Pontificia Universidad Católica del Ecuador

Facultad de Hábitat, Infraestructura y Creatividad

Maestría en Sistemas de Información mención Data Science



Trabajo previo para la obtención del título de Magister en Sistemas de Información mención Data Science

Tema:

Desarrollo de un modelo predictivo para estimar la producción de leche en la finca El Corazón, orientado a mejorar la calidad y cantidad mediante la optimización de recursos.

Autor:

Wilson Santiago García Sánchez

Tutor:

Carlos Julio Mayorga Arias Dr.

Quito - marzo 2026

Agradecimientos

A mi esposa Deysi, por ser mi mayor fuente de fortaleza y motivación. Su apoyo incondicional, su comprensión en los momentos de mayor dedicación a este proyecto y su confianza constante en mis capacidades hicieron posible que este trabajo llegará a su culminación. Su amor y compañía han sido el pilar fundamental que me impulsó a avanzar incluso en los días más exigentes.

Este logro no solo me pertenece, lo comparto con toda mi familia por la unidad y fortaleza que he recibido de su parte en todo momento.

Índice

Capítulo 1 8
1 Introducción 8
1.1 Objetivos 9
1.2 Justificación 10
1.3 Preguntas de Investigación 11
Capítulo 2 12
2 Marco Teórico 12
2.1 Modelos Factoriales 12
2.2 Modelización Matemática y Estadística 14
2.3 Producción lechera y factores determinantes 16
2.4 Metodología 18
2.5 Resultados y Análisis 30
Capítulo 3 48
3 Conclusiones y Recomendaciones 48
3.1 Conclusiones 48
3.2 Limitaciones 49
3.3 Recomendaciones 49
Referencias 51

Índice de Tablas

Tabla 2. 1. Variables operacionales. 22

Tabla 2. 2. Variables categóricas que se analizan para el modelo predictivo. 25

Tabla 2. 3 Indicadores de diferencias significativas (ANOVA → Producción L/día). 32

Tabla 2. 4 Análisis de Tukey HSD. 33

Tabla 2. 5. Análisis de Tukey por estado gestante. 34

Índice de Gráficos

Gráfico 2. 1 Construcción del modelo predictivo. 28

Gráfico 2. 2. Estandarización de la variable. 28

Gráfico 2. 3. Código del modelo predictivo de producción de leche. 28

Gráfico 2. 4. Código del intercepto explicito. 29

Gráfico 2. 5. Código para el ajuste de mínimos cuadrados ordinarios. 29

Gráfico 2. 6. Código para separa el intercepto y betas estandarizadas. 29

Gráfico 2. 7. Mapa de calor de la producción lechera Finca el Corazón. 31

Gráfico 2. 8. Análisis de la producción lechera por raza mediante boxplot. 34

Gráfico 2. 9. Análisis de producción lechera por estado gestante mediante boxplot. ... 35

Gráfico 2. 10. Análisis PCA de las variables productivas y ambientales. 36

Gráfico 2. 11. Código empleado para la limpieza y normalización de datos Vacas_2025. 37

Gráfico 2. 12. Código para el análisis de correlaciones. 38

Gráfico 2. 13. Resultado del análisis de correlación sobre el dataset Vacas_2025. 38

Gráfico 2. 14. Produccion lechera vs Raza dataset Vacas_2025. 39

Gráfico 2. 15. Produccion lechera vs tipo de pasto dataset Vacas_2025. 40

Gráfico 2. 16. Produccion lechera vs días de Gestación dataset Vacas_2025. 41

Gráfico 2. 17. Construcción de características y conjunto de datos para el modelado. 42

Gráfico 2. 18. Análisis de correlación de Pearson para definir las variables predictoras. 42

Gráfico 2. 19. Construcción del modelo predictivo. 44

Gráfico 2. 20. Resultados del modelo predictivo sobre el dataset Vacas_2025. 44

Gráfico 2. 21. Análisis de impacto de las variables y la producción lechera del dataset Vacas_2025. 46

Gráfico 2. 22. Resultado de la predicción para conocer un estimado de producción de leche..... 47

Índice de Formulas

Fórmula 2. 1. Formula de modelo predictivo. 23

Formula 2. 2. Coeficientes de regresión. 24

Formula 2. 3. Bondad de ajuste.....24

Fórmula 2. 4. Error cuadrático médio. 24

Formula 2. 5. Error absoluto medio 24

Formula 2. 6. Ecuación para la construcción del modelo predictivo.....27

Capítulo 1

1 Introducción

La producción lechera constituye una actividad agropecuaria fundamental para el desarrollo económico y social en diversas regiones del país, su aporte más importante es sobre la seguridad alimentaria, la generación de empleo y el fortalecimiento de las cadenas productivas rurales (CIL del Ecuador, 2025). Sin embargo, factores como políticas gubernamentales, las características nutricionales del ganado, el manejo del hato y la disponibilidad de recursos tecnológicos influyen directamente en la calidad y cantidad de leche producida, dificultando la gestión eficiente de pequeñas fincas ganaderas. En este contexto, el uso de herramientas tecnológicas y metodologías basadas en análisis de datos se ha convertido en una estrategia clave para una toma de decisiones informada buscando optimizar los procesos productivos (Nicolás, 2025).

La finca El Corazón, ubicada en el cantón Sigchos de la provincia de Cotopaxi, está ubicada a 2600 msnm, dedicada a la producción bovina de leche, enfrenta el desafío de administrar de manera eficaz los recursos disponibles y anticipar los niveles de producción, con el fin de mejorar la calidad y cantidad de leche, y con ello aumentar la rentabilidad. Actualmente, la ausencia de herramientas tecnológicas y de predicción limita la capacidad para planificar actividades de manejo, alimentación y sanidad del hato, lo que puede generar pérdidas económicas y subutilización de los recursos. Ante esta situación, se vuelve necesario el diseño de un modelo predictivo que permita estimar la producción de leche de forma confiable y oportuna.

El desarrollo de un modelo basado en técnicas estadísticas y de aprendizaje automático ofrece la posibilidad de identificar patrones, variables determinantes y relaciones clave dentro del

sistema productivo. Estos modelos no solo permiten anticipar la producción esperada, sino también proponer acciones orientadas a mejorar la eficiencia, reducir costos y elevar los indicadores de calidad. Así mismo, la incorporación de tecnologías de análisis predictivo contribuye al fortalecimiento de prácticas ganaderas sostenibles y competitivas en el mercado (Hastie, The Elements of Statistical Learning. Springer, 2009).

El presente trabajo de investigación tiene como propósito desarrollar un modelo predictivo para estimar la producción de leche en la finca El Corazón, con el objetivo de optimizar el uso de recursos y mejorar tanto la calidad como la cantidad del producto final. Para ello, se recopilará y analizará información relevante del hato bovino y del entorno productivo, se evaluarán distintas técnicas predictivas y se seleccionará el modelo con mayor precisión y aplicabilidad práctica. El resultado esperado es una herramienta que pueda ser implementada en la gestión diaria de la finca, aportando a la toma de decisiones informadas y a la mejora continua del sistema productivo.

1.1 Objetivos

1.1.1 Objetivo general:

Desarrollar y evaluar un modelo predictivo basado en datos históricos y factores climáticos que permita mejorar la cantidad y calidad de la producción de leche optimizando el uso de recursos e insumos.

1.1.2 Objetivos específicos:

- Recolectar, depurar y procesar datos históricos provenientes de registros de operación de la Finca el Corazón, asegurando su calidad e integridad.
- Recolectar, depurar y procesar factores climáticos como humedad, precipitación y temperatura.
- Identificar y seleccionar variables relevantes, como raza, tamaño, hábitos alimenticios, edad, factores climáticos que influyan en la producción de leche.

- Construir un modelo predictivo utilizando técnicas de aprendizaje automático y estadística, adaptadas al tipo de datos y naturaleza del problema.

1.2 Justificación

En Ecuador, los pequeños productores representan el 80% del total de productores de leche, generando alrededor del 46% de la producción nacional. A pesar de su relevancia, este sector es el que menos ha incorporado tecnologías modernas como la ciencia de datos en la operación de sus fincas (Liz, Ortiz, 2024).

La aplicación de Data Science en ganadería se encuentra en fase emergente. Proyectos como el Observatorio Lácteo del Ecuador han implementado minería de datos, modelos predictivos y biomodelos sostenibles para mejorar la gestión en fincas pequeñas. Estas iniciativas evidencian que la digitalización y el análisis de datos son herramientas clave para enfrentar desafíos como baja eficiencia, variabilidad climática y limitación de recursos (Yazmín, Bustán, 2024).

La introducción de herramientas analíticas en fincas pequeñas constituye un gran desafío, principalmente por la falta de datos históricos que permitan realizar análisis predictivos. Sin embargo, cuanto antes se adopten estas tecnologías en sectores poco explorados, más rápido se obtendrán resultados beneficiosos para este importante segmento de la producción lechera.

El análisis de datos constituye la fase fundamental en el desarrollo de un modelo predictivo orientado a estimar la producción de leche en la finca El Corazón. Este proceso implica examinar de manera sistemática los datos disponibles, identificar patrones y relaciones entre variables, y generar información confiable que sirva como base para la toma de decisiones estratégicas. El objetivo principal es comprender los factores que inciden en la calidad y cantidad de leche producida, optimizando el uso de recursos para mejorar la eficiencia del sistema productivo (Juhn & Kjell, 2013).

1.3 Preguntas de Investigación

¿Qué variables productivas, ambientales y de manejo influyen de manera significativa en la producción de leche en la finca El Corazón?

¿Cómo se comporta la producción de leche en función de dichas variables y qué patrones pueden identificarse mediante técnicas estadísticas y de análisis de datos?

¿Qué tipo de modelo predictivo ofrece el mejor desempeño para estimar la producción de leche en este contexto específico?

¿Cuál es el nivel de precisión del modelo predictivo desarrollado y cómo se compara con los métodos tradicionales de estimación utilizados en la finca?

¿De qué manera la implementación del modelo predictivo puede contribuir a la optimización de recursos y a la mejora de la calidad y cantidad de la producción lechera?

Capítulo 2

2 Marco Teórico

2.1 Modelos Factoriales

Un modelo factorial estudia el efecto de dos o más factores (variables independientes categóricas) sobre una variable respuesta (variable dependiente). Estos factores pueden tener distintos niveles y sus combinaciones generan tratamientos experimentales (Spanish Psychological Databases, 2026).

Elementos principales

- Factores: Variables que influyen en la respuesta (ej.: tipo de forraje, suplemento, raza, estado fisiológico).
- Niveles: Categorías o intensidades de cada factor (ej.: tres tipos de forraje; dos razas).
- Variable respuesta: Resultado que se desea analizar o predecir (ej.: litros de leche/día).

En el contexto agropecuario, especialmente en sistemas de producción lechera, los modelos factoriales permiten comprender cómo variables como alimentación, genética, manejo, clima, salud del hato y condiciones de infraestructura impactan la producción y calidad de la leche (Lacteos Franz, 2024).

Este tipo de modelamiento es clave para desarrollar sistemas predictivos, ya que permite cuantificar la contribución relativa de cada factor y evaluar posibles interacciones que afecten la eficiencia productiva.

2.1.1 Tipos de Modelos Factoriales

El Diseño factorial completo Evalúa todas las combinaciones posibles entre los niveles de cada factor. Es ideal cuando se requiere un análisis detallado y se busca evaluar interacciones complejas (Minitab Support, 2026).

Ejemplo aplicado a producción lechera:

- Factor A: Suplementación (0 kg, 1 kg, 2 kg).
- Factor B: Tipo de pastura (natural, mejorada).
- Combinaciones totales = 3 × 2 = 6 tratamientos.

2.1.2 Diseño factorial fraccionado

Utilizado cuando el número de combinaciones es muy alto o costoso. Permite analizar efectos principales con menos recursos, aunque sacrifica algunas interacciones. Es ideal para fincas agropecuarias pequeñas y medianas que buscan eficiencia en la experimentación (Gutierrez Pulido & Vara Salazar, 2008).

2.1.3 Modelos factoriales mixtos

Son modelos que Incluyen:

- Factores fijos. Ejemplo: manejo, suplementación, etc.
- Factores aleatorios. Ejemplo: vacas individuales, variabilidad genética, etc.

Son comunes en estudios agropecuarios debido a la variabilidad biológica (Tognetti & Perez, 2019).

2.1.4 Interacciones en los Modelos Factoriales

Una interacción ocurre cuando el efecto de un factor depende del valor de otro. En sistemas biológicos como la producción de leche, esto es común.

Ejemplos:

- La producción de leche depende de la combinación de raza × alimentación.

- El efecto del clima puede depender del estado fisiológico de la vaca.
- El suplemento puede ser más efectivo en vacas de alta producción (Gutiérrez Pulido, 2008).

Comprender estas interacciones permite desarrollar modelos predictivos más precisos.

2.1.5 Modelos Factoriales como Base de Modelos Predictivos

De los diseños factoriales se derivan modelos estadísticos, generalmente modelos lineales, que permiten:

- Estimar efectos.
- Evaluar significancia.
- Generar ecuaciones predictivas.

Estas ecuaciones pueden integrarse en:

- Regresión múltiple.
- Modelos mixtos.
- Métodos de machine learning (árboles, random forest, SVM).
- Modelos de optimización de recursos (Montgomery, 2004).

Así, el análisis factorial se convierte en el fundamento para construir un modelo predictivo que estime la producción de leche con base en distintas combinaciones de factores.

2.2 Modelización Matemática y Estadística

La modelización matemática y estadística es el proceso de representar fenómenos reales mediante ecuaciones, funciones o estructuras matemáticas que nos permitan entender, predecir y tomar decisiones (Staff, 2023).

Se usa en prácticamente cualquier industria: marketing, finanzas, ingeniería, economía, logística, ciencia de datos, biomedicina y más (UNIR MEXICO, 2025).

2.2.1 Modelización Matemática

Consiste en traducir un problema real a un lenguaje matemático. Implica:

- Identificar el fenómeno.
- Seleccionar variables relevantes.
- Formular relaciones entre las variables.
- Construir ecuaciones o sistemas de ecuaciones.
- Resolverlas y analizar resultados.

A continuación, se presentan los pasos que se deben seguir para crear modelos matemáticos.

- Definir el problema real.
- Identificar las variables reales.
- Establecer las relaciones entre variables.
- Construir el modelo.
- Calibrar el modelo con datos.
- Validar el modelo (Cervantes Gómez, 2015).

2.2.2 Modelización Estadística

La modelización estadística utiliza datos reales para estimar las relaciones entre variables y evaluar incertidumbre (Team Familiarize, 2025). Incluye:

- Regresiones (lineal, logística, múltiple).
- Series temporales.
- Modelos de clasificación.
- Inferencia (intervalos de confianza, tests de hipótesis).

A continuación, se presentan los pasos que se deben seguir para crear modelos estadísticos.

- Definir el objetivo del modelo.
- Seleccionar las variables reales y recolectar datos.

- Elegir el tipo de modelo estadístico.
- Estimar los parámetros del modelo.
- Evaluar el modelo (James y otros, 2021).

2.3 Producción lechera y factores determinantes

La producción de leche en sistemas ganaderos es un proceso multifactorial influenciado por variables biológicas, ambientales, nutricionales y de manejo. Diversos estudios destacan que la productividad depende principalmente de factores como genética, alimentación, salud del hato, etapa de lactancia, y condiciones ambientales.

La comprensión de estas variables es esencial para desarrollar modelos capaces de predecir el rendimiento lechero y apoyar decisiones de optimización en fincas agropecuarias.

Factores como la nutrición impactan directamente en la producción lechera, se estima que representa hasta el 60% de variabilidad en la producción diaria. Las condiciones ambientales también influyen de forma significativa en la producción lechera, tomando en cuenta que la temperatura y humedad inciden en el estrés calórico del animal. Finalmente tenemos factores como son las prácticas de manejo del hato y actividades al momento de ordeño (Calvache García, 2012).

2.3.1 Modelos predictivos en la producción lechera

El uso de modelos predictivos en el sector agropecuario ha ganado relevancia debido al avance de herramientas estadísticas y técnicas de inteligencia artificial. Estos modelos permiten anticipar resultados productivos mediante el análisis de múltiples variables simultáneas, facilitando la toma de decisiones basadas en datos

El análisis de datos constituye la fase fundamental en el desarrollo de un modelo predictivo orientado a estimar la producción de leche en la finca El Corazón. Este proceso implica examinar de manera sistemática los datos disponibles, identificar patrones y relaciones entre

variables, y generar información confiable que sirva como base para la toma de decisiones estratégicas. El objetivo principal es comprender los factores que inciden en la calidad y cantidad de leche producida, optimizando el uso de recursos para mejorar la eficiencia del sistema productivo (La Nota en Línea, 2024). Entre los métodos más utilizados se encuentran:

- Regresión lineal múltiple: empleada para estimar la influencia de variables continuas como peso, edad o consumo de alimento sobre la producción de leche.
- Árboles de decisión: útiles para clasificar y predecir comportamientos productivos basados en reglas derivadas de datos históricos.
- Redes neuronales artificiales y aprendizaje automático (Machine Learning): destacan por su capacidad para detectar patrones no lineales y complejos en grandes volúmenes de datos (Torres & Medina, 2021).

La producción de leche depende de múltiples factores interrelacionados, algunos no lineales (por ejemplo, la interacción entre dieta y temperatura). Los métodos tradicionales pueden ser insuficientes para capturar estas complejidades. Al aplicar el concepto Machine Learning al modelo predictivo desarrollado permitirá adaptarse a cambios en las condiciones de operación de la finca.

El desarrollo del modelo predictivo es la etapa central del proyecto, en la que se construye una herramienta capaz de estimar la producción de leche en la finca El Corazón a partir de variables relevantes previamente identificadas en el análisis. Este modelo permitirá anticipar resultados, optimizar recursos y mejorar tanto la calidad como la cantidad de leche producida.

2.3.2 Optimización de recursos en sistemas lecheros

La eficiencia en los sistemas ganaderos es esencial para garantizar sostenibilidad económica y ambiental. La optimización de recursos consiste en utilizar de manera estratégica los insumos disponibles, maximizando la producción sin incrementar costos innecesarios.

Entre las estrategias más destacadas se encuentran:

- Optimización de dietas balanceadas para mejorar conversión alimenticia.
- Implementación de tecnologías de monitoreo de salud y comportamiento animal.
- Ajustes en infraestructura para reducir estrés ambiental.
- Uso de modelos matemáticos para definir decisiones de manejo.

Estos enfoques se fortalecen mediante técnicas de análisis de datos, que permiten detectar puntos críticos de ineficiencia y establecer intervenciones precisas. El desarrollo de un modelo predictivo orientado a la optimización de recursos, como se plantea en esta investigación, se alinea plenamente con dichas estrategias (La Nota en Línea, 2024).

En Ecuador, la aplicación de Data Science en ganadería se encuentra en fase emergente. Proyectos como el Observatorio Lácteo del Ecuador han implementado minería de datos, modelos predictivos y biomodelos sostenibles para mejorar la gestión en fincas pequeñas. Estas iniciativas evidencian que la digitalización y el análisis de datos son herramientas clave para enfrentar desafíos como baja eficiencia, variabilidad climática y limitación de recursos (La Nota en Línea, 2024).

2.4 Metodología

2.4.1 Enfoque

El proyecto sigue un enfoque cuantitativo, experimental–predictivo y basado en analítica de datos, cuyo objetivo es desarrollar un modelo capaz de estimar la producción de leche a partir de variables productivas, ambientales, nutricionales y de manejo del hato.

La metodología se organiza en cinco fases principales:

- Recolección y preparación de datos.
- Análisis exploratorio y selección de variables.
- Construcción del modelo predictivo.

- Validación y optimización del modelo.
- Implementación y generación de recomendaciones para la finca.

2.4.2 Alcance Metodológico

El presente estudio está enfocado a la construcción de un modelo capaz de estimar la producción de leche en la finca El Corazón, con el propósito de mejorar la calidad y cantidad del producto mediante la optimización del uso de recursos productivos, alimenticios y ambientales.

2.4.3 Alcance del estudio

El alcance se enmarca en los niveles descriptivo, correlacional y predictivo (Ordoñez Merino, 2023):

Descriptivo: Se caracterizarán las condiciones actuales de producción de leche, las variables del hato, los factores ambientales y los procesos de manejo de la finca.

Correlacional: Se determinarán las relaciones existentes entre variables productivas (edad, raza, días en lactancia, alimentación, salud, entre otras), ambientales (temperatura, humedad, precipitación) y de manejo con la producción de leche.

Predictivo: Se desarrollará un modelo basado en técnicas de analítica de datos y aprendizaje automático que permita estimar la producción de leche futura por vaca y por periodo.

2.4.4 Delimitación metodológica

2.4.4.1 Delimitación espacial

El estudio se llevará a cabo en la finca El Corazón, considerando únicamente su sistema de producción bovino lechero.

2.4.4.2 Delimitación temporal

El análisis utiliza datos históricos de los últimos 12 meses y contempla un periodo de validación y prueba del modelo durante el tiempo del proyecto.

2.4.4.3 Delimitación poblacional

La población objeto del estudio está compuesta por el total de vacas en producción dentro de la finca durante el periodo evaluado.

2.4.5 Alcance en la recolección de datos

La metodología incluirá:

- Obtención de datos productivos: litros de leche por día, días en lactancia, número de partos, etc.
- Datos zootécnicos: edad, raza, condición corporal, historial sanitario y reproductivo.
- Datos de alimentación: tipo de pastura y suplementos.
- Datos ambientales: temperatura, precipitaciones, humedad relativa, obtenidos del INAMHI.
- Datos operativos: personal asignado para el manejo del hato.

Los datos se recopilarán a partir de registros existentes en las plantillas utilizadas por el operador.

2.4.6 Alcance en el análisis de datos

El estudio contemplará:

- Limpieza, integración y normalización de datos.
- Análisis exploratorio para detectar tendencias, variaciones y valores atípicos.
- Cálculo de correlaciones y análisis estadístico para identificar factores determinantes en la producción.
- Selección de variables relevantes para modelamiento predictivo.

2.4.7 Alcance en el desarrollo del modelo predictivo

Se construye y analiza diferentes tipos de modelos, entre ellos:

- Regresión lineal múltiple.

- Redes neuronales.
- Modelos temporales (p.ej., ARIMA o Prophet) según la estructura de los datos.

El alcance incluye:

- Entrenamiento del modelo.
- Validación cruzada.
- Evaluación mediante métricas como MAE, RMSE y R^2.
- Selección final del modelo más preciso.

El modelo final será capaz de predecir la producción de leche por vaca o por día, según disponibilidad de datos.

2.4.8 Alcance en la implementación

El modelo se entregará en forma de una guía de interpretación y uso para el personal de la finca. Y recomendaciones para optimizar recursos alimenticios, sanitarios y operativos basadas en los hallazgos del modelo.

2.4.9 Límites del estudio

El alcance no incluye:

- Modificación física de infraestructura de la finca.
- Implementación de nuevos sistemas de ordeño.
- Análisis económico-financiero detallado.
- Modelamiento genético de animales.

2.4.10 Diseño Metodológico

2.4.10.1 Variables del Estudio.

Variable Dependiente: Producción de leche (vaca/día) y/o por lote y periodo (día/semana/mes).

Variables Independientes:

- Zootécnicas: raza, edad, número de partos.
- Alimentación y manejo: tipo de forraje, suplementación, acceso a sales minerales.
- Sanidad: calendario de vacunación y desparasitación.
- Ambientales: temperatura media diaria, humedad relativa, precipitación.
- Operativas: personal operativo (Ordoñez Merino, 2023).

2.4.11 Operacionalización de Variables

Se presenta un cuadro con algunas de las variables que se consideran en el desarrollo del modelo.

Variable	Tipo	Unidad	Forma de medición	Frecuencia
Producción de leche	Continua	L/vaca/día	Total en cada ordeño por vaca.	Diario
Raza	Categórica	Holstein, Jersey, etc.	Codificación arete	Única/actualizable
Suplementación	Continua	kg/vaca/día	Peso ofrecido/consumo	Diario
Temperatura	Continua	°C	Promedio diario	Diario
Precipitación	Continua	mm	Acumulado diario	Diario

Tabla 2. 1. Variables operacionales.

En la tabla 2.1 se presentan las variables operacionales que se utilizaran en el análisis y posterior construcción del modelo predictivo.

2.4.12 Población, Muestra y Muestreo

Población: Todas las vacas en producción de la finca El Corazón durante el periodo de estudio.

Muestra: Todas las vacas del hato.

Muestreo: No probabilístico.

2.4.13 Criterios de Inclusión

En el presente trabajo se definen los siguientes criterios para la inclusión de elementos, datos y participantes que formaran parte del estudio.

- Animales: Se considera a vacas en producción, que tengan registros históricos y sanas (que no afecten la seguridad alimentaria).
- Registros: Se utiliza registros produtivos diários por vaca.
- Alimentación: Se consideran el tipo de pasto, suplementos y sales minerales.
- Condiciones climáticas: Se usan registros de humedad, temperatura y precipitación diaria.
- Periodo de estudio: Se utilizan los registros tomados diariamente por un año.

2.4.14 Formulas

2.4.14.1 Correlación entre variables.

Para identificar cuáles factores están más relacionados con la producción. Coeficiente de correlación de Pearson.

$$r = \frac{\sum(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum(x_i - \bar{x})^2 \sum(y_i - \bar{y})^2}}$$

Formula 2.1. Correlación entre variables.

2.4.14.2 Modelo predictivo.

Es el modelo que se ajusta al propósito del presente estudio para predicciones productivas en fincas.

Ecuación general

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_n X_n + \varepsilon$$

Fórmula 2. 1. Formula de modelo predictivo.

Donde:

Y= producción de leche (variable dependiente)

X_i= factores predictivos (alimentación, raza, días en leche, clima, etc.)

β_i= coeficientes

ε= error

2.4.14.3 Coeficientes de regresión.

$$\boldsymbol{\beta} = (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{Y}$$

Formula 2. 2. Coeficientes de regresión.

2.4.14.4 Bondad de ajuste (R^2)

$$R^2 = 1 - \frac{SS_{res}}{SS_{tot}}$$

Formula 2. 3. Bondad de ajuste.

2.4.15 Fórmulas de error del modelo

Se utiliza las siguientes técnicas para medir el error del modelo.

2.4.15.1 Error cuadrático medio (RMSE).

$$RMSE = \sqrt{\frac{1}{n}\sum(y_i - \hat{y}_i)^2}$$

Fórmula 2. 4. Error cuadrático médio.

2.4.15.2 Error absoluto medio (MAE)

$$MAE = \frac{1}{n}\sum | \, y_i - \hat{y}_i \, |$$

Formula 2. 5. Error absoluto medio

2.4.16 Variables

2.4.16.1 Variables Categóricas.

Variable	Tipo	Categorías
Raza	Nominal	Holstein, Jersey, Pardo Suizo
Estado sanitario	Nominal	sano, mastitis, ubre despredida
Estado reproductivo	Nominal	gestante, vacía
Número de parto (paridad)	Ordinal	1, 2, 3, 4+
Uso de suplemento	Nominal	sí / no
Tipo de Pasto	Nominal	Kikuyo o Reygrass
Operador	Nominal	Jose Maldonado, Isidro Calama, Patricio Granja

Tabla 2. 2. Variables categóricas que se analizan para el modelo predictivo.

En la Tabla 2.2 se listan las variables categóricas del dataset Vacas_2025 que se utilizan en el modelo predictivo, se define el tipo de variable y posibles categorías.

2.4.17 Técnicas e Instrumentos

El presente estudio se emplean técnicas e instrumentos que permiten recopilar, organizar, medir y analizar los datos necesarios para la construcción del modelo predictivo de producción de leche.

2.4.18 Técnicas de Recolección de Datos

2.4.18.1 Fuentes.

Registros de ordeño, sanidad, reproducción, alimentación, planillas de potreros, estación meteorológica o servicio climático.

2.4.18.2 Instrumentos.

- Plantillas estandarizadas de la producción diaria registradas por el operador.
- Formularios semiestructurados para eventos de salud como: mastitis, enfermedades, cambios de dieta.

- Medidor de litros de leche.
- Registros del INAMHI para las condiciones climáticas.

2.4.19 Técnicas de Análisis de Datos

Estadística Descriptiva: Se aplica para caracterizar los datos mediante:

- Media.
- Desviación estándar.
- Varianza.
- Tablas y gráficos de comportamiento.

Análisis de Correlación: Para determinar la relación entre variables.

Modelado Predictivo: Se empleará para construir el modelo matemático capaz de estimar la producción de leche según las variables independientes consideradas (Coll Serrano, 2023).

2.4.19.1 Instrumentos

Se utiliza software para recolección y análisis de datos como:

- Excel.
- R.
- Python.

2.4.20 Procedimiento y Análisis de datos

A continuación, se describe de manera ordenada y secuencial las actividades necesarias para recopilar, organizar y analizar la información que permite construir el modelo predictivo. También se presenta el análisis de datos con enfoque cuantitativo utilizando técnicas estadísticas y de modelado predictivo, con ello se puede determinar los factores que afectan e impactan en la producción de leche (Bluman, 2018).

2.4.21 Depuración y organización de los datos

Se ingresa la información de las plantillas llenadas por el operador en hojas de cálculo para la construcción del dataset. Sobre el dataset construido se realizan las siguientes acciones:

- Eliminación de duplicados
- Corrección de registros incompletos
- Codificación de variables categóricas.
- Normalización de datos (Garcia y otros, 2015.).

2.4.22 Análisis estadístico descriptivo

Se realiza un análisis exploratorio de datos obteniendo la media, moda, desviaciones estándar, diagramas e histogramas con el afán de encontrar posibles patrones (Triola, 2018).

2.4.23 Análisis de correlación

Se realizan pruebas de correlación para identificar cuáles variables tienen mayor relación con la producción de leche:

- Coeficiente de Pearson.
- V de Cramer o Chi cuadrado (Agresti, 2007).

2.4.24 Construcción del modelo predictivo

Se utiliza regresión lineal múltiple, con la ecuación general (Hastie, An Introduction to Statistical Learning. Springer, 2021):

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_n X_n + \varepsilon$$

Formula 2. 6. Ecuación para la construcción del modelo predictivo.

Donde:

Y= producción de leche.

X_n= variables independientes.

β_n= coeficientes de regresión.

Para la construcción del modelo de predicción de producción de leche se utilizó la regresión lineal múltiple sobre variables estandarizadas. En términos matemáticos se tiene un modelo representado por la siguiente formula:

$$y = \beta_0 + \beta_1 x_1^* + \beta_2 x_2^* + \cdots + \beta_p x_p^* + \varepsilon$$

Gráfico 2. 1 Construcción del modelo predictivo.

Donde cada x_j^* es la versión estandarizada de la variable:

$$x_j^* = \frac{x_j - \bar{x}_j}{\sigma_j}$$

Gráfico 2. 2. Estandarización de la variable.

Tomando en cuenta que la construcción del modelo predictivo se basa en el dataset "Vacas_2025" (ANEXO 1), se tiene la variable dependiente "produccion" y las variables independientes son "raza", "operador", "pasto", "días de gestacion", etc. En la construcción del código la variable dependiente aparece como "Yc" y todas las variables independientes como "Xc".

De acuerdo al fundamento matemático para la elaboración del modelo predictivo en el código se empezó por estandarizar X:

```
X_std = (Xc - Xc.mean())/Xc.std(ddof=0)
X_std = X_std.fillna(0.0)
```

Gráfico 2. 3. Código del modelo predictivo de producción de leche.

En el grafico 2.1 se estandariza cada columna (media 0, varianza 1; usas ddof=0 → desviación típica poblacional). Luego se rellena NaN con 0.

Luego se realiza el intercepto explicito donde se añade una columna de unos.

```
X_design = np.hstack([np.ones((X_std.shape[0],1)), X_std.values])
```

Gráfico 2. 4. Código del intercepto explicito.

En el grafico 2.2 se elabora el código para añadir una columna con el resultado de unos en los interceptos y concatena las variables estandarizadas.

Se realiza el ajuste por mínimos cuadrados ordinarios. Los coef[0] es el intercepto (con y en su escala original), y coef[1:] son los coeficientes sobre variables estandarizadas.

```
coef = np.linalg.lstsq(X_design, yc.values, rcond=None)[0]
```

Gráfico 2. 5. Código para el ajuste de mínimos cuadrados ordinarios.

El grafico 2.3 muestra como se ejecuta el ajuste por mínimos cuadrados ordinarios.

Finalmente, se realiza la separación de intercepto y betas estandarizadas más ordenamiento.

```
intercepto = coef[0]
betas = pd.Series(coef[1:], index=Xc.columns, name='beta_estandarizada') \
        .sort_values(key=lambda s: s.abs(), ascending=False)
intercepto, betas.head(20)
```

Gráfico 2. 6. Código para separa el intercepto y betas estandarizadas.

Se ordenan por magnitud absoluta de mayor a menor, lo que sirve para ver qué variables tienen mayor efecto en valor absoluto sobre y, dado que todas están en la misma escala z-score. Y devuelve el intercepto y el Top 20 de efectos por tamaño absoluto.

2.4.25 Validación del modelo

Se emplean métricas como:

- RMSE Error cuadrático medio.
- MAE Error absoluto medio.
- R^2 coeficiente de determinación (Hyndman & Koehler, 2006).

2.4.26 Interpretación y recomendaciones

Se analizan los coeficientes y resultados del modelo para determinar:

- Variables que aumentan o disminuyen la producción lechera.
- Impacto del manejo, alimentación y clima.
- Ajustes necesarios para optimizar recursos.

2.5 Resultados y Análisis

2.5.1 Resultados

2.5.1.1 Análisis exploratorio de datos.

Se construyó el dataset “Vacas_2025” (ver Anexo 1) a partir de las plantillas registradas diariamente por el operador de la finca. Este conjunto de datos abarca información recopilada entre el 1 de enero y el 31 de diciembre de 2025. Incluye registros operativos sobre la producción diaria de leche por vaca, así como datos relacionados con su estado de salud; por ejemplo, la presencia o ausencia de mastitis; y variables asociadas al manejo, como el tipo de pasto consumido, etc.

El dataset contiene registros diarios sobre la producción de leche y el manejo del hato a lo largo de todo el año. Sin embargo, no se disponía de información de los factores climaticos para todos los días del período. Para completar esos vacíos, se incorporaron valores referenciales obtenidos del INAMHI.

El dataset está compuesto por 138.457 registros. Durante su elaboración, los datos se fueron incorporando y depurando conforme se ingresaba cada plantilla, dado que en varios casos la información original no era completamente legible.

En la construcción del conjunto de datos, se identificó la presencia de algunas celdas vacías; sin embargo, estas correspondían a ausencias legítimas de información asociada al manejo del hato. Por ejemplo, cuando una vaca dejaba de producir leche, la columna de producción aparecía vacía porque no existía un valor que registrar en esos días.

Para empezar el análisis exploratorio de datos se elaboró un mapa de calor que permite visualizar la distribución de la producción de leche a lo largo del tiempo y para cada vaca. Esta representación facilita identificar patrones y tendencias, proporcionando una base sólida para el desarrollo del modelo predictivo

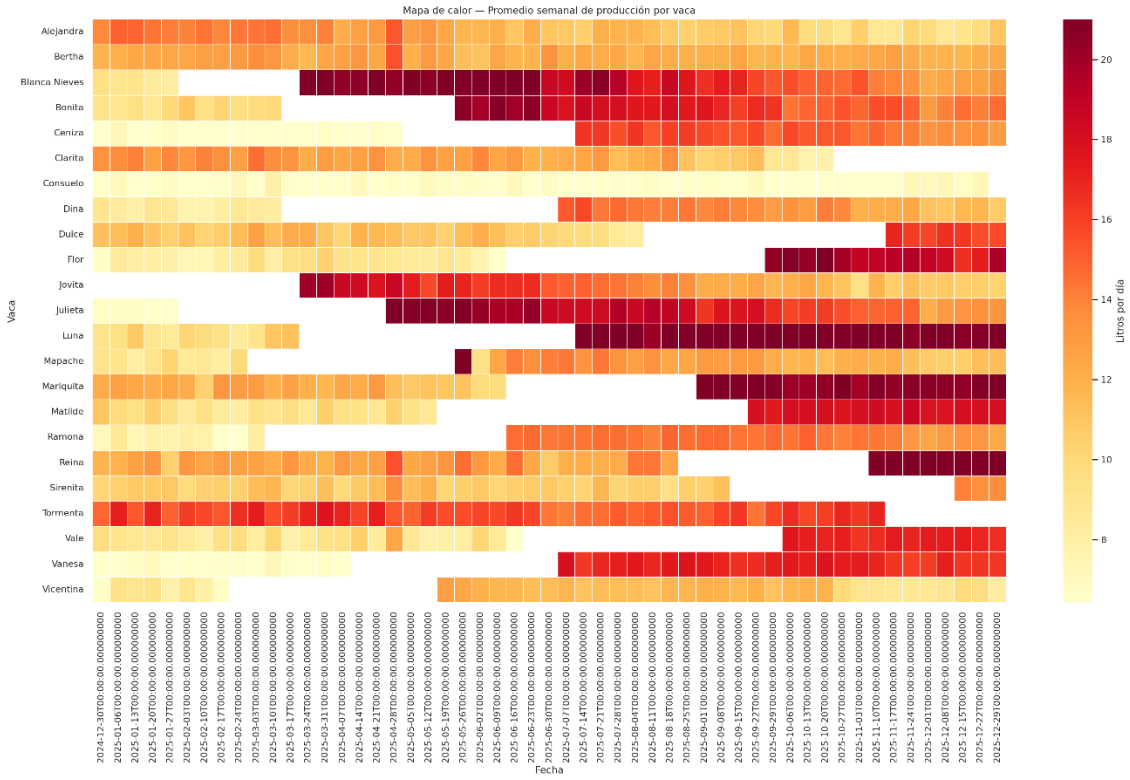


Gráfico 2. 7. Mapa de calor de la producción lechera Finca el Corazón.

En el Gráfico 2.1 se muestra un mapa de calor de la producción lechera que tiene la Finca el Corazón en promedio semanal por cada una de las vacas.

2.5.1.2 Análisis de ANOVA

Se realiza un análisis de la Producción diaria (litros) entre razas y se elabora una tabla de indicadores.

Indicadores de diferencias significativas (ANOVA → Producción L/día).

Factor	Grupos (k)	Muestras (n)	F	p-value	η²
Raza	6	6547	103.516	1.91928e-105	0.0733263
Estado Gestante	2	6453	2756.89	0	0.299405
Partos	7	6546	282.641	8.39912e-323	0.205935

Tabla 2. 3 Indicadores de diferencias significativas (ANOVA → Producción L/día).

La Tabla 2.3 tiene los indicadores de las diferencias significativas de ANOVA realiza sobre la producción por cada día.

De los resultados obtenidos se puede observar:

- F: razón de varianzas. Mientras mayor es F, más evidencia contra H_0.
- p-value: si $p < 0.05$, concluimos diferencias significativas entre al menos dos niveles del factor.
- η² (eta-cuadrado): tamaño del efecto ≈ proporción de varianza de Producción explicada por el factor. Estado Gestante ≈ 0.30, Partos ≈ 0.21 → impacto fuerte; Raza ≈ 0.07 → impacto moderado).
- Estado Gestante: $p \approx 0.000$ → sí hay diferencias claras entre gestantes y no gestantes.
- Partos: $p \ll 0.05$ → sí hay diferencias por número de partos.
- Raza: $p \ll 0.05$ → sí hay diferencias entre razas, aunque con efecto moderado.

2.5.1.3 Análisis de Tukey HSD

Una vez que el ANOVA detecta diferencias, Tukey nos dice qué pares difieren y en cuánto.

Tukey por Raza:

Grupo 1	Grupo 2	Dif. medias	p-ajustada	Límite inf.	Límite sup.	Rechaza H0
Brownsuiss	Holstein	-1.8521	0	-2.4347	-1.2694	True
Brownsuiss	Jersey	-1.4338	0	-2.0883	-0.7794	True
Brownsuiss	Normando	-1.9643	0	-2.6114	-1.3171	True
Brownsuiss	Pizan	3.6476	0	2.6934	4.6019	True
Brownsuiss	Simental	-3.0379	0	-3.765	-2.3107	True
Holstein	Jersey	0.4182	0.0804	-0.0274	0.8639	False
Holstein	Normando	-0.1122	0.9775	-0.5471	0.3227	False
Holstein	Pizan	5.4997	0	4.6745	6.3249	True
Holstein	Simental	-1.1858	0	-1.7327	-0.6389	True
Jersey	Normando	-0.5305	0.0476	-1.0577	-0.0033	True
Jersey	Pizan	5.0815	0	4.2041	5.9588	True
Jersey	Simental	-1.604	0	-2.2268	-0.9812	True

Tabla 2. 4 Análisis de Tukey HSD.

En la Tabla 2.4 se presenta el análisis de Tukey HSD de la producción lechera por raza del dataset Vacas_2025.

Se interpreta que por ejemplo la raza Holstein vs Pizan muestra +5.50 L/día a favor de Holstein (p-ajustada < 0.01), es decir, difieren significativamente. Se puede observar en el siguiente gráfico.

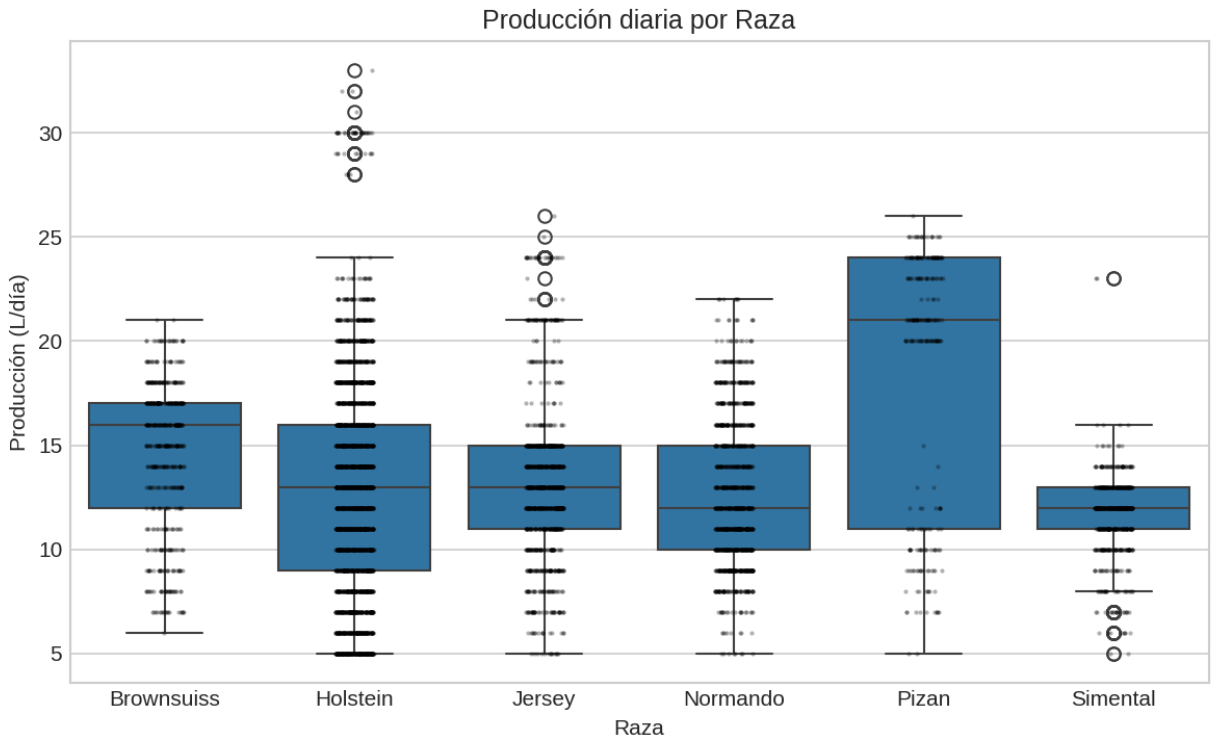


Gráfico 2. 8. Análisis de la producción lechera por raza mediante boxplot.

En el grafico 2.2 se realiza un análisis de la producción lechera por raza mediante diagramas de tipo boxplot para representar la distribución de los datos numéricos del dataset Vacas_2025.

Tukey por Estado Gestante:

Grupo 1	Grupo 2	Dif. medias	p-ajustada	Límite inf.	Límite sup.	Rechaza H0
No	Si	-4.9639	0	-5.1492	-4.7786	True

Tabla 2. 5. Análisis de Tukey por estado gestante.

En la tabla 2.5 Se puede observar que vacas gestantes producen ≈ 4.96 L/día más que las no gestantes ($p < 0.001$). Se puede observar en el siguiente grafico:

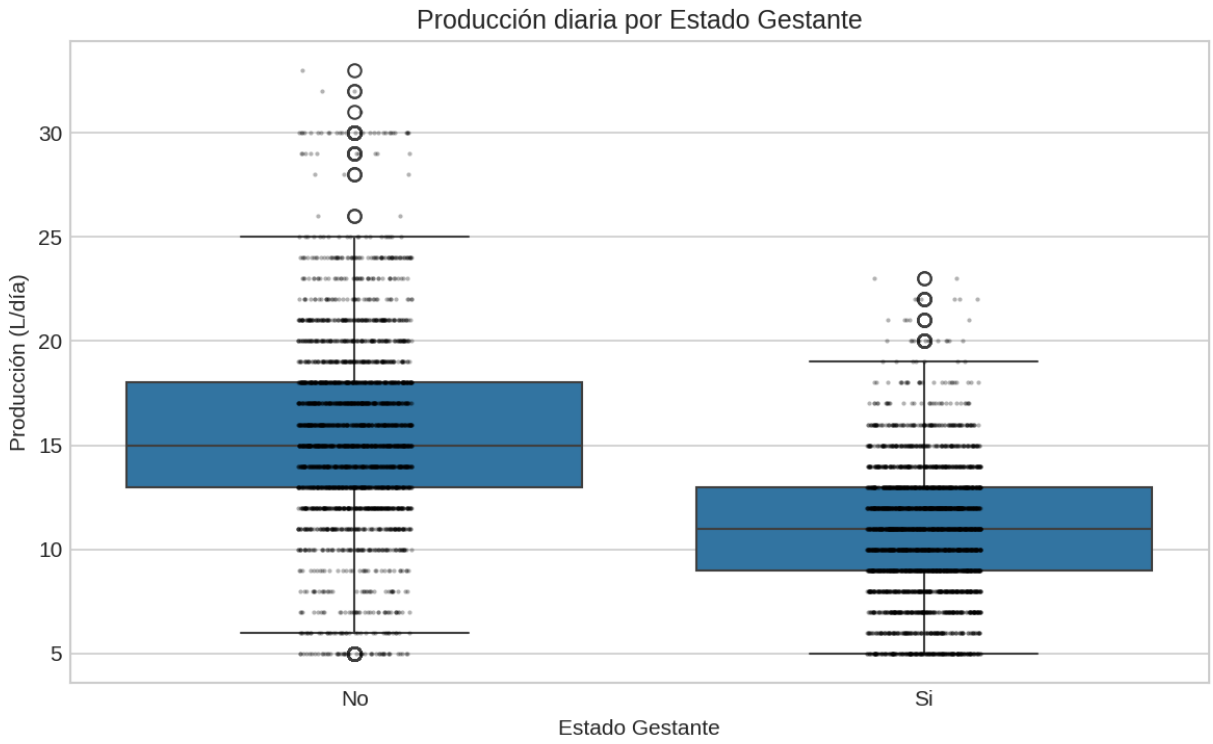


Gráfico 2. 9. Análisis de producción lechera por estado gestante mediante boxplot.

Se presenta el Grafico 2.3 donde se puede observar que las vacas que no están gestantes producen una media de 15 litros mientras las vacas que están en gestación producen aproximadamente 5 litros menos.

2.5.1.4 PCA (Análisis de Componentes Principales) + Gráfica

Para una mirada global, se presenta un PCA con variables numéricas (*Producción, Partos, Humedad, Temperatura, Sales, Lluvia, Vacunación, Desparasitación*, etc.).

- Varianza explicada: PC1 = 23.5%, PC2 = 17.8% (total ≈ 41.3%).
- Color por Raza: ayuda a ver si hay agrupaciones naturales por genética.

Se presenta un gráfico donde los puntos cercanos en el plano PC1–PC2 tienen perfiles similares (producción/ambiente). Y se observan separaciones por raza, consistentes con la diferencia de medias detectada por ANOVA.

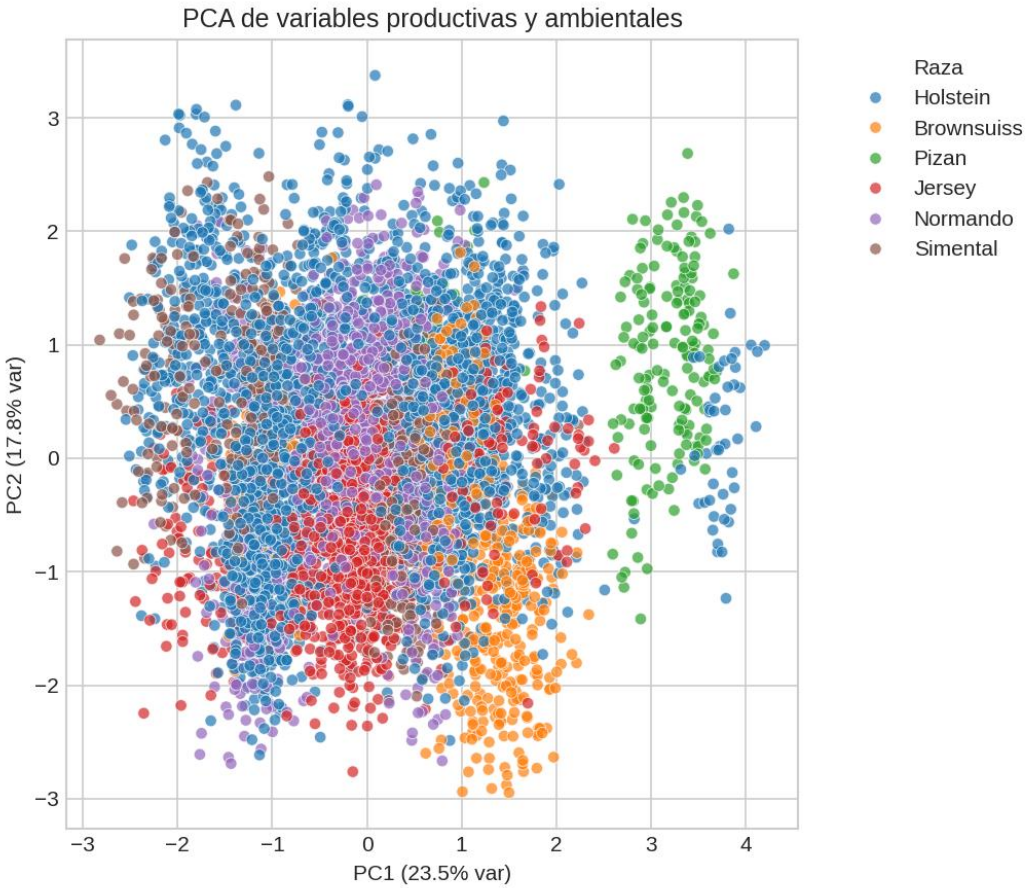


Gráfico 2. 10. Análisis PCA de las variables productivas y ambientales.

Es importante señalar que se aplican análisis basados en ANOVA, pruebas de Tukey y procedimientos de análisis exploratorio de datos como complemento al análisis principal. Estos métodos se emplean con el fin de reforzar la interpretación de los resultados, considerando que el presente estudio se fundamenta en un enfoque estadístico-explicativo, y no corresponde a un ejercicio comparativo entre múltiples algoritmos de machine learning.

2.5.1.5 Limpieza y Normalización de datos

Se normaliza y realiza una limpieza del dataset para que esté listo para el análisis.

Carga de datos y normalización de nombres

```
df = pd.read_excel(PATH, engine='openpyxl')
orig_cols = df.columns.tolist()
df.rename(columns={c: c.strip() for c in df.columns}, inplace=True)

# Mapeo flexible
col_alias = {}
for c in df.columns:
    lc = c.lower()
    if 'produccion' in lc and 'lt' in lc: col_alias['y'] = c
    elif 'numero de partos' in lc:       col_alias['partos'] = c
    elif 'humedad' in lc:                col_alias['humedad'] = c
    elif 'temperatura' in lc:            col_alias['temp'] = c
    elif 'sales' in lc and 'minerales' in lc: col_alias['sales'] = c
    elif 'lluvia' in lc:                 col_alias['lluvia'] = c
    elif 'estado gestante' in lc:        col_alias['gestante'] = c
    elif 'suplement' in lc:              col_alias['suplementos'] = c
    elif 'vacunas' in lc:                col_alias['vacuna'] = c
    elif 'desparasit' in lc:             col_alias['desp'] = c
    elif 'enfermedad' in lc:             col_alias['enfermedad'] = c
    elif 'tipo de pasto' in lc:          col_alias['pasto'] = c
    elif 'raza' in lc:                   col_alias['raza'] = c
    elif 'operador' in lc:               col_alias['operador'] = c
    elif lc == 'fecha':                  col_alias['fecha'] = c
    elif 'fecha de nacimiento' in lc:    col_alias['nac'] = c
    elif 'fecha de preñ' in lc or 'fecha de pre' in lc: col_alias['preniez'] = c

assert 'y' in col_alias, 'No se encontró la columna objetivo: Produccion Diaria lt'
df[col_alias['y']] = pd.to_numeric(df[col_alias['y']], errors='coerce')
orig_cols, col_alias
```

Gráfico 2. 11. Código empleado para la limpieza y normalización de datos Vacas_2025.

El dataset se encuentra listo para empezar el análisis. Se identifican variables numéricas y variables categóricas y se calcula coeficientes de correlación de Pearson entre cada una de las variables predictoras y la variable “Produccion.”

Correlaciones con Producción (Pearson)

```
corrs = {}
for col in Xc.columns:
    try:
        corrs[col] = np.corrcoef(Xc[col], yc)[0,1]
    except Exception:
        corrs[col] = np.nan
corrs_s = pd.Series(corrs, name='corr_con_produccion').sort_values(key=lambda s: s.abs(), ascending=False)
corrs_s.head(20)
```

Gráfico 2. 12. Código para el análisis de correlaciones.

Se observan a las variables tipo de pasto, días de gestación y raza con mayor grado de incidencia sobre la producción lechera, estos resultados nos ayudan a la construcción del modelo predictivo.

	corr_con_produccion
Numero de partos	0.317694
dias_gest	-0.308495
edad_anos	0.278569
Raza_Holstein	-0.231333
enf_ubre	0.225610
enf_ninguna	-0.147847
enf_mastitis	-0.115041
Tipo de Pasto_Raygrass	0.105440
Operador_Jose Maldonado	-0.088518
Humedad	0.060399
Operador_Patricio Granja	0.059234
Raza_Jersey	0.051759
Temperatura C	-0.035084
Raza_Pizan	0.034263
Raza_Simental	0.029762
Raza_Normando	0.029585
Sales Minerales gr	-0.027858
Lluvia	-0.018531
Estado Gestante	NaN

Gráfico 2. 13. Resultado del análisis de correlación sobre el dataset Vacas_2025.

2.5.1.6 Resultados de Correlación

Según los resultados de correlación las variables días de gestación, pasto y raza son factores determinantes en la producción leche. A continuación, se grafican individualmente cada una de las tres variables que inciden en la producción lechera.

La variable raza influye fuertemente en la producción lechera, se observa que las vacas de raza Pizan y Brownsuiss son las que mayor cantidad de litros producen.

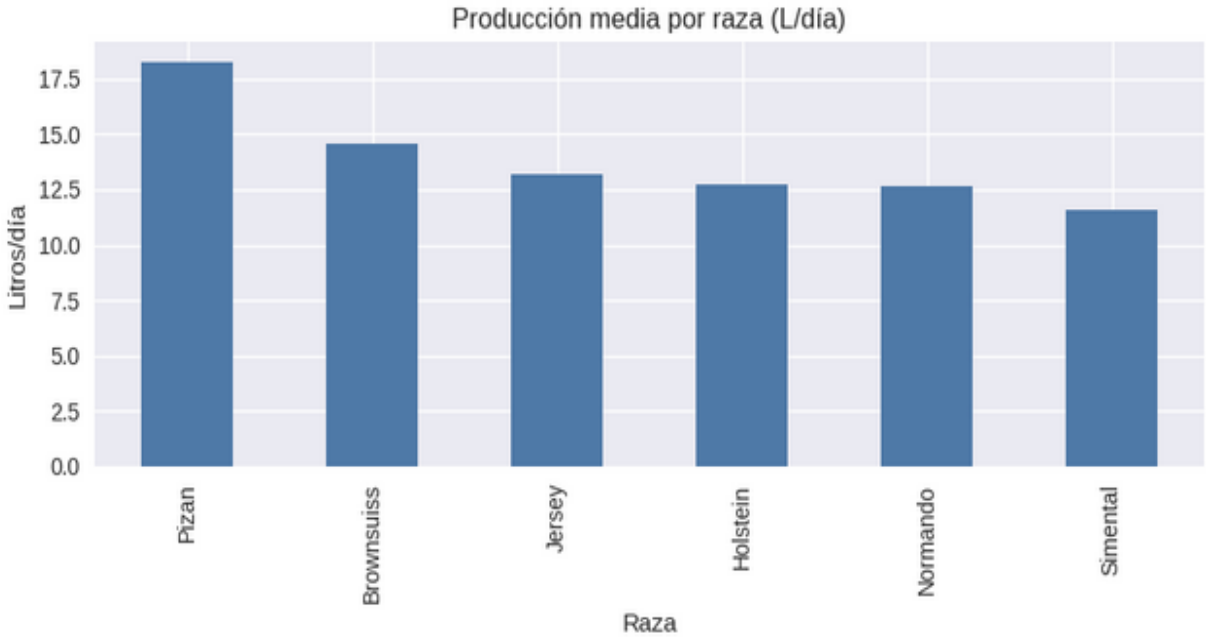


Gráfico 2. 14. Produccion lechera vs Raza dataset Vacas_2025.

En el grafico 2.8 se realiza un diagrama de barras de las razas y la producción lechera, siendo la raza Pizan la que mejor produce.

El siguiente factor que incide en la producción lechera es el tipo de pasto, cuando las vacas consumen el forraje Raygrass incrementan la producción.

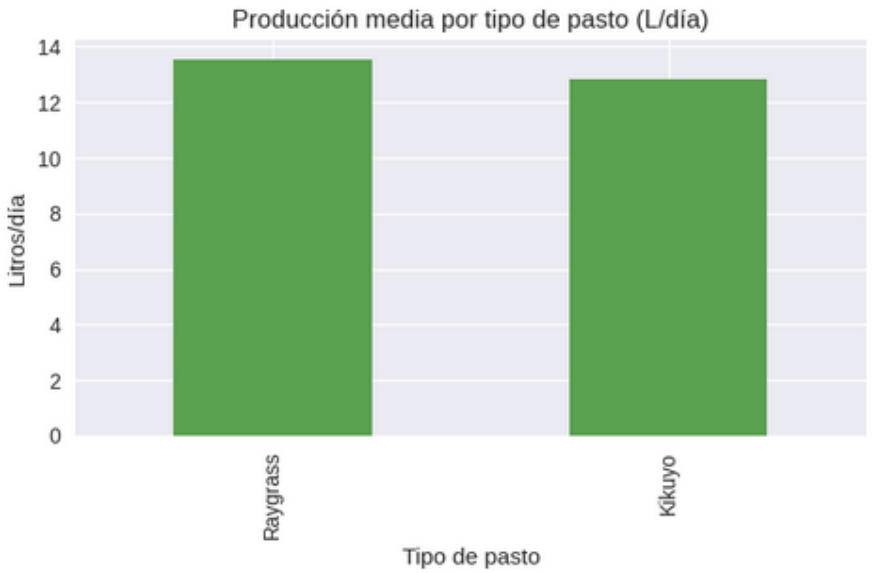


Gráfico 2. 15. Produccion lechera vs tipo de pasto dataset Vacas_2025.

En el grafico 2.9 se analiza mediante un diagrama de barras la producción lechera vs el tipo de pasto, indicando que el Raygrass es mejor.

Finalmente, los días de gestación es un factor que incide gravemente en la producción, mientras avance el estado de gestación la producción decrece.

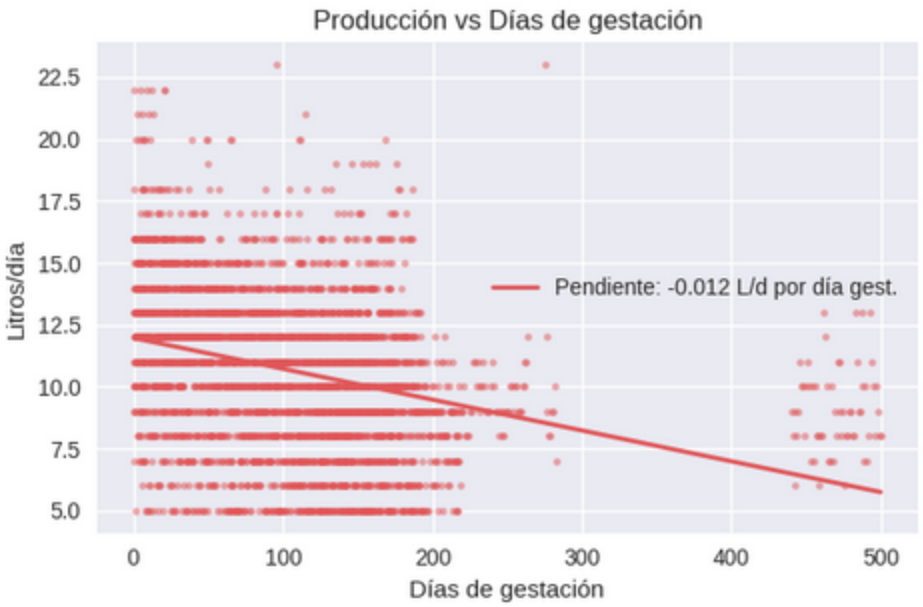


Gráfico 2. 16. Produccion lechera vs días de Gestación dataset Vacas_2025.

En el grafico 2.10 se analiza la producción lechera y los días de gestación de las vacas, se puede ver claramente que la producción decae mientras avanzan los días.

2.5.1.7 Modelo predictivo para producción lechera

Para la construcción del modelo predictivo para la producción lechera se realiza una regresión estandarizada. Se empieza por construir el conjunto de datos con las características en el eje X y para la producción lechera en el eje Y. Posteriormente, se realiza la identificación y conversión de las columnas numéricas. Luego se realiza la identificación de las columnas categóricas como raza, pasto, operador, etc. En los últimos pasos se define la variable objetivo y se construye la matriz de características X. Finalmente se crean los coeficientes Xc e yc para seleccionar filas donde Y no es nulo y ninguna característica en X es nula, así se asegura que el modelo se entrene solo con datos válidos.

```
num_cols = []
for k in ['partos','humedad','temp','sales']:
    c = col_alias.get(k)
    if c:
        df[c] = pd.to_numeric(df[c], errors='coerce')
        num_cols.append(c)

cat_cols = []
for k in ['raza','pasto','operador']:
    c = col_alias.get(k)
    if c:
        cat_cols.append(c)

y = df[col_alias['y']]
X_parts = []
if num_cols: X_parts.append(df[num_cols])
if si_no_cols: X_parts.append(df[si_no_cols])
X_parts.append(df[['enf_mastitis','enf_ubre','enf_otra','enf_ninguna']])
X_parts.append(df[['edad_anos','dias_gest']])
if cat_cols: X_parts.append(pd.get_dummies(df[cat_cols].astype('category'), drop_first=True, dtype=float))

X = pd.concat(X_parts, axis=1).replace([np.inf,-np.inf], np.nan)
X = X.dropna(axis=1, how='all')
X = X.loc[:, X.nunique() > 1]  # quitar varianza cero
mask = (~y.isna()) & (~X.isna().any(axis=1))
Xc, yc = X.loc[mask].copy(), y.loc[mask].copy()
Xc.shape, yc.shape
```

Gráfico 2. 17. Construcción de características y conjunto de datos para el modelado.

En el Grafico 2.11. se realizan los primeros pasos para la construcción del modelo predictivo, se ajustan las características y la variable producción para establecer un conjunto de datos listo para analizar.

A continuación, se realiza un análisis de correlaciones de Pearson entre cada una de las variables predictoras con la variable objetivo, con ellos se puede establecer las variables que mayor impacto tienen sobre la producción.

```
corrs = {}
for col in Xc.columns:
    try:
        corrs[col] = np.corrcoef(Xc[col], yc)[0,1]
    except Exception:
        corrs[col] = np.nan
corrs_s = pd.Series(corrs, name='corr_con_produccion').sort_values(key=lambda s: s.abs(), ascending=False)
corrs_s.head(20)
```

Gráfico 2. 18. Análisis de correlación de Pearson para definir las variables predictoras.

En el grafico 2.12 se realiza el análisis de correlaciones de Pearson para definir las variables predictoras de mayor impacto sobre la producción.

Con la información de correlaciones se construye el modelo predictivo para encontrar los coeficientes MAE, RMSE y R2.

```
from math import sqrt

def metrics_dict(y_true, y_pred):
    y_true = np.asarray(y_true); y_pred = np.asarray(y_pred)
    mae = np.mean(np.abs(y_true - y_pred))
    rmse = sqrt(np.mean((y_true - y_pred)**2))
    ss_res = np.sum((y_true - y_pred)**2)
    ss_tot = np.sum((y_true - y_true.mean())**2)
    r2 = 1 - ss_res/ss_tot if ss_tot>0 else np.nan
    return {'MAE': mae, 'RMSE': rmse, 'R2': r2}

# Split train/test
rng = np.random.default_rng(42)
idx = np.arange(Xc.shape[0])
rng.shuffle(idx)
split = int(0.8*len(idx))
tr_idx, te_idx = idx[:split], idx[split:]
X_train, X_test = Xc.iloc[tr_idx].copy(), Xc.iloc[te_idx].copy()
y_train, y_test = yc.iloc[tr_idx].copy(), yc.iloc[te_idx].copy()

best_model = None; best_name=None; best_pred=None; best_metrics=None; feature_importance=None; model_artifacts = {}

try:
    from sklearn.linear_model import ElasticNet
    from sklearn.ensemble import RandomForestRegressor
    from sklearn.model_selection import GridSearchCV
    import joblib
    SK=True
except Exception:
    SK=False
```

```python
if SK:
    enet = ElasticNet(max_iter=5000, random_state=42)
    grid_enet = {'alpha': [0.001, 0.01, 0.1, 1.0], 'l1_ratio': [0.1, 0.5, 0.9]}
    gs_enet = GridSearchCV(enet, grid_enet, cv=3, n_jobs=-1, scoring='neg_root_mean_squared_error')
    gs_enet.fit(X_train, y_train)
    pred_enet = gs_enet.predict(X_test)
    m_enet = metrics_dict(y_test, pred_enet)

    rf = RandomForestRegressor(random_state=42)
    grid_rf = {'n_estimators': [200, 400], 'max_depth': [None, 12, 20], 'min_samples_leaf': [1, 2, 4]}
    gs_rf = GridSearchCV(rf, grid_rf, cv=3, n_jobs=-1, scoring='neg_root_mean_squared_error')
    gs_rf.fit(X_train, y_train)
    pred_rf = gs_rf.predict(X_test)
    m_rf = metrics_dict(y_test, pred_rf)

    if m_rf['RMSE'] <= m_enet['RMSE']:
        best_model, best_name, best_pred, best_metrics = gs_rf.best_estimator_, 'RandomForest', pred_rf, m_rf
        feature_importance = pd.Series(best_model.feature_importances_, index=X_train.columns).sort_values(ascending=False)
    else:
        best_model, best_name, best_pred, best_metrics = gs_enet.best_estimator_, 'ElasticNet', pred_enet, m_enet
        feature_importance = pd.Series(best_model.coef_, index=X_train.columns).sort_values(key=lambda s: s.abs(), ascending=False)

    try:
        joblib.dump(best_model, 'modelo_produccion.pkl')
        model_artifacts['modelo_guardado'] = 'modelo_produccion.pkl'
    except Exception:
        model_artifacts['modelo_guardado'] = None
else:
    mu = X_train.mean(); sigma = X_train.std(ddof=0).replace(0, 1.0)
    Ztr = (X_train - mu)/sigma
    Zte = (X_test - mu)/sigma
    Xtr = np.hstack([np.ones((Ztr.shape[0],1)), Ztr.values])
    coef = np.linalg.lstsq(Xtr, y_train.values, rcond=None)[0]
    intercepto = coef[0]; beta = coef[1:]
    y_pred = intercepto + Zte.values @ beta
    best_model, best_name, best_pred, best_metrics = {'intercepto': float(intercepto), 'beta': beta.tolist(), 'mu': mu.to_dict(), 'sigma': sigma.to_dict()}, 'OLS_nump
    feature_importance = pd.Series(beta, index=X_train.columns).sort_values(key=lambda s: np.abs(s), ascending=False)
    model_artifacts['modelo_guardado'] = None

best_name, best_metrics
```

Gráfico 2. 19. Construcción del modelo predictivo.

A continuación, se presentan los coeficientes como resultado del modelo:

```
('RandomForest',
 {'MAE': np.float64(1.3083327979845472),
  'RMSE': 1.6786530518228604,
  'R2': np.float64(0.7134197847469818)})
```

Gráfico 2. 20. Resultados del modelo predictivo sobre el dataset Vacas_2025.

MAE: 1.308 indica el error absoluto entre las predicciones y los valores reales, es decir el modelo se equivoca 1.31 unidades, mientras más bajo sea el valor del error, mejor.

RMSE: 1.678 representa al error cuadrático medio y se encarga de penalizar los errores mas grandes. Se interpreta de cuanto se alejan las predicciones del valor real.

R2: 0.713: Se encarga de explicar que tan bien el modelo explica la variabildiad del dato, mientras mas cercano a 1 mejor.

2.5.2 Análisis

2.5.2.1 Análisis del resultado de correlación.

Según los resultados obtenidos el estado de gestación, el tipo de pasto del cual se alimentan las vacas y la raza (genética) son factores determinantes e indicen directamente en la producción leche. A continuación, se realiza un análisis detallado por cada uno de los factores.

En los resultados del análisis exploratorio de datos se puede observar que la raza (genetica) incide directamente en la producción. Las vacas de la raza Pizan y Brownsuiss son las que mayor cantidad de litros producen, entre 17 y 15 litros cada una. Las razas Jersey, Holstein y Normando tienen un nivel de producción similar, aproximadamente 12 litros, cabe destacar que la raza Jersey logra los mismos niveles de producción que razas grandes como la Holstein y Normando, es decir, con menos consumo de pasto logra mejores rendimientos.

El tipo de pasto del cual se alimentan las vacas incide en la producción lechera, el pasto de tipo Raygrass eleva la producción aproximadamente en un litro por vaca, lo cual indica que mientras mejor calidad de forraje consume la vaca mejor producción es capaz de realizar.

El estado y tiempo de gestación es el factor más importante en el análisis, ya que impacta drásticamente la producción diaria de leche. Las vacas tienen sus picos de producción en los primeros 100 días, tiempo en el cual no están en gestación o llevan muy pocos días, a medida que avanza la gestación la producción decrece hasta llegar a niveles mínimos.

2.5.2.2 Análisis de resultados del modelo predictivo

Se realiza un análisis de la importancia entre las variables clave y la producción de leche. Con ello se determina cuanto cambia la producción diaria por cada cambio realizado en las variables clave.

```
feature_importance.head(25) if feature_importance is not None else 'N/A'
```

	0
edad_anos	0.470137
dias_gest	0.194474
Numero de partos	0.091816
Sales Minerales gr	0.072037
Raza_Holstein	0.044086
Humedad	0.026500
enf_ninguna	0.018389
enf_ubre	0.016852
Raza_Normando	0.015227
Temperatura C	0.014456
Raza_Pizan	0.008169
Raza_Jersey	0.006069
Raza_Simental	0.005798
Tipo de Pasto_Raygrass	0.005169
Lluvia	0.005011
Operador_Jose Maldonado	0.003379
Operador_Patricio Granja	0.001429
enf_mastitis	0.001002
Estado Gestante	0.000000

Gráfico 2. 21. Análisis de impacto de las variables y la producción lechera del dataset Vacas_2025.

Ejecutando el modelo predictivo se obtienen valores de producción muy similares a los valores reales, con ello se puede confirmar que el modelo es correcto y puede ser usado para estimar valores futuros de producción.

```
muestra = pd.DataFrame({'y_real': y_test, 'y_pred': best_pred}, index=X_test.index).head(10)
muestra
```

	y_real	y_pred
6582	14.0	15.286389
7054	16.0	15.292867
4911	10.0	10.196980
3624	9.0	9.958920
699	9.0	10.237160
1279	7.0	7.489816
3117	13.0	11.972450
4633	11.0	11.333352
4302	15.0	15.923015
4053	12.0	12.583373

Gráfico 2. 22. Resultado de la predicción para conocer un estimado de producción de leche.

Capítulo 3

3 Conclusiones y Recomendaciones

3.1 Conclusiones

- A partir de las plantillas de registro realizadas por el operador de la finca se logró construir un dataset con información útil y solida de la producción lechera. Al dataset se aplicó técnicas de limpieza y normalización de datos, consiguiendo así un conjunto de datos listo para el análisis.
- En el análisis de correlación las variables independientes días de gestación, pasto y raza son las variables que mas impacta en la producción lechera. Las vacas tienen su pico de producción aproximadamente por 100 días, luego de ello existe un descenso en la producción debido principalmente a que empezó con el estado de gestación. El consumo de pasto Raygrass es mejor que el pasto Kikuyo, cada vaca incrementa su producción hasta en un litro. La raza Pizan presenta los mejores niveles de producción, mientras que la raza Simental muestra los niveles más bajos
- Las variables ambientales y de manejo no inciden en la producción, pese a que la finca esta ubicada en un sitio de alta precipitación y es manejada por varios operadores estos dos factores no afectan la producción diaria de leche.
- El modelo predictivo explica el 71,3% de la variabilidad de la producción; es decir, es capaz de captar bien los patrones de los datos. El coeficiente R^2: 0,713 es un valor aceptable e indica que tan bien ajusta un modelo de regresión.
- Tomando en cuenta que ya se conocen las variables que afectan la producción lechera, la implementación del modelo predictivo permite tomar decisiones informadas, anticipar,

planificar y prevenir; optimizando recursos y mejorando la calidad y cantidad de producción.

3.2 Limitaciones

- Los datos recopilados sobre la operación y el manejo del hato mediante las plantillas se encontraban debidamente estructurados y organizados. Sin embargo, no se disponía de registros relacionados con los factores climáticos. Por esta razón, fue necesario recurrir a fuentes externas, como el Instituto Nacional de Meteorología e Hidrología (INAMHI), para complementar la información requerida.
- No se dispone de información detallada sobre el estado general de salud de las vacas, ya que en las plantillas solo se registraban enfermedades asociadas al sistema mamario. No existen registros sobre otros eventos sanitarios relevantes, tales como episodios de fiebre, diarreas u otras afecciones que podrían influir en la productividad del hato.
- Los registros disponibles en las plantillas abarcan únicamente un periodo de un año, lo cual restringe el análisis histórico y limita la posibilidad de comparar la producción en las mismas etapas de años anteriores.
- La infraestructura tecnológica en la finca El Corazón es prácticamente inexistente, lo que dificulta la implementación de sistemas y herramientas que permitan automatizar y optimizar los procesos operativos.

3.3 Recomendaciones

- Fortalecer la calidad y la recolección de datos mediante la incorporación de infraestructura tecnológica que permita automatizar y optimizar ciertos procesos operativos, como la medición de litros de leche a través de sensores especializados. De esta manera, se contará con registros más precisos, continuos y confiables, lo que

mejorará significativamente la base de datos disponible para el desarrollo y actualización del modelo predictivo.

- Desarrollar un plan de mejora de pasturas, ya que el estudio determinó que la alimentación basada en Ryegrass contribuye a un incremento en la producción de leche. Implementar estrategias de renovación, rotación y manejo adecuado de estas pasturas permitirá optimizar la disponibilidad y calidad del forraje para el hato.
- Implementar protocolos de sincronización de celo en las vacas, con el fin de asegurar al menos una lactancia por año en cada animal. Esto permitirá mantener un mayor número de vacas en sus picos de producción, optimizando así la eficiencia reproductiva y el rendimiento lechero del hato.
- Impulsar un programa de mejoramiento genético en la finca, orientado a reducir progresivamente la presencia de la raza Simental e incrementar el número de vacas Pizán, dada su mayor aptitud lechera. Esta estrategia permitirá optimizar el potencial productivo del hato y mejorar los indicadores de rendimiento a mediano y largo plazo.

Referencias

Agresti, A. (2007). *An Introduction to Categorical Data Analysis*. Wiley.

Bluman, A. (2018). *Elementary Statistics: A Step by Step Approach*. McGraw-Hill.

Calvache García, I. &. (2012). Factores que influyen en la composición nutricional de la leche. *Revista de la Facultad de Ciencias Agropecuarias, Universidad de La Salle*. Retrieved from https://revistas.lasalle.edu.co/index.php/ca/article/view/3609/3587

Cervantes Gómez, L. C. (2015). *Modelización matemática: principios y aplicaciones*. Benemérita Universidad Autónoma de Puebla. Retrieved from https://www.fcfm.buap.mx/assets/docs/publicaciones/Modeliza.pdf

CIL del Ecuador, C. d. (2025). *Sector lácteo ecuatoriano*. Retrieved from https://www.cil-ecuador.org/post/sector-l%C3%A1cteo-ecuatoriano-pilar-estrat%C3%A9gico-para-la-seguridad-alimentaria-y-el-desarrollo-sostenible

Coll Serrano, V. (2023). *Estadística descriptiva. Universitat de València*. Retrieved 02 24, 2026, from Estadística descriptiva. Universitat de València: https://www.uv.es/777EBE4Pzs=/Calculo_y_estadistica/Tema_2_estadistica.pdf

Corporación Financiera Nacional. (2023). Retrieved from https://www.cfn.fin.ec/wp-content/uploads/downloads/biblioteca/2023/fichas-sectoriales-1-trimestre/Ficha-Sectorial-Leche-y-Derivados.pdf

Garcia, S., Julián, L., & Herrera, F. (2015.). *Data Preprocessing in Data Mining. Springer*.

Gutiérrez Pulido, H. &. (2008). *Análisis y diseño de experimentos*. McGraw-Hill/Interamericana Editores.

Gutierrez Pulido, H., & Vara Salazar, R. (2008). *Análisis y diseño de experimentos*. McGraw-Hill.

Hastie, T. T. (2009). *The Elements of Statistical Learning. Springer*.

Hastie, T. T. (2021). *An Introduction to Statistical Learning. Springer*.

Hyndman, R., & Koehler, A. (2006). *Another Look at Measures of Forecast Accuracy*.

James, G., Witten, D., Hastie, T., & Tibshirani, R. (2021). *An Introduction to Statistical Learning. Springer*.

Juhn, M., & Kjell, J. (2013). *Applied Predictive Modeling*. Springer.

La Nota en Línea. (2024). *Observatorio recolecta datos estadísticos de la cadena láctea disponible en Ecuador.* Retrieved from lanotaenlinea: https://lanotaenlinea.com/observatorio-recolecta-datos-estadisticos-de-la-cadena-lactea-disponible-en-ecuador/

Lacteos Franz. (2024). *Factores clave en la producción lechera: genética, nutrición y más*. Retrieved Febrero 22, 2026, from Lacteos Franz: https://lacteosfranz.org/factores-de-la-lecheria/ [novis.com.mx]

Línea, L. N. (2024). *Ecuador ya cuenta con un observatorio para recolectar datos estadísticos de la cadena láctea*. Retrieved 02 27, 2026, from Ecuador ya cuenta con un observatorio para recolectar datos estadísticos de la cadena láctea: https://www.revistazonalibre.ec/2024/09/20/ecuador-ya-cuenta-con-un-observatorio-para-recolectar-datos-estadisticos-de-la-cadena-lactea

Liz, Ortiz. (2024). Sector lácteo en Ecuador: una mirada a los desafíos y oportunidades. *Primicias, Gestion Digital.* Retrieved from https://revistagestion.primicias.ec/analisis-economia-y-finanzas/sector-lacteo-en-ecuador-una-mirada-los-desafios-y-oportunidades/

Minitab Support. (2026). *Diseños factoriales y factoriales fraccionados*. Retrieved 02 21, 2026, from Soporte técnico de Minitab: https://support.minitab.com/es-mx/minitab/help-and-how-to/statistical-modeling/doe/supporting-topics/factorial-and-screening-designs/factorial-and-fractional-factorial-designs

Montgomery, D. (2004). *Diseño y análisis de experimentos.* Limusa Wiley. Retrieved from https://bibliotecadigital.espol.edu.ec/library/publication/diseno-y-analisis-de-experimentos

Nicolás, D. V. (2025, 06 23). *Herramientas, técnicas y metodologías para la aplicación de analytics.* Retrieved Febrero 25, 2026, from InnovaciónDigital360: https://www.innovaciondigital360.com/big-data/analisis-de-datos-tecnicas-y-metodologias-para-la-aplicacion-de-analytics/

Ordoñez Merino, J. I. (2023). *Modelo predictivo aplicando algoritmos de Machine Learning para la producción lechera en la Hacienda El Prado.* Retrieved from https://repositorio.puce.edu.ec/items/3176972a-f2d4-4319-8af4-f1aff1709661

Spanish Psychological Databases. (2026, Febreo 25). *Análisis factorial – factor analysis*. Retrieved from https://spanish.arabpsychology.com/trm/analisis-factorial-factor-analysis

Staff, C. (2023). *¿Qué es un modelado estadístico?* Retrieved 02 23, 2026, from Causera: https://www.coursera.org/mx/articles/statistical-modeling

Team Familiarize. (2025). *Modelado estadístico: técnicas, aplicaciones y tendencias*. Retrieved from Modelado estadístico: técnicas, aplicaciones y tendencias.: https://docs.familiarize.com/es/glossary/statistical-modeling/

Tognetti, P., & Perez, A. (2019). *Introducción a los modelos mixtos*. Universidad de Buenos Aires, Facultad de Agronomía. https://doi.org/ISBN 978-987-3738-22-7.

Triola, M. (2018). *Elementary Statistics*. *Pearson*.

UNIR MEXICO. (2025). *Modelado matemático: qué es, usos y objetivos*. Retrieved 02 16, 2026, from UNIR LA UNIVERSIDAD EN INTERNET: https://mexico.unir.net/noticias/ingenieria/modelado-matematico

Yazmín, Bustán. (2024, Septiembre 20). Ecuador ya cuenta con un observatorio para recolectar datos estadísticos de la cadena láctea. *Revista Zona Libre*. Retrieved from https://www.revistazonalibre.ec/2024/09/20/ecuador-ya-cuenta-con-un-observatorio-para-recolectar-datos-estadisticos-de-la-cadena-lactea/