificar la relación entre el peso y el calibre del arándano.

Se evaluó una población conformada por todos los frutos de arándano maduro en el periodo de tiempo del 21/08/2023 – 06/09/2023 de la Parcela 71. Se empleó la técnica del análisis de regresión lineal. Llegó a concluir que, en cada uno de los 10 momentos evaluados, la correlación superó el 90%. Posteriormente, se determinó que el modelo de regresión lineal utilizado para predecir el peso del arándano a partir de su calibre, en frutos maduros cultivados en la Parcela 71, presenta características de modelo potencial. Además, se comprobó que los coeficientes de este modelo son estadísticamente equivalentes en todas las mediciones realizadas. Finalmente, se verificaron los supuestos del modelo de regresión lineal, lo que permitió confirmar su eficacia para estimar adecuadamente el peso del arándano.

Niedbala et al. (2022) desarrollaron modelos predictivos para estimar el rendimiento del arándano (*Vaccinium corymbosum* L.) utilizando algoritmos de aprendizaje automático, bajo un enfoque aplicado y no experimental. Concluyen que el algoritmo con mejor desempeño fue Extreme Gradient Boosting, alcanzando un error porcentual absoluto medio (MAPE) del 12,48%, nivel considerado adecuado para su uso práctico. Además, resaltó la flexibilidad de los modelos mediante la incorporación de datos recolectados en nuevas campañas, lo que fortalece su desempeño y aplicabilidad en operaciones agrícolas.

Qu et al. (2021) realizaron una investigación cuantitativa, no experimental y longitudinal, con el objetivo de analizar los factores y metodologías asociados a la estimación del peso de los arándanos mediante técnicas modernas. Los autores concluyeron que, según las proyecciones climáticas, un incremento en las condiciones cálidas y húmedas podría reducir el cuajado de frutos del arándano silvestre en un 8%.

Obsie et al. (2020) en su investigación incorporaron a las variables climatológicas y la presencia de abejas en los cultivos de arándanos, que tienen el papel de incrementar el tamaño de los frutos al favorecer una mayor cantidad de semillas. Se utilizaron datos

generados por un modelo de polinización de arándanos silvestres, validado mediante observaciones de campo y datos experimentales recopilados en Maine (EE.UU.) durante los últimos 30 años. Se evaluaron la regresión lineal múltiple (MLR), árboles de decisión reforzados (BDT), bosques aleatorios (RF) y el refuerzo de gradiente extremo (XGBoost). Además, se realizó una selección de predictores, lo que permitió reducir la dimensionalidad de los datos. Se identificó que el tamaño del clon, la abundancia de abejorros, las especies de abejas *Andrena* y *Osmia*, el máximo de los rangos de temperatura superior y el número de días con precipitaciones fueron las variables predictoras más relevantes. Entre los modelos evaluados, XGBoost obtuvo el mejor desempeño, alcanzando un coeficiente de determinación (R^2) de 0,938, un error cuadrático medio (RMSE) de 343,026, un error absoluto medio (MAE) de 206 y un error cuadrático medio relativo (RRMSE) de 5,444%. Este estudio demostró que las predicciones del rendimiento de cultivos pueden basarse eficazmente en conjuntos de datos generados mediante modelado por ordenador, lo cual tiene un impacto significativo, especialmente en escenarios donde la recolección de datos de campo resulta desafiante.

Strik y Vance (2019) realizaron un estudio con el objetivo de determinar la relación entre el número de semillas por baya, el diámetro de la baya, la época de recolección y el peso de la baya. Los autores identificaron una relación lineal entre el peso de las bayas y la cantidad de semillas viables por fruto. En general, no se encontraron diferencias significativas en el peso de la baya por semilla (pendiente de la relación) entre los dos años analizados, con excepción de los cultivares 'Draper' (donde la diferencia fue significativa solo en 2009), 'Legacy' y 'Reka'. El cultivar 'Bluegold' presentó el mayor peso por semilla, mientras que 'Reka' registró el menor. Además, la relación más débil entre el peso de la baya y el número de semillas se observó en 'Draper'. Entre los cultivares que produjeron frutos partenocárpicos de tamaño comercial se incluyeron 'Aurora', 'Bluegold' y 'Liberty'. Por otro

lado, cultivares como 'Duke', 'Legacy', 'Reka' y 'Ozarkblue' generaron bayas comercialmente aceptables (0,75-1,0 g) con menos de siete semillas por fruto.

Jorquera et al. (2017) en su investigación desarrollaron un modelo alométrico para estimar el peso del fruto del arándano a partir de mediciones del diámetro. Se propuso un modelo que relacione el peso fresco con el diámetro del fruto utilizando una función de tipo potencia, con el objetivo de aplicar esta herramienta tanto en estudios fisiológicos como en la industria frutícola, donde el calibre del fruto incide directamente en el valor comercial. El estudio se realizó a lo largo de tres temporadas de crecimiento, recolectando un total de 416 frutos sanos de cuatro cultivares de arándano (Bluegold, Brigitta, Elliot y Legacy) en distintos huertos comerciales ubicados en la Región de La Araucanía, Chile. Se midió semanalmente el diámetro ecuatorial y el peso fresco de los frutos desde el cambio de color hasta el inicio de la cosecha comercial. Los datos se dividieron aleatoriamente en un conjunto de calibración ($n = 200$) y uno de validación externa ($n = 216$). A través de análisis de covarianza sobre los datos transformados logarítmicamente, se evaluó si era necesario construir modelos específicos por cultivar o si era factible usar un modelo común. Al no encontrarse efectos de interacción significativos entre diámetro y cultivar/temporada, se ajustó un único modelo general. El modelo alométrico ajustado mostró una relación funcional de tipo potencia entre el diámetro del fruto (FD) y el peso (FW), con una ecuación representativa del tipo $FW = a \cdot FD^b$, siendo $a = 0.0007$ y $b = 2.84$. La validación externa evidenció un buen ajuste del modelo, con errores relativos de raíz cuadrática media (RRMSE) que oscilaron entre 4.7% y 6.9% según el cultivar. Además, los residuos se distribuyeron aleatoriamente respecto a los valores estimados, lo que indicó homocedasticidad y validez estadística del modelo. Los resultados demostraron que la relación entre el peso y el diámetro del fruto es suficientemente estable entre cultivares y temporadas, lo que justifica el uso de un modelo común.

2.2. Bases teóricas

2.2.1. Arándano

2.2.1.1. Generalidades.

Existen tres variedades principales de arándanos cultivadas con fines comerciales. El arándano Lowbush (*Vaccinium angustifolium*) se produce principalmente en las regiones orientales de Canadá. Por otro lado, el arándano Rabbiteye (*Vaccinium ashei*) es típico de las zonas del sureste de los Estados Unidos. Finalmente, el arándano Highbush (*Vaccinium corymbosum*) es la especie más popular en Estados Unidos y, con el tiempo, se ha expandido a otras áreas del continente americano (Pritts y Hancock, 1992).

2.2.1.2. Origen.

El arándano (*Vaccinium spp.*) es un arbusto originario de América del Norte, clasificado dentro de la familia Ericaceae. Se conocen más de 400 variedades de esta planta distribuidas principalmente en regiones del hemisferio norte (Yadong, et al., 2009).

Las especies silvestres y comestibles de *Vaccinium spp.* han sido recolectadas por las comunidades indígenas durante miles de años. *V. corymbosum* fue domesticado por primera vez en 1908 por Frederick Coville, un investigador del Departamento de Agricultura de los Estados Unidos, quien determinó las condiciones esenciales para su cultivo, como la necesidad de suelos ácidos y bien drenados, la ausencia de pelos radiculares y la necesidad de un período de descanso a bajas temperaturas. Las variedades más comunes de arándano cultivado incluyen *Vaccinium corymbosum* L. (arándano de arbusto alto), *V. virgatum* Ait. (arándano ojo de conejo) y el *V. angustifolium* Ait. (arándano de arbusto bajo). Según sus requerimientos de frío y resistencia al invierno, los cultivares de arbusto alto se clasifican en tres tipos: del norte, del sur y tipos intermedios (Retamales y Hancock, 2018).

Las variedades de arándanos cultivados fueron introducidas en América del Sur en los años 80 para analizar su viabilidad como nuevo cultivo en la región. Para 1993, Chile

contaba con 580 hectáreas dedicadas al cultivo de arándanos, mientras que Argentina solo tenía algunas hectáreas en producción comercial. Desde ese momento, la industria del arándano creció rápidamente en ambos países. En 2004, Chile alcanzó las 2500 hectáreas y exportó 9666 toneladas de fruta, mientras que Argentina tenía 1200 hectáreas y exportaba 900 toneladas. Este crecimiento se explicó principalmente por los altos precios del mercado internacional de fruta fresca fuera de temporada, especialmente en América del Norte y Europa, además de la creciente demanda mundial. La cosecha de arándanos en Argentina y Chile se realiza entre finales de septiembre y abril, y más del 95% de la producción se destina a la exportación como fruta fresca (Bañados, 2006).

2.2.1.3. Importancia económica en Perú.

La producción de arándanos en Perú comenzó recientemente, con las primeras plantaciones establecidas entre 2006 y 2007. Fue en 2007 cuando se introdujo esta fruta, y su cosecha empezó a partir de 2009, aunque inicialmente en pequeñas cantidades. En 2012, el cultivo ocupaba solo 300 hectáreas, y ese mismo año se exportaron 40 toneladas de arándanos, generando ingresos cercanos a los US\$ 400,000, siendo los Países Bajos el principal destino. A partir de 2015, la producción aumentó considerablemente, alcanzando las 10.7 mil toneladas. Este crecimiento continuó a un ritmo acelerado, registrando 52.3 mil toneladas en 2017, 84.8 mil toneladas en 2018 y alcanzando las 145.6 mil toneladas en 2019, un récord. Además, la mayor parte de las hectáreas destinadas al cultivo de arándanos se encuentran en los departamentos de Lambayeque y La Libertad, que en conjunto representan el 97% de la producción total de arándanos en el país (MIDAGRI, 2020).

En los últimos años, Perú ha experimentado un crecimiento significativo en la producción y exportación de arándanos, consolidándose como líder mundial en este sector. Entre 2018 y 2021, la superficie sembrada de arándanos aumentó de 7,884 a 16,850 hectáreas, alcanzando una producción superior a 261,000 toneladas, con un rendimiento

promedio de 19,206 kilogramos por hectárea, el más alto registrado a nivel global. Las exportaciones de arándanos frescos casi se triplicaron en el mismo periodo, pasando de 85,600 a 227,500 toneladas, fortaleciendo la posición de Perú como principal exportador mundial (Agraria, 2021).

En la campaña de 2022, el arándano se consolidó como una de las principales exportaciones de Perú, con un volumen de 270,000 toneladas métricas y un valor FOB de USD 1,435 millones, lo que representó un incremento del 34% en volumen y del 16% en valor respecto al 2021. Sin embargo, se observó una disminución del 13% en el precio promedio, que varió entre USD 4.88 y USD 6.17 por kilo en la campaña anterior. En la campaña 2023, debido a los efectos del fenómeno El Niño costero, hubo un impacto negativo sobre la producción de la variedad Ventura, una variedad que en esa campaña representó el 30% del área sembrada del país. Las temperaturas elevadas, que alcanzaron los 32 °C entre marzo y abril, provocaron un crecimiento vegetativo excesivo, retrasando la floración de seis a ocho semanas y desplazando el pico de producción de setiembre-octubre a octubre-noviembre. Además, se observó una disminución en el rendimiento debido a una menor cantidad de flores por yema y a la activación limitada de yemas por rama (Gómez, 2023).

Durante la campaña 2024-2025, la producción y exportación de arándanos peruanos han continuado mostrando características destacadas, con un crecimiento del 61% en las exportaciones acumuladas hasta la fecha, alcanzando 254,406 toneladas métricas. De este total, el 89% corresponde a producción convencional y el 11% a orgánica. Se proyecta un cierre de campaña con 323,928 toneladas exportadas, superando en un 4.1% las estimaciones iniciales. Los principales destinos incluyen Estados Unidos (53%), Europa (25%), China (14%), Reino Unido (5%) y otros mercados (4%), con el 97% del volumen transportado por vía marítima a través de puertos clave como Philadelphia, Rotterdam y Shenzhen. Las principales regiones productoras son La Libertad (49%), Lambayeque (24%) e Ica (12%),

mientras que las variedades más representativas incluyen Ventura (26%), Biloxi (16%) y Sekoya Pop (14%). Destaca la creciente diversificación genética, con aportes significativos de Fallcreek (32%) y Sekoya (16%), y un cambio en las tendencias de las áreas cultivadas, con una disminución en la participación de Biloxi y un notable aumento en variedades emergentes como Sekoya Pop (Portal Frutícola, 2024).

2.2.1.4. Clasificación taxonómica.

Desde una perspectiva taxonómica, el arándano se clasifica dentro del reino Plantae, en la división Magnoliophyta, perteneciendo a la clase Dicotiledónea. Su orden es Ericales, su familia Ericaceae, y forma parte de la subfamilia Vaccinioideae. Además, se incluye en la tribu Vaccinieae, sección Cyanococcus, dentro del género *Vaccinium* y corresponde a la especie *Vaccinium corymbosum* L. (Retamales y Hancock, 1992).

2.2.1.5. Cultivares.

2.2.1.5.1. Ventura.

Variedad de porte alto que produce frutos grandes y firmes, con piel más gruesa que ‘Springhigh’ y color más claro. Presenta crecimiento erguido y vigoroso, con buen rendimiento tanto en costa como en sierra del Perú. Su maduración es temprana y no requiere un número específico de horas de frío para su desarrollo (Fall Creek, 2021).

2.2.1.5.2. Biloxi.

Se caracteriza por su alto rendimiento en etapas iniciales, con follaje denso y estructura robusta. Produce frutos de tamaño medio, color azul claro, buena firmeza y sabor. Su crecimiento es vertical y vigoroso, con requerimientos cercanos a 400 horas de frío. El tamaño de las bayas puede variar, destacando una particularidad en el núcleo del fruto (San Martin, 2013).

2.2.1.5.3. *Misty.*

Variedad reconocida por la calidad de sus frutos. Presenta alta formación de yemas florales, lo que favorece una producción abundante y menor desarrollo foliar durante la primavera (Espinoza, 2022).

2.2.1.5.4. *Legacy.*

Los tallos destacan por ser fuertes, de crecimiento vertical y con buena rentabilidad. La fruta tiene un tamaño medio, un excelente sabor y un equilibrio ideal entre acidez y dulzura. Sus frutos, similares a los de otras variedades, poseen un color azul claro, son firmes y agradables al gusto. Una de sus características más notables es su pequeña cicatriz en el pedúnculo. Además, cuenta con una prolongada vida útil tras la cosecha, manteniéndose en buenas condiciones por un promedio de nueve semanas bajo ambientes controlados (Espinoza, 2022).

2.2.1.6. ***Morfología.***

2.2.1.6.1. *Raíz.*

El sistema de raíces de la planta es superficial y fibroso, compuesto por raíces finas y cortas. Carece de pelos radiculares, por lo que son las raíces jóvenes las que se encargan de la absorción de nutrientes, lo que implica que su capacidad de absorción es menor en comparación con otras especies vegetales (Buzeta, 1997). El sistema radicular del arándano requiere de una humedad constante, por lo que los suelos con poco drenaje le resultan perjudiciales. Si el suelo llega a acumular demasiada agua, la planta puede morir en un plazo de dos días (Valenzuela, 1988).

El sistema radicular es poco profundo, con el 80% de las raíces ubicadas en los primeros 40 cm del suelo. Las raíces son delgadas y fibrosas, y se distinguen por la ausencia de pelos absorbentes. La corona, situada entre las raíces y la parte aérea, tiene la capacidad de generar brotes basales. Generalmente, se asocia de manera natural con la micorriza

ericoide, lo que favorece una simbiosis que promueve un crecimiento vegetativo más robusto. Además, es sensible al exceso de agua en suelos pesados (García y García, 2010).

2.2.1.6.2. *Hojas.*

Las hojas de esta planta son sencillas y se encuentran dispuestas de manera alternada a lo largo de las ramas. Su tamaño puede variar entre 1 y 8 cm de longitud y su forma es ovalada o lanceolada. En cuanto a su color, predominan tonos verdes suaves o claros, pero en otoño adquieren una tonalidad rojiza. En la parte inferior de las hojas, se encuentran los poros epidérmicos o “estomas”, cuya densidad puede alcanzar hasta 300 por milímetro cuadrado (Buzeta, 1997). Desde la perspectiva anatómica, las hojas presentan una epidermis compuesta por una capa de células que actúan como barrera, mientras que su tejido parenquimático es esponjoso, con numerosos espacios llenos de aire (Muñoz, 1988).

Hojas simples y alternas, con un pedicelo corto, de forma elíptica a lanceolada, alcanzando alrededor de 5 cm de largo. Son caducas y su color varía desde un verde pálido hasta uno muy intenso, dependiendo del cultivar. Presentan bordes ligeramente dentados y están finamente nervadas en su parte inferior (García, 2010).

2.2.1.6.3. *Flor.*

Las flores son completas y se encuentran por encima del nivel del suelo. Se agrupan en racimos que surgen de yemas laterales simples en las ramas, creciendo hacia el lado del brote. Estas flores se desarrollan y se expanden durante el verano, en una dirección basipétala (Gil, 2000).

La inflorescencia más frecuente en los arándanos es un racimo, que suele ser axilar. Estos se distinguen por las yemas terminales de las ramillas, las cuales dejan de crecer cuando comienza el otoño. Las flores tienen una forma de campana, con pétalos fusionados, y contienen entre 8 y 10 estambres que se insertan en la base de la corola (Huamantingo, 2016).

2.2.1.6.4. *Fruto.*

El fruto tiene la forma de una baya redonda, con tamaños que varían entre 0.7 cm y 1.5 cm de diámetro. Su color puede ser azul o negro, dependiendo de la variedad. Posee una secreción serosa, lo que le otorga un aspecto atractivo. Dentro de su endocarpio, puede contener pequeñas semillas, llegando a alcanzar hasta 100. En términos comerciales, se busca que la cicatriz del fruto sea pequeña y seca, y que el fruto sea estable en relación al grosor de su epidermis (Muñoz, 1988 y Buzeta, 1997).

En su parte superior tiene una pequeña corona. La pulpa, de sabor agridulce, tiene un tono vinoso, y en el centro del fruto se encuentran varias semillas (Meléndez y López, 2021).

2.2.1.7. *Fenología.*

El ciclo anual de la planta de arándano (*Vaccinium corymbosum* sp.) abarca las fases vegetativa y reproductiva, las cuales pueden verse alteradas por factores ambientales y las técnicas de manejo empleadas. La fenología conecta el desarrollo y los cambios morfológicos que se observan en la planta con las variaciones en las condiciones climáticas (Rivadeneria, 2008).

La fenología del arándano varía considerablemente debido a los diferentes climas en las áreas productoras. En las regiones de Chincha y Cañete, el ciclo comienza con la poda en los meses de enero y febrero, durante el verano. Después de la poda, se realiza un tratamiento de sanitización, aplicando una pasta de sulfato de cobre para asegurar la cicatrización de las heridas y prevenir infecciones por hongos como *Lasiodiplodia* sp. y *Pestalotia* sp. Tras la poda, las ramas comenzarán a brotar, y el crecimiento de los brotes continuará hasta la floración, que ocurre en abril y mayo. Las nuevas hojas, que surgen de estos brotes, son vulnerables a enfermedades como *Alternaria* sp. y la roya. Los frutos seguirán su proceso de cuajado y pigmentación entre mayo y junio. Este período medio del

año es crítico para el hongo *B. cinérea*, ya que las condiciones de alta humedad favorecen su proliferación. La cosecha comienza a finales de junio y puede extenderse hasta septiembre, e incluso hasta noviembre o diciembre, debido a que el arándano continúa brotando y floreciendo. La fecha de la poda es determinada por los productores para que la cosecha coincida con la temporada comercial más rentable (Estacio, 2019).

2.2.1.8. *Requerimientos edafoclimáticos.*

2.2.1.8.1. *Clima.*

El arándano de arbusto alto puede tolerar temperaturas invernales extremadamente bajas, llegando a -30 °C, lo que minimiza el riesgo frente a heladas. Sin embargo, las temperaturas elevadas, por encima de 28-30 °C, afectan de manera adversa al fruto, provocando deshidratación y quemaduras. Los vientos intensos, especialmente durante los primeros años de desarrollo de la planta, perjudican su crecimiento al dañar las hojas, interferir en la floración y dificultar la polinización realizada por insectos. Además, estos vientos pueden provocar la caída de frutos y generar lesiones en ellos. En lo que respecta al tipo de suelo, este debe ser ligero, bien drenado y con un contenido de materia orgánica superior al 3%, lo cual favorece la retención de humedad, crucial para el crecimiento óptimo de las raíces. El pH del suelo representa un factor crítico para este cultivo, ya que requiere valores ácidos por debajo de 5.5, siendo el rango ideal entre 4.5 y 5.5. (García y Rubio, 2005).

2.2.1.8.2. *Suelo.*

Para el adecuado desarrollo del arándano, es fundamental que el suelo cumpla con ciertas condiciones específicas. Esta planta prospera en terrenos ácidos de textura arenosa, con un excelente sistema de drenaje y un nivel de materia orgánica superior al 3% (Pritts y Hancock, 1992). El nivel de materia orgánica en el suelo puede elevarse mediante la incorporación de turba (peat moss) antes de realizar la siembra. Asimismo, la corteza de

pino, al descomponerse con el tiempo, resulta útil para aumentar la cantidad de materia orgánica. En los cultivos de arándanos donde se ha utilizado cobertura vegetal durante varios años, es común encontrar raíces fibrosas que se desarrollan en la materia orgánica en proceso de descomposición sobre la superficie del suelo (Williamson y Lyrene, 1994). El pH óptimo del suelo es de 4.5, aunque las plantas pueden adaptarse a valores de pH que oscilen entre 3.8 y 5.5, siempre que el nivel de materia orgánica presente sea elevado (Pritts y Hancock, 1992).

Cuando el pH supera 5.8, tanto el desarrollo como la productividad de las plantas comienzan a reducirse progresivamente. Si el pH alcanza un valor de 7, el crecimiento se detiene por completo, llevando eventualmente a la muerte de la planta. En caso de que el suelo presente un pH demasiado alto, es necesario realizar las correcciones correspondientes. El azufre es el producto más comúnmente empleado para este propósito, debiendo aplicarse entre 3 y 4 meses antes de la siembra para prevenir posibles toxicidades. Es importante considerar la conductividad eléctrica (CE), ya que indica el contenido de sales en el suelo. Este valor no debe superar los 0.5 dS/m, debido a la alta sensibilidad del cultivo a la acumulación de sales, lo que podría afectar su desarrollo y la eficiencia de la fertilización. La textura óptima del suelo debe ser ligera, preferiblemente de tipo arenoso o franco arenoso, con un contenido de arcilla inferior al 20% y una baja proporción de limo fino. Esta combinación permite que el suelo conserve humedad mientras garantiza un buen drenaje. Adicionalmente, en caso de que el agua se estanque en los primeros 45 centímetros del suelo durante la temporada de lluvias, es recomendable sembrar las plantas en camas elevadas o canteros para evitar problemas de acumulación de agua (García y García, 2010).

2.2.1.8.3. *Agua.*

La calidad química del agua comprende la cantidad de sales presentes y la proporción de diversos iones disueltos, los cuales pueden influir en las propiedades del suelo y en el

desarrollo de los cultivos. Conocer estas características es esencial para evaluar su idoneidad en actividades agrícolas, ya que desempeña un papel clave en la gestión de la fracción de riego y los procesos de lixiviación (Tartabull y Betancourt, 2016). La salinidad es un factor relevante que influye negativamente en el rendimiento de los cultivos al dificultar la absorción de agua por las plantas. Cuando las concentraciones de sodio son altas, este catión puede reemplazar al calcio y al magnesio o formar precipitados como bicarbonatos de estos elementos. Este proceso provoca la separación de las partículas del suelo, lo que reduce significativamente la velocidad a la que el agua se infiltra (Vivot et al., 2010; Tartabull y Betancourt, 2016).

2.2.1.9. *Requerimientos nutricionales.*

Los arándanos se originan en regiones con suelos arenosos o pantanosos, caracterizados por capas freáticas con baja disponibilidad de nutrientes. Las investigaciones sobre sus necesidades nutricionales han evidenciado que requieren pocos aportes de nutrientes, aunque reaccionan favorablemente a su aplicación siempre que las cantidades sean las adecuadas. Sin embargo, un exceso en la fertilización puede resultar en la muerte de las plantas. Este cultivo prospera en suelos ácidos, condición que mejora la disponibilidad de ciertos elementos esenciales como hierro y manganeso, los cuales son demandados en mayores proporciones por los arándanos en comparación con otras especies frutales (Hirzel y Rodríguez, 2003).

2.2.1.10. *Manejo agronómico.*

2.1.10.1. *Riego.*

El consumo de agua depende de la densidad de la cobertura vegetal (sombra) y del sistema de riego empleado (Holzapfel et al., 2004). Se ha comprobado que el riego localizado genera una variación significativa tanto en la calidad como en la productividad,

esto debido a la poca profundidad que alcanzan las raíces en este tipo de riego (Ramos, 2023).

2.1.10.2. *Fertilización.*

Durante la última década, el método predominante a nivel global para llevar a cabo la fertilización ha sido la fertirrigación, una técnica que integra la irrigación con la aplicación de fertilizantes. Además, se han desarrollado innovaciones tecnológicas, como la automatización mediante el uso de software especializado (Vidal, 2005). El nitrógeno es el elemento clave en los programas de fertilización y el recurso más limitado en el suelo debido a su baja disponibilidad natural. Esta limitación se vuelve especialmente significativa durante la fase de crecimiento vegetativo. Cuando se aplica en cantidades apropiadas, el nitrógeno favorece el desarrollo y fortalece el sistema radicular como de la parte aérea de la planta, además de incrementar la producción de flores y el desarrollo de frutos. Sin embargo, un suministro excesivo de este elemento puede provocar un crecimiento desmedido, mayores problemas fitosanitarios, frutos de consistencia blanda y un aumento en la proliferación de malezas (Hirzel, 2013). En el caso del arándano, es fundamental identificar la forma de nitrógeno que la planta puede absorber. Este nutriente se incorpora mayormente en su modalidad amoniacal (NH_4^+), aspecto que debe tomarse en cuenta al elaborar los planes de fertilización. Por otro lado, el ion nitrato (NO_3^-) tiende a ser lixiviado, lo que limita su utilización efectiva por la planta (Retamales y Hancock, 2012). En el caso del cultivar Biloxi, cuando se utiliza una solución nutritiva con pH 5 y una proporción de $\text{NH}_4^+/\text{NO}_3^-$ de 0/100, se observa un incremento en el desarrollo vegetativo del arbusto, a pesar de ser una condición inusual para los suelos ácidos típicos del arándano, donde predomina la forma amoniacal (NH_4^+) (Crisóstomo et al., 2014).

El fósforo (P) es clave en el desarrollo radicular, la acumulación de nutrientes y la floración. El potasio (K), por su parte, influye directamente en el rendimiento y en la calidad

del fruto, mejorando el tamaño, la firmeza, el sabor y el aroma, además de aumentar la tolerancia al estrés hídrico, bajas temperaturas y enfermedades; sin embargo, en exceso puede provocar la división de los frutos. En cuanto al calcio (Ca), favorece el tamaño y el cuajado, y contribuye a una mayor resistencia a problemas fitosanitarios, así como a una mejor calidad en almacenamiento y postcosecha. Cabe señalar que el arándano, al adaptarse a suelos ácidos, puede acumular aluminio en sus tejidos. Se ha observado que altas concentraciones de este elemento (600 μM) disminuyen de forma considerable la absorción de diversos nutrientes, independientemente de la presencia de micorrizas y del tipo de sustrato utilizado (arena o suelo). Sin embargo, el aluminio se acumula rápidamente en las hojas, lo que mitiga sus efectos negativos en comparación con las plantas que no poseen resistencia a elevados niveles de este elemento (Hirzel, 2013).

2.1.10.3. *Uso de oberturas de suelo.*

El empleo de mulch resulta, en numerosos casos, una práctica ventajosa para el cultivo de arándanos, ya que contribuye a reducir la pérdida de humedad en el suelo, lo cual es crucial debido a la poca profundidad de sus raíces. En particular, la aplicación de mulch de aserrín de pino en suelos de textura arcillo-arenosa ha demostrado incrementar significativamente el rendimiento de las plantas Southern (Ochoa, 2010). Al evaluar factores como la rapidez del crecimiento vegetativo, la altura de las plantas y el volumen de las raíces se determina que los desechos provenientes de actividades agrícolas y procesos de fabricación, tales como subproductos de la elaboración del aceite de oliva, el orujo de uva, el sarmiento triturado y la cascarilla de arroz, representan una opción viable en comparación con la mezcla tradicional de suelo y aserrín (Cruzat y Mancilla, 2010). El empleo de coberturas sintéticas, como las elaboradas con polietileno, resulta más conveniente en contextos comerciales. Se sugiere colocarlas antes de realizar la plantación y una vez que se haya instalado el sistema de riego. Este tipo de cobertura contribuye a incrementar el peso

seco de las raíces en mayor proporción que las plantas de la variedad Ojo de Conejo tratadas con coberturas orgánicas (Ochoa, 2010).

2.2.1.11. *Composición química del arándano.*

El arándano destaca por sus beneficios nutricionales y nutraceuticos, entre los cuales se incluyen su bajo aporte calórico, bajo contenido de grasas y sodio, además de no contener colesterol. Es una fuente rica en fibras, minerales, y presenta altas concentraciones de vitaminas A y C. Su capacidad antioxidante permite neutralizar los radicales libres, lo que reduce el riesgo de desarrollar cáncer y enfermedades cardiovasculares. Adicionalmente, contribuye al incremento del colesterol bueno, ayuda a ralentizar el envejecimiento y fortalece las defensas naturales del cuerpo. También se le atribuyen propiedades antibióticas, antiinflamatorias y desinfectantes, según la FDA (Pino, 2007).

2.2.1.12. *Fructificación y crecimiento del fruto.*

El porcentaje de fructificación puede oscilar entre el 50 % y el 100 %, dependiendo del tipo de cultivar. La caída de los frutos ocurre entre 3 y 4 días después de la floración, aunque es poco frecuente en las variedades del tipo Highbush. Alrededor del 80 % de las células del fruto se desarrollan antes de la floración. Todas las especies comerciales de arándanos siguen un patrón de crecimiento que se representa mediante una curva doble sigmoidea dividida en tres etapas. En la primera etapa, se produce la división celular en el pericarpio, la cual dura entre 25 y 35 días, dependiendo del cultivar y las condiciones ambientales. En este proceso, el fruto gana peso seco y experimenta una notable división celular, alcanzando aproximadamente 7000 células en un corte transversal al inicio de esta fase. En la segunda etapa, que dura entre 30 y 40 días, se da el desarrollo de la semilla, específicamente el crecimiento del endospermo y el embrión. La tercera etapa, que abarca entre 30 y 60 días, se caracteriza por la elongación celular en el mesocarpio, el cambio de color de verde a azul y la acumulación de azúcares. Al final de esta fase, la cantidad de

células alcanza las 9000. El crecimiento del fruto está principalmente relacionado con la expansión de las células, y en menor medida, con la proliferación celular. La maduración del fruto involucra transformaciones específicas en los azúcares y ácidos, y la baya se ablanda debido a la actividad enzimática en la pared celular. Diversos estudios sugieren que el fruto se considera climatérico, particularmente en la tercera fase de su desarrollo (Gil, 2000).

2.2.1.13. *Calidad del fruto.*

Los criterios de calidad que la industria de frutas frescas suele considerar incluyen: el color, el tamaño, la forma, la ausencia de defectos, la firmeza y el sabor. Por lo tanto, todas las acciones llevadas a cabo antes y después de la cosecha deben centrarse en asegurar que el producto que recibe el consumidor sea de alta calidad (Alcalde, 2019).

2.2.1.14. *Peso de arándano.*

El peso del arándano corresponde a la cantidad de materia de cada fruto individual, generalmente expresada en gramos o miligramos según la precisión requerida. Al tratarse de un fruto pequeño, este valor varía según la variedad, las condiciones de cultivo y factores ambientales. Su medición se realiza con una balanza de precisión (Alegría, 2024).

2.2.1.15. *Calibre de arándano.*

El calibre se define como el diámetro ecuatorial del fruto, medido en milímetros. Es un criterio clave para la clasificación y comercialización. Para su medición se emplea un vernier o pie de rey digital (Alegría, 2024).

2.2.1.16. *Relación entre peso y calibre de arándano.*

El peso y el calibre presentan una relación positiva consistente, donde frutos de mayor diámetro tienden a tener mayor peso. Esta relación se mantiene estable en el tiempo, sin diferencias significativas entre calibres, lo que respalda el uso de modelos lineales para su estimación (Alegría, 2024).

2.2.2. Análisis de datos.

El análisis de datos integra procesos de exploración, procesamiento y modelado con el objetivo de identificar patrones y generar información útil para la toma de decisiones. Para ello, combina herramientas estadísticas y técnicas de aprendizaje automático que permiten comprender mejor la estructura de los datos. Mientras los modelos estadísticos se apoyan en distribuciones teóricas para describir su comportamiento, los algoritmos de machine learning permiten capturar relaciones más complejas y realizar predicciones en distintos contextos. Asimismo, técnicas como la reducción de dimensionalidad ayudan a simplificar conjuntos de datos complejos, resaltando la información más relevante y reduciendo el ruido. De forma complementaria, los métodos de resumen facilitan la interpretación de grandes volúmenes de información. En conjunto, el análisis de datos permite transformar datos en conocimiento aplicable, aunque implica desafíos como el manejo de grandes volúmenes de información y el riesgo de identificar patrones no representativos (Mathar et al., 2020).

2.2.3. Marco normativo de la inteligencia artificial.

El desarrollo y la adopción de la inteligencia artificial en el Perú se enmarcan en una estrategia estatal que la reconoce como un factor clave para el desarrollo económico y social, en línea con los Objetivos de Desarrollo Sostenible y la transformación digital del país. En este contexto, la Ley N.º 31814 establece a la inteligencia artificial como una tecnología emergente con capacidad para mejorar la productividad, impulsar la innovación y contribuir al bienestar, siempre bajo principios éticos, de privacidad y respeto a los derechos fundamentales. Asimismo, reconoce que estos sistemas pueden generar predicciones y decisiones con distintos niveles de autonomía, lo que hace necesario contar con mecanismos de supervisión y responsabilidad (Congreso de la República, 2023).

Esta ley también declara de interés nacional el fortalecimiento del talento digital y el uso de tecnologías emergentes en sectores como educación, salud, justicia y seguridad, posicionando a la inteligencia artificial como un eje de la modernización del Estado. Bajo este enfoque, se entiende la sociedad digital como un entorno donde la información, la conectividad y la innovación generan valor, integrando la inteligencia artificial dentro de un ecosistema basado en la confianza y la seguridad (Congreso de la República, 2023).

La conducción de este proceso recae en la Presidencia del Consejo de Ministros, a través de la Secretaría de Gobierno y Transformación Digital, que actúa como ente rector y coordina políticas, lineamientos éticos e infraestructura digital. Esta labor se articula con instrumentos como la Política Nacional de Transformación Digital al 2030, la Estrategia Nacional de Inteligencia Artificial y la Estrategia Nacional de Gobierno de Datos, los cuales orientan la implementación y evaluación de estas tecnologías en el país (Congreso de la República, 2023).

La Política Nacional de Transformación Digital al 2030 concibe este proceso como continuo y estratégico, basado en el uso intensivo de datos y tecnologías digitales para generar valor. En este marco, los servicios predictivos basados en analítica e inteligencia artificial permiten anticipar necesidades, optimizar procesos y reducir riesgos, fortaleciendo un enfoque preventivo en la gestión pública (Presidencia del Consejo de Ministros, 2023).

Por su parte, el Reglamento de la Ley N.º 31814 (D.S. N.º 115-2025-PCM) refuerza estos lineamientos al incorporar principios como la supervisión humana, la sostenibilidad y el uso responsable de la inteligencia artificial. También introduce el concepto de sesgo algorítmico como un riesgo asociado a los datos y al diseño de los modelos, lo que exige implementar mecanismos de control y evaluación en su aplicación (Presidencia del Consejo de Ministros, 2025).

Finalmente, se destaca la importancia de fortalecer las competencias en áreas como ciencia de datos, programación, estadística y ética, con el fin de asegurar un uso adecuado de estas tecnologías. En conjunto, el marco normativo peruano busca aprovechar el potencial de la inteligencia artificial, reduciendo sus riesgos y promoviendo su contribución al desarrollo sostenible (Congreso de la República, 2023; Presidencia del Consejo de Ministros, 2025).

2.2.4. Algoritmo.

Un algoritmo es un conjunto ordenado y finito de instrucciones que permite resolver un problema o realizar un cálculo. Su etimología proviene del matemático persa del siglo IX, Mohammed ibn Musa al-Khwarizmi. En informática, representa una secuencia de pasos que transforma datos de entrada en resultados, siendo la base de los procesos computacionales y de los sistemas inteligentes (Sheikh, Prins & Schrijvers, 2023).

2.2.5. Machine learning.

El aprendizaje automático es una disciplina de la inteligencia artificial (IA) que permite a las máquinas identificar patrones y relaciones en los datos, facilitando la toma de decisiones y la predicción de comportamientos futuros. Su funcionamiento se basa en el uso de datos de entrenamiento, a partir de los cuales los algoritmos aprenden reglas sin necesidad de ser programados de forma explícita (Dalal, 2020; Ponte et al., 2021).

A diferencia de los enfoques tradicionales basados en reglas fijas, el aprendizaje automático tiene la capacidad de adaptarse y mejorar con el tiempo conforme incorpora nueva información, ajustando sus patrones de manera progresiva (Flarance et al., 2018).

2.2.6. Algoritmo de machine learning.

Un algoritmo de aprendizaje automático es un procedimiento que aprende a partir de los datos y mejora su desempeño en una tarea específica (T), evaluada mediante un criterio de rendimiento (P), a medida que adquiere experiencia (E). Este proceso implica ajustar sus

parámetros en función de la información observada, con el fin de optimizar los resultados (Mitchell, 1997).

2.2.7. Modelo.

Un modelo en aprendizaje automático es la función resultante del proceso de entrenamiento de un algoritmo, comúnmente representada como $y(x)$. A partir de los datos de entrada, el modelo aprende patrones que le permiten generar predicciones o clasificaciones sobre nuevos datos. Su finalidad es capturar relaciones en los datos de entrenamiento y generalizarlas a situaciones no observadas (Bishop, 2006).

2.2.8. Tipos de machine learning.

Según Sarker (2021) el aprendizaje automático se clasifica en tres tipos principales:

- Aprendizaje supervisado: utiliza datos etiquetados para entrenar modelos que permiten realizar predicciones, como en la detección de spam.
- Aprendizaje no supervisado: identifica patrones en datos sin etiquetar, organizándolos y estructurándolos para su análisis, con aplicaciones como la segmentación de mercados.
- Aprendizaje por refuerzo: el modelo aprende mediante la interacción con su entorno, ajustando sus acciones para maximizar resultados.

2.2.9. Regresión.

La regresión es una técnica estadística utilizada para analizar la relación entre variables y predecir una variable respuesta (y) a partir de una o más variables independientes (x). Puede ser simple, cuando involucra una sola variable explicativa, o múltiple, cuando incluye varias. Este enfoque es ampliamente utilizado para modelar datos continuos y encontrar patrones relevantes, siendo la regresión lineal y la logística dos de los métodos más comunes (Montgomery, Peck y Vining, 2005; Solano y Caballero, 2021).

2.2.8.1. *Regresión lineal.*

El modelo de regresión lineal establece una relación matemática entre dos variables aleatorias, conocidas como “X” (variable independiente o predictora) e “Y” (variable dependiente o respuesta). Este modelo, denominado simple cuando vincula únicamente una variable independiente con “Y”, se expresa como:

$$Y = \beta_0 + \beta_1 X + \varepsilon$$

Aquí, ε representa el término de error, que introduce variabilidad en la relación entre “X” e “Y”. Esto significa que cada valor de “Y” es igual a $\beta_0 + \beta_1 X$ más una cantidad de error ε , que puede ser positiva o negativa. Este error explica por qué los puntos de datos reales no se alinean exactamente sobre la recta, sino que se distribuyen alrededor de ella, como se observa en un gráfico de dispersión. En esta fórmula, β_0 y β_1 son parámetros desconocidos que representan la ordenada al origen y la pendiente, respectivamente. El modelo se denomina “lineal” porque la dependencia de “Y” con respecto a “X” es una relación lineal, y porque los parámetros β_0 y β_1 no se elevan a potencias ni se combinan de forma no lineal. Este enfoque facilita modelar y analizar las relaciones en los datos (Szretter, 2013).

2.2.8.2. *Regresión no lineal.*

La regresión no lineal es un método que permite modelar y predecir sistemas cuya respuesta no sigue una relación lineal. Este enfoque captura relaciones más complejas entre las variables. Estas no linealidades pueden originarse por el comportamiento de los materiales, la presencia de componentes no lineales o la interacción entre distintos elementos, así como por la aplicación de cargas no lineales. Por ello, resulta fundamental para analizar y diseñar sistemas sometidos a condiciones más exigentes y dinámicas (O’Brien & Silcox, 2024).

2.2.8.3. *Regresión polinomial.*

La regresión polinomial puede entenderse como una extensión de la regresión lineal múltiple, en la que se busca ajustar un polinomio que represente adecuadamente la relación entre las variables (Chanchí y Campo, 2021). Este tipo de modelo se utiliza cuando la variable dependiente presenta un comportamiento curvilíneo o no lineal. La ecuación (1) representa el modelo de regresión polinomial de grado k para una variable independiente, mientras que la ecuación (2) corresponde al modelo ajustado. Los parámetros se estiman mediante el método de mínimos cuadrados, como se muestra en la ecuación (3), donde Y , X y X^T son vectores. Según Astorga (2014), la definición formal de la regresión polinomial es la siguiente:

$$Y = \beta_0 + \beta_1 X + \beta_2 X^2 + \dots + \beta_k X^k + \epsilon \quad (1)$$

$$\hat{Y} = \hat{\beta}_0 + \hat{\beta}_1 X + \hat{\beta}_2 X^2 + \dots + \hat{\beta}_k X^k \quad (2)$$

$$\hat{\beta} = (X^T X)^{-1} X^T Y \quad (3)$$

2.2.10. **Arboles de decisión.**

Un árbol de decisión es un modelo de aprendizaje supervisado no paramétrico, utilizado en tareas de clasificación y regresión. Se estructura de forma jerárquica, con un nodo raíz, ramas, nodos intermedios y nodos terminales. Permite modelar relaciones complejas al identificar interacciones entre múltiples variables, lo que facilita un análisis más detallado de los datos (Quinlan, 1986).

2.2.11. **Random forest.**

El modelo Random Forest consiste en un conjunto de árboles de decisión generados a partir de muestras aleatorias de los datos. Este enfoque permite que el error de generalización se estabilice a medida que aumenta el número de árboles, dependiendo tanto de la precisión individual de cada árbol como de la correlación entre ellos. Para reducir esta correlación, el algoritmo selecciona aleatoriamente subconjuntos de variables en cada

división, logrando altos niveles de precisión comparables a otros métodos como AdaBoost (Breiman, 2001).

En este modelo, cada árbol divide los datos en grupos cada vez más específicos según reglas definidas. Sin embargo, cuando los árboles son muy profundos, pueden sobreajustarse al conjunto de entrenamiento, capturando incluso el ruido. Random Forest reduce este problema al combinar múltiples árboles entrenados sobre distintos subconjuntos de datos y variables, lo que mejora la capacidad de generalización y la robustez del modelo (Kumar & Jain, 2020).

2.2.12. R programming.

El lenguaje de programación R es una herramienta gratuita ampliamente utilizada en el aprendizaje estadístico que destaca por su capacidad para manejar datos y desarrollar modelos, lo que ha facilitado su adopción en diversas disciplinas. Ha contribuido a que el análisis estadístico pase a ser una herramienta clave en aplicaciones como el modelamiento supervisado, no supervisado y el pronóstico (James, Witten, Hastie, & Tibshirani, 2023).

2.3. Bases conceptuales

La variable respuesta corresponde al peso promedio semanal de la baya de arándano (gramos), estimado a partir del pesaje de muestras en campo. Como variables predictoras se consideran: (i) el calibre de la baya, variable numérica continua que representa el diámetro de la baya, obtenido de la misma muestra usada para el peso de baya; y (ii) los factores estacionales, representados mediante variables indicadoras (*dummies*) de mes o trimestre que capturan la variación temporal del peso. Estas variables buscan explicar la variabilidad del peso semanal y mejorar la precisión del pronóstico.

Según la tabla 1, la operacionalización de variables muestra a las siguientes variables de investigación:

Variable Dependiente:

- Pronóstico del peso de bayas de arándano.

Variable Independiente:

- Factores estacionales del peso de las bayas de arándano.
- Calibre de baya de arándano.

Tabla 1.

Operacionalización de variables.

Variable	Definición conceptual	Definición operacional	Indicador	Escala de Medición	Fuente de Datos
Pronóstico del peso de bayas de arándano (Variable Dependiente)	Peso estimado en gramos por cada baya de arándano en un cultivo específico.	Promedio semanal del peso de bayas (g).	Peso promedio semanal de las bayas de arándano.	Continua (gramos)	Registros de pesaje de bayas en la operación agrícola.
Factores estacionales del peso de las bayas de arándano (Variable Independiente)	Fluctuaciones en el peso de las bayas asociadas a patrones estacionales (como meses o trimestres).	Variables indicadoras por mes o trimestre (1 = pertenece, 0 = no pertenece)	Dummies estacionales que capturan efectos de meses/trimestres en el peso de las bayas.	Categórica (Variables Dummies)	Registros internos y datos históricos de producción.
Calibre de baya de arándano (Variable Independiente)	Diámetro ecuatorial de las bayas de arándano.	Promedio del diámetro de las bayas medidas en campo.	Calibre de baya	Continua (Variable numérica)	Observaciones y clasificaciones en campo.

III. DISEÑO METODOLÓGICO

3.1. Tipo y nivel de investigación.

- Enfoque Cuantitativo. El estudio adopta un enfoque cuantitativo, empleando tanto datos medibles como indicadores fenológicos susceptibles de cuantificación y analizando sus resultados mediante técnicas estadísticas (Hernández & Mendoza, 2018).
- Tipo Aplicada. Se enmarcó dentro de la investigación aplicada, pues se utilizan teorías y procedimientos ya validados por diversos autores para proyectar el calibre de frutos de arándano en contextos agrícolas reales.
- Nivel Explicativa. El nivel explicativo se alcanza al aplicar estadística inferencial para cumplir con los objetivos planteados (Hernández & Mendoza, 2018). Más allá de describir el comportamiento agronómico y fenológico del cultivo, se busca desarrollar modelos predictivos válidos que permitan estimar el calibre de frutos del arándano.

3.2. Diseño de investigación.

La estrategia utilizada para alcanzar los objetivos fue:

- Análisis primario de datos. Los registros provinieron de una empresa productora de arándano en Olmos, Lambayeque (Perú). Se construyó un preprocesamiento reproducible para garantizar la calidad e integridad de los datos, que incluyó limpieza, transformación y partición en conjuntos de entrenamiento y validación. A continuación, estos datos se emplearon en el desarrollo, ajuste de hiperparámetros y validación de modelos predictivos basados en machine learning, comparándolos con modelos lineales simples para seleccionar la técnica más precisa y robusta.

3.3. Población y muestra.

3.3.1. Población.

La población de la investigación estuvo constituida por el conjunto completo de datos contenidos en el reporte de calidad agronómica de la campaña año 2023 - 2024 de una empresa agroindustrial de arándanos, ubicada en Olmos (región Lambayeque, Perú). Estos registros incluyeron mediciones de calibre y peso de cada baya muestreada durante la campaña de cosecha 2023 - 2024, sin restricciones de fecha, lote o variedad, y sirvieron como base para el desarrollo de los modelos predictivos.

3.3.2. Muestra.

La muestra fue censal, es decir, incluyó las 405 observaciones disponibles de la población definida. Para la construcción y evaluación de los modelos predictivos se aplicó una partición aleatoria inicial con proporción 75%–25%. De este modo, se obtuvo:

Muestra de entrenamiento (75%): 304 observaciones, utilizadas para ajustar y optimizar los parámetros de los modelos.

Muestra de validación (25%): 101 observaciones, empleadas para evaluar el desempeño predictivo de los modelos seleccionados.

3.4. Técnicas e instrumentos de recolección de datos.

3.5.1. Regresión.

En la presente investigación se utilizó la técnica de regresión para pronosticar el peso de las bayas de arándano. Se emplearon datos de fuentes primarias que se relacionan con las variables predictoras, como factores estacionales y el calibre de las bayas.

3.5.2. Elaboración de instrumentos.

Se empleó como instrumento un datalake que se construyó a partir de dos conjuntos de datos: registros históricos de peso y clasificaciones de bayas de arándano, obtenidos de bases de datos internas de la operación agrícola y observaciones en campo. Estos datos

fueron procesados y transformados para ser utilizados en la construcción de modelos predictivos.

3.5. Procedimientos.

Se creó la base de datos a partir de registros históricos sobre el peso promedio semanal y el calibre de las bayas de arándano en un lote de la operación agroindustrial de arándano. Estos datos fueron organizados con una frecuencia semanal durante el periodo de estudio.

3.6. Plan de procesamiento y análisis de datos.

Se llevó a cabo el procesamiento de los datos con el propósito de calcular el peso promedio semanal de las bayas y su calibre medio. Para ello, se utilizó el lenguaje de programación R versión 4.5.1; a continuación, se generaron tablas, se elaboraron gráficos y se efectuó el análisis estadístico correspondiente.

Entre las opciones consideradas para el modelamiento predictivo estuvieron:

- Modelo lineal general (regresión lineal - lm).
- Modelo lineal general polinómica (regresión polinomial - pm).
- Árbol de regresión (árbol de decisión - dt).
- Random Forest (bosque aleatorio - rf).

Durante la creación de los algoritmos predictivos, se estableció una partición inicial de los datos con 75 % para entrenamiento y 25 % para prueba, y se aplicó validación cruzada de K-Fold repetido con cinco particiones y veinte repeticiones. Para el ajuste, la validación y la prueba de los modelos se empleó R versión 4.5.1 junto con los paquetes tidyverse y tidymodels.

En la calibración de hiperparámetros de los modelos basados en árboles se implementó un procedimiento bayesiano mediante la función *tune_bayes* del paquete *tune*,

seleccionándose el conjunto de hiperparámetros óptimo según la métrica RMSE (Error cuadrático medio).

Finalmente, en la fase de validación se seleccionó el mejor algoritmo evaluando en conjunto las métricas RMSE, MAE, MAPE y R^2 ; acto seguido, se llevó a cabo un ajuste final y se evaluó el desempeño predictivo en la fase de prueba. A continuación, se detallan los conceptos de las métricas empleadas:

- Error cuadrático medio (RMSE): se definió como la desviación estándar de los errores predictivos de un modelo, midiendo el nivel de dispersión de dichos errores.
- R^2 (RSQUARE): representó la proporción de la varianza de la variable dependiente explicada por el modelo.
- Error medio absoluto (MAE): midió la magnitud del error predictivo del modelo en la misma unidad que la variable dependiente, calculándose como la media de los errores absolutos.
- Error medio porcentual absoluto (MAPE): midió la magnitud del error predictivo en porcentaje absoluto del modelo.

3.7. Matriz de consistencia.

En la Tabla 2 se detalla la matriz de consistencia de la investigación.

Tabla 2.

Matriz de consistencia.

Título	Problema	Objetivos	Hipótesis	Variables	Metodología
Modelos de machine learning para el pronóstico del peso de bayas de arándano <i>Vaccinium corymbosum</i> L.,	¿Cuál modelo de machine learning permite predecir el peso de baya (en g) del arándano en función de su calibre?	Objetivo general: Desarrollar modelos predictivos basados en técnicas de machine learning para pronosticar con precisión el peso de bayas de arándano en función de su calibre.	Hipótesis general: Es posible predecir con precisión el peso de las bayas de arándano en función de su calibre utilizando modelos de machine learning.	Variable Independiente: Factores estacionales. Calibre de baya.	Enfoque: Cuantitativo Tipo: Aplicada Nivel: Explicativo Diseño: No Experimental Población: La población estuvo constituida por los datos agronómicos de calibre y peso de bayas registrados en el reporte de calidad de la campaña 2023 - 2024 de una empresa agroindustrial de arándanos en Olmos (Lambayeque, Perú). Muestra: La muestra fue censal, incluyendo las 405 observaciones disponibles, las cuales se dividieron aleatoriamente en un 75 % (304 datos) para entrenamiento y un 25 % (101 datos) para validación de los modelos predictivos.
		Objetivos específicos: – Preprocesar los datos históricos de calibre y peso de bayas de arándano para mejorar la calidad de los datos y el desempeño de los modelos predictivos. – Evaluar el desempeño de los modelos predictivos basados en técnicas de machine learning con los modelos estadísticos tradicionales para identificar la técnica más precisa y robusta. – Seleccionar el modelo predictivo con mejor desempeño mediante la evaluación de diferentes algoritmos de machine learning, optimizando sus hiperparámetros.	Hipótesis específicas: – El preprocesamiento de los datos históricos del calibre y peso de las bayas de arándano mejorará el desempeño de los modelos predictivos. – Los modelos predictivos basados en técnicas de machine learning tendrán un desempeño más preciso y robusto que los modelos estadísticos tradicionales al analizar e identificar patrones en los datos utilizados. – Existe un modelo predictivo basado en técnicas de machine learning que tiene un mejor desempeño en la predicción del peso de bayas de arándano en función de su calibre que los modelos basados en técnicas estadísticas tradicionales.		

IV. RESULTADOS

El presente capítulo detalla los hallazgos obtenidos en la investigación, los cuales están orientados al cumplimiento del objetivo general: predecir los calibres (en mm) por semana de los frutos del cultivo de arándano empleando datos fenológicos y algoritmos de aprendizaje automático. En primer lugar, el subcapítulo *4.1. Análisis exploratorio de datos* presenta la identificación de patrones clave en los datos, así como las correlaciones entre las variables fenológicas y temporales con la variable objetivo *fruit_size*. En la sección *4.2. Preprocesamiento de los datos e ingeniería de variables*, se describen las técnicas aplicadas para optimizar el conjunto de datos y asegurar la adecuada selección de variables para el modelado. Luego, en el subcapítulo *4.3. Prueba de los algoritmos*, se evalúa el desempeño de los modelos mediante métricas como MAE, MAPE y MPE, considerando validación cruzada, testeo interno y externo. Esto permite comparar de forma integral los algoritmos Random Forest (RF), modelos aditivos generalizados mixtos (GAMM) y regresión lineal múltiple (LM).

A continuación, se presentan los principales resultados del análisis exploratorio, junto con las metodologías empleadas para abordar el modelamiento del calibre de las bayas de arándano, así como las variables consideradas (ver Tabla 3).

Tabla 3.*Diccionario de variables.*

Variable	Nombre descriptivo	Escala / Tipo de dato	Descripción
berry_weight	Peso de la baya	Cuantitativa continua (g)	Peso individual promedio de la baya de arándano medido en gramos. Es la variable respuesta utilizada para evaluar las relaciones con las variables predictoras.
berry_size	Calibre de la baya	Cuantitativa continua (mm)	Medida del tamaño o calibre del fruto de arándano. Se relaciona directamente con el peso de la baya y presentó la mayor correlación con esta variable.
date_num	Número de fecha codificado	Cuantitativa discreta	Representación numérica de la fecha dentro del periodo de observación, utilizada para capturar la tendencia temporal a lo largo de la campaña agrícola.
date	Fecha de registro	Temporal (fecha)	Fecha exacta en la que se registró la observación o medición del peso de la baya. Permite analizar la evolución del peso del fruto durante la campaña.
week	Semana calendario	Cuantitativa discreta	Número de semana del calendario correspondiente al registro de la observación. Representa la posición temporal del dato dentro del año agrícola.
month_num	Número de mes	Cuantitativa discreta	Representación numérica del mes del año (1–12) utilizada como predictor temporal para analizar patrones estacionales en el peso de la baya.
month_name	Nombre del mes	Cualitativa nominal	Identificación textual del mes en que se registró la observación (por ejemplo: julio, agosto, septiembre). Se usa para analizar variaciones estacionales del peso de la baya.
berries_in_plant	Número de bayas por planta	Cuantitativa discreta	Conteo total de bayas presentes en una planta al momento de la medición. Representa la carga de frutos y su posible relación con el peso promedio de la baya.

4.1. Análisis exploratorio

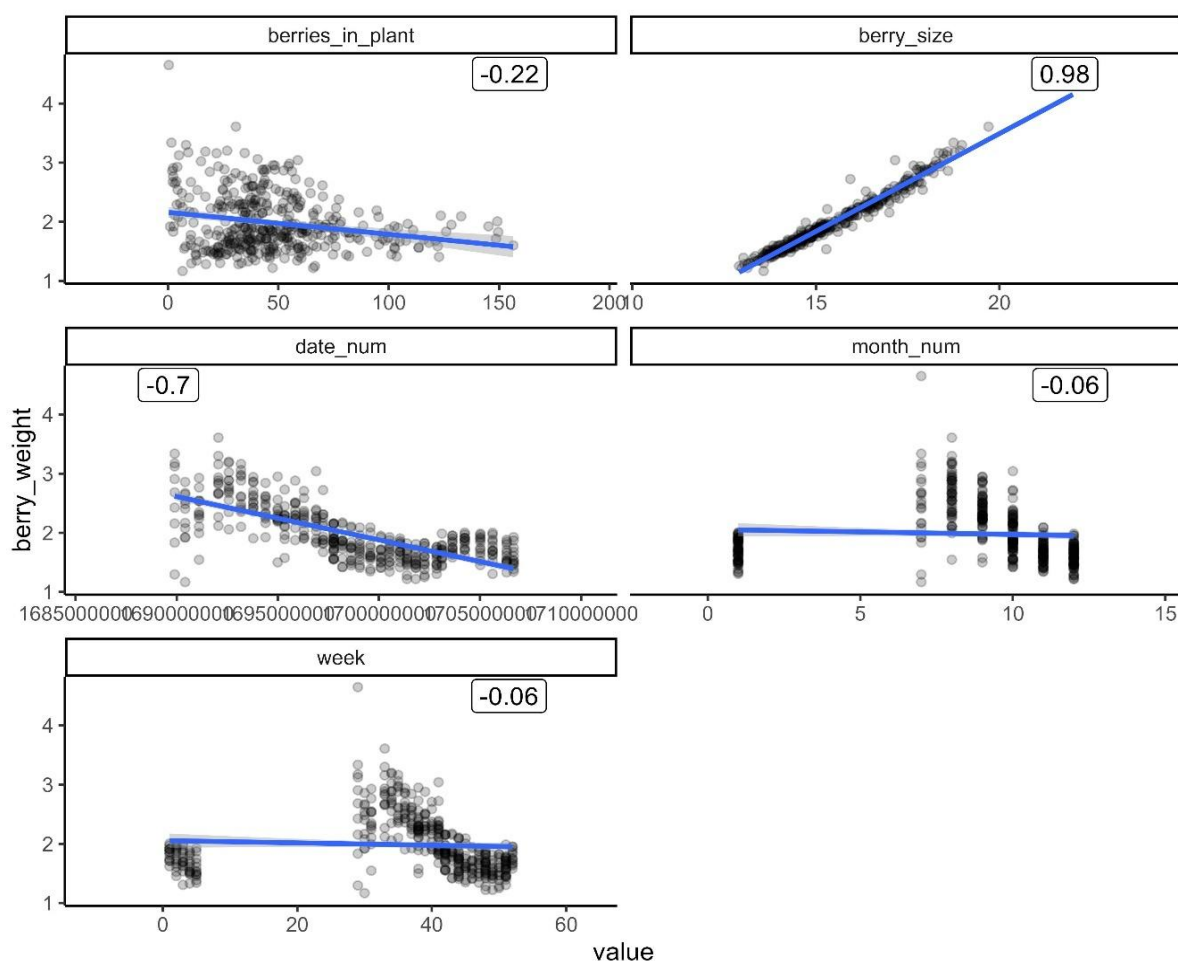
4.1.1. Relación entre peso de bayas y variables cuantitativas

El análisis de correlación reveló que la variable con mayor asociación lineal con el peso de la baya fue el calibre (*berry_size*), con un coeficiente de Pearson de 0.98, lo que denotó una relación casi perfecta y directamente proporcional. El número de fecha codificado (*date_num*) mostró una fuerte correlación negativa ($R = -0.70$), evidenciando que conforme avanzó la campaña el peso promedio se fue reduciendo de manera bastante continua. El conteo de bayas por planta (*berries_in_plant*) presentó una correlación moderada y negativa ($R = -0.22$), sugiriendo que mayores densidades de fruto en la planta

tendieron a asociarse con bayas ligeramente más livianas. Por su parte, los predictores temporales calendario (*month_num*) y semanales (*week*) registraron coeficientes muy próximos a cero ($R = -0.06$), indicando relaciones prácticamente nulas con el peso y posibles comportamientos no lineales o influencias de otros factores no capturados en estas series temporales, por lo que se podría sugerir el uso de variables predictoras como *week* y *month_num* luego de una ingeniería de variables, debido a que tienen una relación no lineal con el peso de baya (ver Figura 1).

Figura 1.

Relación entre el peso de baya y las variables cuantitativas.

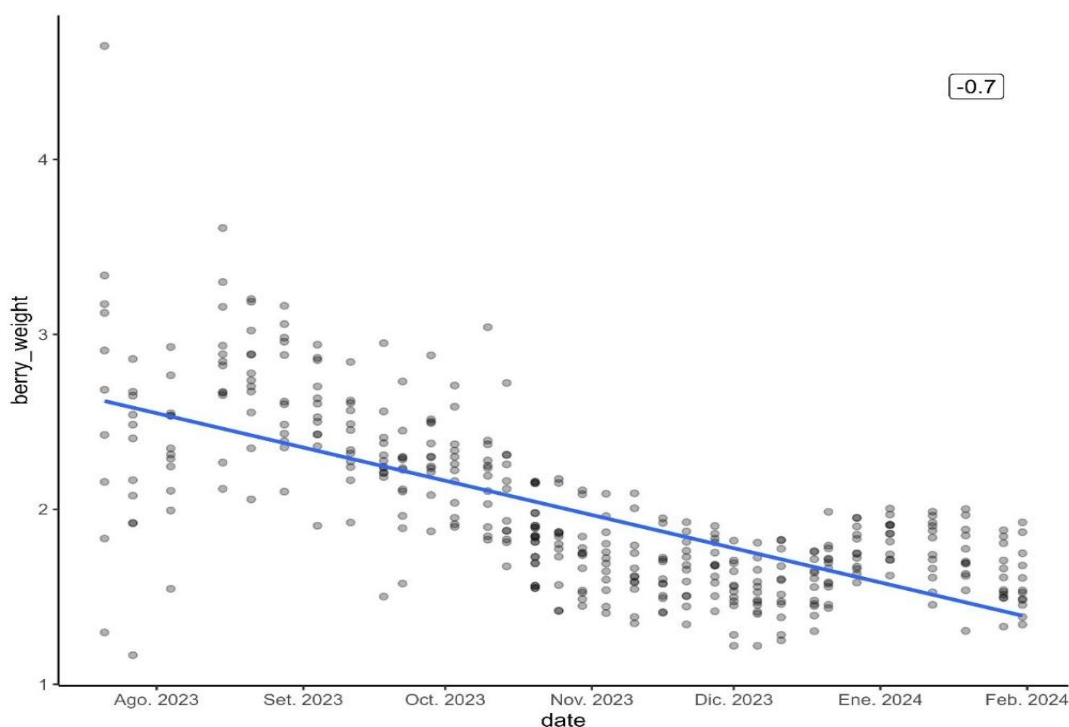


4.1.2. Relación entre peso de bayas y variables temporales

El análisis de la relación entre fecha de cosecha y peso de la baya mostró una tendencia negativa marcada (Pearson $R = -0,70$).

Figura 2.

Relación entre el peso de baya y la variable fecha.



A lo largo del periodo estudiado, desde agosto de 2023 hasta febrero de 2024, el peso promedio descendió de aproximadamente 2,5 g a valores cercanos a 1,5 g. Aunque al inicio de la campaña existieron bayas aisladas de más de 4 g, la mayoría de los datos se concentró por debajo de esa cota y la línea de regresión evidenció un descenso continuo conforme avanzó el calendario. Esta disminución sostenida sugiere que, los frutos posteriores al pico de crecimiento maduraron con menor masa promedio, reflejando efectos estacionales sobre la fisiología del cultivo (ver Figura 2).

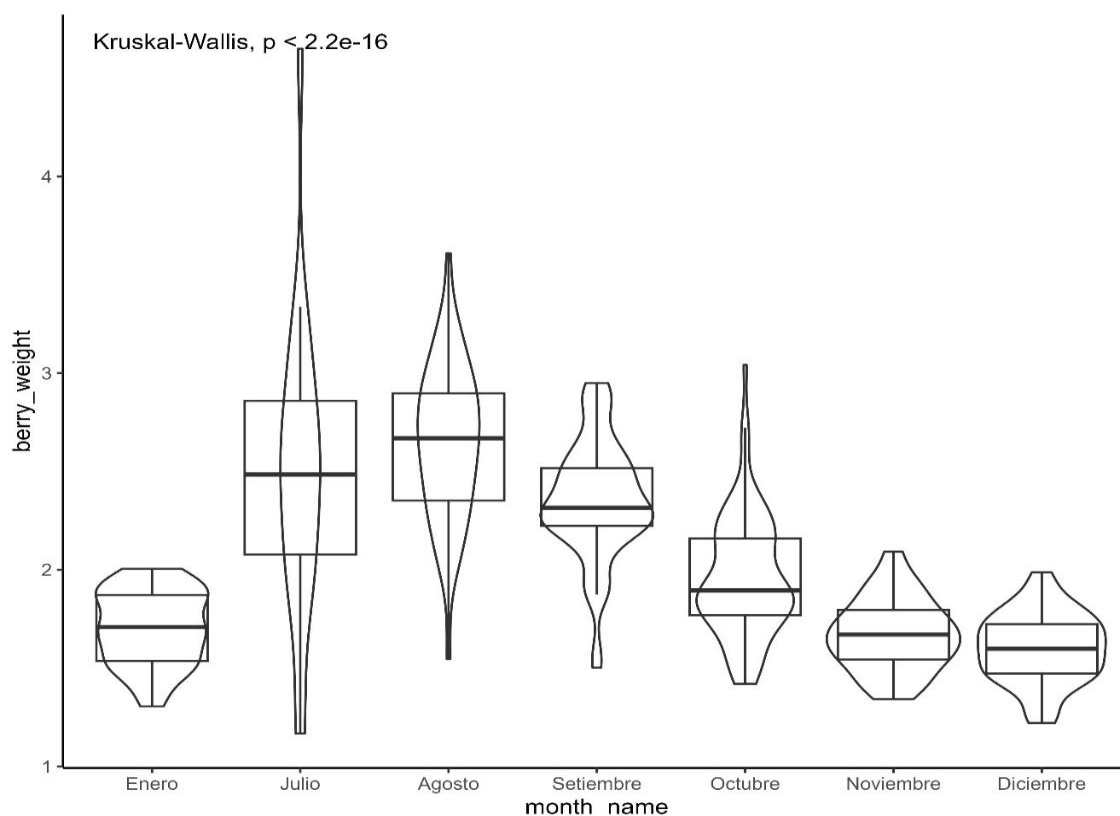
El análisis de la distribución del peso de las bayas por mes puso de manifiesto diferencias estacionales significativas (prueba de Kruskal–Wallis, $p < 2.2 \times 10^{-16}$).

La Figura 3 corresponde a un gráfico de violín con caja y bigotes, que permite visualizar simultáneamente la distribución y la variabilidad de los datos. El violín muestra la densidad de los valores del peso de la baya, de modo que las zonas más anchas indican mayor concentración de observaciones. En el interior se encuentra el diagrama de caja y

bigotes, donde la caja representa el rango intercuartílico (50 % central de los datos) y la línea central corresponde a la mediana. Los bigotes se extienden hacia los valores mínimo y máximo observados, mostrando la dispersión de los datos en cada mes.

Figura 3.

Relación entre el peso de baya y la variable mes.



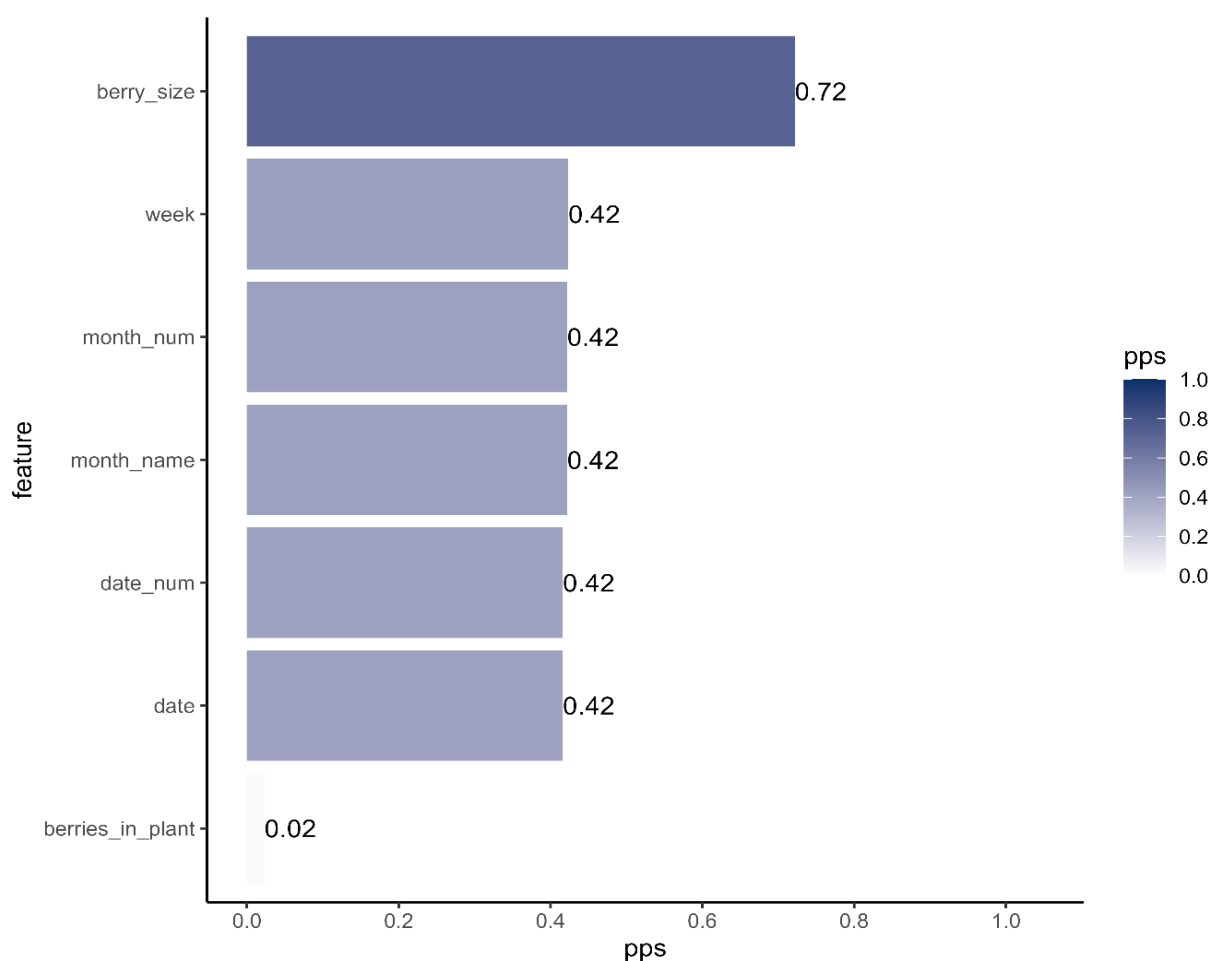
Durante julio y agosto de 2023 se registraron los mayores valores medianos de peso, superando los 2.5 g en promedio, y presentaron además una mayor variabilidad, como lo evidencian los ensanchamientos de los violines y la longitud de sus bigotes (ver Figura 3). En septiembre el peso promedio descendió levemente, ubicándose alrededor de 2.3 g, mientras que en octubre se observó un nuevo pero moderado descenso a aproximadamente 1.9 g. Hacia noviembre y diciembre de 2023 las bayas alcanzaron sus valores más bajos, con valores medianos (cuartil 50) próximos a 1.7 g y menor dispersión. El mes de enero de 2024 mostró una reducción adicional, consolidando la tendencia decreciente iniciada tras el pico de crecimiento (ver Figura 3).

4.1.3. Métrica de poder predictivo de las variables exógenas

El análisis de poder predictivo medido mediante el score de poder predictivo (PPS) mostró que el calibre de la baya (*berry_size*) aportó la mayor capacidad de predicción, con un PPS de 0.72. Las variables temporales, las cuales fueron semana calendario (*week*), número de mes (*month_num*), nombre de mes (*month_name*), fecha codificada (*date_num*) y fecha en formato original (*date*), registraron un PPS uniforme de 0.42, lo que indicó un poder predictivo moderado y muy similar entre ellas. Por su parte, el conteo de bayas por planta (*berries_in_plant*) presentó un PPS prácticamente nulo (0.02), confirmando su escasa utilidad como predictor independiente (ver Figura 4).

Figura 4.

Métricas de poder predictivo de las variables exógenas.



Estos resultados del poder predictivo mostraron que el calibre de la baya aportó la mayor información predictiva, mientras que el conteo de bayas por planta resultó prácticamente irrelevante. También, se podría enriquecer la capacidad predictiva de los modelos añadiendo interacciones entre el calibre y la temporalidad o aplicar polinomios de segundo grado sobre el calibre para capturar no linealidades. Ante esta evidencia, se pudo asumir que, añadiendo al menos una variable temporal, en combinación con el uso del calibre de baya, se lograría un algoritmo con alto poder predictivo. Bajo esta hipótesis, se recurrió a la ingeniería de variables, detallado en el subcapítulo *Pre-procesamiento de los datos*.

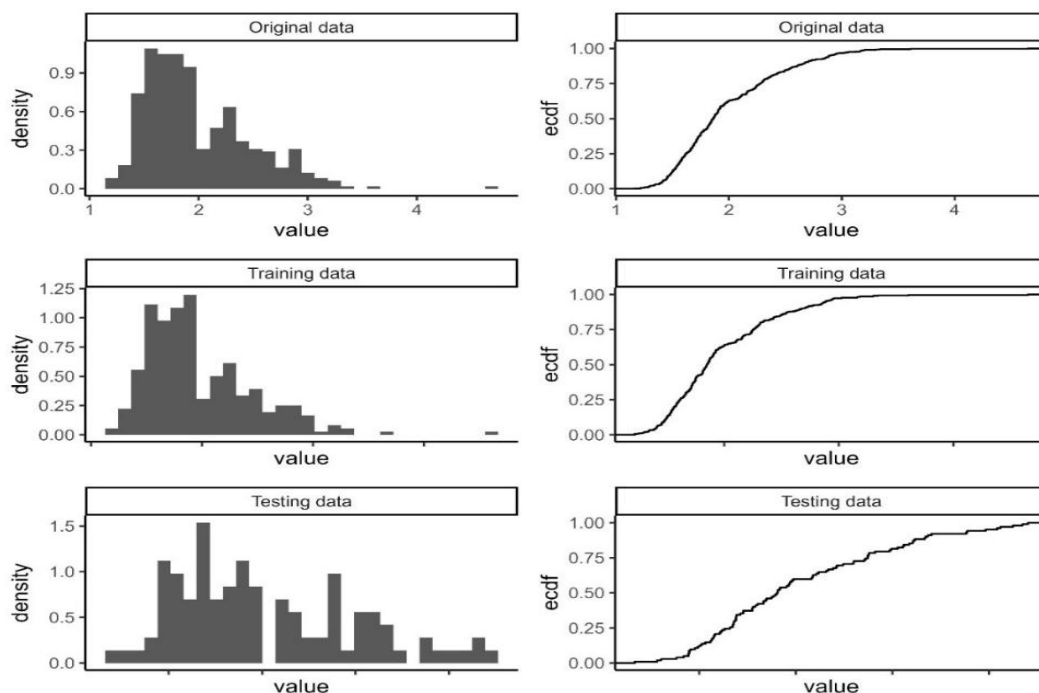
4.2. Pre-procesamiento de los datos

4.2.1. Comparación de distribuciones.

La Figura 5 muestra la comparación entre la distribución original de los pesos de baya y las de los subconjuntos de entrenamiento y prueba.

Figura 5.

Distribución y función de probabilidad acumulada del peso de bayas en los conjuntos de datos original, de entrenamiento y de prueba.



En el histograma y la función de distribución empírica (ECDF) de los datos originales se observó una distribución asimétrica hacia la derecha, con un máximo de densidad alrededor de 1,8 g y cola prolongada hasta valores superiores a 4 g. Al comparar el subconjunto de entrenamiento, tanto el histograma como la ECDF reprodujeron fielmente la forma general de la distribución original, confirmando que la partición aleatoria mantuvo la misma dispersión y sesgo de la población completa. El conjunto interno de prueba también conservó el patrón general, aunque mostró una ligera concentración adicional de observaciones en el rango superior (alrededor de 2,5 g–3,0 g), lo cual se reflejó en una ECDF más gradual entre el percentil 50 y el 90. Los resultados validaron que la división 75 % / 25 % generó subconjuntos representativos de la distribución de peso de las bayas, asegurando que tanto el entrenamiento como la prueba capturaran adecuadamente la variabilidad observada en la campaña.

4.2.2. Flujo reproducible de datos.

Previo al entrenamiento de algoritmos, se diseñó e implementó un flujo de trabajo reproducible para garantizar la calidad, consistencia e integridad de los datos históricos de calibre y peso de bayas de arándano utilizando el paquete `recipes` del ecosistema `tidymodels`, como requisito previo al desarrollo de los modelos predictivos.

En primer lugar, se definió la variable respuesta como el peso de la baya (*berry_weight*) y se asignaron roles específicos a las variables de fecha, de modo que la columna `date` fuera reconocida como variable temporal. A continuación, se seleccionaron de forma explícita aquellas columnas consideradas relevantes: la fecha en su formato original, el número de día en el año (`date_num`), el nombre del mes, el tamaño de la baya (*berry_size*), la semana de cosecha y el número de bayas por planta. Sobre este subconjunto se aplicó una transformación logarítmica ($\log(x+1)$) a todas las variables numéricas distintas del peso de la baya y de las coordenadas temporales, con el objetivo de estabilizar la varianza y mejorar

la aproximación a la normalidad de los predictores. El objeto receta fue luego “preparado” (*prep*) y aplicado (*bake*) de forma idéntica a los conjuntos de entrenamiento y prueba, garantizando así la consistencia en todas las fases del análisis.

4.2.3. Selección de predictores.

Para identificar las variables de mayor contribución predictiva se implementó un modelo de regresión Lasso usando *glmnet* dentro de un flujo de trabajo (workflow) de *tidymodels*. El parámetro de penalización se calibró mediante búsqueda en cuadrícula sobre 100 combinaciones, evaluadas con validación cruzada repetida (5 folds $\times$ 20 repeticiones) y optimizadas con múltiples métricas de error. De cada ajuste se extrajeron los coeficientes y, filtrando aquellos distintos de cero, se obtuvo un vector de predictores finales, obteniéndose como predictores finales al calibre y al mes (*month_name*). Esta selección resultó fundamental para simplificar las especificaciones posteriores y para diseñar modelos más parsimoniosos sin sacrificar precisión.

Sobre la base de los predictores seleccionados por Lasso se definieron dos modelos lineales. El primero incluyó únicamente los efectos lineales de las variables retenidas (predictores finales por Lasso), mientras que el segundo incorporó un término polinómico de segundo grado sobre el tamaño de la baya para capturar posibles relaciones no lineales. Ambos modelos reutilizaron la receta original de preprocesamiento, lo que aseguró que las transformaciones y escalados permanecieran consistentes y reproducibles a lo largo de los experimentos. La validación cruzada aplicada a estas especificaciones permitió obtener estimaciones precisas de los errores RMSE, MAE y MAPE, así como de la estabilidad de los coeficientes en las distintas particiones.

4.2.4. Pre-procesamiento para modelos basados en árboles de decisión.

Para los modelos basados en árboles de decisión (decision tree) y en bosques aleatorios (random forest) se enriqueció el conjunto de variables añadiendo características

temporales derivadas de la fecha mediante la función *step_timeseries_signature*. Esta etapa generó decenas de nuevas variables como día del mes, día del año y variables cíclicas de estación. Para evitar la sobrecarga y el ruido en el modelo, se eliminaron los componentes horarios, las variables con varianza cercana a cero y aquellas de alta colinealidad, y se codificaron las variables categóricas mediante variables dummy (variables binarias que representan cada categoría con valores 0 o 1). El bosque aleatorio, en particular, aprovechó esta ingeniería de variables sin necesidad de especificar fórmula, pues su mecanismo interno de permutación realizó la selección y ponderación de las características más relevantes durante el ajuste.

De esta manera, el preprocesamiento e ingeniería de variables combinó transformaciones paramétricas, extracción de información temporal y selección automática de predictores, conformando una base sólida y reproducible para comparar de manera rigurosa el desempeño de modelos estadísticos tradicionales frente a técnicas de machine learning. El diseño modular del preprocesamiento permitió además incorporar fácilmente futuros refinamientos de las variables o de los algoritmos, manteniendo el control completo de la calidad e integridad de los datos a lo largo de todo el proceso analítico.

4.3. Validación de modelos

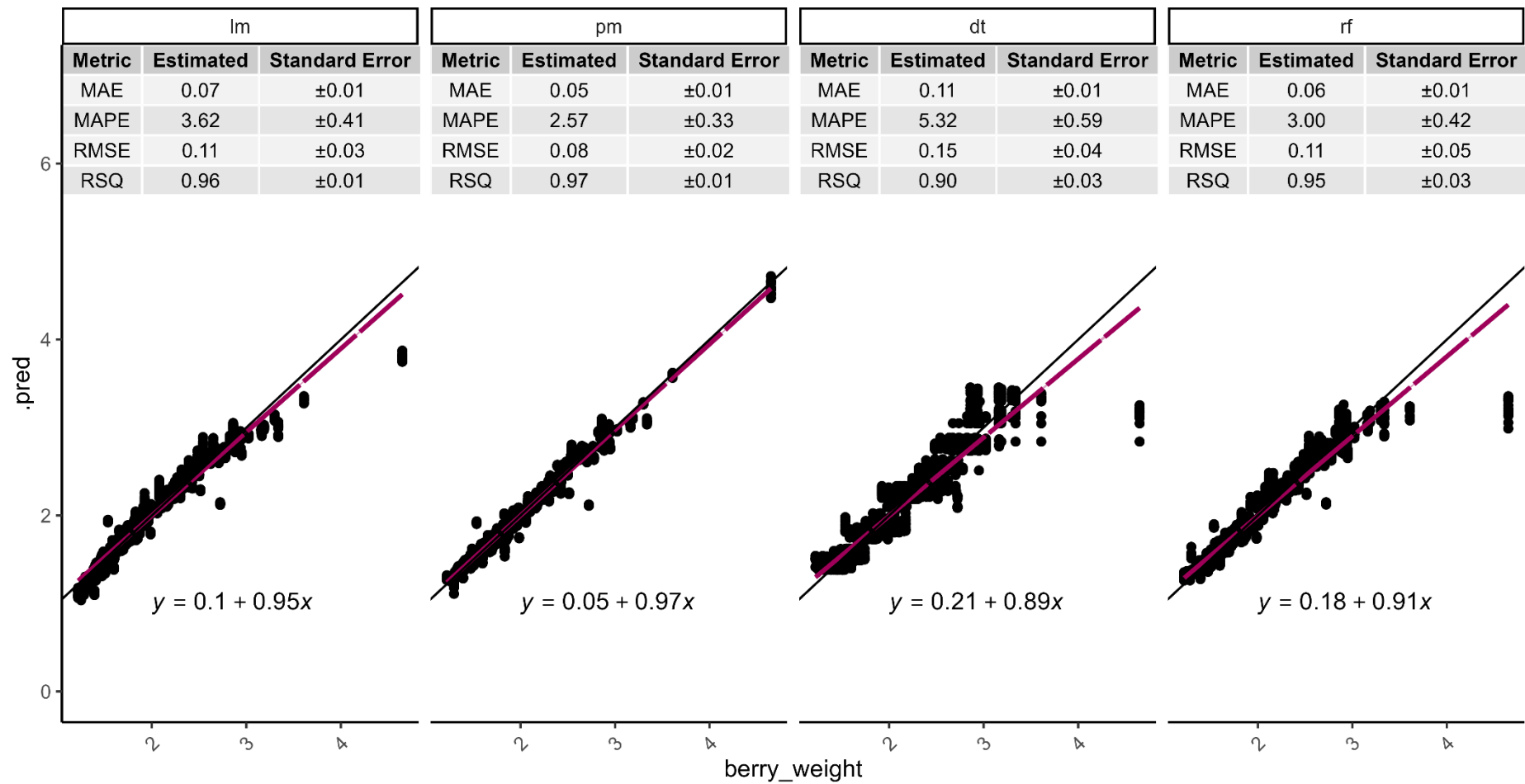
4.3.1. Validación cruzada de los modelos.

La evaluación mediante validación cruzada reveló diferencias claras en el desempeño de los cuatro enfoques (ver Figura 6). En la validación cruzada, el modelo polinómico obtuvo el menor MAE (0,05 g), lo que denotó un ajuste muy preciso de las diferencias absolutas entre predicciones y valores observados. La regresión lineal presentó un MAE ligeramente mayor (0,07 g), seguido por Random Forest (0,06 g). El árbol de decisión registró el MAE más alto (0,11 g), reflejando su menor capacidad para ajustar de forma consistente los valores individuales.

Figura 6.

Comparación de valores predichos y observados en validación cruzada para los modelos.

Cross-Validation: Predicted value vs real value



El MAPE puso de manifiesto la precisión relativa de los enfoques. El modelo polinómico alcanzó el menor MAPE (2,57 %), seguido por Random Forest (3,00 %) y regresión lineal (3,62 %). El árbol de decisión exhibió el MAPE más elevado (5,32 %), indicando que sus errores relativos fueron sustancialmente mayores, especialmente en rangos de peso bajo y alto.

El RMSE destacó la dispersión de los errores al penalizar las desviaciones grandes. El modelo polinómico volvió a liderar con un RMSE de 0,08 g, mientras que la regresión lineal obtuvo 0,11 g. Random Forest presentó un RMSE de 0,11 g, muy similar al lineal, y el árbol de decisión se ubicó en el peor lugar con 0,15 g, mostrando su vulnerabilidad a valores atípicos. La proporción de varianza explicada fue máxima para el modelo polinómico ($R^2 = 0,97$), seguido por la regresión lineal (0,96). Random Forest alcanzó un valor de 0,95, lo que denotó alta capacidad explicativa pese a su ligera subestimación en los extremos. El árbol de decisión obtuvo el R^2 más bajo (0,90), confirmando que explicó menos variabilidad del peso de la baya que los demás métodos.

Durante la validación cruzada los enfoques basados en regresiones lineales capturaron de forma muy precisa la relación entre calibre y peso, mientras que las técnicas de árbol resultaron menos exactas en este caso particular.

4.3.2. Testing de los modelos.

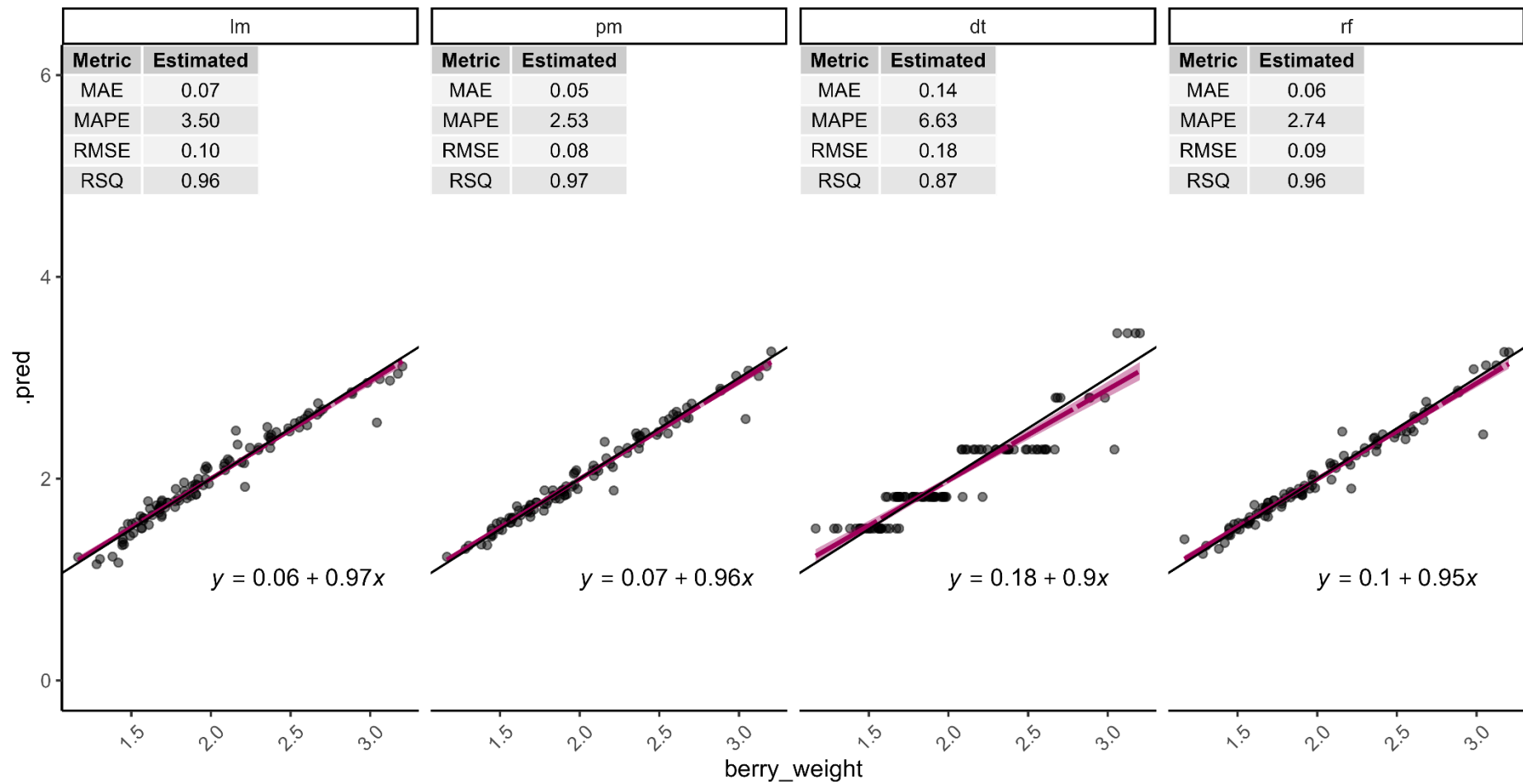
4.3.2.1. *Desempeño global por métricas.*

Para una visión global del desempeño de cada modelo en el conjunto de prueba, se compararon las cuatro métricas clave que aparecen en la Figura 7. En el caso del error medio absoluto, el modelo polinómico obtuvo el menor valor (0,05 g), lo que reflejó su capacidad superior para minimizar la discrepancia promedio entre las predicciones y las observaciones reales.

Figura 7.

Comparación de valores predichos y observados en el conjunto de prueba para los modelos.

Test: Predicted value vs real value



La regresión lineal presentó un MAE ligeramente mayor (0,07 g), mientras que Random Forest quedó en un punto intermedio (0,06 g). Por su parte, el árbol de decisión mostró un MAE de 0,14 g, evidenciando su menor precisión al reproducir individualmente los pesos de las bayas. Al analizar el error medio porcentual absoluto, se constató que el modelo polinómico lideró nuevamente la tabla, con un MAPE de 2,53 %, lo que indicó que sus desviaciones relativas respecto al peso observado fueron las más reducidas. Random Forest registró un MAPE de 2,74 %, mientras que la regresión lineal alcanzó 3,50 %. El árbol de decisión, con un MAPE de 6,63 %, exhibió la mayor variabilidad relativa en las predicciones, especialmente en los extremos del rango de pesos.

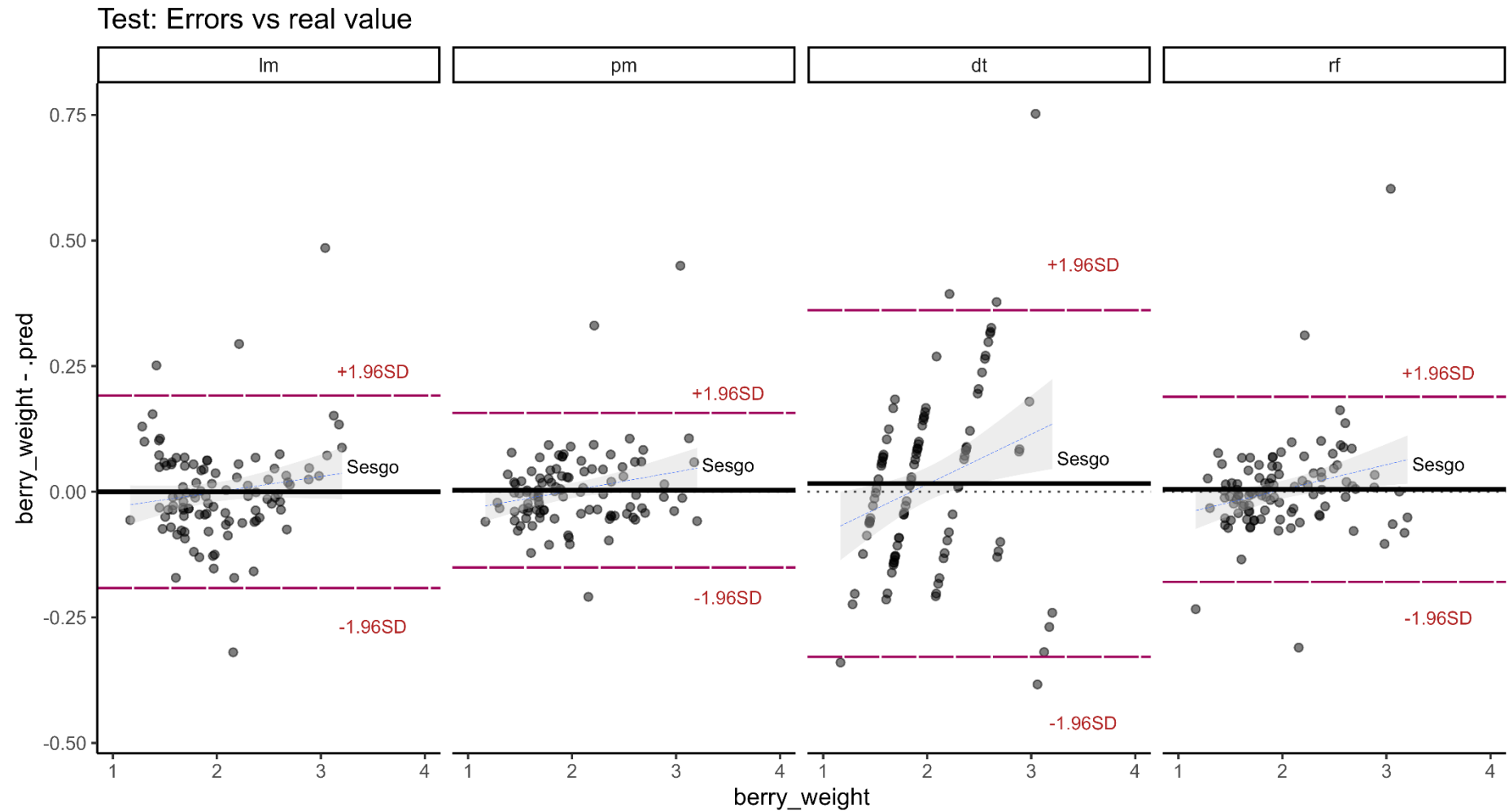
La raíz del error cuadrático medio puso de manifiesto el grado de dispersión de los errores, penalizando de forma más severa las desviaciones elevadas. El modelo polinómico obtuvo un RMSE de 0,08 g, que contrastó con el valor de 0,10 g de la regresión lineal y el de 0,09 g de Random Forest. El árbol de decisión alcanzó un RMSE de 0,18 g, confirmando su vulnerabilidad ante predicciones atípicas y su menor robustez en datos no vistos. El coeficiente de determinación destacó que tanto el modelo polinómico como la regresión lineal explicaron el 97 % y el 96 % de la variabilidad del peso de las bayas, respectivamente. Random Forest igualó el R^2 de la regresión lineal (0,96), lo que evidenció su casi equivalente capacidad explicativa, a pesar de sus ligeras subestimaciones en valores altos. En contraste, el árbol de decisión alcanzó un R^2 de 0,87, lo que indicó que dejó de capturar un 13 % de la variabilidad observada, confirmando su menor adecuación en este contexto.

4.3.2.2. *Distribución de errores.*

Para evaluar la distribución de los errores en función del peso real de las bayas, se graficaron las diferencias entre los valores observados y predichos junto con los límites de $\pm 1,96$ desviaciones estándar (ver Figura 8).

Figura 8.

Distribución de errores frente al peso real de la baya en el conjunto de prueba.



En el modelo lineal simple, los errores se dispersaron de manera simétrica alrededor de cero, con pocos valores excediendo los límites de confianza y una ligera tendencia al aumento de sesgo conforme crecía el peso real, aunque dentro de márgenes reducidos. El modelo polinómico mostró un comportamiento similar, presentando una nube de puntos más concentrada y una pendiente del sesgo apenas positiva, lo que indicó estabilidad en su ajuste a lo largo de todo el rango de pesos.

En contraste, el árbol de decisión mostró un sesgo más pronunciado, con varios errores positivos cuando el peso superó los 2,0 g y algunos valores fuera de los límites de confianza, lo que evidencia dificultades para generalizar en rangos extremos. Por su parte, Random Forest presentó un mejor comportamiento, con errores más dispersos pero controlados, manteniendo la mayoría de las observaciones dentro de los límites y con un sesgo mínimo. En conjunto, estos resultados indican que los modelos lineales y de ensamblaje generan errores más estables, mientras que el árbol de decisión es más sensible a la variabilidad de los datos.

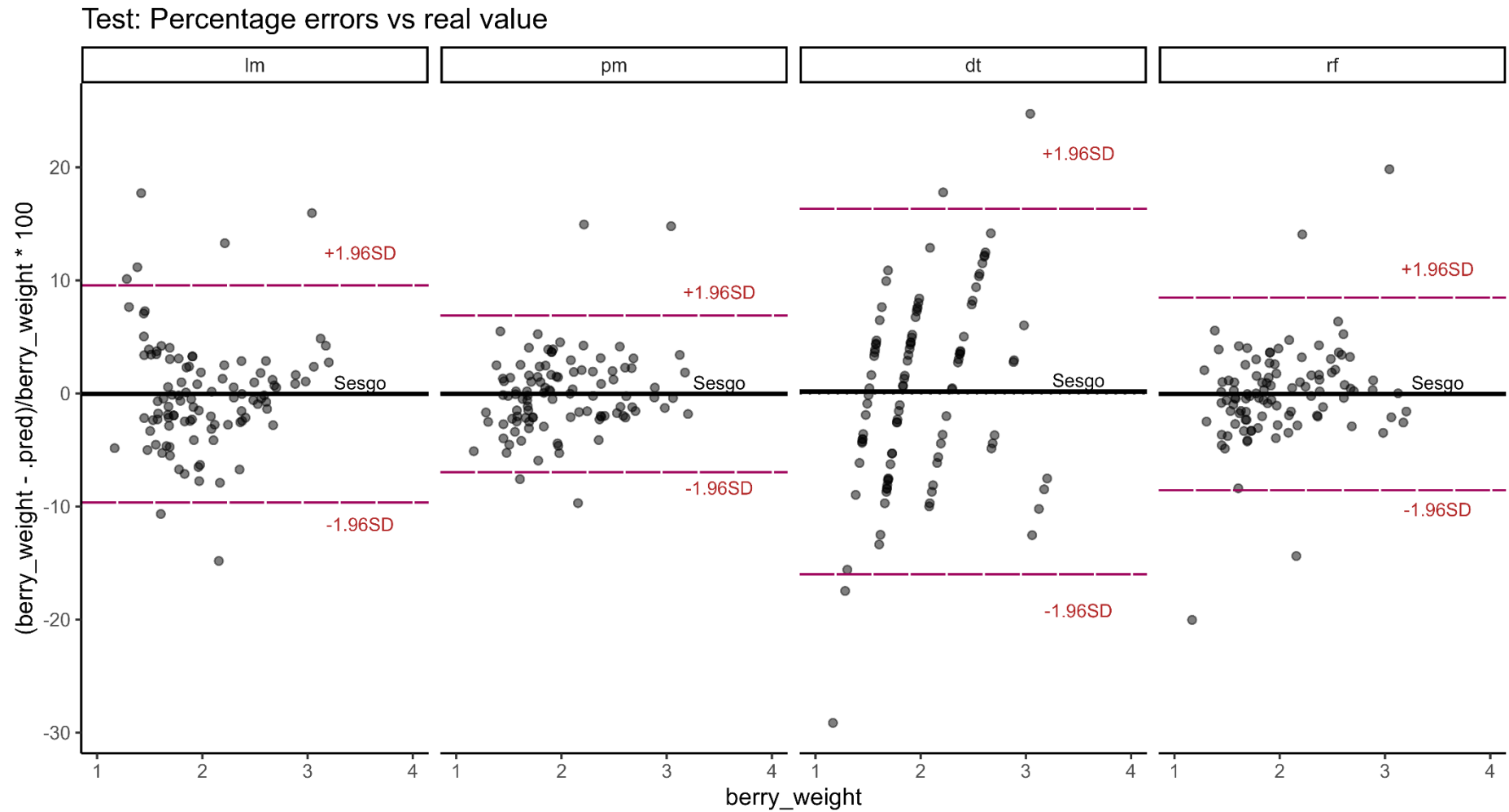
Desde el punto de vista operativo, es fundamental que los errores de predicción del peso de las bayas se mantengan bajos, ya que esto permite una mejor planificación de la cosecha, una clasificación más precisa por calibre y una estimación más confiable de los ingresos. Modelos con errores controlados facilitan además la gestión de la recolección y el manejo poscosecha, reduciendo pérdidas económicas.

4.3.2.3. *Distribución de errores porcentuales.*

Para evaluar la variación relativa de los errores a lo largo del rango de pesos, se graficaron los porcentajes de error en función del peso real de las bayas, incluyendo los límites de $\pm 1,96$ desviaciones estándar (ver Figura 9).

Figura 9.

Distribución de errores porcentuales frente al peso real de la baya en el conjunto de prueba.



En el modelo lineal, la mayoría de los errores porcentuales se concentró cerca de cero, con unos pocos casos secundarios que alcanzaron hasta el 20 %, pero sin una tendencia marcada a aumentar con el peso. El modelo polinómico presentó una dispersión aún más contenida, manteniendo casi todos los puntos dentro de $\pm 10\%$ y exhibiendo un sesgo imperceptible a lo largo de toda la escala de pesos. Por el contrario, el árbol de decisión mostró una nube de errores porcentuales muy amplia, superando en varios casos el 20 %, especialmente para bayas de peso intermedio, lo que reflejó su falta de consistencia en rangos particulares. Random Forest combinó una dispersión moderada de los errores con la mayoría de observaciones dentro de $\pm 10\%$, demostrando un sesgo muy ligero y la mejor estabilidad relativa entre todos los enfoques.

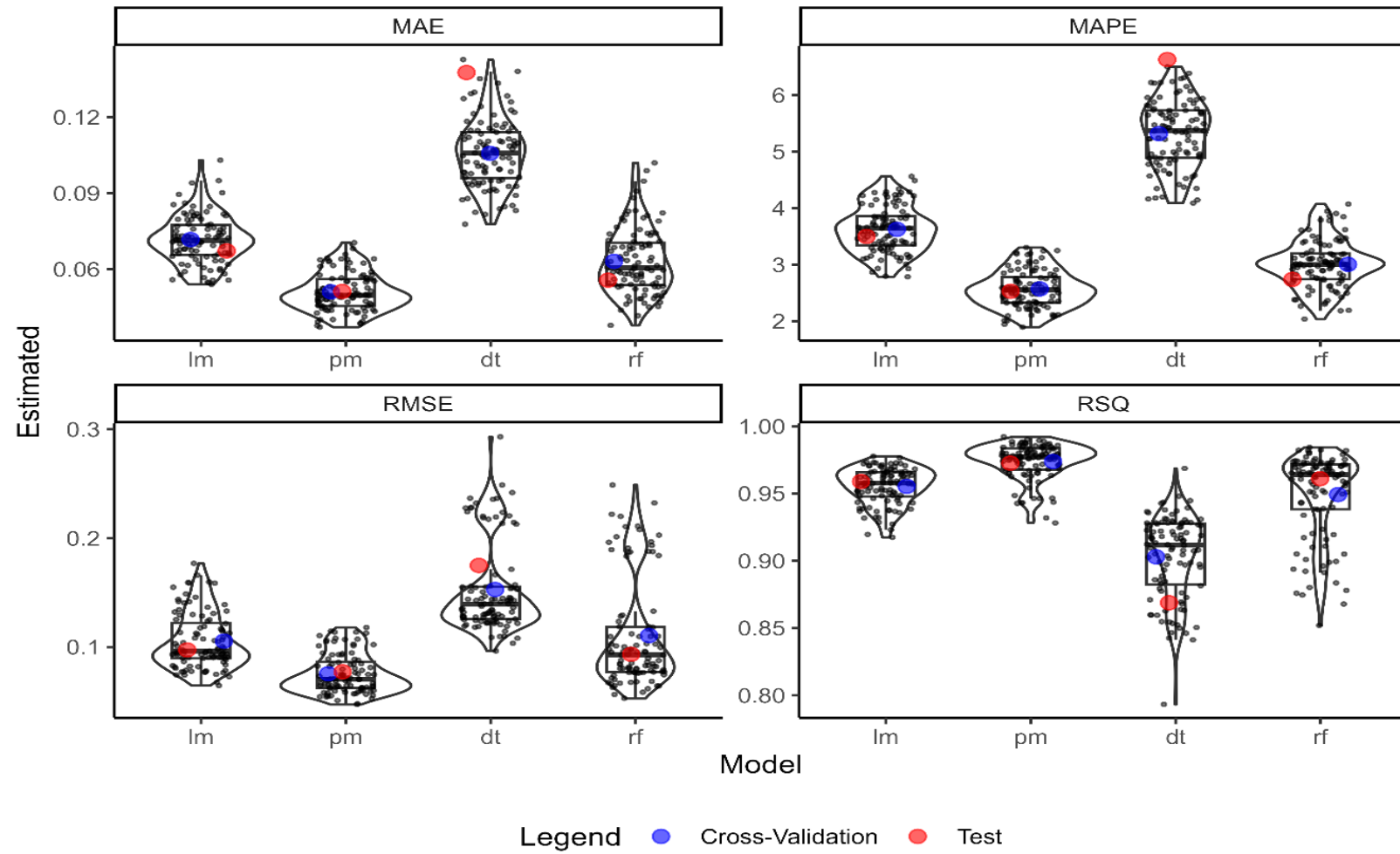
Mantener los errores porcentuales del peso de las bayas dentro de márgenes estrechos resulta esencial para asegurar una correcta clasificación de calibres, estandarizar precios de venta y optimizar los procesos de selección y empaque. Un modelo que minimice las desviaciones relativas permite calibrar con mayor precisión las estimaciones de rendimiento y facilitar la planificación logística, reduciendo mermas y mejorando la consistencia de la calidad ofertada.

4.3.3. Comparativa general de los modelos

Para comparar la estabilidad y la generalización de cada enfoque, se graficaron las distribuciones de las cuatro métricas tanto en validación cruzada (puntos azules) como en prueba externa (puntos rojos) (ver Figura 10).

Figura 10.

Distribución de métricas de desempeño en validación cruzada y prueba para los modelos evaluados.



En cuanto al error medio absoluto, el modelo polinómico exhibió la distribución más concentrada en torno a valores bajos, con medianas cercanas a 0,05 g en ambos escenarios, lo que evidenció su consistencia. La regresión lineal presentó medianas ligeramente superiores (alrededor de 0,07 g) y una variabilidad moderada, mientras que Random Forest mostró un MAE próximo a 0,06 g con una dispersión algo mayor en la prueba. El árbol de decisión registró la mayor dispersión y medianas alrededor de 0,13 g, indicando su menor robustez frente a los demás métodos.

En el error medio porcentual absoluto, los valores del modelo polinómico se concentraron alrededor de 2,6 % tanto en validación cruzada como en prueba, confirmando su precisión relativa. Random Forest se mantuvo en torno a 3 %, con ligeras diferencias entre los dos escenarios, y la regresión lineal alcanzó medianas cercanas a 3,5 %. El árbol de decisión presentó un amplio rango de MAPE, con medianas próximas al 6 % y casos extremos que superaron el 7 %, lo que subrayó su inestabilidad en términos relativos.

La raíz del error cuadrático medio mostró que el polinómico mantuvo valores reducidos y muy consistentes, con medianas cercanas a 0,08 g y una dispersión mínima. La regresión lineal y Random Forest obtuvieron medianas próximas a 0,10 g, aunque Random Forest presentó algo más de variabilidad en la prueba. El árbol de decisión, en contraste, alcanzó medianas alrededor de 0,17 g y una distribución muy amplia, reflejo de su sensibilidad a errores puntuales.

Finalmente, el coeficiente de determinación confirmó que el modelo polinómico explicó alrededor del 97 % de la varianza en ambos escenarios, con muy poca variación entre validación y prueba. La regresión lineal y Random Forest alcanzaron medianas de R^2 cercanas al 96 %, mostrando solo un ligero descenso en la prueba externa. El árbol de decisión obtuvo la mayor dispersión en R^2 , con medianas alrededor de 0,92 y valores que

disminuyeron hasta cerca de 0,85 en algunos pliegues de prueba, lo que ratificó su menor capacidad de generalización.

Desde un punto de vista técnico, la adopción de un modelo de complejidad moderada, como la regresión polinómica de grado bajo, ofreció un excelente balance entre desempeño y simplicidad. Al mantener la estructura del modelo transparente, resultó posible validar de forma consistente sus métricas en cada pliegue de la validación cruzada y comprobar su generalización en el conjunto de prueba, reduciendo drásticamente el riesgo de sobreajuste. Por el contrario, los enfoques demasiado flexibles, como el árbol de decisión, mostraron variaciones pronunciadas en MAE, MAPE, RMSE y R^2 entre validación y prueba, lo que dificulta contrastar resultados y compromete la reproducibilidad. Random Forest, pese a su mayor complejidad interna, logró estabilizar sus métricas gracias al ensamblaje, aunque a costa de una interpretabilidad reducida. En aplicaciones agrícolas, donde los tomadores de decisión requieren modelos que puedan auditarse y entenderse con facilidad, elegir una técnica de baja complejidad, pero con rendimiento robusto en múltiples particiones facilita la comparación de métricas y asegura una implementación confiable en entornos productivos.

V. DISCUSIÓN DE RESULTADOS

Objetivo específico 1: Preprocesar los datos históricos de calibre y peso de bayas de arándano para mejorar la calidad de los datos y el desempeño de los modelos predictivos.

Hipótesis específica: El preprocesamiento de los datos históricos del calibre y peso de las bayas de arándano mejorará el desempeño de los modelos predictivos.

Para asegurar la calidad e integridad de los datos históricos se diseñó un preprocesamiento reproducible basado en el paquete *recipes* del ecosistema *tidymodels*, que incluyó la asignación explícita de roles a las variables, transformaciones logarítmicas para estabilizar la varianza, generación de firmas temporales y selección automática de predictores mediante penalización Lasso. Este flujo de trabajo se preparó y aplicó de forma idéntica a los subconjuntos de entrenamiento y prueba, garantizando la misma estructura de transformaciones en todas las fases del análisis. Como resultado, los modelos construidos sobre esta base mostraron errores reducidos y coeficientes estables en validación cruzada, y reprodujeron fielmente la distribución original de los datos en el conjunto de prueba, lo que confirmó que un preprocesamiento riguroso mejora el desempeño predictivo.

Estos hallazgos complementaron y ampliaron lo reportado por Dávila y Vásquez (2025), quienes emplearon One-Hot Encoding, natural splines y validación cruzada temporal, formalizando el flujo de trabajo documentado asegurando la trazabilidad de cada etapa. De manera similar, Alegría (2024) y Jorquera et al. (2017) aplicaron particiones aleatorias y regresión lineal, pero no integraron transformaciones parametrizadas ni selección automática de variables en un único preprocesamiento, lo que pudo introducir variabilidad entre pruebas. En contraste, el diseño modular de *recipes* permitió controlar la consistencia de las transformaciones, optimizar la selección de predictores y mantener la integridad de los datos a lo largo de todo el proceso, validando así la hipótesis de que un

preprocesamiento reproducible y estructurado mejora significativamente el rendimiento de los modelos predictivos.

Objetivo específico 2: Evaluar el desempeño de los modelos predictivos basados en técnicas de machine learning con los modelos estadísticos tradicionales para identificar la técnica más precisa y robusta. **Hipótesis específica:** Los modelos predictivos basados en técnicas de machine learning tendrán un desempeño más preciso y robusto que los modelos estadísticos tradicionales al analizar e identificar patrones en los datos utilizados.

La comparación de desempeño entre los cuatro enfoques reveló que, si bien Random Forest superó al árbol de decisión en todas las métricas y mostró una leve ventaja frente a la regresión lineal simple, no alcanzó al modelo polinómico de grado dos, el cual obtuvo los menores MAE y RMSE y la mayor proporción de varianza explicada tanto en validación cruzada como en prueba externa. De este modo, los modelos de machine learning evaluados no ofrecieron en conjunto una superioridad contundente sobre los métodos estadísticos tradicionales; únicamente el ensamblaje de árboles mantuvo una precisión comparable a la regresión lineal, mientras que el árbol único resultó claramente menos robusto. Estos resultados difirieron de lo reportado por Dávila y Vásquez (2025), quienes encontraron que Random Forest superaba con claridad a la regresión lineal múltiple y a los modelos aditivos mixtos generalizados al predecir calibre fenológico, confirmando allí la hipótesis de que las técnicas de machine learning muestran mayor precisión. En cambio, coincidieron con los hallazgos de Alegría (2024) y Jorquera et al. (2017), que mostraron cómo modelos de regresión bien formulados podían explicar más del 90 % de la variabilidad del peso de la baya con un error relativo menor al 7 %, gracias a una adecuada ingeniería de variables y un diseño experimental riguroso. Asimismo, los altos R^2 y bajos errores absolutos del modelo polinómico validaron que, en contextos con relaciones casi lineales y sin gran dimensionalidad de predictores, los métodos clásicos siguen siendo la opción más sencilla,

interpretable y, en este caso, incluso más precisa. En consecuencia, la hipótesis específica solo fue parcialmente confirmada: el ensamblaje de árboles demostró robustez similar a la de los métodos tradicionales, pero no logró superar al modelo polinómico optimizado, lo que sugiere que la elección técnica debe fundamentarse también en la complejidad del problema y en la necesidad de interpretabilidad y se recomienda usar el modelo polinómico de grado dos para resolver el problema de pronóstico de peso de baya basado en calibres.

Objetivo específico 3: Seleccionar el modelo predictivo con mejor desempeño mediante la evaluación de diferentes algoritmos de machine learning, optimizando sus hiperparámetros. **Hipótesis específica:** Existe un modelo predictivo basado en técnicas de machine learning que tiene un mejor desempeño en la predicción del peso de bayas de arándano en función de su calibre que los modelos basados en técnicas estadísticas tradicionales.

Para cumplir con el tercer objetivo, se aplicó una búsqueda bayesiana de hiperparámetros a los modelos basados en árboles y se comparó su desempeño con los enfoques de regresión. Tras optimizar el número de árboles, la profundidad máxima y la selección aleatoria de variables de Random Forest, este algoritmo alcanzó un error medio absoluto de 0,06 g, un RMSE de 0,09 g, un MAPE de 2,74 % y un R^2 de 0,96 en el conjunto de prueba, superando claramente al árbol de decisión ($MAE = 0,14$ g; $R^2 = 0,87$) y presentándose como la mejor alternativa dentro de las técnicas de machine learning evaluadas. Estos resultados coincidieron con los hallazgos de Dávila y Vásquez (2025), quienes igualmente identificaron a Random Forest como la técnica más robusta para predecir calibre fenológico, y con Niedbala et al. (2022), que reportaron el éxito de XGBoost frente a otros algoritmos de aprendizaje automático. No obstante, cabe destacar que, en este estudio, la regresión polinómica optimizada también mostró un desempeño excepcional ($MAE = 0,05$ g; $R^2 = 0,97$), lo que situó al enfoque paramétrico al mismo nivel o ligeramente

por encima de Random Forest. Este hallazgo refuerza lo observado por Alegría (2024) y Jorquera et al. (2017), quienes demostraron que modelos de regresión correctamente calibrados pueden rivalizar con técnicas más complejas cuando la relación entre calibre y peso es casi lineal. En consecuencia, aunque la hipótesis específica se confirmó al identificar un modelo de machine learning (Random Forest con hiperparámetros afinados) de desempeño excepcional, el análisis también evidenció que, bajo las condiciones de este problema y con ingeniería de variables adecuada, un modelo estadístico tradicional como la regresión polinómica de grado moderado puede alcanzar niveles de precisión iguales o superiores. Esto subraya la importancia de evaluar la complejidad del modelo frente al beneficio real de mejoras marginales en la exactitud predictiva.

Objetivo general: Desarrollar modelos predictivos basados en técnicas de machine learning para pronosticar con precisión el peso de bayas de arándano en función de su calibre. **Hipótesis general:** Es posible predecir con precisión el peso de las bayas de arándano en función de su calibre utilizando modelos de machine learning.

En esta investigación se desarrollaron y compararon cuatro enfoques predictivos, regresión lineal simple, regresión polinómica de grado dos, árbol de decisión y Random Forest, para estimar el peso de las bayas de arándano a partir de su calibre. El modelo polinómico de segundo grado se destacó por su precisión, alcanzando en validación cruzada un error medio absoluto de 0,05 g, un MAPE de 2,57 % y un coeficiente de determinación de 0,97; en el conjunto de prueba externa, mantuvo valores similares (MAE = 0,05 g; MAPE = 2,53 %; $R^2 = 0,97$). Estos resultados indicaron que, con una ingeniería de variables adecuada y sin recurrir a estructuras complejas, fue posible capturar la relación casi perfecta entre calibre y peso. Los hallazgos de esta tesis corroboraron y, en algunos casos, mejoraron los resultados reportados en la literatura. Alegría (2024) determinó mediante regresión lineal correlaciones superiores al 90 % entre diámetro y peso, comparable al 96 %–97 % de

varianza explicada por nuestros modelos lineales y polinómicos. Li et al. (2025) observaron una correlación superior al 96 % al predecir peso con YOLOv5 modificado sobre imágenes, un desempeño muy parecido al $R^2 = 0,97$ logrado aquí, pero con la ventaja de no requerir procesamiento de imagen. Dávila y Vásquez (2025) emplearon Random Forest para predecir calibre fenológico y reportaron un MAPE de $2,44 \% \pm 0,13 \%$; en nuestra prueba, Random Forest alcanzó un MAPE de 2,74 %, mientras que la regresión polinómica mantuvo el error relativo por debajo del 2,6 %, sugiriendo que un modelo más sencillo puede igualar el rendimiento del ensamblaje sin sacrificar interpretabilidad. Asimismo, el modelo alométrico de Jorquera et al. (2017) obtuvo RRMSE (Raíz del error cuadrático medio relativo) entre 4,7 % y 6,9 %, claramente superiores al MAPE de 2,5 % logrado aquí. Estas comparaciones confirman que la aplicación de técnicas de machine learning “interpretables” puede ofrecer precisión y robustez equivalentes o superiores a enfoques más complejos, permitiendo desarrollar modelos predictivos fiables y contrastables en distintos esquemas de validación, confirmando que la hipótesis general es verdadera.

VI. CONCLUSIONES

Se lograron desarrollar modelos predictivos basados en técnicas de machine learning que pronosticaron con alta precisión el peso de las bayas de arándano en función de su calibre. El modelo polinómico de segundo grado evidenció un desempeño superior con un error medio absoluto de 0,05 g, un error medio porcentual absoluto cercano al 2,5 % y un coeficiente de determinación de 0,97 tanto en validación cruzada como en prueba externa, lo que confirmó su capacidad para capturar las relaciones no lineales entre el calibre y el peso de la baya.

El preprocesamiento reproducible implementado con el paquete *recipes* garantizó la calidad e integridad de los datos históricos de calibre y peso, al validar que las particiones de entrenamiento y prueba conservaron fielmente la distribución original y al aplicar de forma idéntica las transformaciones logarítmicas y de escalamiento en todas las fases del análisis, asegurando consistencia en el preprocesamiento.

La comparación entre técnicas de machine learning y modelos estadísticos tradicionales mostró que la regresión polinómica de segundo grado superó a la regresión lineal, al árbol de decisión y a Random Forest en métricas como RMSE, MAE y R^2 . Esto evidencia la importancia de capturar relaciones no lineales para lograr un mejor desempeño en datos no vistos.

La selección del modelo se realizó mediante la evaluación de distintos algoritmos y la optimización de sus hiperparámetros. En este proceso, el modelo polinómico de bajo grado ofreció el mejor equilibrio entre interpretabilidad y precisión, lo que facilita su aplicación en la planificación de cosecha y en la clasificación de frutos.

VII. RECOMENDACIONES

Para profundizar en la capacidad predictiva, se propone incorporar arquitecturas ligeras de aprendizaje profundo y métodos de ensamblaje ponderado, con el fin de mejorar la generalización y la robustez frente a variaciones en el calibre de las bayas. Asimismo, resulta necesario ampliar la variedad de algoritmos evaluados más allá de los enfoques lineales y de ensamblaje tradicionales, buscando mejoras en el desempeño predictivo en distintos escenarios.

Con el objetivo de asegurar la integridad y trazabilidad de los datos, futuras investigaciones deberían incorporar validaciones automáticas de calidad y sistemas de versionado que registren cada transformación aplicada, permitiendo rastrear los cambios desde las fuentes originales hasta el preprocesamiento.

Para obtener comparaciones más sólidas entre modelos de machine learning y métodos estadísticos, se recomienda aplicar pruebas de significancia sobre las distribuciones de error en validación cruzada y prueba externa, complementadas con un análisis conjunto de métricas como error absoluto, error porcentual y varianza explicada.

Al priorizar simultáneamente la precisión y la interpretabilidad, es útil adoptar un criterio de selección del modelo óptimo que integrara tanto la precisión predictiva como la parsimonia y la interpretabilidad. Se considera adecuado implantar un esquema de validación cruzada anidada junto con optimización bayesiana de hiperparámetros, facilitando así la selección de un modelo polinómico de bajo grado que ofreciera el mejor equilibrio entre exactitud predictiva y simplicidad estructural.

VIII. REFERENCIAS

- Agraria.pe (2021). Fall Creek: *Perú tiene el rendimiento productivo promedio más alto de arándanos en el mundo*. Recuperado de <https://agraria.pe/noticias/fall-creek-peru-tiene-el-rendimiento-productivo-promedio-de--30244>
- Alcalde, K. (2019). *Estimación de pérdidas causadas por plagas en la calidad postcosecha de Vaccinium corymbosum “arándano”* (tesis de pregrado). Universidad Privada Antenor Orrego, Trujillo, Perú.
- Alegría, G.H. (2024). *Modelo de regresión linealizable para estimar el peso en función del calibre de arándano* (tesis de pregrado). Universidad Nacional de Trujillo, Trujillo, Perú.
- Arreguín, J.G. (2019). *Desarrollo tecnológico en agricultura de precisión: Perspectivas para México* (tesis de maestría). Instituto Politécnico Nacional, Ciudad de México, México.
- Astorga, J.M. (2014). Aplicación de modelos de regresión lineal para determinar las armónicas de tensión y corriente Ingeniería Energética. 35(3), 223-241. Recuperado de: <https://www.redalyc.org/pdf/3291/329132445008.pdf>.
- Bañados, M. P. (2006). Blueberry production in South America. *Acta Horticulturae*, 715, 165–172. <https://doi.org/10.17660/ActaHortic.2006.715.24>
- Bishop, C. M. (2006). *Pattern Recognition and Machine Learning*. Springer.
- Breiman, L. (2001). Random Forests. *Machine Learning* 45, 5–32. Doi: <https://doi.org/10.1023/A:1010933404324>
- Buzeta, A. (1997). *Chile: Berries para el 2000* (1.^a ed.). Fundación Chile.
- Calisaya, D.A. y Espinoza, G.E. (2014). *Desarrollo de un protocolo para el establecimiento y multiplicación in vitro de Vaccinium corymbosum L. (arándano de arbusto alto) variedad Misty, a partir de segmentos nodales en un Reactor de Inmersión Temporal* (tesis de pregrado). Universidad Católica de Santa María, Arequipa, Perú.
- Carrero, M. (2018). *Análisis de los factores que influirían en la adopción de la agricultura de precisión en el departamento de Lambayeque* (tesis de pregrado). Universidad Nacional Pedro Ruiz Gallo, Lambayeque, Perú.
- Chanchí, G.E. y Campo, W.Y. (2020). Aplicación de la regresión polinomial para la caracterización de la curva del COVID-19, mediante técnicas de machine learning. 8(2). Recuperado de: <https://revistas.unisimon.edu.co/index.php/innovacioning/article/view/4103>.
- Congreso de la República. (2023). *Ley N.º 31814: Ley que promueve el uso de la inteligencia artificial en favor del desarrollo económico y social del país*. Diario Oficial El Peruano.
- Crisóstomo, M. N. et al. (2014). *Relaciones amonio/nitrato en soluciones nutritivas ácidas y alcalinas para arándano*. *Revista Mexicana de Ciencias Agrícolas*, 5(3), 525–532. <https://doi.org/10.29312/remexca.v5i3.956>
- Dalal, K. R. (2020). *Analysing the role of supervised and unsupervised machine learning in IoT*. In *Proceedings of the 2020 International Conference on Electronics and Sustainable Communication Systems (ICESC)* (pp. 75–79). IEEE. <https://doi.org/10.1109/ICESC48915.2020.9155761>
- Dávila, F. R. y Vásquez, C. R. A. (2025). *Predicción de calibres de bayas del arándano Vaccinium corymbosum L. empleando variables fenológicas y técnicas de aprendizaje*

- automático* (tesis de pregrado). Universidad Nacional Pedro Ruiz Gallo, Lambayeque, Perú.
- Defilippi, B., & Ulloa, P. (Eds.). (2023). Calidad en arándanos: Bases fisiológicas y tecnológicas para el manejo de cosecha y postcosecha [Boletín INIA N° 477]. *Instituto de Investigaciones Agropecuarias (INIA)*. Recuperado de: <https://hdl.handle.net/20.500.14001/68950>
- Doi, K. Inoue, R. y Iwasaki, N. (2021). Seed weight mediates effects of pollen on berry weight, ripening, and anthocyanin content in highbush blueberry. *Scientia Horticulturae*. Recuperado de: <https://www.sciencedirect.com/science/article/abs/pii/S0304423821004209?via%3DiHub>.
- Espinoza, J.G. (2022). *Dosis de quitosano en el rendimiento y parámetros de calidad de arándanos (Vaccinium Corymbosum L. cv. Biloxi), en condiciones climáticas del Valle De Salas, Ica, 2020* (tesis de pregrado). Universidad Nacional Hermilio Valdizán, Huánuco, Perú.
- Estacio, C.I. (2019). *Identificación de especies fúngicas asociadas al decaimiento de plantas en el cultivo de arándano (Vaccinium corymbosum) en la Región Ica y el valle de Cañete* (tesis de pregrado). Universidad Nacional San Luis Gonzaga, Ica, Perú.
- Fall Creek (2021). *Fall Creek® U.S. & Canada 2021–2022 catalog*. Calaméo. Recuperado de: <https://www.calameo.com/books/0072208404d630574647b>
- Flarence, A. R., Bethu, S., Sowmya, V., Anusha, K., & Sankara Babu, B. (2018). Importance of supervised learning in prediction analysis. *Periodicals of Engineering and Natural Sciences*, 6(1), 201–214. <https://doi.org/10.21533/pen.v6i1.288>
- García, G., & Rubio, J. (2005). *El cultivo de arándanos en Asturias*. SERIDA.
- García, J. C., & García, G. (2010). *Guía de cultivo: Orientaciones para el cultivo del arándano*. Proyecto de cooperación “Nuevos Horizontes”, Ministerio de Medio Ambiente y Medio Rural y Marino; Servicio Regional de Investigación y Desarrollo Agroalimentario (SERIDA). Recuperado de: https://frutales.files.wordpress.com/2011/05/guia_del_arandano.pdf
- Gil, G. (2000). *Fruticultura: El potencial productivo: Crecimiento vegetativo y diseño de huertos y viñedos* (4.ª ed.). Ediciones Universidad Católica de Chile.
- Gómez, J. L. (2023). *Exportación de arándanos: Una fruta que resiste los embates de El Niño en el Perú*. Recuperado de: <https://www.esan.edu.pe/conexion-esan/exportacion-de-arandanos-una-fruta-que-resiste-los-embates-de-el-nino-en-el-peru>
- Gonzáles, L.T.; Vargas, L.R.; Avila, J.D. y Molina, S. (2022). Implementación de la agricultura de precisión a través del desarrollo de sistemas productivos en áreas protegidas o de conservación para optimizar la producción de cultivos. Una revisión sistemática de literatura. *Cuaderno Activa*, 14, 65-77. Recuperado de: https://www.researchgate.net/publication/373154040_Implementacion_de_la_agricultura_de_precision_a_traves_del_desarrollo_de_sistemas_productivos_en_areas_prot egidas_o_de_conservacion_para_optimizar_la_produccion_de_cultivos_Una_revisio n_sistemica_de.
- Hancock, J. F., Retamales, J., Lyrene, P., Moggia, C., & Lolas, M. (1992). *Blueberry culture in Chile—Current status, future prospects*. *HortTechnology*, 2(3), 310–315. <https://doi.org/10.21273/HORTTECH.2.3.310>
- Hirzel, J. (2013). *Fertilización en arándano*. En P. Undurraga & S. Vargas (Eds.), *Manual del arándano* (Boletín INIA N.º 263, pp. 31–42). Instituto de Investigaciones Agropecuarias (INIA). Recuperado de: <https://hdl.handle.net/20.500.14001/7627>

- Hirzel, J., & Rodríguez, N. (2003). *Acidificación de suelos para plantación de arándanos*. Informativo Agropecuario - INIA.
- Hirzel, J.; Retamales, J.; Balbontín, C, Muñoz, P. y Moya, E. (2023). Effect of load regulation by winter pruning on fruit quality attributes of Duke and Legacy blueberry cultivars. *Chilean Journal of Agricultural Research*, 83(4), 418-431. Recuperado de: <https://www.scielo.cl/pdf/chiljar/v83n4/0718-5839-chiljar-83-04-418.pdf>.
- Holzapfel, E. A., Hepp, R. F., & Marino, M. A. (2004). Effect of irrigation on fruit production in blueberry. *Agricultural Water Management*, 67(3), 173–184. <https://doi.org/10.1016/j.agwat.2004.03.002>
- Huamantingo, J.A. (2016). *Evaluación del crecimiento de plantines de dos variedades de arándano (Vaccinium corymbosum L.) en tres pisos altitudinales a condiciones de vivero en Abancay – Apurímac* (tesis de pregrado). Universidad Tecnológica de los Andes, Abancay, Apurímac.
- James, G., Witten, D., Hastie, T., & Tibshirani, R. (2023). *An Introduction to Statistical Learning* (Springer Texts in Statistics). Springer.
- Jorquera, E., Génard, M., Ribera, A., & Franck, N. (2017). A simple allometric model for estimating blueberry fruit weight from diameter measurements. *Scientia Horticulturae*, 219, 131–134. <https://doi.org/10.1016/j.scienta.2017.03.009>
- Kumar, A. & Jain, M. (2020). *Ensemble learning for AI developers*. Berkeley, California: BAPress. Doi: <https://doi.org/10.1007/978-1-4842-5940-5>
- Kumarihami, H. M. P. C.; Park, H.G.; Kim, S.M.; Park, J.I.; Lee, E.J.; Kim, H.L.; & Kim, J. G. (2021). Flower and leaf bud density manipulation affects fruit set, leaf-to-fruit ratio, and yield in southern highbush “Misty” blueberry. *Scientia Horticulturae*, Recuperado de: <https://www.sciencedirect.com/science/article/abs/pii/S0304423821006373>.
- Li, X., Ru, S., He, Z., Spiers, J. D., & Xiang, L. (2025). High-throughput phenotyping tools for blueberry count, weight, and size estimation based on modified YOLOv5s. *Fruit Research*, 5, e012. <https://doi.org/10.48130/frures-0025-0006>
- Mathar, R., et al. (2020). *Fundamentals of Data Analytics: With a View to Machine Learning*. Springer. Doi: <https://doi.org/10.1007/978-3-030-56831-3>
- Maticorena, M.F. y Escobedo, J.A. (2024). Cinco tipos de poda en arándano (*Vaccinium Corymbosum* L. Cv. Biloxi) y su influencia en determinados parámetros productivos. *Ciencia Latina*, 8(3), 4020-4035. Recuperado de: <https://dialnet.unirioja.es/servlet/articulo?codigo=9610613>
- Meléndez, M. R., & López, M. (2021). Caracterización física y química de frutos de arándano (*Vaccinium spp.*). *Revista Agropecuaria*, 12(1), 109–118.
- Ministerio de Desarrollo Agrario y Riego (MIDAGRI). (2020). *Estacionalidad de las exportaciones peruanas de arándanos frescos y el mercado norteamericano* (Repositorio Institucional). <https://hdl.handle.net/20.500.13036/1516>
- Mitchell, T. M. (1997). *Machine Learning*. McGraw-Hill.
- Montgomery, D. C., Peck, E. A., & Vining, G. G. (2005). *Introducción al análisis de regresión lineal*. CECSA.
- Muñoz, C. (1988). *Arándano: Antecedentes generales* (Serie Carillanca, No. 02, pp. 5–13). En *Seminario: El cultivo del arándano*. Instituto de Investigaciones Agropecuarias (INIA), Estación Experimental Carillanca, Temuco, Chile.
- Muñoz, P. S. (2024). Modelo predictivo basado en algoritmos de *machine learning* para la estimación del peso de racimos de banana en una hacienda. *Revista Latinoamericana de Ciencias Sociales y Humanidades*, 5(6). <https://doi.org/10.56712/latam.v5i6.3061>

- Naroui, M.R. et al. (2015). Application of Artificial Neural Networks to predict the final fruit weight and random forest to select important variables in native population of melon (*Cucumis melo*. Pahlavan). *Scientia Horticulturae*, 181, 108–112. Doi: <https://doi.org/10.1016/j.scienta.2014.10.025>
- Niedbala, G.; Kurek, J.; Świdorski, B., Wojciechowski, T., Antoniuk, I., & Bobran, K. (2022). Prediction of Blueberry (*Vaccinium corymbosum* L.) Yield Based on Artificial Intelligence Methods. *Agriculture*, 12(12), 2089. Recuperado de: <https://doi.org/10.3390/agriculture12122089>.
- O'Brien, T. E., & Silcox, J. W. (2024). *Nonlinear regression modelling: A primer with applications and caveats*. *Bulletin of Mathematical Biology*, 86(4), Article 40. <https://doi.org/10.1007/s11538-024-01274-4>
- Obsie, E.Y. et al. (2020). Wild blueberry yield prediction using a combination of computer simulation and machine learning algorithms. *Computers and Electronics in Agriculture*, 178, 105778. Doi: <https://doi.org/10.1016/j.compag.2020.105778>
- Ochoa, S. (2010). *Rentabilidad del negocio del arándano, establecimiento del cultivo y primeros cuidados*. Consultora MS.
- Pino, C. (2007). *Descripción del desarrollo vegetativo y de las características físicas y químicas de los frutos de cuatro clones de arándano alto (Vaccinium corymbosum L.)* (tesis de pregrado). Universidad Austral de Chile, Valdivia, Chile.
- Ponte, D., Espinoza, A., Gibeaux, S., Gonzáles, C. (2021). Estado actual del aprendizaje automatizado aplicado al internet de las cosas para automatizar procesos agrícolas. *Plus Economía*, 9(2), 4-11. Recuperado de: <https://revistas.unachi.ac.pa/index.php/pluseconomia/article/view/497/429>.
- Portal Frutícola. (2024). *Campaña de arándanos peruanos ha sido más lenta, pero con volúmenes superiores*. Recuperado de: <https://www.portalfruticola.com/noticias/2024/12/13/campana-arandanos-peruanos>
- Presidencia del Consejo de Ministros. (2023). *Decreto Supremo N.º 085-2023-PCM que aprueba la Política Nacional de Transformación Digital al 2030*. Diario Oficial El Peruano.
- Presidencia del Consejo de Ministros. (2025). *Decreto Supremo N.º 115-2025-PCM que aprueba el Reglamento de la Ley N.º 31814*. Diario Oficial El Peruano.
- Pritts, M. P., & Hancock, J. F. (Eds.). (1992). *Highbush blueberry production guide* (NRAES-55). Northeast Regional Agricultural Engineering Service.
- Qu, H.; Xiang, R.; Obsie, E.Y.; Wei, D. y Drummond, F. (2021). Parameterization and Calibration of Wild Blueberry Machine Learning Models to Predict Fruit-Set in the Northeast China Bog Blueberry Agroecosystem. *Agronomy*, 11(9), 1736. Recuperado de: <https://www.mdpi.com/2073-4395/11/9/1736>.
- Quinlan, J. R. (1986). *Induction of decision trees*. *Machine Learning*, 1, 81–106. <https://doi.org/10.1007/BF00116251>
- Ramos, D.L. (2023). *Efecto de diferentes dosis y frecuencia de aplicación de CITO-GIB® (giberelina – citoquinina) sobre el rendimiento, calibre y firmeza de la fruta de arándano (Vaccinium corymbosum L.) var. Emerald en Caraz – Ancash* (tesis de pregrado). Universidad Nacional de Ucayali, Pucallpa, Perú.
- Retamales, J. B., & Hancock, J. F. (2012). *Blueberries* (Crop Production Science in Horticulture, Vol. 21). CABI.
- Retamales, J. B., & Hancock, J. F. (2018). *Blueberries* (2nd ed.). CABI.
- Rivadeneira, M. F. (2008). *Etapas fenológicas en arándanos durante las campañas 2006–2007. Avances en Cultivos No Tradicionales* (pp. 1–10). Facultad de Agronomía, Universidad de Buenos Aires (FAUBA).

- Roman, P. et al. (2014). Forecasting Fruit Size and Caliber by Means of Diffusion Processes. Application to “Valencia Late” Oranges. *Journal of Agricultural, Biological, and Environmental Statistics*, 19(2), 292–313. Doi: <http://dx.doi.org/10.1007/s13253-014-0172-3>
- Salvo, S. et al. (2012). An estimate of potential blueberry yield using regression models that relate the number of fruits to the number of flower buds and to climatic variables. *Scientia Horticulturae*, 133, 56-63. Doi: <https://doi.org/10.1016/j.scienta.2011.10.020>
- San Martín, J. (2013). *Situación varietal en arándano*. En P. Undurraga & S. Vargas (Eds.), *Manual del arándano* (Boletín INIA N.º 263, pp. 15–22). Instituto de Investigaciones Agropecuarias (INIA). <https://hdl.handle.net/20.500.14001/7627>
- Sarker, I. H. (2021). *Machine learning: Algorithms, real-world applications and research directions*. *SN Computer Science*, 2(3), 160. <https://doi.org/10.1007/s42979-021-00592-x>
- Sheikh, H., Prins, C., & Schrijvers, E. (2023). Artificial intelligence: Definition and background. En *Mission AI: The new system technology* (pp. 15–41). Springer.
- Solano, C.A. y Caballero, I.P. (2021). *Modelo de predicción de plagas en el cultivo de palto utilizando metodología de aprendizaje automático supervisado, empresa Virú S.A., 2019-2021* (tesis de pregrado). Universidad Privada Antenor Orrego, Trujillo, Perú.
- Strik, B.C. y Vance, A.J. (2019). Highbush Blueberry Cultivars Differ in the Relationship between Seed Number and Berry Weight during the Harvest Season. *HortScience*, 54(10), 1728-1736. Recuperado de: <https://journals.ashs.org/hortsci/view/journals/hortsci/54/10/article-p1728.xml>.
- Szretter, M. E. (2013). *Apunte de regresión lineal*. Buenos Aires, Argentina.
- Tartabull, T., & Betancourt, C. (2016). La calidad del agua para el riego: Principales indicadores de medida y procesos que la impactan. *Revista Científica Agroecosistemas*, 4(1), 47–61. Recuperado de: <https://www.researchgate.net/publication/321951119>
- Tinoco, C. J., Zambrano, L. M., Roque, O., Chávez, R. W., Maquiña, B. M., & Espejo, J. W. (2023). Los arándanos, generalidades y desarrollo en el mercado mundial: Una revisión de literatura. *Paideia XXI*, 13(1), 125–140.
- Valenzuela, B. (1988). *Requerimientos agroclimáticos de las especies de arándanos*. En W. Lobos A. (Ed.), *Seminario: El cultivo del arándano* (pp. 15–23). Instituto de Investigaciones Agropecuarias (INIA), Estación Experimental Carillanca, Temuco, Chile.
- Vidal, I. R. (2019). *Fertirrigación: desde la teoría a la práctica*. Universidad de Concepción, Editorial Universidad de Concepción.
- Vivot, E., Rugnaa, C. M., Gieco, A. M., Sánchez, C. I., Ormaechea, M. V., & Sequin, C. J. (2010). *Calidad del agua subterránea para usos agropecuarios en el departamento Villaguay, Entre Ríos*. AUGMDOMUS. Recuperado de: <https://revistas.unlp.edu.ar/domus/article/view/96/159>
- Williamson, J. G., & Lyrene, P. M. (1994). *Guía para el cultivo de los arándanos en Florida* (Circular 1192). Florida Cooperative Extension Service, Institute of Food and Agricultural Sciences, University of Florida.
- Zapata, L., Heredia, A., Malleret, A., Quinteros, F., Cives, H., & Carlazara, G. (2013). Evaluación de parámetros de calidad que ayuden a definir la frecuencia de recolección de bayas de arándanos. *Revista Iberoamericana de Tecnología Postcosecha*, 14(2), 186–194. Recuperado de: <https://www.redalyc.org/pdf/813/81329290013.pdf>