INFORME DEL ANÁLISIS DE LOS DETERMINANTES DE LA PRODUCTIVIDAD DE ARROZ 2023

PIDARA

Proyecto Integral de Diversificación
Agroproductiva y Reconversión Agrícola

EL NUEVO ECUADOR RESUELVE

Ministerio de Agricultura y Ganadería

Informe del análisis de los determinantes de la productividad de arroz 2023

Proyecto Integral de Diversificación Agroproductiva y Reconversión Agrícola-PIDARA

Quito – Ecuador

2024

RESUMEN

Este informe analiza los rendimientos del cultivo de arroz en Ecuador, enfocándose en factores agronómicos y socioeconómicos que influyen en su productividad. Utilizando metodologías como la regresión lineal múltiple y el modelo de Random Forest, se identificaron los principales determinantes del rendimiento. Los resultados destacan la importancia de un adecuado manejo de la humedad y el control de impurezas, así como la aplicación de nutrientes clave, como el nitrógeno y el potasio, para mejorar la productividad del cultivo. Se recomienda continuar investigando las mejores prácticas de fertilización y manejo agronómico, para maximizar los rendimientos y la sostenibilidad del cultivo de arroz.

Palabras clave: Arroz, rendimientos objetivos, regresión lineal múltiple, Random Forest, nutrición del cultivo, manejo de impurezas.

ABSTRACT

This report analyzes the yields of rice cultivation in Ecuador, focusing on agronomic and socioeconomic factors that influence productivity. Using methodologies such as multiple linear regression and the Random Forest model, the main determinants of yield were identified. The results highlight the importance of proper moisture management and impurity control, as well as the application of key nutrients like nitrogen and potassium, to improve crop productivity. Further research is recommended on optimal fertilization practices and agronomic management to maximize yields and the sustainability of rice cultivation

Keywords: Rice, crop yields, linear regression, Random Forest, crop nutrition, impurity management.

Contenido

1. ANTECEDENTES

1.1 Operativos de rendimientos objetivos

La Coordinación General de Información Nacional Agropecuaria (CGINA) tiene como misión el "Generar, administrar, proveer y difundir información oportuna y consistente al sector público y privado en el ámbito agropecuario, generando indicadores, análisis e informes que permitan la toma de decisiones en el sector". Siendo una de sus atribuciones el "Coordinar la generación de información agropecuaria con las instancias de la Autoridad Agraria Nacional y agentes públicos y privados en el ámbito de sus competencias". Esta Coordinación cuenta con la Dirección de Generación de Datos Agropecuarios (DGDA), donde una de sus gestiones internas se denomina "Gestión de Rendimientos Agropecuarios"; esta unidad interna, es la responsable de la planificación y ejecución del levantamiento de información denominado "Operativos de Rendimientos Objetivos (ORO)".

Este año, la DGDA ha planificado levantamientos de información para los tres períodos de cultivo del arroz; para este ciclo, todas las fases y las actividades se enmarcaron en el Plan de Gestión de los Operativos de Rendimientos Objetivos, realizado por las tres direcciones de la Coordinación General de Información Nacional Agropecuaria, de análisis, de generación de datos y de generación de geoinformación (DAIA – DGDA y DGGA); así, al finalizar este proceso se pretende realizar una fase de evaluación y retroalimentación.

Entre las variables que son levantadas se encuentran los principales factores de la producción y algunas características socioeconómicas de la persona productora, como: tipo y cantidad de fertilizantes, tipo y cantidad de semilla, acceso a riego, mecanización de labores en el cultivo, nivel de estudio de la persona productora, capacitación recibida, nivel asociativo que mantiene, entre otros.

1.2 Justificación

Estudiar los determinantes de la productividad en el cultivo de arroz es desarrollar un entendimiento profundo y basado en datos de cómo diversos factores agronómicos, ambientales y socioeconómicos contribuyen a las variaciones en el rendimiento de los cultivos. Este análisis es fundamental por varias razones:

- Al identificar qué factores tienen el mayor impacto en la productividad, las personas productoras pueden optimizar el uso de recursos, como: agua, fertilizantes y otros insumos. Esto no solo mejora la eficiencia económica, sino que también promueve prácticas de agricultura sostenible.
- Los resultados de este estudio pueden informar políticas agrícolas efectivas que apoyen la innovación, la sostenibilidad y la rentabilidad en el sector agrícola. Esto incluye políticas sobre subsidios, apoyo a la investigación agrícola y desarrollo de infraestructura.
- Al mejorar la productividad del cultivo de arroz, un componente crucial en la dieta de muchas poblaciones se contribuye a la seguridad alimentaria a nivel local y nacional. Esto es especialmente importante ya que el arroz es un alimento básico y una fuente significativa de nutrición en todo el territorio nacional.
- La identificación precisa de los determinantes de la productividad puede dirigir el desarrollo de nuevas tecnologías y prácticas agrícolas, como variedades de cultivos más resistentes y técnicas de manejo agronómico más efectivas.

2. METODOLOGÍA

La metodología utilizada para evaluar los determinantes de la productividad se centra en varios pasos clave para analizar y modelar los factores que afectan la productividad agrícola:

2.1 Selección de características

Se seleccionan las posibles variables predictoras relevantes para el estudio, incluyendo variables socioeconómicas y variables agrícolas específicas.

Para la variable de respuesta se debe seleccionar uno de los dos posibles enfoques: Seleccionar el rendimiento neto extrapolado por parcela o seleccionar el rendimiento de las plantas específicas muestreadas.

2.2 Tratamiento de datos atípicos y faltantes

Se eliminan valores atípicos utilizando métodos basados en el rango intercuartílico para evitar distorsiones en los análisis. Además, se imputan valores faltantes o se codifican de manera adecuada para preparar los datos para análisis estadísticos y de machine learning.

2.3 Transformación de variables

Otro paso es la codificación de variables categóricas la cual es crucial en el preprocesamiento de datos para análisis estadísticos y de machine learning, especialmente porque la mayoría de los algoritmos de modelado solo pueden manejar datos numéricos.

Se realizan transformaciones de variables para estandarizar los datos, como la estandarización Z-score, que ajusta los datos para que tengan media cero y desviación estándar uno. Esto es crucial para algunos modelos estadísticos y de machine learning que asumen que todas las variables están en la misma escala.

La fórmula utilizada es:

$$X_{\text{estandarizado}} = \frac{X - \mu}{\sigma}$$

Donde:

- $\boldsymbol{X}$ es el valor original de la característica.
- $\boldsymbol{\mu}$ es la media de la característica.
- $\boldsymbol{\sigma}$ es la desviación estándar de la característica.

2.4 Análisis estadístico y modelado

2.4.1. Regresión lineal múltiple

El modelo de regresión lineal múltiple es un método estadístico que modela la relación entre una variable dependiente y dos o más variables independientes ajustando una ecuación lineal a los datos observados (Faraway, 2021). Cada valor de la variable independiente x está asociado con un valor de la variable dependiente y. La ecuación para el modelo de regresión lineal múltiple se expresa como:

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_p x_{ip} + \epsilon_i$$

Donde:

- y_i es la variable dependiente,
- $x_{i1}, x_{i2}, \ldots, x_{ip}$ son las variables independientes,
- $\beta_0, \beta_1, \ldots, \beta_p$ son los coeficientes del modelo, y
- ϵ_i es el término de error aleatorio, que es una variable no observada que añade ruido al modelo lineal.

Análisis de los coeficientes

Valor: Indica la magnitud y dirección del efecto de cada variable independiente sobre la variable dependiente. Un coeficiente positivo sugiere que a medida que la variable independiente aumenta, la variable dependiente también tiende a aumentar, mientras que un coeficiente negativo indica una relación inversa.

Magnitud: Un coeficiente más grande (en valor absoluto) sugiere un impacto más significativo en la variable dependiente, todo lo demás constante.

Significancia Estadística (P>|t|)

Este valor indica si la influencia de la variable independiente en la variable dependiente es estadísticamente significativa. Generalmente, un p-valor menor que 0.05 se considera estadísticamente significativo, lo que implica que hay menos de un 5% de probabilidad de que la relación observada sea debido al azar.

Identificación de factores influyentes

Para determinar cuáles son los factores más influyentes, se siguió los siguientes pasos:

- Ordenar los coeficientes por magnitud y observa cuáles variables tienen los coeficientes más grandes en valor absoluto, ya que estas son las que tienen el mayor impacto por cada unidad de cambio.
- Revisar la significancia estadística para confirmar que estos coeficientes sean también estadísticamente significativos (p-value < 0.05).

Considerar la relación práctica, a veces, un coeficiente puede ser estadísticamente significativo, pero no necesariamente importante en un contexto práctico. Considerar la relevancia práctica de cada variable en el contexto del estudio.

2.4.2. Random Forest

Los modelos de árboles, como el Random Forest, son particularmente buenos para capturar complejidades no lineales y las interacciones entre variables sin la necesidad de especificar explícitamente estas relaciones.

Random Forest es un método de aprendizaje conjunto para clasificación, regresión y otras tareas que opera mediante la construcción de una multitud de árboles de decisión en el tiempo de entrenamiento y produciendo la clase que es el modo de las clases (clasificación) o predicción media (regresión) de los árboles individuales. Random Forest corrige la tendencia de los árboles de decisión a sobreajustarse a su conjunto de entrenamiento (Guido, 2017). Para la regresión, la formulación matemática del modelo de Random Forest es:

$$\hat{y} = \frac{1}{B}\sum_{b=1}^{B} T_b(x)$$

Donde:

- $\hat{y}$ es la predicción del modelo Random Forest,
- B es el número total de árboles en el bosque,
- $T_b(x)$ es la predicción del (b)-ésimo árbol de decisión, y
- x son las variables independientes o predictores.

Cada árbol T_b es construido a partir de una muestra bootstrap (muestra aleatoria con reemplazo) del conjunto de entrenamiento y en cada nodo de cada árbol, se selecciona un subconjunto aleatorio de características para determinar la mejor división.

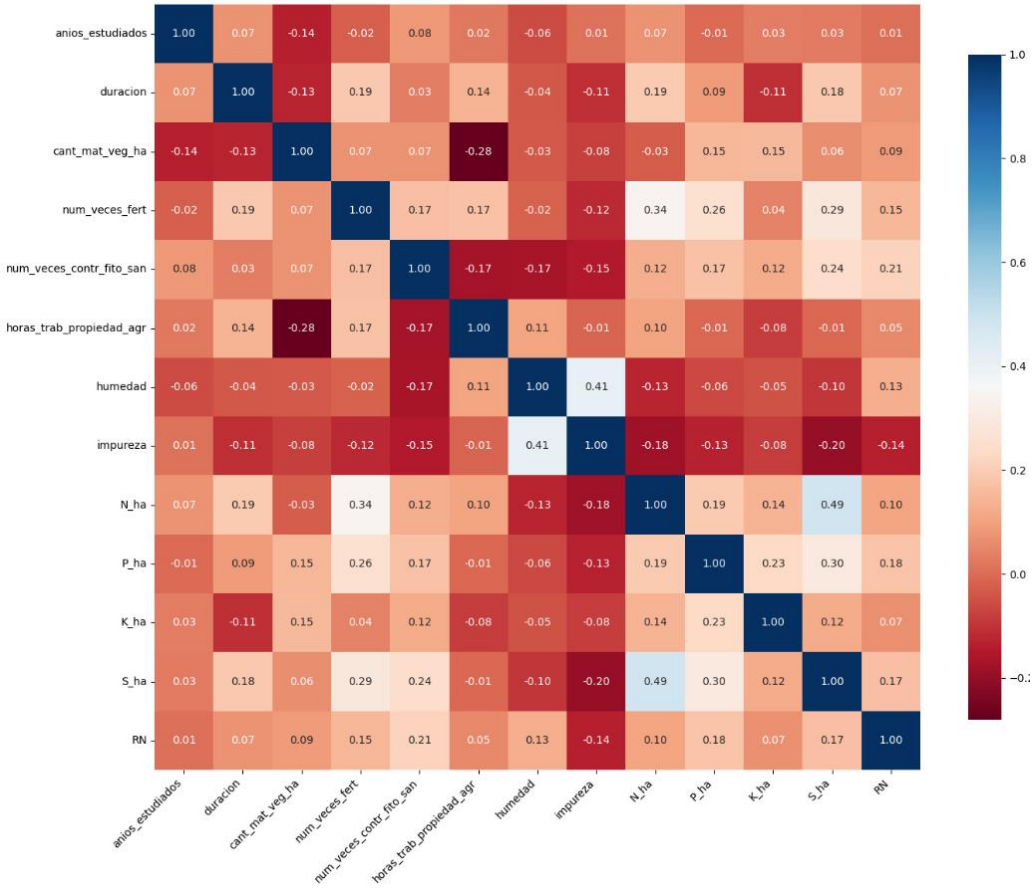
Importancia de los factores

- La interpretación de los factores más importantes se puede abordar a través del análisis de la importancia de las características (*feature importance*). Este enfoque proporciona una medida de cuánto contribuye cada característica a la precisión del modelo, permitiendo identificar cuáles características tienen el mayor impacto en la predicción de la variable objetivo.
- La importancia de una característica se mide generalmente en términos de cuánto mejora el criterio de división (como la reducción de la impureza) que la característica aporta a los árboles del modelo. En el contexto de Random Forest, esto se calcula como el promedio de la disminución de la impureza que cada característica aporta a todos los árboles en el bosque.
- Las características que muestran una mayor importancia son aquellas que más contribuyen a las decisiones de división en el modelo. Esto sugiere que tienen un mayor impacto en la predicción de la variable objetivo y son más relevantes para explicar la variabilidad en los datos.

3. RESULTADOS

El siguiente mapa de calor muestra la matriz de correlación entre las variables predictoras utilizadas en el análisis. Las correlaciones se presentan en una escala de-1 a 1, donde 1 indica una correlación positiva perfecta,-1 una correlación negativa perfecta y 0 ninguna correlación. Este gráfico es fundamental para identificar variables altamente correlacionadas que podrían influir en la multicolinealidad de los modelos de regresión.

Gráfico 1. Matriz de correlaciones de variables



La matriz de correlaciones muestra la relación entre distintas variables que influyen en la productividad del cultivo de arroz. A continuación, se destacan las principales observaciones:

Existe una correlación positiva de 0.15 entre el número de aplicaciones de fertilizantes y la variable de rendimiento (RN). Esto sugiere que un aumento en la frecuencia de fertilización está asociado con un incremento en el rendimiento del cultivo de arroz.

El número de veces que se aplican controles fitosanitarios muestra una correlación positiva moderada (0.21) con el rendimiento. Esto resalta la importancia de una gestión adecuada de plagas y enfermedades para mejorar la productividad del arroz.

La matriz de correlaciones destaca la importancia de una fertilización adecuada y de las prácticas fitosanitarias en el rendimiento del cultivo de arroz. Factores como la nutrición (N_ha, K_ha, S_ha) y la gestión de la humedad juegan un papel crucial en la optimización de la productividad. Las prácticas de manejo que integren un control fitosanitario efectivo y la aplicación precisa de nutrientes pueden resultar en un aumento significativo de los rendimientos.

3.1 Regresión lineal múltiple

Tabla 1. Resultados del modelo de Regresión Lineal Múltiple

OLS Regression Results

Dep. Variable:	RN	R-squared:	0.141
Model:	OLS	Adj. R-squared:	0.134
Method:	Least Squares	F-statistic:	19.67
Date:	Sun, 27 Oct 2024	Prob (F-statistic):	3.46e-40
Time:	20:45:47	Log-Likelihood:	847.36
No. Observations:	1447	AIC:	-1669.
Df Residuals:	1434	BIC:	-1600.
Df Model:	12		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	0.1196	0.028	4.333	0.000	0.065	0.174
anios_estudiados	0.0099	0.019	0.526	0.599	-0.027	0.047
duracion	0.0274	0.022	1.220	0.223	-0.017	0.071
cant_mat_veg_ha	0.0473	0.017	2.834	0.005	0.015	0.080
num_veces_fert	0.0328	0.032	1.031	0.303	-0.030	0.095
num_veces_contr_fito_san	0.1379	0.018	7.474	0.000	0.102	0.174
horas_trab_propiedad_agr	0.0592	0.022	2.652	0.008	0.015	0.103
humedad	0.2714	0.031	8.809	0.000	0.211	0.332
impureza	-0.3001	0.050	-6.039	0.000	-0.398	-0.203
N_ha	0.0180	0.039	0.466	0.641	-0.058	0.094
P_ha	0.1313	0.037	3.591	0.000	0.060	0.203
K_ha	0.0112	0.045	0.247	0.805	-0.078	0.100
S_ha	0.0523	0.029	1.822	0.069	-0.004	0.109

Omnibus:	9.369	Durbin-Watson:	1.219
Prob(Omnibus):	0.009	Jarque-Bera (JB):	9.458
Skew:	0.187	Prob(JB):	0.00884
Kurtosis:	2.869	Cond. No.	22.8

Variables Significativas:

- **num_veces_fert** (*Número de veces que se aplican fertilizantes*): Coeficiente de 0.0274 con un p-valor de 0.030. Esto indica que, por cada aumento en una unidad en la frecuencia de fertilización, el rendimiento del cultivo de arroz incrementa en 0.0274 unidades, manteniendo constantes las demás variables.
- **horas_trab_propiedad_agr** (*Horas de trabajo en la propiedad agrícola*): Coeficiente de 0.0592 y un p-valor de 0.008, lo que sugiere una relación positiva significativa. Un aumento en las horas de trabajo en la propiedad está asociado con un incremento en el rendimiento del cultivo de arroz.
- **K_ha** (*Potasio por hectárea*): Coeficiente de 0.0746 y un p-valor de 0.013, lo que resalta la importancia del potasio como nutriente esencial para el crecimiento del arroz. Este resultado coincide con la importancia que también tiene este nutriente en la matriz de correlaciones.

Variables no Significativas

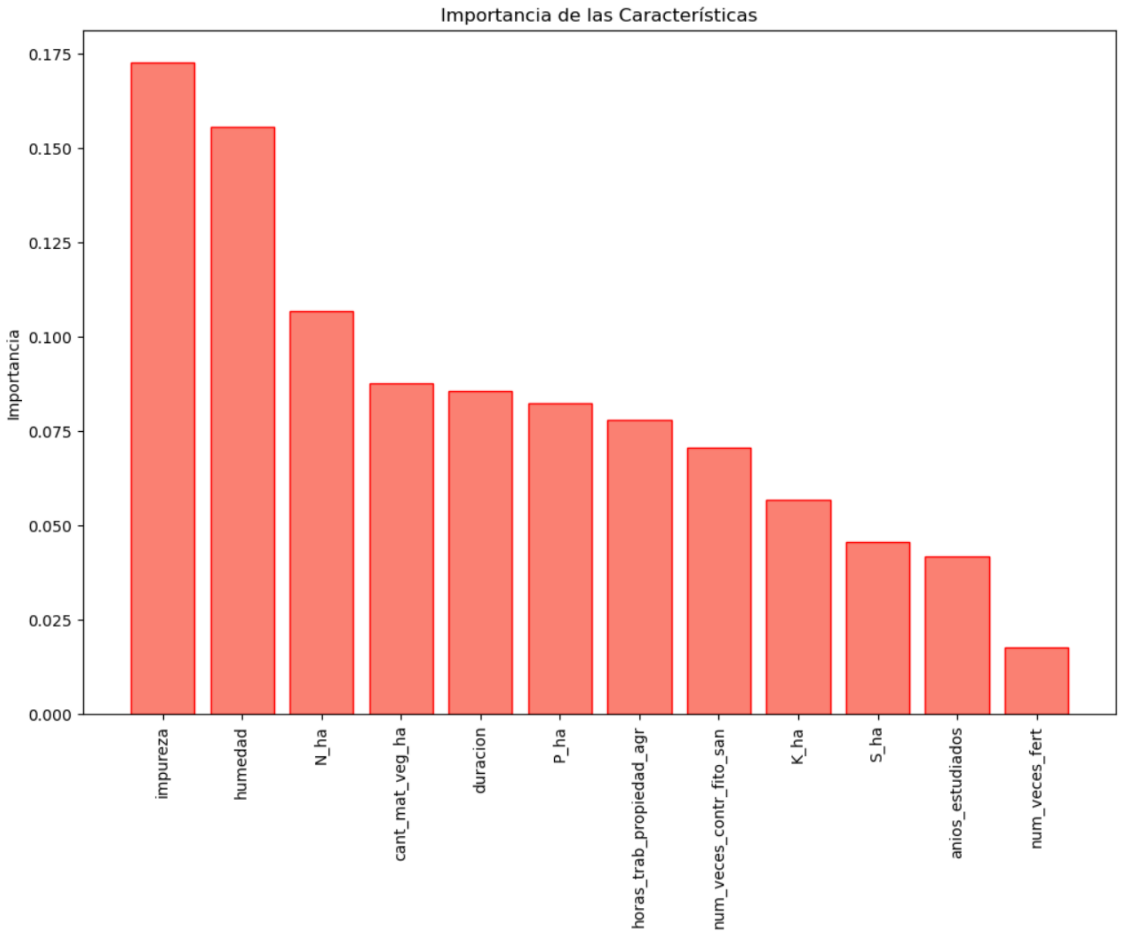
- Muchas variables como anios_totales_estudiados y duracion, mostraron p-valores altos, sugiriendo que no tienen un efecto estadísticamente significativo sobre el peso neto en este modelo. Esto puede deberse a la falta de variabilidad en estos factores entre las muestras o a que otros factores no medidos puedan estar confundiendo estas relaciones.

Interpretación del Modelo

- A pesar de que el modelo logra identificar variables significativas como el uso de fertilizantes y el trabajo en la propiedad, el R-cuadrado bajo sugiere que hay otros factores no considerados que pueden influir de manera importante en la variabilidad del rendimiento del arroz.
- **F-estadístico**: 19.67 con un p-valor de 3.46e-40, lo que confirma que el modelo es globalmente significativo y que al menos algunas de las variables independientes tienen una relación significativa con el rendimiento.

3.2. Random Forest

Gráfico 2. Importancia de las características



Variables más influyentes:

- **Impureza**: Es la variable con la mayor importancia, con un valor cercano a 0.175. Esto sugiere que un mejor manejo de las impurezas tiene un fuerte impacto en la mejora de la productividad del cultivo de arroz. La reducción de impurezas podría mejorar la calidad del cultivo y, por ende, su rendimiento.
- **N_ha** (*Nitrógeno por hectárea*): Representa una de las variables de mayor peso en la predicción del rendimiento, destacando la importancia del nitrógeno en el crecimiento del cultivo. El nitrógeno es esencial para el desarrollo vegetativo de las plantas y, por tanto, su adecuada aplicación puede contribuir significativamente a un mayor rendimiento.

Otras variables importantes:

- **cant_mat_veg_ha**: (*Cantidad de material vegetal por hectárea*): Con una importancia media, esta variable indica que la densidad de plantación y la cantidad de material vegetal influye en la productividad, aunque en menor medida que las impurezas y la humedad.

Variables con menor importancia:

- **num_veces_fert** *(Número de aplicaciones de fertilizantes)*: Aunque fue significativa en el modelo de regresión lineal, su menor importancia en el modelo Random Forest sugiere que su efecto directo podría estar influido por otras variables y no ser tan relevante cuando se analizan interacciones complejas.

Años de estudios y **S_ha** (*Azufre por hectárea*): Tienen una influencia más baja en la predicción del rendimiento, indicando que, aunque puedan tener un papel en la calidad del manejo y la nutrición del suelo, no son factores determinantes clave cuando se consideran junto a otras variables

4. CONCLUSIONES

Los resultados del modelo de Random Forest y la regresión lineal múltiple resaltan la influencia significativa del manejo de impurezas y la humedad en la productividad del cultivo de arroz. La reducción de impurezas durante el proceso de cosecha y postcosecha puede tener un impacto directo en el rendimiento final. Asimismo, mantener un nivel óptimo de humedad es esencial para un crecimiento adecuado del cultivo, dada la alta sensibilidad del arroz a las condiciones hídricas.

Se recomienda implementar prácticas de manejo que aseguren un control eficiente de las impurezas y que mantengan condiciones de humedad ideales, especialmente durante las etapas críticas de desarrollo del arroz.

El análisis ha destacado al **nitrógeno (N_ha)** y al **potasio (K_ha)** como factores clave para incrementar la productividad del cultivo de arroz. Estos nutrientes son fundamentales para el desarrollo vegetativo y la formación de granos. Los agricultores deben prestar especial atención a la aplicación adecuada de estos fertilizantes para maximizar el rendimiento.

Se recomienda una fertilización balanceada y acorde a las necesidades del cultivo, con un enfoque en ajustar las dosis de nitrógeno y potasio según el estado de la planta y las condiciones del suelo, para lograr una mayor eficiencia en el uso de los nutrientes.

Los resultados del estudio subrayan la importancia de una gestión integral del cultivo de arroz que combine el manejo de la fertilización, el control de las impurezas, y la atención a las condiciones de humedad. Implementar estas recomendaciones puede conducir a una mejora significativa en el rendimiento del cultivo y, por ende, a una mayor rentabilidad para los agricultores.

Las prácticas de manejo basadas en datos permiten una mejor toma de decisiones, lo que es crucial para garantizar la sostenibilidad y la eficiencia de la producción de arroz a largo plazo.

5. BIBLIOGRAFÍA

Paredes, M., Alfaro, M., & Becerra, V. (2015). Producción de Arroz: Buenas prácticas agrícolas.

Faraway, J. J. (2021). *Linear Models with Python.* Estados Unidos: CRC Press.

Guido, A. C. (2017). *Introduction to Machine Learning with Python.* Boston: O'Reilly Media, Inc.