INFORME DEL ANÁLISIS DE LOS DETERMINANTES DE LA PRODUCTIVIDAD DE PAPA 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 papa 2023

Proyecto Integral de Diversificación Agroproductiva y Reconversión Agrícola-PIDARA

Quito – Ecuador

2024

RESUMEN

Este informe analiza los rendimientos objetivos del cultivo de papa (Solanum tuberosum) en Ecuador, abordando factores agronómicos y socioeconómicos. El estudio se basa en datos recolectados en el año 2023 en siete provincias, utilizando metodologías como la regresión lineal múltiple y el modelo de Random Forest para identificar los principales determinantes de la productividad. Los resultados destacan la importancia de las prácticas de manejo fitosanitario y la aplicación adecuada de fertilizantes, especialmente el potasio, como factores críticos para incrementar el rendimiento. Se recomienda una mayor investigación en combinaciones óptimas de prácticas agrícolas para mejorar la productividad y sostenibilidad del cultivo.

Palabras clave: Papa, rendimientos objetivos, regresión lineal múltiple, Random Forest

ABSTRACT

This report analyzes the target yields of potato (Solanum tuberosum) cultivation in Ecuador, addressing agronomic and socioeconomic factors. The study is based on data collected in 2023 in seven provinces, using methodologies such as multiple linear regression and the Random Forest model to identify the main determinants of productivity. The results highlight the importance of phytosanitary management practices and proper fertilization, especially potassium, as critical factors for increasing yield. Further research is recommended on optimal combinations of agricultural practices to improve crop productivity and sustainability.

Keywords: Potato, crop yields, linear regression, Random Forest

Contenido

1. ANTECEDENTES 5
1.1. Operativos de rendimientos objetivos 5
1.2. Justificación 5
2. METODOLOGÍA 6
2.1. Selección de características 6
2.2. Tratamiento de datos atípicos y faltantes 6
2.3. Transformación de variables 6
2.4. Análisis Estadístico y Modelado 7
3. RESULTADOS 9
3.1. Regresión Lineal Múltiple 10
3.2. Random Forest 12
4. CONCLUSIONES 13
5. BIBLIOGRAFÍA 14
ANEXOS: ..... ¡Error! Marcador no definido.

1. ANTECEDENTES

1.1. Operativos de rendimientos objetivos

La papa (*Solanum tuberosum)* comprende un cultivo de gran importancia económica para el Ecuador; en la actualidad, se cultiva en casi todos los países y se considera un alimento básico de consumo mundial. En el Ecuador, aproximadamente el 81 % de la producción se comercializa para consumo en fresco y el resto es utilizado por la industria de procesamiento.

Teniendo en cuenta esta relevancia, el Ministerio de Agricultura y Ganadería viene realizando, desde el año 2015, los operativos de rendimientos objetivos ORO, actualmente en 9 provincias del Ecuador: al norte Carchi e Imbabura, al centro Pichincha, Cotopaxi, Bolívar, Tungurahua y Chimborazo; y, al sur Cañar y Azuay, donde se siembra entre los 2800 hasta los 3500 metros sobre el nivel del mar.

La importancia de contar con información radica en su diversidad, pues alrededor de 30 variedades mejoradas son cultivadas en las 3 regiones identificadas; además, su composición nutricional destaca por el contenido de macro y micronutrientes. Una papa de tamaño grande (200 g), con cáscara, aporta el 26 % del requerimiento diario de cobre, de 17 a 18 % de potasio, fósforo y hierro; de 5 a 13 % de Calcio, Zinc, Magnesio y Manganeso; y hasta el 50 % del requerimiento diario de vitamina C. (Cuesta, 2022)

Con la finalidad de evitar un sesgo de información, por motivos de percepción de la persona productora al momento del levantamiento en campo, del rendimiento, fue necesario incorporar a este levantamiento de información la captura de muestras del cultivo mediante un protocolo definido; posteriormente, estas muestras pasan por una fase de laboratorio, donde se obtiene otro tipo de información (peso, humedad e impurezas). Información con la cual, se logra realizar un cálculo del rendimiento; el mismo que se ha denominado como “rendimiento objetivo”.

1.2. Justificación

Estudiar los determinantes de la productividad en el cultivo de papa es desarrollar un entendimiento profundo, 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 cuáles son los factores que tienen el mayor impacto en la productividad, los agricultores 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 generar 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 papa, un componente crucial en la dieta de muchas poblaciones, se contribuye a la seguridad alimentaria a nivel local y nacional. Esto es especialmente importante en regiones donde la papa es un alimento básico y una fuente significativa de nutrición.
- 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. Por lo regular, 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:

- Ordenar los coeficientes por magnitud y observar 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

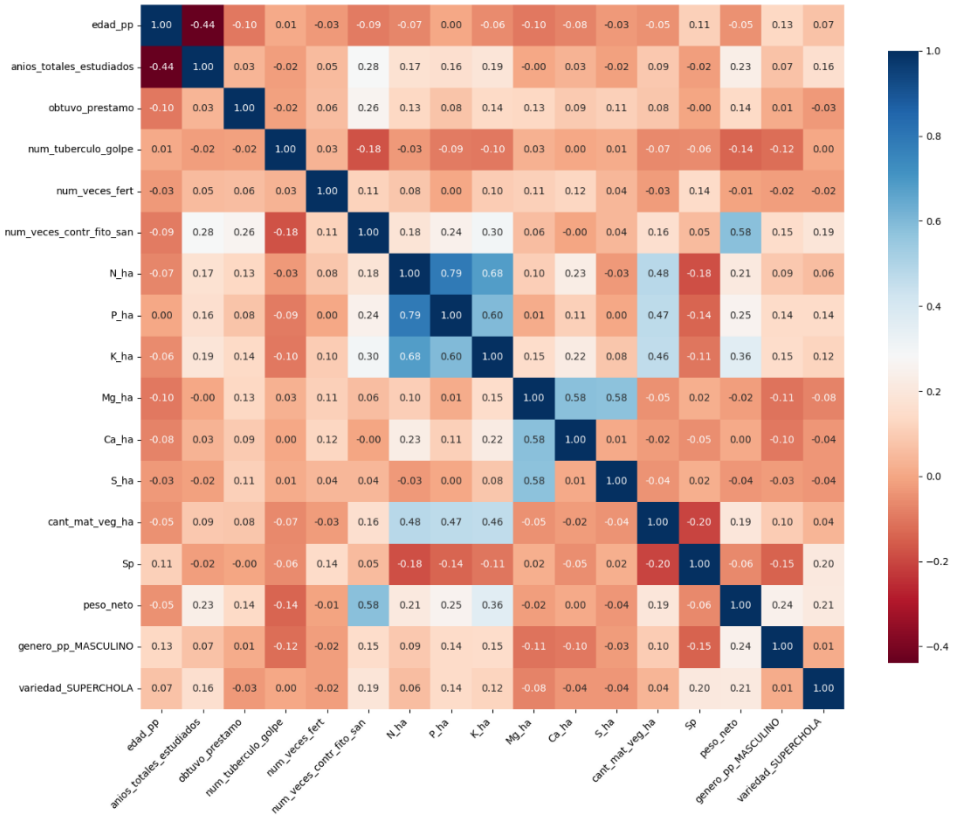
- La interpretación de los factores más importantes se puede abordar a través del análisis de la importancia de las características (*feature importance*). Este enfoque proporciona una medida de cuánto contribuye cada característica a la precisión del modelo, permitiendo identificar cuáles características 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



Observamos una correlación positiva importante entre el número de controles fitosanitarios y el rendimiento de las plantas muestreadas, sugiriendo que un aumento de este tipo de controles podría estar asociado con un incremento en la productividad.

3.1. Regresión Lineal Múltiple

Tabla 1. Resultados del modelo de Regresión Lineal Múltiple

Dep. Variable:	peso_neto	R-squared:	0.419
Model:	OLS	Adj. R-squared:	0.402
Method:	Least Squares	F-statistic:	24.35
Date:	Tue, 30 Apr 2024	Prob (F-statistic):	1.09e-53
Time:	23:57:24	Log-Likelihood:	-639.06
No. Observations:	557	AIC:	1312.
Df Residuals:	540	BIC:	1386.
Df Model:	16		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	1.067e-16	0.033	3.25e-15	1.000	-0.064	0.064
edad_pp	0.0002	0.038	0.006	0.995	-0.075	0.075
anios_totales_estudiados	0.0391	0.039	0.999	0.318	-0.038	0.116
obtuvo_prestamo	-0.0025	0.035	-0.073	0.942	-0.071	0.066
num_tuberculo_golpe	-0.0175	0.034	-0.516	0.606	-0.084	0.049
num_veces_fert	-0.0652	0.034	-1.914	0.056	-0.132	0.002
num_veces_contr_fito_san	0.4853	0.038	12.640	0.000	0.410	0.561
N_ha	-0.0732	0.062	-1.172	0.242	-0.196	0.049
P_ha	0.0083	0.057	0.146	0.884	-0.103	0.120
K_ha	0.2292	0.049	4.670	0.000	0.133	0.326
Mg_ha	0.0007	0.058	0.012	0.990	-0.113	0.114
Ca_ha	-0.0106	0.048	-0.219	0.826	-0.105	0.084
S_ha	-0.0665	0.047	-1.413	0.158	-0.159	0.026
cant_mat_veg_ha	-0.0005	0.040	-0.013	0.990	-0.079	0.078
Sp	-0.0697	0.036	-1.941	0.053	-0.140	0.001
genero_pp_MASCULINO	0.1204	0.035	3.441	0.001	0.052	0.189
variedad_SUPERCHOLA	0.0977	0.035	2.767	0.006	0.028	0.167

Omnibus:	13.108	Durbin-Watson:	1.394
Prob(Omnibus):	0.001	Jarque-Bera (JB):	13.954
Skew:	0.323	Prob(JB):	0.000933
Kurtosis:	3.428	Cond. No.	4.47

Variables Significativas:

- **num_veces_contr_fito_san:** Con un coeficiente de 0.4853 y un p-valor significativo ($p < 0.001$), esta variable indica un aumento fuerte en el peso neto por cada unidad adicional en el número de veces que se aplican controles fitosanitarios, sugiriendo que las prácticas de manejo de la salud de las plantas son cruciales para la productividad.
- **K_ha (potasio por hectárea):** Tiene un coeficiente de 0.2292 con un p-valor < 0.001, mostrando un impacto positivo significativo en el peso neto. Esto puede indicar la importancia de una nutrición adecuada, específicamente del potasio, en el rendimiento de los cultivos.
- **genero_pp_MASCULINO:** Un coeficiente de 0.1204 y p-valor de 0.001, sugiere que los datos de hombres asociados con la agricultura tienen una productividad mayor en el cultivo de papa comparado con las mujeres, lo que podría reflejar diferencias en las prácticas agrícolas o en las parcelas gestionadas entre géneros.
- **variedad_SUPERCHOLA:** Con un coeficiente de 0.0977 y un p-valor de 0.006, indica que esta variedad específica tiene un mejor rendimiento en términos de peso neto, comparada con otras variedades.

Variables no Significativas

- Muchas variables como anios_totales_estudiados, obtuvo_prestamo, num_tuberculo_golpe, N_ha, P_ha, Mg_ha, Ca_ha, S_ha y cant_mat_veg_ha, mostraron p-valores altos, sugiriendo que no tienen un efecto estadísticamente significativo sobre el peso neto en este modelo. Lo que 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

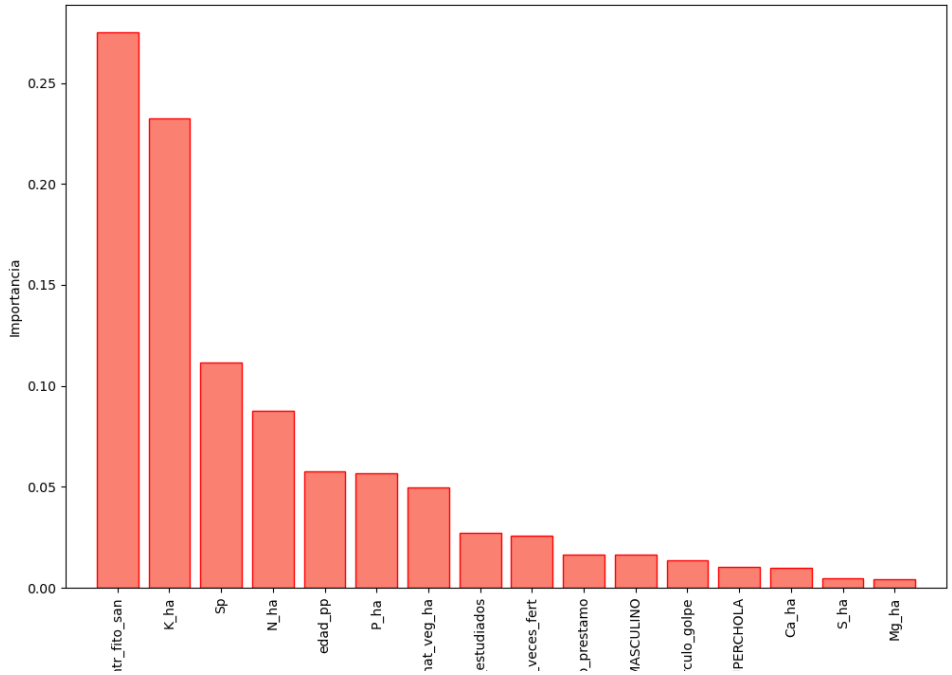
- R-cuadrado (0.419): El valor indica que, el modelo explica aproximadamente el 41.9 % de la variabilidad en el peso neto de los tubérculos. Aunque es una cantidad considerable, también sugiere que hay otros factores no incluidos en el modelo, que están afectando la productividad.
- F-estadístico (24.35): El valor de la prueba F y su probabilidad asociada (1.09e-53) denotan que el modelo es globalmente significativo, lo que significa que, al menos, algunas de las variables independientes están relacionadas significativamente con la variable dependiente.

Diagnósticos del Modelo

- Durbin-Watson (1.394): Este valor está cerca de 1, lo que podría indicar una posible autocorrelación positiva en los residuos del modelo. Esto es importante para validar en estudios futuros, ya que puede afectar la confiabilidad de los coeficientes estimados.
- Condiciones de No. (4.47): Un número de condición menor a 10 sugiere que no hay problemas significativos de multicolinealidad en el modelo.

3.2. Random Forest

Gráfico 2. Importancia de las características



Variables más influyentes:

- **num_veces_contr_fito_san**: Esta es la característica más importante, según el modelo Random Forest, con la mayor puntuación de importancia. Esto sugiere que la frecuencia de las medidas fitosanitarias tiene un impacto significativo en la productividad de los tubérculos. La gestión efectiva de plagas y enfermedades parece ser crucial para optimizar el rendimiento del cultivo.
- **K_ha**: El potasio por hectárea también muestra una alta importancia. Como nutriente esencial, el potasio influye significativamente en la salud de las plantas y su capacidad para producir una cosecha abundante. Esto es coherente con los resultados de la regresión lineal, que también mostraban una relación positiva significativa con el peso neto.

Otras variables importantes:

- **Sp y N_ha**: Estas variables tienen importancias moderadas y están relacionadas con aspectos de manejo agronómico y nutricional del suelo. Sp se refiere a la superficie por planta, que es un parámetro crucial en la gestión del espacio y densidad de plantación, afectando directamente el acceso de cada planta a recursos como luz, agua y nutrientes. N_ha alude al nitrógeno aplicado por hectárea, otro nutriente fundamental para el crecimiento vegetal. La adecuada gestión de estos factores puede optimizar el rendimiento del cultivo, al asegurar que cada planta tenga los recursos necesarios para desarrollarse plenamente.

Variables con menor importancia:

- Las características como edad_pp, P_ha (fósforo por hectárea), y aspectos socioeconómicos como anios_totales_estudiados y obtuvo_prestamo tienen una importancia relativamente baja en este modelo. Esto indica que su influencia directa en la productividad del cultivo es menor, comparada con las medidas fitosanitarias y de fertilización.
- Sorprendentemente, variedad_SUPERCHOLA y genero_pp_MASCULINO, a pesar de ser significativas en la regresión lineal, aparecen con baja importancia en este modelo. Lo que puede sugerir que, su efecto es más específico o que otras variables en el modelo Random Forest están capturando parte de su influencia.

4. CONCLUSIONES

Al examinar y comparar los resultados de ambos modelos de análisis, el de regresión lineal múltiple y el Random Forest, podemos consolidar varias conclusiones clave sobre los determinantes de la productividad, específicamente en el contexto del rendimiento de las plantas:

- Ambos modelos subrayan la importancia significativa de las prácticas de manejo fitosanitario (num_veces_contr_fito_san) y el uso adecuado de fertilizantes, en particular el potasio (K_ha), como factores críticos que influyen positivamente en la productividad del cultivo. Estos resultados sugieren que, mejorar la eficacia y la frecuencia de estas prácticas podría conducir a un aumento significativo en el rendimiento de los cultivos.
- Mientras que, el modelo de Random Forest identificó otras variables como el nitrógeno (N_ha) y ciertos parámetros del suelo (Sp) como relativamente importantes, la regresión lineal proporcionó una perspectiva diferente, destacando variables socioeconómicas y de variedad de cultivo como significativas. Esta variabilidad en la importancia atribuida a diferentes variables entre los modelos, sugiere que las interacciones complejas entre múltiples factores influyen en la productividad.
- Se recomienda una inversión en investigación agronómica que explore más a fondo, cómo las combinaciones óptimas de prácticas fitosanitarias y regímenes de fertilización pueden maximizar los rendimientos. Además, sería beneficioso investigar las razones detrás de las diferencias en la importancia de las variables socioeconómicas y de manejo entre diferentes modelos para entender mejor sus efectos indirectos o moderadores en la productividad.
- La integración de los hallazgos de ambos modelos podría ofrecer un enfoque más robusto para la toma de decisiones en la agricultura. Utilizar un enfoque híbrido o conjunto, que combine las fortalezas del análisis lineal y los modelos basados en árboles, podría proporcionar predicciones más precisas y estrategias de gestión más efectivas.

La consolidación de estos modelos destaca la importancia crítica de las prácticas de manejo fitosanitario y nutricional, junto con la necesidad de considerar factores socioeconómicos y agronómicos en la planificación y optimización de la producción agrícola. Al prestar atención a estas prácticas y explorar más a fondo las interacciones complejas entre las variables, los productores pueden lograr mejoras significativas en la productividad y sostenibilidad de los cultivos.

Este análisis destaca la importancia de ciertas prácticas agrícolas y características del cultivo en la productividad de los tubérculos. Además, identifica áreas donde la investigación adicional podría ser necesaria para entender mejor otros factores que afectan la productividad y cómo mejorar las estrategias de gestión agrícola.

Considerando las limitaciones del modelo, como la explicación parcial de la variabilidad y la posible autocorrelación, sería prudente realizar análisis adicionales y considerar la inclusión de más variables o datos de seguimiento para futuros modelos.

El análisis del modelo Random Forest ofrece una visión complementaria a la obtenida por la regresión lineal, destacando la importancia de ciertas prácticas agrícolas sobre otras variables más generales o socioeconómicas. Esto debería guiar futuras decisiones de manejo y enfoques de investigación en la agricultura, para optimizar el rendimiento de los cultivos.

5. BIBLIOGRAFÍA

Cuesta, X. a. (2022). *Catálogo de variedades de papa.* Quito: INIAP.

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.