Evaluación productiva y genética de hatos lecheros del trópico bajo colombiano mediante análisis multivariado integrado: segmentación con K-means y Random Forest

Juan Felipe Marin Grajales

Asesor

Edith Johana Morales Liberato

Universidad Nacional Abierta y a Distancia UNAD

Escuela de Ciencias Básicas, Tecnología e Ingeniería ECBTI

Especialización en Ciencia de Datos y Analítica

2025

Resumen

El presente estudio evaluó el desempeño productivo y genético de vacas lecheras en el trópico bajo colombiano, mediante la integración de técnicas de ciencia de datos y aprendizaje automático. Se analizaron 6.712 lactancias registradas entre 2010 y 2024, aplicando un enfoque basado en el modelo CRISP-DM. A partir de la limpieza, transformación y análisis exploratorio de los datos, se seleccionó la variable Leche día/IEP como indicador integral de productividad ajustada. El análisis reveló que las vacas con genética balanceada (50 % Bos taurus / 50 % Bos indicus), producto del cruzamiento Holstein × Gyr, presentaron el mayor rendimiento promedio. Adicionalmente, las vacas nacidas por transferencia de embriones superaron consistentemente a las obtenidas por reproducción natural, destacando la efectividad de esta biotecnología en condiciones tropicales. Mediante K-means, se identificaron clústeres de progenitores (padres y madres genéticas) con alto rendimiento productivo, mientras que el modelo Random Forest permitió estimar la importancia relativa de cada progenitor en la predicción del desempeño lechero. La combinación metodológica reveló líneas genéticas emergentes con alto potencial de replicación, incluso en animales con pocas lactancias, lo cual representa una oportunidad estratégica para acelerar el mejoramiento genético. Los hallazgos permiten fundamentar decisiones reproductivas con base en evidencia cuantitativa, optimizando la selección de donadoras y reproductores, y fortaleciendo la sostenibilidad de sistemas lecheros tropicales a través de un manejo reproductivo basado en datos.

***Palabras claves*:** producción lechera; genética bovina; K-means; Random Forest; trópico bajo.

Abstract

This study evaluated the productive and genetic performance of dairy cows in Colombia's lowland tropics through the integration of data science techniques and machine learning algorithms. A total of 6,712 lactation records collected between 2010 and 2024 were analyzed using the CRISP-DM methodology. After thorough data cleaning, transformation, and exploratory analysis, the variable Milk per Day/Calving Interval was selected as a comprehensive indicator of adjusted productivity. The analysis demonstrated that cows with balanced genetics (50% Bos taurus / 50% Bos indicus), primarily from Holstein × Gyr crossbreeding, exhibited the highest average performance. Additionally, embryo-derived cows consistently outperformed those from natural reproduction, highlighting the effectiveness of this reproductive biotechnology under tropical conditions. Using K-means clustering, progenitors (sires and dams) were grouped based on the average productivity of their offspring, revealing distinct high-performance clusters. Meanwhile, Random Forest modeling estimated the relative importance of each progenitor in predicting milk yield. This combined approach revealed emergent genetic lines with high replicative potential, even among animals with limited lactation history, representing a strategic opportunity to accelerate genetic improvement. These findings provide a data-driven foundation for reproductive decision-making, enabling the optimization of donor and sire selection, and enhancing the sustainability of tropical dairy systems through precision herd management.

Keywords: dairy production; bovine genetics; K-means; Random Forest; lowland tropics.

Tabla de Contenido

Lista de Tablas

Lista de Figuras

Lista de Apendices

Introducción

La ganadería lechera constituye una de las principales actividades económicas de Colombia, desempeñando un papel esencial en la seguridad alimentaria y en el desarrollo socioeconómico de las zonas rurales. La producción de leche no solo contribuye significativamente al Producto Interno Bruto agropecuario (Ríos-Utrera et al., 2012), sino que también brinda sustento a numerosas familias campesinas. Sin embargo, en regiones del trópico bajo que son caracterizadas por un clima cálido y húmedo, junto a condiciones ambientales adversas; se han evidenciado importantes limitaciones en el potencial productivo y en la sostenibilidad del hato bovino.

La gestión tradicional, basada en prácticas empíricas y en una toma de decisiones poco estructurada, ha resultado en una baja productividad y rentabilidad en el sector (Morales-Cardoso et al., 2020a). La escasa incorporación de enfoques basados en datos productivos y reproductivos impide la identificación oportuna de estrategias de mejoramiento que integren aspectos genéticos y de manejo eficiente. En paralelo, aunque los sistemas de cruzamiento genético se utilizan para potenciar la adaptabilidad del ganado a las condiciones tropicales, la falta de estudios rigurosos sobre su impacto en la producción limita la toma de decisiones informadas (Dijkinga, 2023).

Ante este escenario, el presente trabajo de grado se propone aplicar técnicas avanzadas de ciencia de datos para optimizar la gestión de hatos lecheros en el trópico bajo colombiano. Mediante la integración de métodos de aprendizaje automático y análisis de clústeres en particular, con la segmentación con K-means y el modelado predictivo mediante Random Forest se buscará analizar de forma exhaustiva los registros productivos y reproductivos. Este enfoque permitirá identificar patrones subyacentes, optimizar la selección genética y formular

recomendaciones prácticas orientadas a mejorar indicadores clave, como la producción acumulada de leche y los intervalos entre partos.

La utilización de estos modelos predictivos y de segmentación se alinea con la tendencia global hacia la ganadería de precisión, en la que la integración de técnicas multivariadas y el análisis de grandes volúmenes de datos son fundamentales para la modernización del sector. En particular, la posibilidad de identificar de manera objetiva a los progenitores con mayor desempeño productivo, especialmente en hatos que utilizan biotecnologías de reproduccion como transferencia de embriones e inseminación artificial; ofrece a los productores herramientas decisivas para la selección y mejora genética, potenciando así la eficiencia y sostenibilidad del sistema ganadero.

Esta investigación, al incorporar un enfoque basado en análisis de datos robusto y técnicas de machine learning, pretende contribuir a la modernización del sector lechero colombiano, mejorando tanto los procesos productivos como reproductivos y garantizando una mayor rentabilidad y sostenibilidad a largo plazo.

Descripción del Problema

La producción lechera en el trópico bajo colombiano enfrenta una serie de desafíos estructurales que impactan negativamente su eficiencia y competitividad. Entre los principales problemas se encuentra la falta de integración de herramientas tecnológicas avanzadas para la captura y análisis de datos, lo que limita la capacidad de los ganaderos para tomar decisiones informadas basadas en la evidencia de su propio contexto productivo. La mayoría de las explotaciones dependen de datos genéricos provenientes de otros sistemas productivos que no reflejan las condiciones particulares del trópico bajo, caracterizado por altas temperaturas, humedad relativa elevada y una menor disponibilidad de forrajes de alta calidad (Ríos-Utrera et al., 2012).

El manejo ineficiente de los recursos genéticos es otro problema crítico. A pesar del uso extendido de cruzamientos genéticos, persiste una notable falta de estudios que permitan evaluar su impacto en los indicadores productivos y reproductivos clave, como el intervalo entre partos, la tasa de concepción y la producción de leche por lactancia. La ausencia de este conocimiento impide la adopción de estrategias de mejoramiento genético que se adapten a las condiciones ambientales locales (Dijkinga, 2023).

Además, la gestión reproductiva inadecuada, reflejada en largos intervalos entre partos y baja eficiencia en la conversión alimenticia, contribuye a la disminución de la rentabilidad del sector. La falta de un análisis detallado y sistemático de los datos históricos de producción dificulta la implementación de estrategias efectivas para mejorar la sostenibilidad del hato.

Planteamiento del Problema

A pesar del potencial inherente al sector lechero en el trópico bajo colombiano, su desarrollo se ve comprometido por la ausencia de estrategias de gestión fundamentadas en el

análisis riguroso y sistemático de datos. En la actualidad, la toma de decisiones en muchos hatos se basa en prácticas tradicionales que carecen de soporte analítico, lo que resulta en problemas como intervalos prolongados entre partos, alta variabilidad en la producción de leche y un uso ineficiente de los recursos genéticos (Morales-Cardoso et al., 2020b; Spositto Osvaldo et al., 2020).

Esta situación se agrava en el entorno del trópico bajo, donde las condiciones ambientales adversas como las altas temperaturas y la humedad añaden complejidades adicionales a la gestión productiva y reproductiva del ganado. La carencia de herramientas analíticas basadas en metodologías modernas, como el aprendizaje automático, impide la identificación de patrones específicos de desempeño en función de las características productivas y reproductivas propias del entorno (Dijkinga, 2023). En consecuencia, los productores se ven limitados a decisiones empíricas, lo que disminuye la eficiencia y la rentabilidad global del sistema.

Además, la falta de integración de estas metodologías representa una oportunidad desaprovechada. La aplicación de técnicas como K-means y Random Forest permitiría identificar grupos homogéneos de animales, facilitando la implementación de estrategias diferenciales de manejo y selección genética. Esta segmentación no solo optimizaría la selección de reproductores, sino que también ofrecería una base objetiva para la planificación de la gestión del hato, adaptada a la realidad del trópico bajo colombiano.

Por lo tanto, se plantea la necesidad de desarrollar un modelo analítico basado en ciencia de datos, que integre técnicas de segmentación y modelado predictivo, para evaluar de manera objetiva el desempeño productivo y reproductivo de los hatos lecheros. Con este enfoque, se podrá optimizar la selección de reproductores y mejorar la eficiencia del sistema, impulsando la sostenibilidad y rentabilidad del sector lechero en condiciones desafiantes.

Sistematización del Problema

Para abordar de manera estructurada la problemática identificada en el sector lechero del trópico bajo colombiano, resulta fundamental desglosar el problema general en interrogantes específicas que permitan delimitar la investigación y orientar la metodología analítica. Esta sistematización no solo clarifica qué aspectos se deben investigar, sino también cómo los diferentes componentes, tanto productivos como reproductivos, los cuales interactúan en la eficiencia del hato, y de qué manera la aplicación de técnicas avanzadas de ciencia de datos como el aprendizaje automático, K-means y Random Forest, puede aportar a la optimización de la gestión ganadera. Dentro de este marco, se plantean las siguientes preguntas de investigación:

- ¿Cuáles son los factores productivos y reproductivos que más influyen en la eficiencia del hato lechero en el trópico bajo colombiano? Esta pregunta permite identificar y priorizar los indicadores clave como la producción acumulada de leche, los intervalos entre partos y otros parámetros reproductivos que determinan el desempeño general del sistema.
- ¿De qué manera el análisis de datos históricos de producción y reproducción puede contribuir a la optimización de la gestión del hato? Aquí se explora cómo el uso de registros históricos y el análisis sistemático de los mismos pueden ofrecer una visión objetiva y detallada para la toma de decisiones, mejorando tanto la eficiencia operativa como la planeación a largo plazo.
- ¿Cuáles son los patrones productivos y reproductivos que pueden identificarse mediante técnicas de aprendizaje automático y análisis de clústeres? Esta interrogante se centra en la aplicación de metodologías como K-means para segmentar la población animal y Random

Forest para el modelado predictivo, facilitando la detección de comportamientos homogéneos y la identificación de oportunidades de mejora.

- ¿Cómo se puede mejorar la toma de decisiones en la selección genética del hato mediante el uso de herramientas de ciencia de datos? Aquí se evalúa la capacidad de los algoritmos de aprendizaje para identificar de manera objetiva a los progenitores con mejor desempeño, optimizando así los esquemas de selección genética y aumentando la eficiencia productiva.
- ¿Qué estrategias de manejo pueden derivarse de los resultados obtenidos para mejorar la sostenibilidad y rentabilidad del sistema productivo?; esta pregunta busca traducir los hallazgos analíticos en recomendaciones prácticas de manejo que puedan ser implementadas para maximizar la rentabilidad y sostenibilidad del hato, atendiendo a las particularidades del entorno tropical.

La respuesta a estas interrogantes permitió no solo estructurar el análisis de los datos de manera efectiva, sino también identificar áreas críticas de intervención. Al integrar enfoques estadísticos y modelos predictivos con análisis descriptivo, se podrá proponer soluciones prácticas que beneficien a los productores, mejoren la selección genética y optimicen la gestión integral del sistema lechero.

Justificación

La ganadería lechera en el trópico bajo colombiano enfrenta desafíos críticos relacionados con la baja productividad, el uso ineficiente de recursos genéticos y la falta de integración tecnológica en la toma de decisiones productivas y reproductivas. Estos problemas, exacerbados por las condiciones climáticas adversas y la dependencia de métodos tradicionales, limitan el potencial de los hatos y afectan directamente la competitividad y sostenibilidad del sector. Este proyecto responde a la necesidad de abordar estas brechas mediante la aplicación de herramientas avanzadas de aprendizaje automático y análisis de clústeres.

El análisis de registros productivos y reproductivos entre 2000 y 2024 permitirá identificar patrones complejos que los métodos convencionales no pueden detectar. Estas técnicas no solo mejorarán la precisión en la predicción de indicadores clave, como la producción de leche y el intervalo entre partos, sino que también facilitarán la evaluación comparativa de razas, cruzamientos y grupos genéticos. Al optimizar la selección de reproductores y evaluar el impacto de los manejos genéticos, el proyecto busca ofrecer soluciones prácticas y replicables para mejorar la eficiencia del sector.

Además, la implementación de este enfoque analítico tiene implicaciones socioeconómicas significativas, al promover estrategias basadas en datos que incrementen la rentabilidad de los productores y refuercen la sostenibilidad del sector ganadero. Este proyecto no solo aborda un problema técnico relevante, sino que también genera conocimiento aplicable que contribuye al desarrollo del sector agropecuario colombiano, fortaleciendo su papel en la economía rural y en la seguridad alimentaria del país.

Este trabajo, al integrar ciencia de datos y mejoramiento genético, se posiciona como una iniciativa innovadora para transformar el manejo productivo y reproductivo de los hatos lecheros en el trópico bajo, contribuyendo al progreso técnico y social del sector ganadero.

Objetivos

Objetivo General

Aplicar técnicas de bosques aleatorios y análisis de clústeres para analizar los registros productivos y reproductivos de un hato lechero en el trópico bajo colombiano, recopilados entre los años 2000 y 2024, con el fin de identificar patrones, optimizar la selección genética y mejorar la productividad y sostenibilidad del sistema ganadero.

Objetivos Específicos

Realizar la limpieza, transformación y análisis exploratorio de la base de datos mediante el análisis estadístico descriptivo de las variables, para garantizar la calidad e integridad de la información y establecer una base confiable para el desarrollo de modelos predictivos en el estudio.

Determinar si la composición genética y el origen reproductivo de las vacas inciden significativamente en la productividad del hato lechero, mediante el análisis comparativo de diferentes categorías genéticas y reproductivas, para fundamentar estrategias de mejoramiento y manejo.

Identificar a los progenitores genéticos asociados a un alto rendimiento productivo en la descendencia, integrando técnicas de agrupamiento (K-means) y modelación predictiva (Random Forest) para fundamentar estrategias de selección y manejo reproductivo.

Integrar los hallazgos del análisis de datos, la incidencia de la composición genética y el origen reproductivo en la productividad, y la identificación de progenitores sobresalientes, para fundamentar estrategias de mejoramiento reproductivo en el hato lechero.

Marco de Referencia

El avance en las tecnologías de análisis de datos, inteligencia artificial y aprendizaje automático ha generado un cambio estructural en los sistemas productivos agropecuarios. En particular, la ganadería de leche ha adoptado progresivamente estas herramientas para analizar con mayor precisión grandes volúmenes de datos generados en campo, con el objetivo de mejorar la eficiencia productiva, la toma de decisiones y la sostenibilidad del sistema (Brum Luciano Moraes da Luz et al., 2019; Fernando Johann et al., 2023)

Los sistemas tradicionales de gestión del hato, que dependían de promedios poblacionales, observación empírica o registros limitados, están siendo reemplazados por modelos matemáticos y algoritmos inteligentes capaces de predecir comportamientos productivos y reproductivos a partir de múltiples variables, incluso en condiciones complejas y no lineales (Dongre & Gandhi, 2016; Sarkar et al., 2018; Sharma Shivangi et al., 2023). Esta transformación se basa en diversas técnicas de la ciencia de datos como desarrollo de redes neuronales artificiales, árboles de decisión, algoritmos bayesianos, técnicas de agrupamiento no supervisado y sistemas expertos adaptados al entorno agropecuario (Chaturvedi Shailesh et al., 2013; Perdigon Llanes Rudibel & Gonzales Benitez Neilys, 2020; Spositto Osvaldo et al., 2020).

En el contexto de la producción bovina, se han documentado aplicaciones exitosas de modelos predictivos para la estimación del valor genético de animales (Dijkinga, 2023), predicción de producción diaria de leche (Zea et al., 2022), eficiencia reproductiva y diagnóstico de mastitis subclínica (Estrada Carvaja Verny et al., 2019), así como para la clasificación y descarte técnico de animales mediante árboles de decisión (Morales-Cardoso et al., 2020a). De igual forma, se ha validado el uso de algoritmos no supervisados en la identificación de perfiles

productivos a partir del análisis multigeneracional, como ocurre en estudios con terneros de raza Angus (Spósitto Osvaldo et al., 2019; Sposítto Osvaldo et al., 2020).

Paralelamente, desde una perspectiva teórica y metodológica, se ha fortalecido el uso de modelos matemáticos para describir fenómenos biológicos como el crecimiento o la curva de lactancia y se ha promovido el desarrollo de metodologías de optimización aplicadas a la nutrición y la eficiencia del sistema lechero (Fernández Chuairey et al., 2017). Adicionalmente, se ha planteado la necesidad de adaptar y simplificar estos modelos para el entorno productivo real, como lo demuestra el uso de redes neuronales artificiales y regresiones múltiples para predecir el rendimiento de vacas bajo condiciones tropicales (Chaturvedi Shailesh et al., 2013; Perdigón Llanes & González Benítez, 2022), tambien es el caso de la adaptación de métodos de aprendizaje supervisado para estimar las producciones de leche en ganaderías tropicales como la raza Gyr y sus cruzamientos en sistemas de pastoreo (Zea et al., 2022).

Asi las cosas, la literatura reciente ha resaltado el papel estratégico de herramientas de inteligencia de negocios y almacenamiento de datos en el sector agrario, permitiendo integrar, procesar y visualizar datos de multiples fuentes que dan soporte a decisiones técnicas y estratégicas en ganadería (Brum Luciano Moraes da Luz et al., 2019). Así mismo, la revisión bibliográfica de Lachman y López (2018) identifica el surgimiento de nuevas habilidades y áreas de conocimiento en sectores como la ganadería de precisión, promoviendo la formación de perfiles interdisciplinares con competencias en ciencia de datos, inteligencia artificial y tecnologías de la informacion.

Finalmente, investigaciones como las de Flores et al. (2006) y Perdigón-Llanes y González-Benítez (2020) han documentado cómo las técnicas de minería de datos y redes bayesianas pueden complementar o incluso anticipar decisiones que tradicionalmente se obtenían

con el modelo BLUP, es cual es un método estadístico utilizado en mejoramiento genético para estimar valores genéticos de individuos a partir de información fenotípica y genealógica; en este sentido la implementación de estas tecnologías hace más eficiente el proceso de selección genética.

En conjunto, estos antecedentes evidencian una transición del enfoque empírico al enfoque analítico en la producción animal, donde el uso de modelos inteligentes, sustentados en ciencia de datos, representa no solo una mejora en la capacidad de diagnóstico y predicción, sino una vía para transformar integralmente la toma de decisiones en sistemas ganaderos tropicales.

Estado del Arte

La transformación digital en el sector agropecuario ha generado un escenario propicio para la incorporación de tecnologías basadas en inteligencia artificial y aprendizaje automático, especialmente en sistemas de producción animal. En la ganadería bovina lechera, estas tecnologías se han consolidado como herramientas clave para mejorar la predicción del rendimiento individual, anticipar problemas sanitarios, optimizar procesos reproductivos y apoyar la toma de decisiones técnicas con base en datos confiables y en tiempo real.

Uno de los avances más significativos ha sido el desarrollo de modelos predictivos basados en redes neuronales artificiales (Zea et al., 2022) ; demostraron que, a partir de registros parciales de ordeño, es posible predecir con alta precisión la producción diaria de vacas de raza Gyr utilizando redes neuronales artificiales. Este enfoque ha sido reforzado por Chaturvedi Shailesh et al. (2013), quienes implementaron modelos retroalimentación para estimar la producción vitalicia de vacas, logrando valores elevados de correlación y mínimos errores de predicción. Perdigón Llanes & González Benítez (2022), en una revisión sistemática, concluyen que los modelos de redes neuronales artificiales superan ampliamente a las regresiones

tradicionales, y destacan la creciente eficacia de arquitecturas como las redes NARX (modelo auto regresivo no lineal con entrada exógena) y CNN (red neuronal convolucional).

Además de las redes neuronales artificiales, el uso de algoritmos de clasificación y agrupamiento no supervisado ha permitido identificar perfiles productivos y genéticos a partir de variables complejas. En esta línea, Spositto Osvaldo et al. (2019, 2020) aplicaron técnicas como K-means, SOM (Self-Organizing Maps) y EM (Expectation-Maximization) para clasificar terneros Angus según su peso al nacer, considerando datos multigeneracionales. Sus hallazgos evidencian que características como la edad de la madre o la genética del abuelo materno pueden agrupar animales con comportamientos productivos similares, incluso sin necesidad de etiquetado previo.

El análisis de la literatura científica también evidencia una tendencia creciente hacia el uso de árboles de decisión para apoyar decisiones estratégicas en el manejo del hato. Morales-Cardoso et al. (2020a) propusieron la metodología M3S, la cual se vale de la ciencia de datos y análisis inteligente para guiar proyectos de minería de datos de forma estructurada y metodológica, similar a CRISP-DM o SEMMA, pero con un enfoque más adaptado a ambientes de modelado inteligente y aprendizaje automático, en este sentido la utilizaron para la identificación de vacas descartables, combinando inteligencia artificial con reglas técnicas establecidas por expertos. Este enfoque permitió automatizar decisiones de descarte de animales con base en variables como abortos, enfermedades, duración de lactancia y producción acumulada. De forma complementaria, Estrada Carvaja Verny et al. (2019) compararon modelos de regresión clásica con algoritmos ML para detectar mastitis subclínica, encontrando que los modelos inteligentes alcanzaron mayores niveles de precisión, especificidad y sensibilidad.

Desde un enfoque más amplio, se han realizado diversos estudios de revisión para analizar el impacto global de estas tecnologías. Dongre & Gandhi (2016) destacan la aplicación de las redes neuronales artificiales en múltiples áreas de la ganadería, incluyendo reproducción, salud, alimentación y calidad de productos. (Sarkar et al., 2018; Sharma Shivangi et al., 2023) por su parte, han documentado el uso de la inteligencia artificial en la agricultura desde una perspectiva integral, resaltando su impacto positivo en sostenibilidad, monitoreo inteligente, automatización y control de variables críticas. Estas aplicaciones también han sido exploradas en el sector lechero mediante el uso de sensores, visión computacional, análisis multivariante y sistemas de apoyo a la toma de decisiones.

A nivel estratégico, Brum Luciano Moraes da Luz et al. (2019) analizaron el uso de herramientas de business intelligence (BI) y data warehouse en sistemas agropecuarios, destacando la relevancia de tecnologías de código abierto para el análisis y visualización de información. Este enfoque es especialmente útil en contextos donde se requiere integrar múltiples fuentes de datos para evaluar el desempeño productivo y económico. En esa misma línea, Jeremias Lachman & Andres lopez. (2018) resaltan que la transformación tecnológica del sector demanda perfiles profesionales con competencias en análisis de datos, IA y tecnologías digitales, lo que plantea nuevos retos para la formación de talento en ganadería de precisión.

Finalmente, el mapeo sistemático realizado por Fernando Johann et al. (2023) ofrece una visión general sobre las aplicaciones de IA y ML en el sector agropecuario, clasificando los estudios en cuatro grandes categorías: producción animal, vegetal, sostenibilidad y aplicaciones generales. Esta revisión confirma que las tecnologías inteligentes no solo están siendo adoptadas ampliamente, sino que tienen un papel fundamental en la construcción de sistemas agropecuarios más rentables, resilientes y ambientalmente responsables.

En conjunto, el estado actual del conocimiento demuestra que la ciencia de datos, aplicada a la ganadería lechera, ha dejado de ser una tendencia emergente para consolidarse como un componente esencial en la innovación y gestión del hato, abriendo nuevas oportunidades para optimizar el rendimiento productivo individual y colectivo en condiciones reales de campo.

La Tabla 1 resume los principales antecedentes identificados en la literatura, organizados por región, técnica utilizada, hallazgos y relevancia directa para este estudio, en síntesis permite posicionar el presente trabajo frente al estado actual del conocimiento.

Tabla 1

Comparativo del Estado del Arte

Autor y año	Región / país	Técnica principal	Hallazgo clave	Relevancia para este estudio
Morales-Cardoso et al. (2020)	Colombia	K-means, Random Forest	Identificación de clústeres de progenitores productivos	Base metodológica para segmentación genética
Zea et al. (2022)	Colombia	ML aplicado a Holstein × Gyr	Brechas en validación de modelos	Justifica aplicabilidad en condiciones locales
Brum et al. (2019)	Brasil	Big Data, BI	Integración de BI en agroindustria	Muestra viabilidad tecnológica
Johann et al. (2023)	Internacional	IA aplicada al agro	Clasificación de aplicaciones por sector	Confirma tendencia global en ciencia de datos

Marco Contextual

La presente investigación se enmarca en el contexto de la ganadería bovina especializada en producción de leche en sistemas tropicales del trópico bajo colombiano, específicamente en una finca ubicada en el Valle del Cauca. Estos sistemas productivos enfrentan condiciones

ambientales altamente variables, caracterizadas por altas temperaturas, elevada humedad relativa, y estacionalidad marcada en la disponibilidad de pasturas, lo cual impacta significativamente en la producción y en el manejo del ganado (Morales-Cardoso et al., 2020a). Tradicionalmente, la toma de decisiones técnicas en estas ganaderías ha dependido de la experiencia empírica del productor, con escasa sistematización de los registros productivos individuales, lo cual restringe considerablemente las oportunidades de mejora en los indicadores de rendimiento (Zea et al., 2022).

En Colombia, gran parte de la ganadería lechera se concentra en condiciones tropicales similares a las del hato objeto de estudio, representando un porcentaje significativo del inventario bovino nacional. Los sistemas lecheros tropicales suelen combinar diferentes razas con el objetivo de equilibrar la productividad y la adaptabilidad a las condiciones climáticas adversas, propias del trópico bajo. Estas condiciones generan desafíos importantes relacionados con la regulación térmica, estrés por calor, disminución del consumo voluntario de alimento y mayor incidencia de enfermedades tropicales, factores que afectan negativamente la eficiencia productiva y reproductiva de los hatos.

El hato en estudio cuenta con aproximadamente 500 vacas en producción y se caracteriza por un alto nivel de tecnificación; con más de 40 años de implementación de programas de mejoramiento genético, inicialmente se enfocaron en cruzamientos entre razas lecheras como Holstein, Jersey, razas criollas como Hartón del Valle y algunas razas tipo indicus, buscando composiciones genéticas diversas con adaptabilidad al clima tropical. Desde aproximadamente el año 2010, se estableció como objetivo el cruzamiento Holstein × Gyr, debido a sus beneficios en productividad y adaptación al estrés térmico, mediante técnicas de monta natural e inseminación artificial con toros élite comerciales.

En los últimos diez años, se ha incorporado la transferencia de embriones para acelerar el mejoramiento genético, manteniendo un núcleo de vacas puras Gyr como donadoras de embriones. La selección de estas vacas Gyr puras es estratégica debido a su comprobada resistencia y adaptabilidad a condiciones tropicales, así como por su potencial productivo y reproductivo. Estas características genéticas las hacen ideales para producir embriones viables que al combinarse con toros Holstein de alto valor genético, identificados y seleccionados comercialmente por su elevado potencial productivo lechero, generan una progenie híbrida con alta capacidad para producir leche en condiciones de trópico bajo, equilibrando la producción y la adaptación ambiental.

Este contexto es ideal para evaluar metodologías de modelado predictivo basadas en ciencia de datos, dado que el hato dispone de registros detallados por lactancia, control riguroso de la producción lechera, antecedentes genéticos conocidos y prácticas avanzadas de manejo reproductivo. Estudios previos han demostrado que el análisis sistemático de datos históricos en sistemas ganaderos similares puede convertirse en una herramienta estratégica para optimizar los procesos productivos, mediante la identificación temprana de patrones asociados a la producción acumulada, días en lactancia, eventos reproductivos y problemas sanitarios (Morales-Cardoso et al., 2020a).

En concordancia con lo planteado por Brum Luciano Moraes da Luz et al. (2019), se destaca la necesidad de contextualizar soluciones tecnológicas al entorno productivo real, integrando plataformas analíticas con herramientas de visualización adaptadas para usuarios no especializados, facilitando así la aplicabilidad y utilidad de la información generada en campo. En este sentido, la utilización de algoritmos y modelos matemáticos basados en aprendizaje automático (Machine Learning) para la estimación precisa de valores genéticos y productivos ha

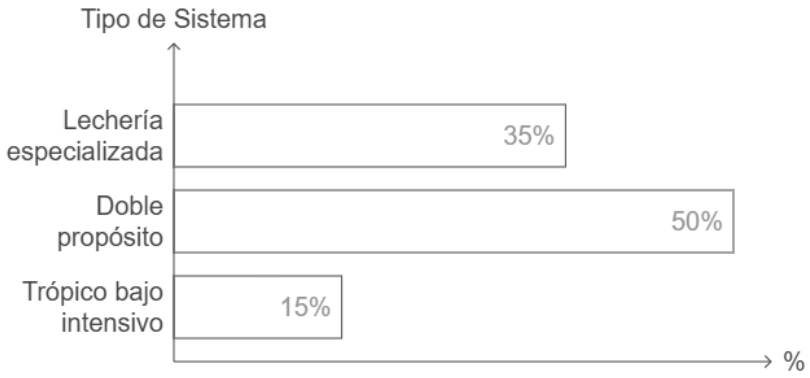
demostrado resultados promisorios, al ofrecer predicciones robustas incluso en condiciones de datos parciales o incompletos (Dijkinga, 2023).

De esta forma, el objetivo del presente trabajo es validar metodologías analíticas basadas en la ciencia de datos que permitan, a partir de registros históricos de este hato especializado, anticipar comportamientos productivos y apoyar decisiones estratégicas fundamentadas en evidencia científica, lo cual contribuirá significativamente a mejorar la eficiencia y sostenibilidad del sistema productivo estudiado.

En este marco, es relevante contextualizar la posición del hato estudiado dentro de la estructura ganadera nacional. Según datos reportados por la Federación Colombiana de Ganaderos (Fedegan, 2023), los sistemas de producción lechera en Colombia se distribuyen principalmente entre el doble propósito (50 %), la lechería especializada (35 %) y los sistemas intensivos de trópico bajo (15 %). Cabe destacar que tanto el sistema de doble propósito como el de trópico bajo intensivo operan bajo las condiciones propias del trópico bajo, caracterizadas por altas temperaturas y desafíos ambientales específicos. Esta composición sectorial no solo permite dimensionar la representatividad del sistema analizado en el presente estudio, sino que también subraya la necesidad de aplicar enfoques analíticos adaptados a las condiciones particulares del trópico bajo, donde los desafíos productivos y genéticos requieren soluciones específicas. La siguiente gráfica presenta de forma resumida esta distribución porcentual, sirviendo como referencia para comprender la pertinencia del enfoque metodológico propuesto.

Figura 1

Distribución Porcentual de Sistemas Productivos Lecheros en Colombia



Nota. Adaptado del Balance y perspectivas sector ganadero 2023 – 2024. *Fuente.* (Fedegan, 2023).

Marco Teórico

La ciencia de datos es un campo multidisciplinar que combina elementos de estadística, programación, ingeniería y conocimiento del dominio para extraer patrones y conocimiento a partir de grandes volúmenes de información (Dongre & Gandhi, 2016). Su evolución se ha visto impulsada por el crecimiento exponencial de la capacidad de cómputo, el abaratamiento del almacenamiento de datos y el desarrollo de metodologías más avanzadas de análisis.

En este marco, el aprendizaje automático o Machine Learning (ML) constituye un conjunto de técnicas orientadas a que los sistemas aprendan de forma automática a partir de los datos (Dongre & Gandhi, 2016; Sarkar et al., 2018). Entre las categorías principales de ML se destacan:

- Aprendizaje supervisado: el modelo se entrena con datos etiquetados para predecir valores futuros (p. ej., redes neuronales para estimar la producción de leche).

- Aprendizaje no supervisado: se buscan patrones en datos sin etiquetas (p. ej., algoritmos de agrupamiento para identificar perfiles productivos en el hato).
- Aprendizaje por refuerzo: el sistema aprende mediante recompensa o castigo a partir de la interacción con un entorno (Sharma Shivangi et al., 2023).

La plataforma Python ha tomado especial relevancia para el desarrollo de proyectos de ciencia de datos, gracias a la amplia disponibilidad de bibliotecas como NumPy, Pandas, scikit-learn, TensorFlow y Keras, que facilitan el flujo completo de análisis: recolección, limpieza, modelado y validación (Kane Frank, 2017). De igual manera, otras referencias como Aurén Géron (2023) subrayan la importancia de buenas prácticas de ingeniería de características, particionado de datos y optimización de hiperparámetros para lograr modelos robustos.

La ganadería de precisión surge como una respuesta a la necesidad de optimizar los recursos productivos y mejorar la eficiencia de los sistemas pecuarios, incorporando tecnologías digitales para la toma de decisiones en tiempo real (Morales-Cardoso et al., 2020a). Tal proceso se enmarca en la denominada transformación digital del sector agropecuario, respaldada por la adopción de sensores, herramientas de Big Data, sistemas de trazabilidad y analítica avanzada (Brum Luciano Moraes da Luz et al., 2019).

En la ganadería de bovinos para leche, la transformación digital ha mostrado resultados positivos al permitir la automatización de tareas como la medición de la producción, el monitoreo del bienestar animal y la detección temprana de enfermedades (Dongre & Gandhi, 2016; Sharma Shivangi et al., 2023). Este enfoque, también llamado agricultura 4.0, aprovecha el análisis masivo de datos para generar alertas y recomendaciones, incrementando la competitividad y la sostenibilidad de la producción (Sarkar et al., 2018; Zea et al., 2022).

Sin embargo, la implementación de estas herramientas exige nuevas competencias en los profesionales del sector. Jeremias Lachman & Andres lopez. (2018) señalan que la formación interdisciplinar, abarcando ciencia de datos, inteligencia artificial y manejo de tecnologías de la información, se vuelve esencial para potenciar la ganadería de precisión. Esto implica la necesidad de un enfoque integral que combine conocimientos de zootecnia, estadística y programación (Brum Luciano Moraes da Luz et al., 2019).

En la producción bovina, los modelos estadísticos y matemáticos cumplen un papel fundamental para comprender y predecir fenómenos biológicos. Uno de los ejemplos más tradicionales es el modelo BLUP (Best Linear Unbiased Prediction), ampliamente utilizado para la estimación del valor genético de los animales a partir de información fenotípica y genealógica (Dijkinga, 2023; Flores M Julia et al., 2006). No obstante, estudios recientes demuestran que técnicas de minería de datos, redes bayesianas y algoritmos de inteligencia artificial pueden complementar o incluso anticipar las decisiones tomadas con BLUP (Perdigón Llanes & González Benítez, 2022).

Por otra parte, la curva de lactancia es un elemento central en la planeación y predicción productiva. Modelos matemáticos como los propuestos por Wood y Brody permiten describir la evolución de la producción láctea en función del tiempo, determinando fases de ascenso, pico máximo de lactancia y la posterior disminución (Fernández Chuairey et al., 2017). Variables como la raza, la nutrición, el estrés calórico y las condiciones ambientales influyen de manera significativa en la forma de la curva y, por ende, en la eficiencia productiva del hato (Chaturvedi Shailesh et al., 2013; Morales-Cardoso et al., 2020a).

Las condiciones tropicales añaden complejidad, dada la variabilidad en la disponibilidad de forrajes, las altas temperaturas y la humedad relativa (Zea et al., 2022). En ese contexto, la

selección de razas o cruzamientos (como Holstein × Gyr) cobra relevancia para equilibrar la productividad y la adaptabilidad a estrés térmico (Morales-Cardoso et al., 2020a). Así mismo, se ha implementado la transferencia de embriones y la selección genética de donadoras puras Gyr para reforzar la resiliencia en el trópico bajo (Dijkinga, 2023).

Aplicaciones de Aprendizaje Automático en Ganadería de Leche

Árboles de Decisión y Bosques Aleatorios (Random Forest)

Constituyen una técnica de aprendizaje automático supervisado ampliamente utilizada en el ámbito agropecuario, debido a que su estructura jerárquica resulta sencilla de interpretar y adaptar a la dinámica de la producción animal (Morales-Cardoso et al., 2020a). Un árbol de decisión descompone el espacio de atributos (características del animal, datos productivos, eventos sanitarios, etc.) mediante reglas sucesivas, hasta llegar a una clasificación final o a una predicción, dependiendo de la variable objetivo (Dongre & Gandhi, 2016).

Aplicaciones en ganadería de leche:

- *Identificación de vacas potencialmente descartables:* considerando variables productivas (producción acumulada, días en lactancia) y reproductivas (abortos, intervalo entre partos).
- *Diagnóstico temprano de enfermedades como la mastitis:* comparando métricas de salud con variables de producción para definir umbrales de alerta (Estrada Carvaja Verny et al., 2019).
- *Definición de estrategias de alimentación y suplementación individualizada:* con base en características como días en producción y condición corporal.

Una extensión de los árboles de decisión es el Random Forest, que consiste en un conjunto o bosque de árboles construidos sobre diferentes muestras aleatorias del conjunto de

datos, combinando posteriormente sus predicciones (bagging o bootstrap aggregation). Esta técnica ofrece las siguientes ventajas (Aurén Géron, 2023; Kane Frank, 2017):

- *Mejor precisión:* al promediar la salida de múltiples árboles, se reducen el sobreajuste (overfitting) y la varianza.
- *Robustez:* maneja eficientemente datos con ruido y variables irrelevantes, ya que cada árbol toma una muestra aleatoria de características.
- *Importancia de características:* el algoritmo estima qué variables son más determinantes, facilitando el entendimiento del modelo (p. ej., producción acumulada, número de partos, edad de la vaca, etc.).

En el contexto de la ganadería de leche tropical, el uso de Random Forest puede ayudar a priorizar aquellas variables ambientales (temperatura, humedad relativa), genéticas (raza, genealogía) o productivas (pico de lactancia, servicios por concepción) que tengan un mayor impacto en el rendimiento del hato (Morales-Cardoso et al., 2020a). Esto resulta especialmente relevante en sistemas con cruzamientos Holstein × Gyr, donde se busca optimizar simultáneamente la adaptación al trópico y la productividad (Zea et al., 2022).

Algoritmos de Clasificación y Agrupamiento No Supervisado

En muchas situaciones, la información de la cual se dispone no está etiquetada, o bien se pretende explorar la estructura interna de los datos para descubrir patrones y grupos de individuos con características similares (Sharma Shivangi et al., 2023). En este caso, se aplican algoritmos no supervisados, cuyos principales objetivos son:

- *Agrupamiento (clustering):* separar el conjunto de datos en grupos o clústeres que tengan alta homogeneidad interna y alta heterogeneidad externa (Spositto Osvaldo et al., 2020).

- *Reducción de dimensionalidad:* simplificar la información, extrayendo factores o componentes principales que capturen la mayor varianza de los datos.

Algoritmos de Clustering: K-means

Es uno de los algoritmos más utilizados, que particiona el conjunto de datos en k grupos basándose en la minimización de distancias internas (Dongre & Gandhi, 2016). En ganadería, se ha empleado para identificar subpoblaciones de vacas con comportamiento productivo o reproductivo similar (Zea et al., 2022).

Estos métodos han demostrado su eficacia para segmentar grupos de animales en función de variables como peso al nacer, edad de la madre, genética del padre, número de servicios, etc., facilitando la elaboración de planes de manejo más específicos (Morales-Cardoso et al., 2020a). En Python, librerías como scikit-learn ofrecen implementaciones robustas de estos algoritmos, así como herramientas de validación y visualización (Kane Frank, 2017).

Clasificación No Supervisada para Detección de Patrones

Otra vertiente de los algoritmos no supervisados radica en la detección de anomalías o patrones inusuales en la producción de leche, índice de grasa, comportamiento reproductivo, entre otros (Estrada Carvaja Verny et al., 2019). Identificar individuos que se desvíen significativamente de la tendencia general puede ayudar a realizar diagnósticos tempranos de problemas sanitarios, nutricionales o de adaptación al trópico (Fernando Johann et al., 2023).

Además, los algoritmos no supervisados facilitan la caracterización de subgrupos de alto rendimiento (alto pico de lactancia, mayor persistencia en producción, etc.), diferenciándolos de subgrupos con menor adaptación o rentabilidad, que potencialmente podrían ser descartados o requerir manejos específicos (Morales-Cardoso et al., 2020a).

La capacidad explicativa y predictiva de los árboles de decisión y bosques aleatorios permite priorizar las variables más influyentes en la productividad y, al mismo tiempo, brindar interpretaciones claras para los tomadores de decisiones (Brum Luciano Moraes da Luz et al., 2019). Por su parte, los algoritmos de agrupamiento no supervisado amplían la comprensión de la dinámica del hato, al segmentar y analizar perfiles de producción y eficiencia reproductiva sin requerir datos previamente etiquetados (Spósitto Osvaldo et al., 2019).

La aplicación de la ciencia de datos en la ganadería lechera no solo apunta a incrementar la producción, sino también a optimizar el uso de recursos. En el trópico bajo, la capacidad de los algoritmos para gestionar la variabilidad climática y ofrecer recomendaciones de manejo personalizadas se traduce en una mayor resiliencia. El mejoramiento genético a través de la identificación de animales superiores en términos de productividad y adaptación al calor, complementado con las metodologías de IA, contribuye a una ganadería más competitiva y sostenible (Morales-Cardoso et al., 2020a).

Los aportes teóricos revisados confirman la relevancia de la ciencia de datos y el aprendizaje automático en la evolución de la ganadería de leche tropical. El tránsito de un manejo empírico hacia un manejo analítico se ve fortalecido por la adopción de metodologías de ML que aprovechan al máximo los datos generados en campo (Morales-Cardoso et al., 2020a; Perdigón Llanes & González Benítez, 2022). Además, la integración de estas tecnologías con sistemas de BI y data warehouse ofrece una vía para la toma de decisiones basadas en evidencia, crucial en un entorno de alta competitividad y variabilidad climática (Brum Luciano Moraes da Luz et al., 2019).

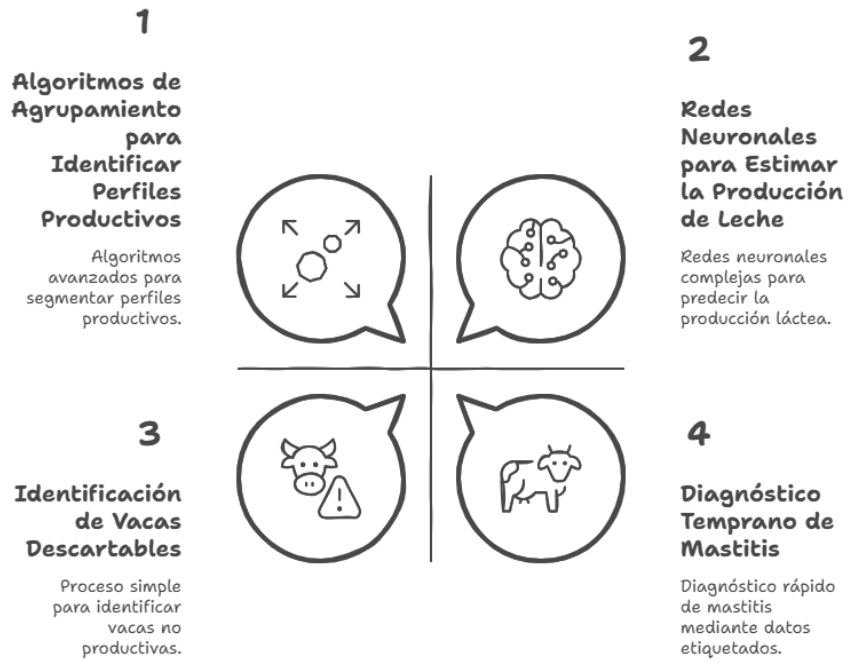
No obstante, se advierte una brecha del conocimiento en cuanto a la sistematización de los datos históricos y la aplicación práctica de algoritmos de última generación en los hatos

tropicales, especialmente en lo referente a cruzamientos Holstein × Gyr, transferencia de embriones y la validación rigurosa de modelos predictivos (Zea et al., 2022). El presente estudio busca precisamente abordar esta brecha, demostrando cómo el uso de metodologías analíticas basadas en ML puede anticipar comportamientos productivos y mejorar la eficiencia y sostenibilidad de sistemas ganaderos en condiciones de trópico bajo.

En ese sentido, el marco teórico aporta la base conceptual y metodológica para la construcción e implementación de modelos de predicción, justificando la importancia de estudiar la curva de lactancia, la evaluación genética, la detección temprana de enfermedades y el uso de técnicas específicas (redes neuronales, árboles de decisión, algoritmos de agrupamiento). Así, se consolida la pertinencia de integrar técnicas de ciencia de datos con el conocimiento zootécnico, garantizando tanto la validez científica como la aplicabilidad práctica de los resultados esperados.

Figura 2

Aprendizaje Automático Aplicado a la Ganadería de Leche



En línea con lo expuesto, la Figura 2 sintetiza las principales aplicaciones del aprendizaje automático en la ganadería de leche, destacando su papel en la identificación de perfiles productivos, la estimación de la producción, la selección de vacas descartables y el diagnóstico temprano de enfermedades. Esta representación visual resume de manera práctica los enfoques metodológicos revisados y refuerza la pertinencia de aplicar técnicas de ciencia de datos en contextos de producción bovina bajo condiciones tropicales.

Marco conceptual

El análisis productivo y genético en ganadería bovina lechera requiere de la comprensión e integración de diversos conceptos que permiten interpretar con mayor profundidad las variaciones en el desempeño de los animales. Este estudio se fundamenta en un enfoque basado en ciencia de datos, genética aplicada y evaluación productiva individual, con énfasis en vacas lecheras del trópico bajo colombiano; algunos de los conceptos clave abordados en el presente trabajo son:

- Ciencia de Datos: es una disciplina que combina estadística, programación y conocimientos del dominio para analizar grandes volúmenes de información (Dongre & Gandhi, 2016). En el sector agropecuario, esta disciplina permite recopilar y examinar registros productivos y reproductivos de manera sistemática, identificando patrones y tendencias que mejoran la toma de decisiones. En este trabajo de grado, la ciencia de datos aporta métodos de análisis para procesar la información de lactancias y evaluar la eficiencia del hato lechero.
- Aprendizaje Automático (Machine Learning - ML): Es una rama de la inteligencia artificial enfocada en crear algoritmos capaces de detectar patrones y hacer predicciones con base en datos históricos (Sarkar et al., 2018). En ganadería, estos algoritmos ayudan a anticipar la producción de leche, clasificar animales según su desempeño y optimizar el manejo

reproductivo. Para este proyecto, se aplican técnicas de aprendizaje supervisado (Random Forest) y no supervisado (K-means), con el fin de estimar y agrupar la eficiencia productiva de las vacas.

- Ganadería de Precisión: Consiste en manejar cada animal de forma individual, a partir de registros específicos sobre su producción, salud y nutrición (Morales-Cardoso et al., 2020a). Con dispositivos y sistemas de medición, el productor puede tomar decisiones más acertadas y oportunas, mejorando la rentabilidad y el bienestar animal. En el contexto de este trabajo, la ganadería de precisión sustenta la recolección detallada de datos y el uso de herramientas analíticas para optimizar la eficiencia de las vacas lecheras.

- Agricultura 4.0: Integra tecnologías digitales (Big Data, sensores, robótica, Internet de las Cosas) en los procesos productivos agropecuarios (Sarkar et al., 2018). En la ganadería lechera, esto se traduce en sensores de ordeño, softwares de gestión y análisis de grandes bases de datos. Gracias a estas herramientas, se facilita el control de variables ambientales, nutricionales y reproductivas, posibilitando la adopción de modelos predictivos que potencien la productividad y la sostenibilidad del hato.

- Transformación Digital en el Sector Agropecuario: Implica la adopción de soluciones tecnológicas que modernizan la producción en el campo (Brum Luciano Moraes da Luz et al., 2019). Esto abarca desde sistemas de registro electrónico de la producción hasta plataformas de análisis en la nube. En ganadería de leche, la transformación digital acelera la recopilación de datos sobre cada vaca, mejora la trazabilidad y permite implementar metodologías de ciencia de datos y aprendizaje automático de manera eficiente.

- Modelos Estadísticos aplicados a la Ganadería (BLUP y otros): Los modelos estadísticos clásicos, como el Best Linear Unbiased Prediction (BLUP), se han usado

tradicionalmente para estimar el valor genético de los animales (Dijkinga, 2023; Flores M Julia et al., 2006). Estos métodos combinan información fenotípica (características productivas) y genealógica (pedigrí) para predecir la capacidad reproductiva y lechera de cada bovino. En la actualidad, técnicas de ciencia de datos amplían y, en ocasiones, superan el alcance de los modelos tradicionales, proporcionando análisis más detallados y versátiles en entornos con mayor variabilidad (Zea et al., 2022).

- Lactancia Bovina: Es el periodo en el que una vaca produce leche tras un parto, con una duración usual de 270 a 305 días, aunque puede variar según la raza y el manejo (Chaturvedi Shailesh et al., 2013). A lo largo de su vida, la vaca tiene varias lactancias, cada una con distintos niveles de producción. En este estudio, se analizan los datos de cada lactancia para conocer la eficiencia y rentabilidad de los animales, teniendo en cuenta factores como la edad de la vaca, la nutrición y las condiciones ambientales.

- Intervalo entre Partos (IEP): Es el lapso entre un parto y el siguiente. Un intervalo cercano a 12 o 13 meses se considera ideal, pues así la vaca mantiene una producción estable y reduce costos improductivos (Flores M Julia et al., 2006). Si el IEP se alarga, la vaca pasa más tiempo sin producir a plena capacidad, lo cual afecta la rentabilidad del sistema. En este proyecto, el IEP es un indicador clave que se relaciona con la eficiencia reproductiva y productiva de cada vaca.

- Indicador Leche/día / IEP: Este indicador integra la producción láctea con la eficiencia reproductiva. Se obtiene dividiendo la leche producida por día (o durante la lactancia) entre los días que dura el intervalo entre partos (Morales-Cardoso et al., 2020a). Un valor elevado evidencia vacas que generan más leche en un menor lapso reproductivo, lo cual se

traduce en mayor rentabilidad. En cambio, un valor bajo puede indicar problemas de producción, fertilidad o manejo, brindando señales tempranas para ajustes o descartes.

- Tipos raciales: Taurus e indicus; en bovinos, se distinguen dos grandes grupos:
  - Taurus (por ej. Holstein, Jersey): Razas típicamente europeas o de climas templados, con alta productividad lechera, pero menos tolerancia al calor.
  - Indicus (por ej. Gyr, Brahman): Razas adaptadas a climas cálidos, con mayor resistencia a parásitos y estrés térmico, aunque su nivel productivo suele ser menor (Zea et al., 2022).

Este concepto es relevante para planificar los cruzamientos que equilibren productividad y adaptación al trópico.

- Lechería de Trópico Bajo: Es la actividad de produccion de leche que se desarrolla en zonas de baja altitud (menos de 1.000 msnm), con altas temperaturas y humedad (Morales-Cardoso et al., 2020a). A diferencia de la lechería de trópico alto, donde el clima es más fresco, en el trópico bajo las vacas enfrentan mayor estrés calórico y un forraje con variaciones estacionales más marcadas. Estas condiciones exigen razas o cruces más resistentes y estrategias de manejo enfocadas en reducir el impacto del calor y las enfermedades típicas de climas cálidos.
- Variables Productivas y Reproductivas Clave: En la ganadería lechera, variables como la producción diaria de leche, los días en lactancia, los servicios por concepción, la tasa de abortos, la condición corporal y la edad al primer parto son esenciales para valorar el rendimiento de cada vaca (Estrada Carvaja Verny et al., 2019). Este conjunto de indicadores permite diagnosticar ineficiencias y orientar decisiones de manejo (nutrición, reproducción, sanidad), siendo la base de los algoritmos de análisis empleados en esta investigación.

- Cruzamientos Holstein × Gyr: El cruce Holstein × Gyr busca combinar la alta productividad de la Holstein (raza taurina) con la resistencia al calor y enfermedades de la Gyr (raza indicus). Este cruce es especialmente valioso en el trópico bajo, ya que aprovecha la heterosis (efecto híbrido) para producir vacas más rentables y adaptadas (Zea et al., 2022). La investigación analiza los datos de vacas que provienen de este cruzamiento, evaluando su desempeño productivo y reproductivo.

- Transferencia de Embriones y Selección Genética: La transferencia de embriones acelera el mejoramiento genético al implantar embriones de vacas donadoras élite en vacas receptoras. En el trópico, se suelen usar donadoras Gyr puras combinadas con toros Holstein de alto valor genético (Dijkinga, 2023). De este modo, se incrementa la aparición de animales con rasgos deseables (producción y adaptación). Para este trabajo, comprender la dinámica de la transferencia de embriones ayuda a contextualizar la variabilidad genética y los resultados productivos en el hato además sus objetivos de selección y mejoramiento de sus animales.

- Árboles de Decisión: Son métodos de aprendizaje supervisado donde se generan reglas “si...entonces” para clasificar o predecir un resultado (Dongre & Gandhi, 2016). Cada pregunta sucesiva en el árbol segmenta los datos según la variable que mejor separa los casos (por ejemplo, la producción lechera o el número de partos). Este enfoque es intuitivo, facilita la explicación de las decisiones tomadas y resulta útil para la selección de vacas descartables o para la priorización de manejos reproductivos.

- Random Forest: Consiste en un conjunto (bosque) de árboles de decisión entrenados con diferentes muestras de datos (Morales-Cardoso et al., 2020a). Sus principales ventajas son:

- Menor sobreajuste que un solo árbol, ya que al combinar múltiples modelos se reduce el sesgo y la varianza.
- Evaluación de importancia de variables, lo que permite conocer qué factores inciden más en la productividad o en la probabilidad de descarte.
- Robustez ante datos ruidosos o incompletos, algo común en registros ganaderos.

En este estudio, Random Forest se utiliza para predecir la eficiencia productiva y clasificar a las vacas y sus progenitores según sus resultados productivos y reproductivos.

- K-means: Es un algoritmo no supervisado de agrupamiento que busca dividir el conjunto de datos en k grupos de manera que cada vaca quede en el clúster con el centroide más cercano (Dongre & Gandhi, 2016; Spositto Osvaldo et al., 2020). El proceso de iteraciones ajusta los centroides hasta estabilizar la clasificación. En ganadería lechera, K-means permite identificar subgrupos de vacas con perfiles productivos semejantes (por ejemplo, alta producción, bajo número de servicios por concepción) y trazar estrategias de manejo específicas para cada grupo.

Marco Normativo

El desarrollo e implementación de modelos predictivos aplicados al análisis genético y productivo de bovinos lecheros se sustenta en un entorno regulatorio que cubre aspectos sanitarios, trazabilidad, bienestar animal, bioética, biotecnología, ciencia de datos y uso responsable de tecnologías emergentes como la inteligencia artificial. Este marco legal e institucional, vigente tanto en el contexto colombiano como a nivel internacional, respalda y orienta el empleo de herramientas analíticas avanzadas en la producción ganadera.

Colombia

En Colombia, el Instituto Colombiano Agropecuario (ICA) es la entidad encargada de reglamentar la trazabilidad y sanidad animal. La Resolución 2341 de 2007, que crea el Sistema de Identificación e Información de Ganado Bovino (SINIGAN), establece la identificación individual del ganado y sirve de base para la gestión de datos productivos y genéticos. De igual forma, la Resolución 7520 de 2007 regula las Buenas Prácticas Ganaderas (BPG), exigiendo registros técnicos sistemáticos para la certificación de hatos productores de leche (Instituto Colombiano Agropecuario - ICA, 2025).

La Ley 1581 de 2012, sobre protección de datos personales, aplica al manejo de bases de datos en ganadería que incluyan información que pueda vincularse directa o indirectamente con personas o productores. Asimismo, la Ley 1774 de 2016, al reconocer a los animales como seres sintientes, establece que cualquier estrategia técnica, incluyendo programas de mejoramiento genético o selección basada en IA, debe alinearse con principios de bienestar animal.

Normativas y Marcos Regulatorios Internacionales

- Brasil: Ha desarrollado un sólido sistema para el mejoramiento genético bovino a través del Programa Nacional de Melhoramento Genético de Bovinos de Leite (PNMGL), coordinado por la Empresa Brasileña de Investigación Agropecuaria (Embrapa). Este programa, articulado con el Ministerio de Agricultura, Ganadería y Abastecimiento (MAPA), regula la evaluación genética y el uso de toros y vacas reproductoras mediante índices oficiales, como la diferencia esperada en la progenie (DEP) (Empresa Brasileira de Pesquisa Agropecuária (Embrapa), 2025). Asimismo, integra la normativa sobre trazabilidad y biotecnologías aplicadas, entre ellas la inseminación artificial y la transferencia de embriones, sujetas al Sistema Brasileño de Identificación y Certificación de Origen Bovina y Bufalino.

- Unión Europea: El Reglamento (UE) 2016/1012 regula la cría y comercialización de animales reproductores, estableciendo requisitos para la inscripción de programas de mejora genética y fomentando la estandarización de registros, la transparencia en las evaluaciones genéticas y la realización de auditorías (Reglamento (UE) 2016/1012 Del Parlamento Europeo y Del Consejo, de 8 de Junio de 2016, 2016). A través de iniciativas como SmartAgriHubs y programas como Horizon Europe, la UE financia y promueve la integración de IA y ciencia de datos en procesos de selección genética, trazabilidad e innovación agropecuaria, asegurando la compatibilidad con principios éticos y medioambientales.
- Estados Unidos: El Council on Dairy Cattle Breeding (CDCB) regula y gestiona el sistema nacional de evaluación genética. Bajo estándares unificados, se promueve el uso de información de genealogía, rendimiento, salud y genómica. Para la selección de animales, se emplean índices como el "Net Merit" (NM$) y "El Lifetime Net Merit", de uso obligatorio en las asociaciones ganaderas acreditadas (Council on Dairy Cattle Breeding, 2025). El National Animal Germplasm Program del Departamento de Agricultura de los Estados Unidos respalda la conservación y utilización estratégica del recurso genético bovino, además de regular las tecnologías reproductivas. En cuanto a la gestión de datos e IA, las políticas del National Institute of Food and Agriculture (NIFA) fijan pautas para el uso ético, equitativo y transparente de modelos predictivos aplicados al sector agro.
- Canadá: El Canadian Dairy Network (CDN) y Lactanet han instaurado uno de los sistemas más integrados del mundo para la evaluación genética de bovinos. Dichos programas, regidos por Agriculture and Agri-Food Canada (AAFC), fomentan la inclusión de indicadores de salud, eficiencia y longevidad, a la vez que se regulan las aplicaciones de IA para mejorar la precisión de la predicción genética (Lactanet Canada, 2025).

- Nueva Zelanda y Australia: El Animal Evaluation (NZAEL) y en Australia, el Australian Breeding Values (ABVs), integran la evaluación genética con la producción en sistemas pastoriles y el bienestar animal como elementos centrales del control reproductivo. Ambos países regulan y promueven el uso de sistemas automatizados para analizar la producción láctea y la genética, a través de agencias nacionales de innovación agrícola.

Análisis Comparado y Perspectiva para Colombia

La revisión normativa comparada evidencia que la ciencia de datos y el aprendizaje automático se están incorporando de forma progresiva en los programas de mejoramiento genético bovino a nivel internacional. Estados Unidos, Canadá, Brasil, Nueva Zelanda y los países miembros de la Unión Europea han implantado marcos regulatorios sólidos y sistemas institucionales que incentivan, auditan y financian el uso de tecnologías avanzadas para la evaluación genética, la selección de animales de alta calidad y la toma de decisiones basada en evidencia.

El machine learning, aplicado al análisis multivariado de indicadores productivos y reproductivos, supone un avance natural y necesario frente a los métodos tradicionales de predicción genética. Su capacidad de modelar relaciones no lineales, identificar patrones complejos y trabajar con bases de datos incompletas o desbalanceadas lo convierte en una herramienta estratégica para los sistemas ganaderos que buscan incrementar la eficiencia, la precisión y la sostenibilidad. Al integrarse con sistemas de trazabilidad, biotecnologías reproductivas y plataformas digitales, la ciencia de datos deja de ser un recurso aislado para convertirse en el eje estructural de la ganadería del futuro.

En contraste, en Colombia el marco regulatorio y la adopción sistemática de estas tecnologías en el sector ganadero aún se encuentran en una fase inicial. Pese a la existencia de

lineamientos clave como los programas SINIGAN, BPG y las leyes de protección de datos, no se ha consolidado un marco legal específico que regule o promueva de manera directa la aplicación de modelos de inteligencia artificial, sistemas de evaluación genética informatizados o plataformas de análisis predictivo en el ámbito pecuario.

Asimismo, se identifican brechas institucionales y técnicas entre productores, gremios ganaderos, entidades públicas y universidades, lo que dificulta la creación de bases de datos estandarizadas, abiertas y de calidad para el entrenamiento de modelos confiables. Esto conlleva una reducida disponibilidad de información individualizada y un limitado aprovechamiento de herramientas tecnológicas de alto potencial transformador. Colombia, por otra parte, tampoco cuenta con programas nacionales de evaluación genética obligatorios y fiscalizados, a diferencia de lo que ocurre en Brasil o en la Unión Europea.

No obstante, el país posee condiciones excepcionales para consolidarse como un referente regional en la aplicación de ciencia de datos al mejoramiento genético bovino. Su diversidad genética, los sistemas de producción en entornos y características diversos, la creciente digitalización del sector agropecuario y la formación emergente de capital humano en ciencia de datos configuran un entorno propicio para el desarrollo de un marco regulatorio moderno, interoperable y fundamentado en la evidencia científica. Lograr este propósito demanda un compromiso conjunto entre el Estado, la academia, los gremios ganaderos y los desarrolladores tecnológicos, a fin de articular políticas públicas orientadas a la transformación digital del sector, con énfasis en la equidad, la ética algorítmica y la sostenibilidad.

Metodología

La presente sección describe la estructura metodológica que se empleará en este proyecto de grado, tomando como base el modelo CRISP-DM (Cross Industry Standard Process for Data Mining). Esta metodología se caracteriza por proporcionar una secuencia estructurada de fases, abarcando desde la comprensión de los objetivos y la naturaleza de los datos hasta la generación y evaluación de modelos analíticos (Chapman Pete et al., 2000). Su enfoque iterativo y flexible permite adaptar las estrategias conforme se profundiza en el entendimiento del problema y de la información disponible.

En el contexto de este proyecto, la adopción de CRISP-DM facilita una visión integral de cada paso, pues se parte de la identificación de las necesidades y preguntas de investigación, pasando por el escrutinio minucioso de la calidad de los datos relacionados con lactancia de las vacas del hato en cuestión, hasta llegar a la construcción y evaluación de modelos analíticos que sirvan para describir o predecir comportamientos relevantes en dicha población de estudio. La iteratividad del modelo ofrece la posibilidad de volver sobre las fases previas ante hallazgos inesperados o ajustes necesarios, asegurando así la trazabilidad de cada decisión metodológica. De esta forma, se procura que los resultados finales no solo respondan a los objetivos planteados, sino que también aporten recomendaciones fundamentadas para la mejora de la práctica y la generación de nuevo conocimiento en el ámbito de la produccion de leche en un hato del trópico bajo colombiano.

A continuación, se presenta una tabla con las fases principales del modelo CRISP-DM y sus conceptos esenciales:

Tabla 2

Fases del Modelo CRISP-DM

Fase del modelo CRISP-DM	Concepto clave
1. Entendimiento del negocio	Entendimiento del problema desde una perspectiva operativa o productiva; definición de objetivos.
2. Comprensión de los datos	Revisión, exploración inicial, evaluación de calidad y relaciones entre las variables del conjunto.
3. Preparación de los datos	Limpieza, transformación, integración y estructuración del conjunto de datos para su análisis.
4. Modelado	Selección y entrenamiento de algoritmos para segmentación o predicción según el objetivo del estudio.
5. Evaluación	Validación del modelo con respecto a los objetivos definidos; análisis de resultados y patrones.
6. Implementación	Generación de productos analíticos, documentación y utilidad práctica de los resultados.

Nota. Adaptado de la estructura del modelo CRISP-DM ampliamente referenciado en literatura sobre ciencia de datos *(Chapman Pete et al., 2000).*

Entendimiento del Negocio

En los sistemas de producción ganadera del trópico bajo colombiano, una problemática constante radica en la baja eficiencia reproductiva y productiva de los hatos, condicionada principalmente por factores genéticos, manejo reproductivo tradicional y condiciones ambientales desafiantes. La diversidad genética, resultado de décadas de cruzamientos entre razas adaptadas y especializadas, genera complejidad al intentar identificar claramente qué combinaciones genéticas o qué métodos reproductivos favorecen una mayor productividad y rentabilidad económica. Esto representa una limitación crítica en la gestión operativa y

estratégica de los sistemas de producción lechera, ya que dificulta la toma de decisiones fundamentadas en datos precisos y oportunos para mejorar la productividad y sostenibilidad del negocio.

Adicionalmente, aunque las técnicas reproductivas avanzadas como la transferencia de embriones (TE) permiten acelerar el mejoramiento genético, existe incertidumbre respecto al impacto real de estos métodos en términos de productividad comparada con métodos tradicionales. Los productores necesitan conocer con precisión qué animales y qué estrategias reproductivas representan inversiones seguras y rentables. En este contexto, la selección informada de animales y métodos reproductivos adecuados se convierte en un factor decisivo para el éxito económico y productivo de los hatos lecheros.

Objetivos del Estudio y Preguntas de Investigación

Para enfrentar la problemática descrita, se formuló como objetivo general que busca evaluar el impacto de la genética y los grupos raciales sobre la productividad lechera en vacas de hatos ubicados en el trópico bajo colombiano, aplicando técnicas analíticas integradas de segmentación (K-means) y predicción (Random Forest).

Contexto y relevancia: La selección genética es un componente crucial para mejorar la productividad de los sistemas ganaderos. La adecuada elección de toros y vacas con los mejores indicadores productivos resulta fundamental para optimizar aspectos clave como la producción de leche, la eficiencia reproductiva y la adaptación al entorno. En regiones como el trópico bajo colombiano, caracterizadas por condiciones climáticas adversas, este proceso adquiere mayor relevancia, al determinar la rentabilidad y sostenibilidad de las explotaciones ganaderas.

El uso de biotecnologías reproductivas avanzadas, como la transferencia de embriones, ha tomado gran relevancia debido a que facilita acelerar la multiplicación de individuos

genéticamente superiores, acortando los ciclos generacionales. Esto permite aprovechar rápidamente las ventajas genéticas, incrementando la productividad del hato y favoreciendo la eficiencia económica del sistema productivo (Ríos-Utrera et al., 2012).

Este análisis es pertinente dado que aborda directamente la necesidad de identificar patrones y diferencias claves en la productividad asociada con variables genéticas y reproductivas, permitiendo esclarecer interrogantes que actualmente limitan la toma efectiva de decisiones estratégicas y operativas en los hatos lecheros.

Comprensión de los Datos

La fuente de datos empleada en este estudio comprende registros detallados de lactancias provenientes de un hato bovino lechero ubicado en el trópico bajo colombiano, recopilados entre los años 2010 y 2024. La base inicial incluyó 8.438 registros de lactancias y 40 variables relacionadas con la lactancia cursada, las cuales abarcan aspectos productivos, reproductivos, sanitarios y genéticos. Las variables de mayor relevancia identificadas fueron: intervalo entre partos (IEP), días en leche, producción acumulada y ajustada a 305 días, incidencia de mastitis y cojeras, así como variables genéticas relativas a la proporción racial Taurus e Indicus y el método reproductivo empleado (natural o transferencia de embriones). La calidad y fiabilidad de estos datos radican en su recolección sistemática diaria en el hato mediante el uso de plataformas tecnológicas especializadas, garantizando así precisión y consistencia.

Procesos de Recolección y Revisión

El proceso de entendimiento de los datos implicó varias etapas secuenciales que permitieron consolidar, depurar y seleccionar las variables esenciales para el análisis. Inicialmente se llevó a cabo una integración preliminar de diversas bases de datos disponibles en la finca, con la intención de identificar las variables relevantes para el estudio.

Posteriormente, se profundizó en un análisis exploratorio exhaustivo a nivel individual de cada lactancia. Este paso permitió reconocer patrones productivos y reproductivos claros, identificar variables críticas que requerían ajustes y determinar aquellas que aportaban mayor valor para los objetivos del estudio.

Finalmente, se realizó una integración definitiva de todos los datos relevantes, estructurando un conjunto robusto que combinaba aspectos genéticos, productivos y reproductivos. Este proceso de revisión rigurosa implicó procedimientos para asegurar la calidad del conjunto final de datos, mediante revisiones exhaustivas, detección y tratamiento de inconsistencias, y eliminación de registros duplicados, asegurando así un dataset consistente y válido para el análisis estadístico y modelado predictivo posterior.

Verificación de la Calidad de los Datos

La verificación y depuración de los datos fue un proceso crítico, centrado en el tratamiento riguroso de valores faltantes, valores atípicos e inconsistencias en formatos. Se identificaron valores faltantes en variables esenciales como intervalo entre partos (IEP), días hasta el primer servicio, días hasta la concepción y producción láctea. Estos valores fueron imputados utilizando métodos estadísticos apropiados (mediana o media, según la distribución específica de cada variable).

Asimismo, mediante técnicas gráficas y estadísticas se identificaron valores atípicos, los cuales fueron cuidadosamente analizados para determinar su inclusión o exclusión justificada, asegurando que los datos finales representaran adecuadamente la realidad productiva del hato y fueran robustos para el análisis estadístico y la modelación predictiva posterior. El conjunto de datos final obtenido después de estos procedimientos garantiza la integridad, fiabilidad y representatividad necesarias para sustentar los resultados del estudio.

Análisis Exploratorio Inicial

Durante el análisis exploratorio inicial se utilizaron técnicas estadísticas descriptivas acompañadas de visualizaciones gráficas (histogramas, boxplots, gráficos de dispersión y matrices de correlación). Los resultados de este análisis fueron fundamentales para seleccionar las variables finales del estudio y ajustar aquellas que requerían correcciones metodológicas, estableciendo así una base sólida para los análisis posteriores.

Análisis estadístico descriptivo: A continuación, se presenta un resumen estructurado de los hallazgos obtenidos en el análisis estadístico descriptivo de las variables seleccionadas. Estas variables fueron priorizadas por su relevancia analítica y su contribución directa al cumplimiento de los objetivos del estudio.

Tabla 3

Variables Seleccionadas para el Estudio

Variable	Tipo	Descripción
Leche día/IEP	Numérica	Producción diaria ajustada por intervalo entre partos (variable dependiente)
Grupo racial	Categórica	Composición genética detallada (Taurus vs Indicus en porcentajes)
Grupo racial simplificado	Categórica	Agrupación en cinco clases raciales según proporción de sangre
Tipo	Categórica	Método reproductivo: Embrión o No embrión
Madre genética	Categórica	Identificador de la madre biológica
Padre	Categórica	Identificador del toro que fecundó el embrión
Abuelo materno	Categórica	Identificador del padre de la madre genética
# lact.	Categórica	Número de lactancia del animal

Variable	Tipo	Descripción
Año parto	Categórica	Año en que ocurrió el parto

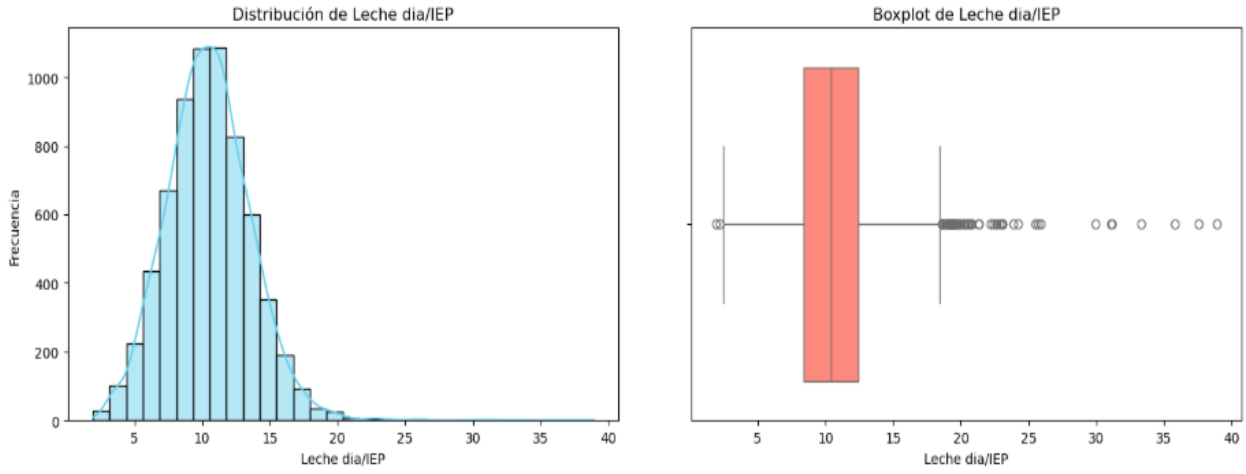
Nota. resume las variables principales empleadas en los modelos y análisis exploratorios desarrollados en el estudio. *Fuente*. Autor.

Hallazgos Relevantes por Variable

- Leche día/IEP: Esta variable fue seleccionada como métrica principal por integrar componentes reproductivos y productivos. Presentó una media de 10.52 kg/día, con valores extremos que alcanzan hasta los 38.91 kg/día. Se observó una distribución asimétrica, con mayor concentración de vacas entre 7 y 12 kg/día. Estos hallazgos permitieron detectar subpoblaciones con alto desempeño productivo y diferenciar perfiles productivos clave para la modelación posterior.

Figura 3

Distribución y Diagrama de Caja de la Variable Leche Día/IEP

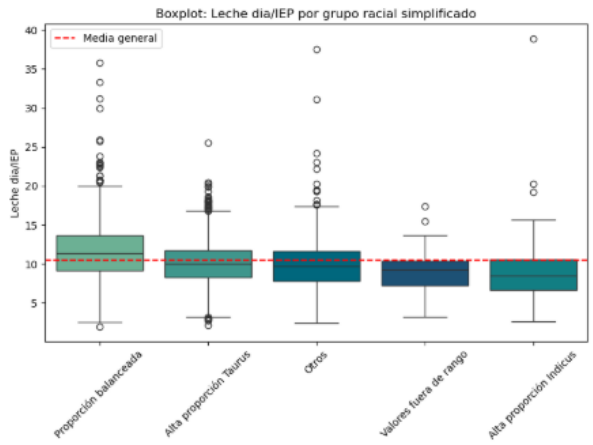


Nota. A la izquierda se muestra el histograma de frecuencia de la variable Leche día/IEP, ajustada mediante una curva de densidad. A la derecha, el diagrama de caja.

- Grupo racial: Aunque esta variable no se utilizó directamente para análisis estadísticos, permitió calcular la proporción genética y construir la variable grupo racial simplificado, que sintetiza la complejidad racial del hato en clases manejables.
- Grupo racial simplificado: Esta variable mostró diferencias notables en productividad. Las vacas con composición 50% Taurus y 50% Indicus (grupo balanceado) evidenciaron el mayor rendimiento (11.37 kg/día). En contraste, las vacas con alta proporción de sangre Indicus presentaron menor desempeño (9.12 kg/día). Esta diferencia no solo es numérica, sino también estadísticamente significativa y relevante en contextos de selección genética. En la figura 4 la línea roja punteada indica la media general de la variable; se observa que las vacas con proporción balanceada (Taurus 50% / Indicus 50%) tienden a presentar una mediana de productividad por encima del promedio general. En contraste, los grupos con alta proporción Indicus presentan menores niveles de producción ajustada; la presencia de múltiples valores atípicos sugiere heterogeneidad interna dentro de cada grupo genético.

Figura 4

Distribución de la Productividad Leche Día/IEP por Grupo Racial Simplificado



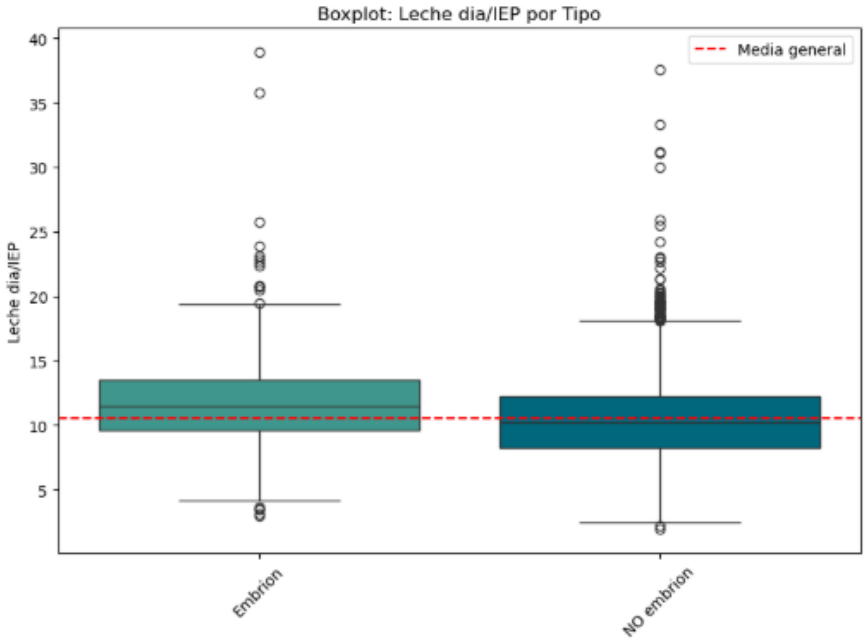
Nota. El diagrama de caja muestra la variabilidad de la variable Leche día/IEP entre los diferentes grupos raciales simplificados.

- Tipo (Embrión vs No embrión): Las vacas tipo embrión superaron

consistentemente a las de origen natural en productividad (11.64 vs 10.33 kg/día). Esta diferencia se mantuvo estable a lo largo de los años y fue especialmente evidente en lactancias medias, reforzando la hipótesis sobre el impacto positivo de la transferencia de embriones. La figura 5 compara la productividad diaria ajustada por intervalo entre partos (Leche día/IEP) entre vacas nacidas por transferencia de embriones y aquellas nacidas mediante métodos reproductivos convencionales. La línea punteada representa la media general. Se observa que ambos grupos presentan una mediana similar, aunque las vacas tipo embrión exhiben una mayor concentración de valores productivos en los rangos superiores, lo cual podría indicar un mayor potencial genético. No obstante, la dispersión y la presencia de valores extremos también es más notoria en este grupo.

Figura 5

Comparación de la Productividad Leche Día/IEP Según Tipo de Nacimiento



Nota. Diagrama de caja de la productividad diaria ajustada por el intervalo entre partos (Leche día/IEP) según su tipo. Fuente.

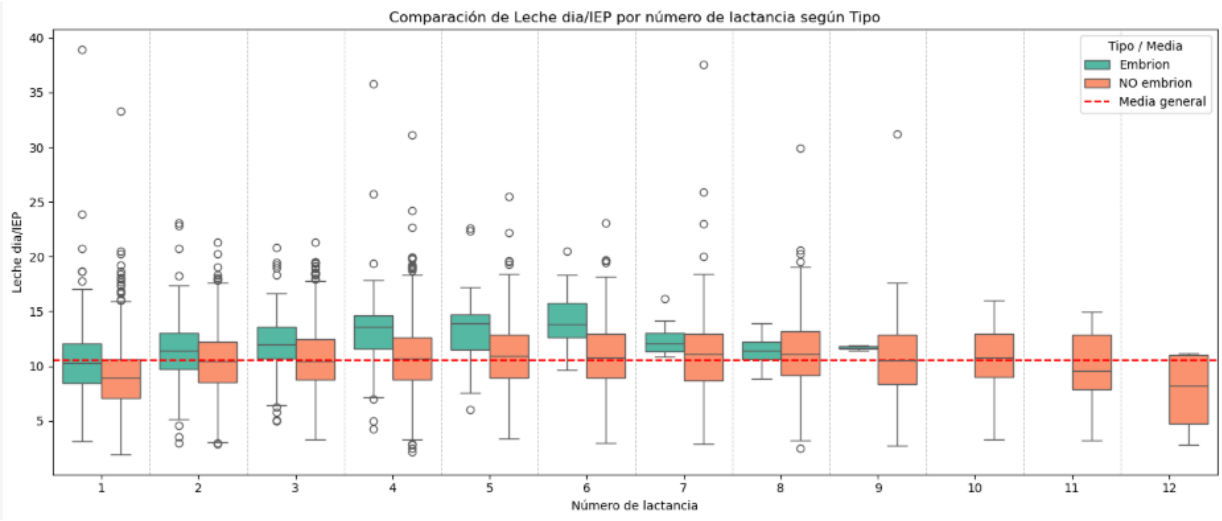
- Madre genética y Padre: Se identificaron individuos que transmiten consistentemente mayores niveles de productividad a su descendencia. En vacas tipo embrión, los análisis mostraron progenitores (padres y madres) con medias significativamente más altas de Leche día/IEP, lo que permitió establecer rankings de eficiencia genética y orientar criterios para la selección de reproductores.
- Abuelo materno: Aunque menos relevante de forma aislada, esta variable permitió verificar consistencia productiva en líneas maternas específicas. Algunos abuelos maternos se asociaron con descendencias de mayor rendimiento, sirviendo como soporte secundario en la validación genealógica.
- # lact.: El número de lactancia mostró una curva productiva característica: las primeras lactancias tienen menor producción, se alcanza un pico entre la tercera y sexta, y posteriormente hay una leve disminución. Este comportamiento valida su inclusión como variable de control y permite comparar eficiencia productiva a lo largo del ciclo vital de la vaca.
- Año parto: La productividad mostró una tendencia creciente en los años más recientes, especialmente en animales tipo embrión. Esta evolución puede atribuirse a mejoras en la selección genética y el manejo técnico del hato. Las diferencias por tipo se acentuaron desde 2016 en adelante.
- Relaciones específicas:
  - En vacas tipo embrión, se identificaron madres y padres genéticos asociados con producciones significativamente superiores.

- El cruce entre análisis por número de lactancia y tipo reproductivo evidenció que, en todas las lactancias, las vacas embrión fueron más productivas, con diferencias marcadas en la tercera a sexta lactancia.
- También se observó que, año a año, la ventaja productiva de las vacas tipo embrión respecto a las No embrión se ha consolidado progresivamente.

La figura 6, ilustra la diferencia entre vacas nacidas por transferencia de embriones (verde) y por método convencional (naranja). La línea punteada roja indica la media general de la productividad; se evidencia que en las primeras lactancias (1 a 5), las vacas tipo embrión tienden a superar consistentemente la productividad media, mientras que esta diferencia se atenúa a partir de la sexta lactancia. Esta visualización sugiere un posible efecto del tipo reproductivo en la expresión temprana del potencial genético.

Figura 6

Comparación de la Productividad Leche Día/IEP por Número de Lactancia y Tipo de Nacimiento

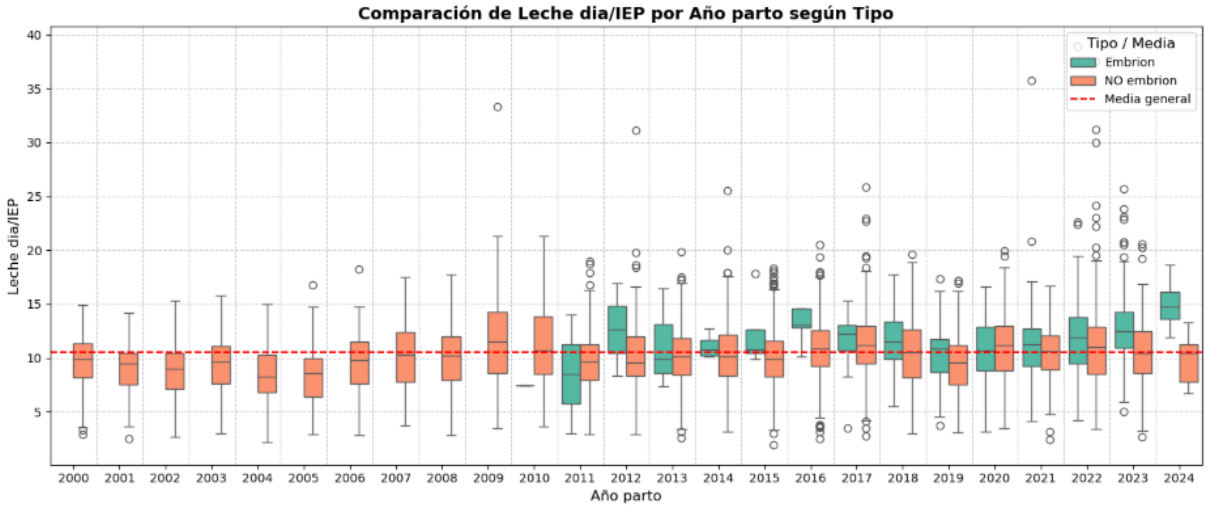


Nota. Distribución de la variable Leche día/IEP desagregada por número de lactancia.

De igual manera la evolución temporal de la productividad de las vacas objeto de estudio se evidencia la diferencia entre vacas nacidas por transferencia de embriones (verde) y por métodos convencionales (naranja); a línea punteada roja representa la media general. A partir de 2012, se observa un incremento progresivo en la presencia de registros tipo embrión, los cuales tienden a concentrarse por encima de la media general en varios periodos, especialmente entre 2016 y 2023; esta tendencia podría asociarse a la implementación más sistemática de biotecnologías reproductivas en los últimos años, así como a una mejora en la selección de progenitores.

Figura 7

Comparación de la Productividad Leche Día/IEP por Año de Parto Según tipo de Nacimiento



Nota. evolución temporal de la productividad Leche día/IEP desde el año 2000 hasta 2024.

Estos hallazgos no solo caracterizan el comportamiento general del hato, sino que proporcionan argumentos sólidos para la segmentación posterior y el desarrollo de modelos predictivos. La secuencialidad lógica entre origen genético, tipo reproductivo, historial individual (lactancia) y productividad permite estructurar modelos robustos que integren genética, manejo y eficiencia animal.

Preparación de los Datos

La selección de variables se fundamentó en los hallazgos obtenidos durante el análisis exploratorio y estadístico descriptivo. Se priorizaron aquellas variables que contribuyen directamente a los objetivos del estudio, es decir, explicar la variabilidad en la productividad lechera ajustada (Leche día/IEP) en función de factores genéticos, reproductivos y productivos. Se excluyeron variables con alta proporción de datos faltantes, redundantes o con baja capacidad explicativa.

Las variables seleccionadas reflejan información clave sobre el origen genético (grupo racial, tipo reproductivo, genealogía), la trayectoria productiva individual (número de lactancia) y el contexto temporal (año de parto). Asimismo, se definió la variable dependiente principal como Leche día/IEP, al combinar eficiencia reproductiva y rendimiento lechero en un solo indicador.

Limpieza, Depuración y Transformación de Datos

El proceso de limpieza incluyó la eliminación de registros con valores faltantes en variables esenciales como IEP, producción de leche y genealogía. Se imputaron datos ausentes en variables numéricas utilizando medidas de tendencia central (mediana) y se aplicaron filtros para asegurar la coherencia en la codificación de categorías.

Para las variables categóricas, se realizó la recodificación de nombres inconsistentes y la homogenización de etiquetas. Las variables de proporción racial fueron transformadas en categorías interpretables (grupo_racial_simplificado), lo cual permitió reducir la complejidad dimensional sin perder información genética relevante. La variable Leche día/IEP fue evaluada por su distribución y se identificaron valores atípicos, algunos de los cuales fueron conservados

al corresponder a vacas de muy alto rendimiento, tras verificar su coherencia con el resto de los datos asociados.

Pruebas Estadísticas

Aunque el enfoque principal del estudio está orientado al análisis multivariado y al modelado predictivo, se realizaron pruebas estadísticas exploratorias para validar la pertinencia de las variables categóricas incluidas. Estas pruebas permitieron identificar diferencias significativas en la variable Leche día/IEP asociadas a los factores genéticos y reproductivos, lo que fortaleció su inclusión en las fases posteriores del análisis. Entre los resultados más relevantes se destacan:

- Se identificaron diferencias significativas en la producción ajustada entre los grupos raciales simplificados, con mayor rendimiento promedio en el grupo "Proporción balanceada".
- Las vacas tipo embrión mostraron, de manera consistente, una productividad significativamente superior frente a las de reproducción natural.
- Se observaron patrones de incremento productivo según el número de lactancia, destacándose un mejor desempeño entre la tercera y sexta lactancia.
- En las comparaciones cruzadas por tipo de reproducción y año de parto, así como por número de lactancia, las vacas tipo embrión mantuvieron su superioridad productiva en diversos contextos.

Estos hallazgos confirman la influencia de la genética y del tipo reproductivo sobre la variable dependiente y respaldan la estructura del conjunto final de datos para el desarrollo de los modelos analíticos posteriores.

Consolidación del Conjunto Final de Datos

Como resultado de los procesos anteriores, se estructuró un conjunto de datos definitivo compuesto por 6.712 registros individuales de lactancias, cada uno con las variables limpias, transformadas y validadas. Este conjunto fue optimizado para garantizar su compatibilidad con modelos de análisis estadístico y aprendizaje automático, eliminando redundancias, inconsistencias y datos incompletos.

La integración de genealogía, grupo racial, tipo reproductivo, rendimiento y control temporal asegura una base robusta para la implementación de algoritmos de segmentación y predicción en las fases siguientes.

Herramientas y Entornos de Desarrollo

El desarrollo del proceso de preparación de datos se llevó a cabo en el lenguaje Python (versión 3.11) utilizando el entorno de trabajo Jupyter Notebook. Se emplearon las siguientes bibliotecas:

- Pandas y NumPy para la manipulación de datos, transformación de variables y limpieza.
- Scipy.stats y statsmodels para las pruebas estadísticas.
- Matplotlib y seaborn para la visualización de patrones, análisis de distribuciones y detección de valores atípicos.

Este entorno permitió mantener un control riguroso y reproducible del flujo de trabajo, facilitando la trazabilidad de cada decisión metodológica tomada en esta fase del estudio.

Modelado

La fase de modelado constituye una etapa crítica en este trabajo, al permitir construir una base analítica sólida para identificar progenitores estratégicamente valiosos dentro del programa

de mejoramiento genético del hato. Las técnicas aplicadas, K-means (agrupamiento no supervisado) y Random Forest Regressor (modelo supervisado) fueron seleccionadas no solo por su robustez y capacidad de adaptación a datos categóricos, sino también como respuesta metodológica a los hallazgos previos obtenidos mediante análisis estadísticos exploratorios e inferenciales.

A través de dichas pruebas estadísticas, se comprobó que el grupo racial balanceado (vacas con composición genética aproximada de 50% Taurus y 50% Indicus), correspondiente principalmente al cruce Holstein × Gyr, presenta los mayores niveles de productividad lechera (Leche día/IEP). De igual forma, se demostró que las vacas nacidas por transferencia de embriones superan significativamente en rendimiento a las nacidas por reproducción natural. Esta diferencia se acentúa dentro del grupo balanceado, donde las vacas tipo “embrión” exhiben el mejor desempeño productivo del conjunto analizado.

A partir de estos hallazgos, se definió como población objetivo para el modelado exclusivamente a las vacas tipo embrión, en quienes se centraron los análisis para determinar qué combinaciones genéticas de padres y madres han generado descendencia más productiva. Esta delimitación responde tanto a criterios biológicos como analíticos, al considerar que esta subpoblación representa el núcleo más valioso del hato desde el punto de vista genético y reproductivo. En este contexto, se aplicaron dos enfoques analíticos complementarios:

- *K-means:* para segmentar a los progenitores según el comportamiento promedio de su descendencia, y así identificar patrones consistentes de productividad mediante agrupamientos no supervisados.

- *Random Forest:* para construir un modelo predictivo que estimara el desempeño lechero ajustado (Leche día/IEP) en función del origen genético, permitiendo además cuantificar la importancia relativa de cada progenitor en la productividad observada.

La articulación de ambos modelos ofrece una visión integral: mientras el agrupamiento permite descubrir estructuras latentes y clústeres de interés, el modelado supervisado permite evaluar el aporte individual de los progenitores dentro de un sistema predictivo robusto, facilitando así la priorización de decisiones reproductivas basadas en evidencia.

Selección de Técnicas de Modelado

Para abordar el análisis genético y productivo en vacas tipo embrión, se adoptó una estrategia dual de aprendizaje automático: K-means, como técnica no supervisada para la segmentación de progenitores, y Random Forest Regressor, como modelo supervisado orientado a la predicción del desempeño individual.

El algoritmo K-means fue seleccionado por su eficacia para agrupar individuos según características comunes, sin requerir etiquetas. Su principio se basa en la partición del espacio de datos en k clústeres, buscando minimizar la distancia intragrupo y maximizar la Inter grupo. Esta técnica es ampliamente usada en estudios donde se busca identificar patrones ocultos y reducir la complejidad del análisis sin sacrificar estructura significativa (Bishop Christopher M., 2016).

Por su parte, el modelo Random Forest, basado en la agregación de múltiples árboles de decisión permite manejar relaciones no lineales, datos categóricos y estimar la importancia relativa de las variables predictoras (Raschka Sebastian & Mirjalili Vahid, 2017). Estas características lo hacen particularmente adecuado para contextos como la genética bovina, donde la interacción entre múltiples factores puede condicionar el resultado productivo.

Ambos modelos fueron implementados en Python utilizando la biblioteca scikit-learn, una de las más reconocidas para la ejecución de algoritmos de machine learning por su robustez, flexibilidad y respaldo comunitario (Aurén Géron, 2023).

Preparación Específica del Conjunto de Datos

El conjunto de datos se depuró rigurosamente para asegurar la validez del modelado. Se seleccionaron únicamente vacas tipo embrión con información completa de producción (Leche día/IEP), padre, madre genética, año de parto y número de lactancia. Para mejorar la estabilidad del análisis, se aplicó un filtro que excluyó progenitores con menos de cinco hijas registradas.

Las variables categóricas fueron transformadas mediante codificación one-hot, lo cual permite al modelo supervisado tratar correctamente los identificadores sin introducir orden artificial ni supuestos lineales (Aurén Géron, 2023). En el caso del modelo no supervisado (K-means), las variables numéricas fueron estandarizadas para evitar que las diferencias de escala afectaran el cálculo de distancias.

Para el modelo Random Forest, se realizó una partición estratificada de los datos en un conjunto de entrenamiento (80%) y uno de prueba (20%), garantizando representatividad y posibilidad de validación externa. Estas etapas siguieron lineamientos metodológicos ampliamente aceptados en procesos de preparación de datos para modelado predictivo (Grus Joel, 2019).

Modelado No Supervisado: K-means

El algoritmo K-means fue aplicado con el fin de segmentar a los progenitores del grupo de vacas tipo embrión según el desempeño productivo medio de su descendencia, medido como Leche día/IEP, y su nivel de representación, medido como el número de lactancias registradas. La combinación de estas dos variables permitió construir un perfil estructurado del mérito

genético observado de los reproductores, sin necesidad de emplear una variable dependiente para el proceso de agrupamiento. El agrupamiento se realizó tanto para padres como para madres genéticas, utilizando dos variables cuantitativas como criterios de segmentación:

- El número total de lactancias registradas por cada progenitor.
- El promedio de leche día/IEP obtenido por su descendencia.

Esta combinación permitió evaluar simultáneamente la intensidad de uso del reproductor y su efectividad productiva, ofreciendo una base objetiva para decisiones estratégicas de selección.

Número óptimo de clústeres: Para determinar el número óptimo de clústeres, se aplicó el método del codo, observando la inercia intra-clúster frente a diferentes valores de k. Como se evidencia en la Figura 8 y Figura 9, tanto para padres como para madres genéticas, tres clústeres resultaron ser la mejor opción de segmentación, por lo que se adoptó ese número para el análisis posterior.

Figura 8

Determinación del Número Óptimo de Clústeres Mediante el Método del Codo (Padres)

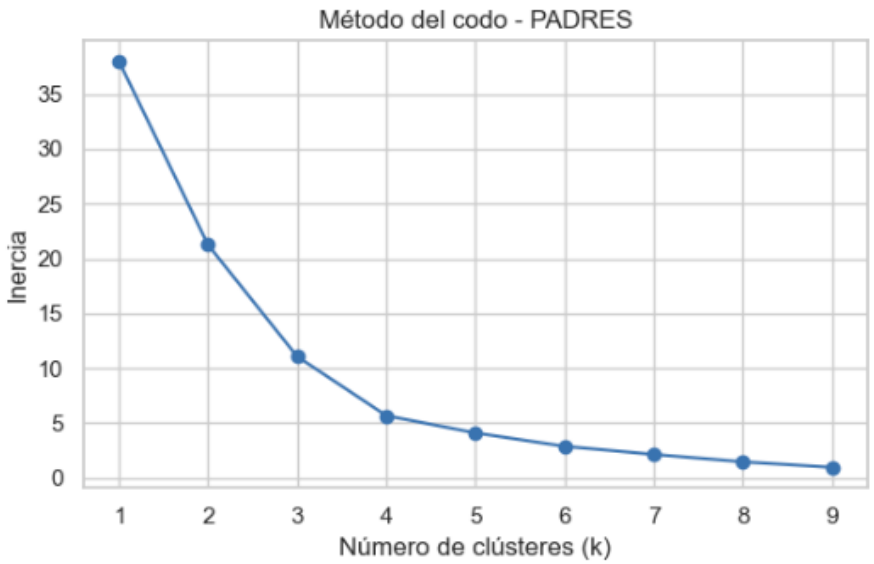
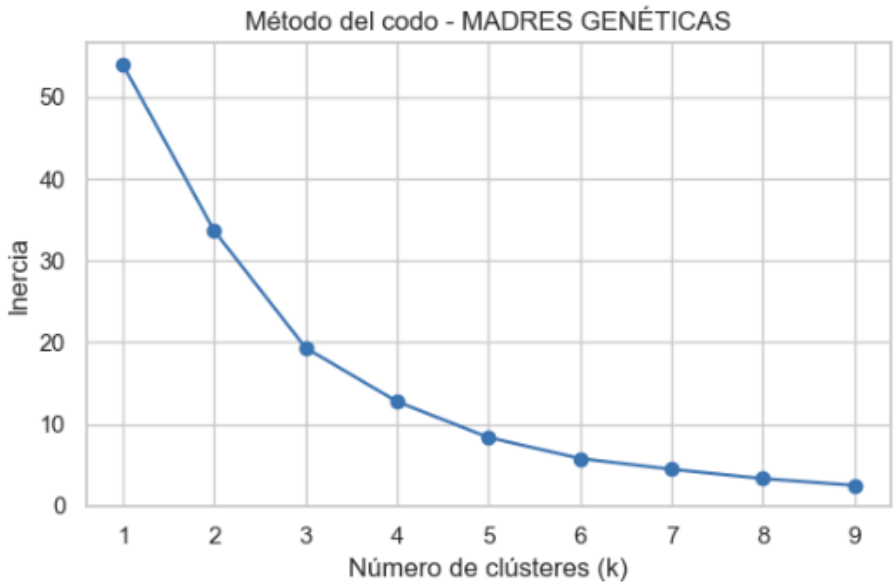


Figura 9

Determinación del Número Óptimo de Clústeres Mediante el Método del Codo (Madres Genéticas)



Agrupamiento de padres genéticos: En la Tabla 4, se presentan los resultados del agrupamiento de los 19 padres genéticos incluidos en el análisis. Los clústeres fueron generados mediante el algoritmo K-means, agrupando a los toros (padres) según la productividad promedio de sus hijas tipo embrión. El clúster 0 agrupa los padres con mejor desempeño, con un promedio de 12,40 litros diarios por IEP, mientras que el clúster 2 incluye solo dos toros con un número elevado de lactancias asociadas, pero menor rendimiento promedio. Estas agrupaciones permiten discriminar linajes genéticos con mayor potencial productivo.

Tabla 4

Clústeres obtenidos para los toros (Padres).

Clúster	N.º Padres	Promedio Lactancias	Promedio Leche	Min. Leche	Máx. Leche
0	10	39.00	12.40	12.06	12.86
1	7	39.57	11.13	10.76	11.62
2	2	121.00	11.07	10.45	11.67

Nota. Resumen descriptivo de los clústeres obtenidos para los toros (Padres) según productividad de sus crías. *Fuente:* Autor.

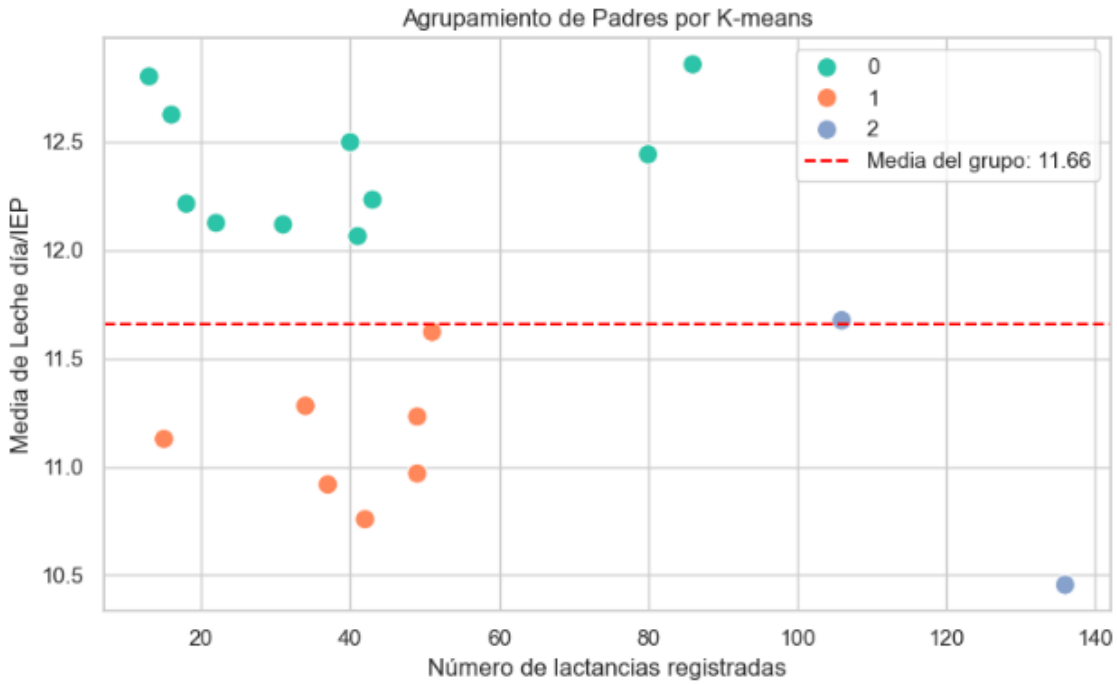
En la figura 10 se muestra la distribución de los padres por clúster, visualizando sus promedios de producción y volumen de registros. Esta visualización refuerza la utilidad del modelo para detectar reproductores con desempeño sobresaliente y otros con uso frecuente pero rendimiento moderado. Esta segmentación metodológica permite una caracterización objetiva de los padres, identificando perfiles productivos distintos, sin emitir aún juicios sobre la conveniencia de su uso, lo cual se desarrollará en la fase de evaluación. El gráfico representa la segmentación de los toros utilizados como padres genéticos mediante el algoritmo K-means, usando como variables el número de lactancias registradas y la producción promedio de sus crías (Leche día/IEP). Se identificaron tres clústeres:

- *Clúster 0 (verde):* toros con alta productividad promedio (>12 Lt/IEP) y número moderado de lactancias, considerados de alto valor genético.
- *Clúster 1 (naranja):* toros con menor productividad y menor número de lactancias.
- *Clúster 2 (azul):* toros con alto número de lactancias pero productividad promedio similar o inferior a la media (línea roja = 11.66 L/IEP).

Esta segmentación permite identificar toros de alto potencial para estrategias de mejoramiento genético

Figura 10

Agrupamiento de toros (Padres) mediante K-means según número de lactancias y productividad promedio



Agrupamiento de madres genéticas: El análisis de K-means también se aplicó a las madres genéticas; se aplicó el modelo a 27 madres genéticas, cuyos resultados se presentan en la Tabla 5. El clúster 0 agrupa a las madres de mayor rendimiento (12.46 kg/día), con un número moderado de lactancias (14.79), mientras que el clúster 2 representa madres con menor productividad (9.93 kg/día). Los clústeres fueron generados mediante el algoritmo K-means, agrupando a las madres genéticas según el promedio de producción Leche día/IEP de su progenie. El clúster 0 representa el grupo con mayor productividad promedio, mientras que el

clúster 2 agrupa madres con menor desempeño productivo. El clúster 1 se distingue por concentrar vacas con un alto número de lactancias asociadas, posiblemente con un rol consolidado como donadoras, pero con valores de productividad más conservadores.

Tabla 5

Clústeres Obtenidos para las Madres Genéticas.

Clúster	N.º Madres	Promedio Lactancias	Promedio Leche	Min. Leche	Máx. Leche
0	14	14.79	12.46	11.31	15.33
1	8	79.00	11.63	10.54	12.29
2	5	14.00	9.93	8.41	10.88

Nota. Resumen descriptivo de los clústeres obtenidos para las madres genéticas según productividad de sus crías. Fuente. Autor.

La figura 11 muestra cómo las madres se distribuyen según su productividad promedio y número de lactancias. Esta visualización contribuye al reconocimiento de patrones colectivos de mérito materno sin requerir evaluación individual de desempeño. Estos clústeres permiten estructurar el universo de progenitoras en grupos homogéneos, proporcionando una base técnica útil para fases posteriores de priorización reproductiva.

El gráfico muestra la segmentación de las madres genéticas a través del algoritmo K-means, utilizando como variables el número de lactancias registradas y la producción promedio de su progenie (Leche día/IEP):

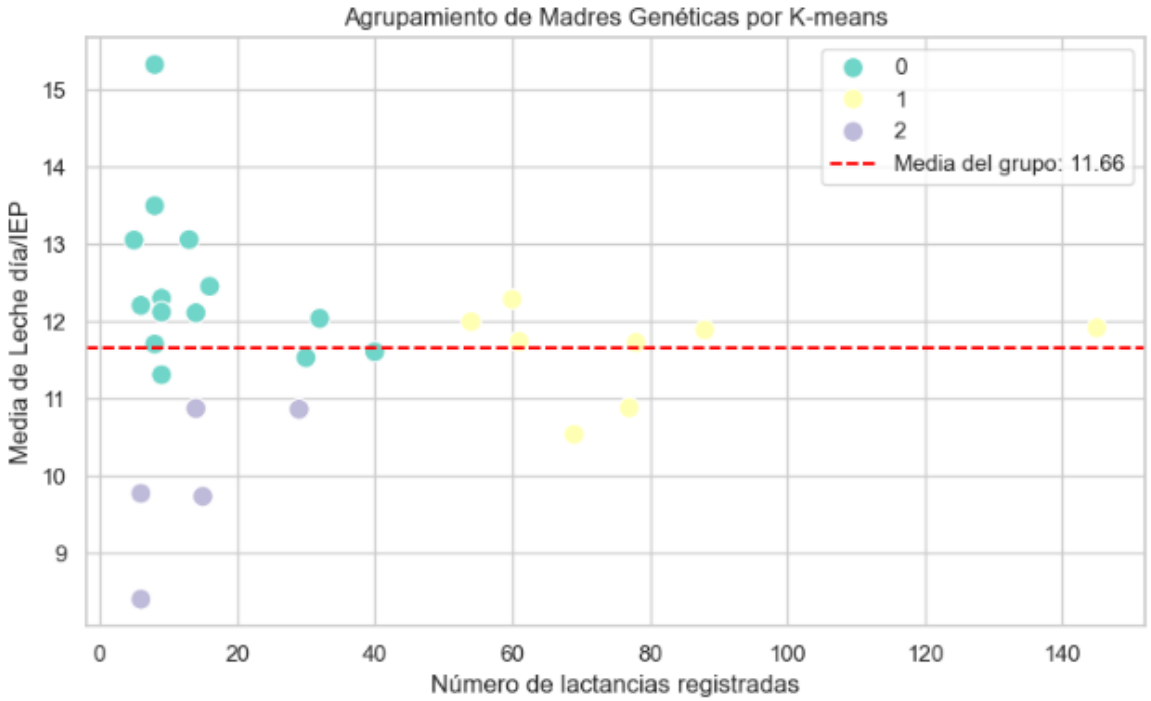
- Clúster 0 (verde): agrupa madres con alta productividad promedio y número moderado de lactancias, perfiladas como reproductoras de alto valor.
- Clúster 1 (amarillo): incluye madres con alto número de lactancias pero con productividad promedio cercana a la media general (línea roja: 11.66 L/IEP).

- Clúster 2 (morado): compuesto por madres con baja productividad, a pesar de un número bajo a moderado de registros.

Esta segmentación permite orientar estrategias de selección y aprovechamiento de donadoras de embriones en función de su rendimiento genético demostrado

Figura 11

Agrupamiento de Madres Genéticas Mediante K-means Según Número de Lactancias y Productividad Promedio



Modelado Supervisado: Random Forest Regressor

Como complemento al agrupamiento, se utilizó un modelo supervisado de regresión basado en Random Forest, con el objetivo de estimar la productividad ajustada de las vacas tipo embrión en función del origen genético (padre y madre genética) y de variables de control como el número de lactancia y el año de parto. Random Forest es un algoritmo de ensamble basado en

árboles de decisión que utiliza el principio de Bootstrap Aggregating o bagging para mejorar la estabilidad y reducir el sobreajuste. Además, permite calcular la importancia relativa de cada predictor en la construcción del modelo, lo que lo convierte en una herramienta apropiada para evaluar relaciones complejas y jerarquizar variables categóricas (Raschka Sebastian & Mirjalili Vahid, 2017). Además, el modelo permite:

- Disminuir la varianza y el riesgo de sobreajuste.
- Estimar la importancia relativa de cada predictor.
- Manejar eficientemente datos categóricos codificados.

Tabla 6

Variables Utilizadas en el Modelo Random Forest

Variable	Tipo	Descripción
Padre	Categórica (one-hot)	Identificador del padre genético
Madre genética	Categórica (one-hot)	Identificador de la madre genética
# lact.	Numérica continua	Número de lactancia de la vaca
Año parto	Numérica discreta	Año del parto asociado
Leche día/IEP	Variable objetivo	Producción diaria ajustada por IEP

Este conjunto de variables permitió incorporar tanto factores genéticos como efectos de etapa productiva y contexto temporal, mejorando la capacidad explicativa del modelo. Las variables categóricas de los progenitores se codificaron mediante one-hot encoding para su inclusión en los modelos predictivos. La variable Leche día/IEP se utilizó como variable objetivo para evaluar el impacto de los efectos genéticos y reproductivos sobre la eficiencia productiva.

Resultados del Modelo: La evaluación del modelo en el conjunto de prueba mostró un desempeño limitado, con un $R^2 = -0.1204$, un RMSE de 2.21 y un MAE de 1.72. Estos

resultados, indican que el modelo no logra realizar predicciones precisas, pero sí permite explorar la contribución relativa de cada variable.

Importancia de Variables: El valor principal de este modelo reside en su capacidad para jerarquizar la relevancia de las variables predictoras. Tal como se observa en el gráfico de importancia de variables, el modelo identifica progenitores con mayor peso estadístico en la explicación de la productividad ajustada, independientemente del promedio observado.

Figura 12

Importancia Relativa de los Toros (Padres) en el Modelo Random Forest

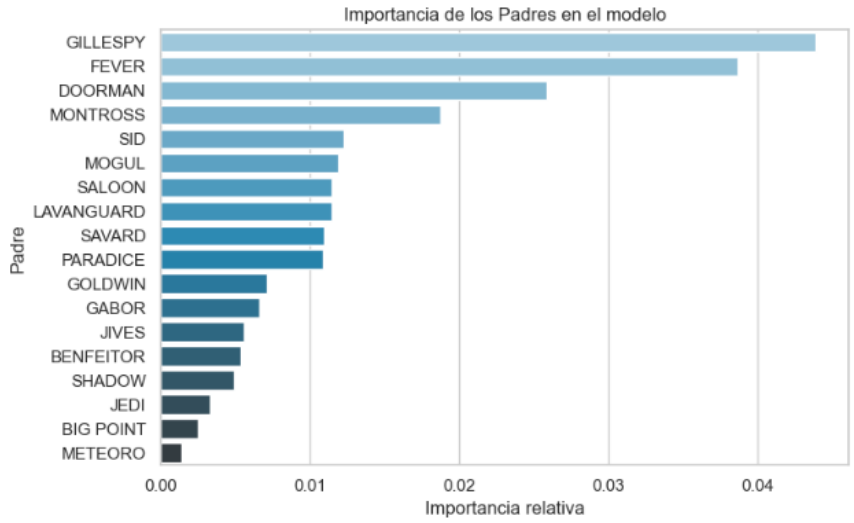
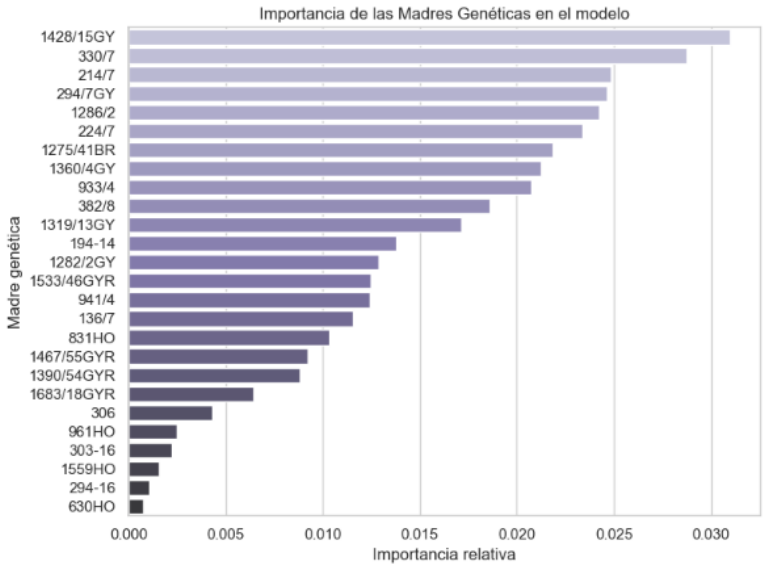


Figura 13

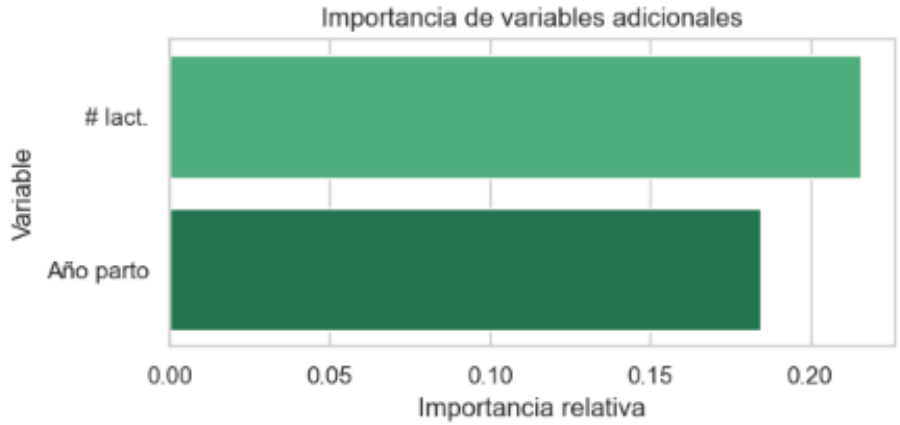
Importancia Relativa de las Madres Genéticas en el Modelo Random Forest



La figura 14 muestra la importancia relativa de las variables adicionales número de lactancia y año de parto en el modelo Random Forest utilizado para predecir la productividad Leche día/IEP. Ambas variables demostraron un aporte considerable a la capacidad predictiva del modelo, destacándose el número de lactancia como el factor más relevante, con una importancia relativa superior al 20 %. Este hallazgo respalda la inclusión del ciclo productivo del animal como una variable crítica en la interpretación de los resultados productivos, particularmente en el contexto del análisis de progenie.

Figura 14

Importancia Relativa de Variables N° de Lactancia y Año del Parto en el Modelo Random Forest



Evaluación

La fase de evaluación permitió examinar en detalle los resultados obtenidos a partir de los modelos de agrupamiento no supervisado (K-means) y regresión supervisada (Random Forest), con el objetivo de valorar la utilidad metodológica de estos enfoques en la segmentación, análisis y jerarquización de progenitores bovinos, dentro de una población de vacas nacidas por transferencia de embriones. Los análisis se centraron en las características de los grupos generados, sus comportamientos agregados y su contribución metodológica al problema de investigación.

Evaluación del Agrupamiento de Progenitores Mediante K-means.

La aplicación del algoritmo K-means permitió clasificar a los progenitores (padres y madres genéticas) en tres grupos diferenciados según dos criterios objetivos: el número de lactancias registradas por su descendencia y el promedio de producción ajustada (Leche día/IEP).

Esta segmentación reveló una estructura clara y funcional que posibilitó el ordenamiento del conjunto total en clústeres de alta, media y baja productividad.

Resultados por Tipo de Progenitor

La siguiente tabla presenta el resumen estadístico general de los clústeres para padres y madres genéticas, evidenciando diferencias sustanciales en el rendimiento promedio entre grupos, así como variaciones importantes en su representatividad.

Tabla 7

Estadísticas por Clúster y Tipo de Progenitor

Tipo de progenitor	Clúster	N.º individuos	Promedio lactancias	Promedio Leche día/IEP
Padre	0	9	39.00	12.39
Padre	1	7	39.57	11.13
Padre	2	2	121.00	11.07
Madre genética	0	13	14.79	12.23
Madre genética	1	8	79.00	11.63
Madre genética	2	5	14.00	9.93

Los clústeres mostraron diferencias no solo en los niveles medios de productividad, sino también en la cantidad relativa de progenitores que los conforman. En particular, el clúster de mayor productividad estuvo compuesto por un número reducido de individuos, con registros de lactancias relativamente bajos pero con promedios productivos elevados. Esta configuración incluye casos donde los registros corresponden a las primeras lactancias de algunos progenitores, lo que permite caracterizar perfiles con alta producción inicial. Esta condición adquiere relevancia si se considera que la productividad tiende a incrementarse en lactancias sucesivas, especialmente entre la cuarta y la sexta, por lo que los grupos con menor número de lactancias y

alta producción ofrecen una estructura útil para el análisis prospectivo del comportamiento productivo.

Evaluación del Modelo Supervisado Random Forest

El modelo de regresión basado en Random Forest fue construido para estimar la productividad de las vacas tipo embrión a partir de su origen genético y contexto productivo. Aunque su capacidad predictiva resultó limitada, con un coeficiente de determinación negativo ($R^2 = –0.1204$), y errores absolutos medios moderados (RMSE = 2.21, MAE = 1.72) el modelo permitió generar información de valor desde una perspectiva explicativa.

Lo más relevante metodológicamente fue la estimación de la importancia relativa de las variables predictoras. El modelo permitió jerarquizar factores como el año del parto, número de lactancia y origen genético, mostrando que las variables contextuales (temporales y fisiológicas) contribuyeron en mayor proporción al modelo que las estrictamente genéticas. Esta observación resalta la complejidad multifactorial de la productividad lechera, e indica que si bien el origen genético influye, su efecto se expresa en interacción con el entorno productivo y no puede interpretarse de forma aislada.

Integración Metodológica de Resultados: K-means + Random Forest Regressor

La articulación entre los enfoques de aprendizaje no supervisado (K-means) y supervisado (Random Forest Regressor) permitió construir una visión multiescalar y complementaria del desempeño genético y productivo de los progenitores incluidos en el estudio. Esta integración se desarrolló como una etapa de síntesis que articuló las salidas estadísticas y clasificatorias generadas previamente, consolidando un marco metodológico que unifica agrupamiento estructural e importancia predictiva en torno a la variable objetivo Leche día/IEP.

Estructura metodológica de la integración: La estrategia de integración consistió en unir las salidas de ambos modelos para cada progenitor evaluado, combinando:

- El clúster asignado por K-means según la productividad promedio y número de lactancias de su descendencia.
- La importancia relativa estimada por el modelo Random Forest, calculada como la contribución del progenitor a la reducción del error cuadrático medio del modelo.

Tabla 8

Variables para la Integración K-means + Random Forest

Variable	Descripción
Tipo	Padre o Madre genética
Número de lactancias	Total de registros asociados
Promedio Leche día/IEP	Producción media de la descendencia
Clúster (K-means)	Agrupación según productividad observada
Importancia relativa (Random Forest)	Contribución estadística del progenitor al modelo predictivo

Criterios de clasificación integrada: La combinación de los dos enfoques dio origen a una clasificación estructurada de los grupos de progenitores, sin enfocarse en individuos particulares. Se definieron cuatro categorías metodológicas:

- Alta productividad e influencia: progenitores ubicados en clústeres de alto rendimiento y con alta importancia relativa.
- Productividad moderada con alta influencia: progenitores con rendimiento medio o bajo, pero con elevada importancia estadística.

- Alta productividad con baja influencia: individuos con buen desempeño medio pero bajo peso explicativo en el modelo.
- Baja productividad y escasa importancia: progenitores con bajo rendimiento e influencia reducida.

Estas representaciones evidencian que los clústeres de mayor productividad no siempre concentran las mayores influencias dentro del modelo, y que, por el contrario, algunos grupos con productividad intermedia tienen mayor peso predictivo.

La figura 15 integra los resultados del modelo Random Forest y del algoritmo de agrupamiento K-means, visualizando la media de producción Leche día/IEP frente al número de lactancias registradas para padres (×) y madres genéticas (●). El tamaño de cada punto refleja la importancia relativa del progenitor en el modelo predictivo, mientras que el color indica el clúster al que pertenece según su comportamiento productivo. La línea roja punteada representa la media general de la variable objetivo (11.66 L/IEP). Este análisis cruzado permite identificar progenitores de alto valor genético (gran tamaño, clúster 0) y priorizarlos en estrategias de mejoramiento. Los progenitores de clúster 2 (azul) presentan bajo desempeño y escasa relevancia predictiva, lo cual sugiere una menor prioridad en programas reproductivos futuros.

Figura 15

Integración de la Importancia Relativa y Segmentación por Clúster Según Tipo de Progenitor

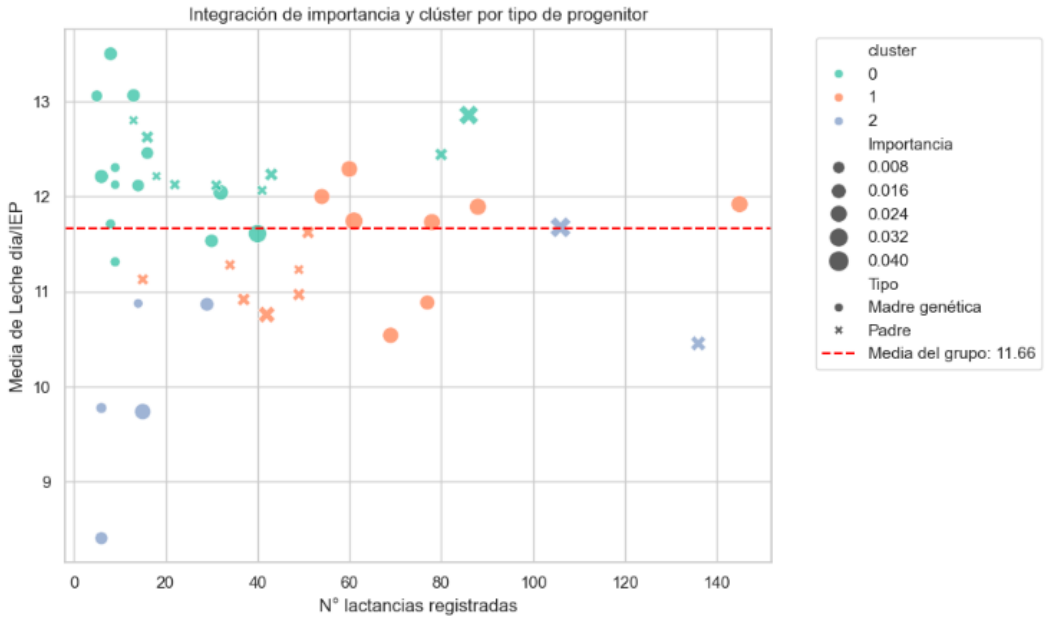
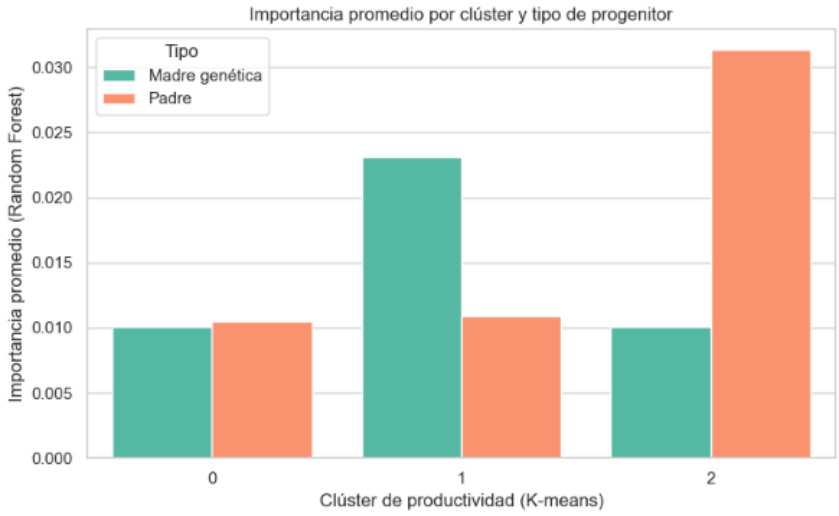


Figura 16

Importancia Promedio por Clúster de Productividad y Tipo de Progenitor en el Modelo Random Forest



El gráfico 16 compara la importancia promedio de padres y madres genéticas en el modelo Random Forest, agrupados según su pertenencia a los clústeres de productividad obtenidos por K-means. Se observa que las madres genéticas del clúster 1, de productividad intermedia, presentan una mayor importancia relativa promedio, mientras que los padres clasificados en el clúster 2 (de bajo desempeño productivo) son los que aportan más información al modelo. Esta distribución sugiere que algunos padres de bajo rendimiento general pueden tener un efecto predictivo más marcado sobre ciertas combinaciones genéticas específicas, mientras que las madres relevantes tienden a concentrarse en los clústeres de mayor productividad. El análisis resalta la necesidad de considerar tanto el desempeño como la relevancia predictiva en decisiones de selección genética.

Análisis Comparativo Entre Modelos

El cruce de información permitió identificar patrones que no serían visibles mediante un solo modelo. Por ejemplo:

- Los clústeres de mayor productividad media no siempre coincidieron con los de mayor importancia estadística.
- Algunos clústeres con menor productividad agregada mostraron mayor dispersión en la variable de importancia relativa, lo que podría estar asociado a mayor representación o variabilidad genética.
- En madres genéticas, los clústeres intermedios presentaron en algunos casos mayor peso predictivo promedio que los de alta productividad.

Tabla 9

Estadísticas Integradas por Tipo de Progenitor y Clúster

Tipo	Clúster	N.º Progenitores	Promedio Leche día/IEP	Promedio Importancia
Madre genética	0	13	12.23	0.0101
Madre genética	1	8	11.63	0.0231
Madre genética	2	5	9.93	0.0101
Padre	0	9	12.39	0.0105
Padre	1	7	11.13	0.0109
Padre	2	2	11.07	0.0313

La anterior tabla muestra la integración de los resultados para el análisis de clústeres de productividad (K-means) y la importancia relativa en el modelo Random Forest para padres y madres genéticas; se observa que, si bien el clúster 0 concentra a progenitores con mayor productividad media, no necesariamente coincide con los valores más altos de importancia predictiva. Los padres del clúster 2, a pesar de tener un rendimiento inferior al clúster 0, presentan la mayor importancia relativa en el modelo, lo que sugiere un efecto predictivo individual significativo sobre la variable objetivo Leche día/IEP.

Valor metodológico de los modelos*:* La Tabla 10 sintetiza los aportes técnicos y las limitaciones metodológicas de cada modelo utilizado; en este sentido el K-means permitió clasificar progenitores en grupos de desempeño productivo homogéneo sin requerir una variable objetivo, mientras que Random Forest asignó importancia estadística a cada predictor en la estimación de la productividad Leche día/IEP; la combinación de ambos métodos permitió un análisis complementario: uno exploratorio y otro explicativo.

Tabla 10

Comparación Entre K-means y Random Forest

Enfoque	Tipo	Aporta	Limita
K-means	No supervisado	Agrupa individuos por desempeño observable	No considera peso estadístico
Random Forest	Supervisado	Jerarquiza importancia de predictores	Sensible a variables de contexto y ruido

Nota. Comparación entre los enfoques K-means y Random Forest en el análisis de progenitores.

La integración metodológica permitió compensar las limitaciones individuales de cada modelo, logrando una evaluación robusta, si dependencia de un único criterio estadístico o técnico.

El proceso de evaluación estructurado a partir de los resultados de K-means y Random Forest Regressor demuestra que es posible organizar y jerarquizar conjuntos complejos de progenitores mediante métodos cuantitativos complementarios. Los resultados confirman que no existe una correspondencia directa entre frecuencia de uso, productividad promedio e importancia estadística, lo que valida el empleo conjunto de técnicas para obtener representaciones más completas y útiles en el análisis genético-productivo del hato.

Implementación

Los resultados obtenidos mediante el uso combinado de los modelos de agrupamiento no supervisado (K-means) y regresión supervisada (Random Forest Regressor) conforman una base técnica que permite su incorporación directa en procesos de análisis y toma de decisiones reproductivas dentro del sistema productivo analizado. Esta fase se concentra en el traslado

operativo de los hallazgos del modelo analítico a una estructura útil para la organización del material genético, la priorización de progenitores y la planificación reproductiva fundamentada en evidencia cuantitativa.

A partir de la segmentación generada por K-means, fue posible clasificar a los progenitores según su rendimiento productivo medio, conformando clústeres que agrupan perfiles homogéneos de comportamiento. De forma complementaria, la jerarquización generada por Random Forest permitió identificar la relevancia estadística de cada progenitor como variable predictiva dentro del modelo de productividad ajustada. La combinación de estas dos aproximaciones proporcionó un marco de referencia sólido para el diseño de estrategias reproductivas con fundamento analítico.

La articulación entre productividad observada, representación numérica y relevancia estadística posibilita una estructura flexible para su uso práctico, aplicable tanto en decisiones de conservación genética como en esquemas de multiplicación selectiva. Asimismo, esta información se integra con los sistemas de gestión de datos productivos y reproductivos, favoreciendo una visión sistemática y reproducible del desempeño del hato. Como parte de esta fase, se estructuraron una serie de apéndices temáticos, derivados directamente de los análisis realizados, que permiten la consulta detallada de la clasificación de progenitores. Estos listados se organizan por tipo de modelo aplicado (K-means o Random Forest), tipo de progenitor (padres o madres) y criterio de ordenamiento (productividad o importancia). Su incorporación como anexos responde a la necesidad de mantener la trazabilidad, transparencia y aplicabilidad de los resultados obtenidos.

Conclusiones

El presente trabajo de investigación permitió demostrar que un enfoque analítico sustentado en ciencia de datos, específicamente mediante el uso de técnicas de aprendizaje automático como K-means y Random Forest, constituye una herramienta poderosa para caracterizar, segmentar y optimizar el manejo genético y productivo de hatos lecheros en condiciones tropicales. Este enfoque integrador permitió identificar patrones relevantes en la productividad de las vacas, particularmente en aquellas nacidas por transferencia de embriones, dentro del marco de una estrategia de mejoramiento genético basada en cruzamientos estratégicos entre razas Bos taurus y Bos indicus.

La transformación y análisis de una base de datos compuesta por más de 6.700 lactancias, sometida a un riguroso proceso de limpieza y depuración, permitió consolidar un conjunto confiable de registros que integran dimensiones genéticas, reproductivas y productivas. Como resultado de la recategorización genética de los individuos, se definieron grupos raciales compuestos y simplificados que facilitaron una comparación estadística robusta. En este contexto, se evidenció que el grupo racial balanceado (Taurus 50% / Indicus 50%) presentó el mayor desempeño promedio en productividad diaria ajustada por intervalo entre partos (Leche día/IEP), superando incluso a los grupos con alta proporción taurina o indicus, y mostrando una destacada consistencia interna. Este hallazgo es particularmente relevante si se considera que dicho grupo corresponde, en su mayoría, al cruzamiento Holstein × Gyr, el cual representa una estrategia zootécnica ampliamente adoptada en sistemas lecheros tropicales para maximizar producción sin sacrificar adaptación. Así, los resultados obtenidos confirman que la composición genética balanceada no solo es viable en términos de producción, sino que también favorece la estabilidad y la sostenibilidad del sistema.

Por otra parte, el análisis exclusivo de vacas nacidas por transferencia de embriones permitió focalizar la investigación en el subconjunto de mayor interés estratégico del hato. En esta población se aplicaron las técnicas de K-means para identificar clústeres de progenitores (padres y madres genéticas) según patrones de rendimiento en la descendencia, y Random Forest para estimar la importancia relativa de cada progenitor en la predicción del desempeño productivo. Esta secuencia metodológica respondió a una necesidad práctica de la empresa ganadera: seleccionar con mayor precisión a las vacas Gyr donadoras de embriones, optimizando así los esquemas de cría de hembras Holstein × Gyr.

La integración de ambos modelos permitió validar la consistencia de los resultados y generar insumos estratégicos de alta aplicabilidad en campo. En particular, se produjeron listados detallados por tipo de progenitor, organizados por importancia predictiva, rendimiento promedio y agrupación genética. Estos insumos fueron sistematizados en los apéndices A – F del trabajo y constituyen una herramienta clave para la toma de decisiones reproductivas y la planificación genética del hato.

Recomendaciones y Limitaciones

Esta investigación está delimitada a la evaluación de datos específicos provenientes de registros productivos, reproductivos y genéticos de un hato lechero del trópico bajo colombiano. Entre los alcances principales se encuentra la capacidad para analizar estadísticamente la influencia de variables como grupo racial, tipo reproductivo y progenitores sobre la productividad láctea.

No obstante, existen limitaciones inherentes a la calidad y precisión de los datos disponibles. La presencia de registros incompletos o sesgados, especialmente relacionados con variables reproductivas y sanitarias, puede afectar la precisión y generalización de los resultados. Además, factores ambientales no controlados y la complejidad biológica de los animales pueden limitar la interpretación directa y absoluta de los hallazgos.

Finalmente, esta investigación no considera otros factores potencialmente relevantes como variables nutricionales o condiciones detalladas de manejo, lo que podría influir en la productividad, y por ende deben tenerse en cuenta al momento de extrapolar o aplicar estos resultados a otros contextos productivos.

Aplicaciones Prácticas e Interpretativas de los Modelos K-means, Random Forest

La implementación conjunta de algoritmos de aprendizaje no supervisado (K-means) y supervisado (Random Forest Regressor) en el presente trabajo permitió identificar patrones genéticos y productivos con una alta capacidad explicativa sobre el desempeño de vacas nacidas por transferencia de embriones. Esta integración metodológica facilitó una lectura profunda de los datos, con implicaciones directas para la toma de decisiones reproductivas y de mejoramiento genético en el hato evaluado, y con alta aplicabilidad en otros sistemas de producción lechera

tropical. A continuación, se resumen las principales conclusiones interpretativas y aplicaciones prácticas de estos modelos.

Interpretación Aplicada del Modelo K-means

Este modelo permitió segmentar a los progenitores (padres y madres genéticas) en clústeres homogéneos según el rendimiento promedio de su descendencia. Este enfoque fue útil para descubrir subpoblaciones con patrones productivos similares, incluso sin una clasificación genética preestablecida.

Se identificaron clústeres con alto rendimiento productivo (leche día/IEP) que agrupan progenitores cuyas crías muestran eficiencia sobresaliente en las condiciones específicas del hato.

Uno de los aportes más significativos del modelo fue evidenciar que animales con pocas lactancias acumuladas pero agrupados en clústeres de alto rendimiento pueden representar líneas genéticas emergentes con alto potencial de replicación. Este hallazgo es de gran valor estratégico, ya que permite realizar una selección genética temprana, acelerando así los ciclos de mejora en el hato.

El número de lactancias debe ser incorporado como criterio de interpretación dentro del análisis de los clústeres, para diferenciar animales consolidados de aquellos con proyección genética promisoria.

Aplicación práctica en el hato:

Uso de clústeres de alto desempeño para estructurar núcleos de mejoramiento, seleccionando como donadoras de embriones aquellas vacas con alto rendimiento y valor predictivo.

Seguimiento individualizado a vacas jóvenes destacadas, con el fin de validar su desempeño temprano y proyectarlas como futuras reproductoras clave.

Clasificación de progenitores por desempeño observado en descendencia, fortaleciendo las decisiones de apareamiento estratégico y rotación de toros.

Interpretación Aplicada del Modelo Random Forest

Este modelo permitió estimar la importancia relativa de cada progenitor en la predicción de la productividad, integrando variables genéticas, reproductivas y categóricas en un solo sistema de evaluación robusto.

Se identificaron progenitores (toros y vacas donadoras) con una alta capacidad predictiva individual, lo cual permite jerarquizar su uso dentro de los programas de inseminación artificial y transferencia de embriones.

Este modelo no solo identifica a los animales más influyentes, sino que también ofrece un marco para construir índices empíricos de valor genético, basados en el rendimiento real observado en campo.

Al combinar esta importancia relativa con el número de lactancias de cada cría, se puede priorizar el análisis de progenie joven altamente productiva, como base para decisiones anticipadas de conservación y multiplicación genética.

Aplicación práctica en el hato:

Construcción de un sistema de monitoreo y selección basado en datos reales del hato, complementando o incluso reemplazando criterios tradicionales basados únicamente en pedigrí.

Exclusión progresiva de progenitores con baja importancia relativa, reduciendo el riesgo de perpetuar líneas de bajo desempeño.

Recomendación de apareamientos fundamentados en resultados de campo y no únicamente en estándares externos.

Aplicación en Otros Hatos del Trópico Bajo Colombiano

La metodología desarrollada es altamente replicable en hatos con características similares, que cuenten con registros de genealogía, producción y reproducción estructurados. Esta aproximación permite identificar líneas genéticas adaptadas y eficientes bajo condiciones tropicales, reduciendo la dependencia de genética foránea no probada en el entorno local; igualmente la identificación de progenie joven de alto desempeño mediante análisis de clústeres y predicción es especialmente valiosa para hatos que trabajan con transferencia de embriones o cría dirigida, ya que permite tomar decisiones anticipadas con base en datos reales.

Impacto Esperado en Otros Sistemas:

Disminución del intervalo entre generaciones de alta productividad.

Implementación de sistemas de evaluación genético-productiva con autonomía técnica local.

Generación de evidencia para justificar inversiones en programas de biotecnología reproductiva con enfoque en eficiencia comprobada.

En síntesis, la combinación estratégica de K-means y Random Forest, complementada con la evaluación del número de lactancias, representa una herramienta innovadora y técnica para acelerar el mejoramiento genético de hatos lecheros tropicales, utilizando datos propios de cada hato y estructuras analíticas replicables. Esta metodología fortalece el proceso de toma de decisiones, reduce la incertidumbre en la selección de reproductores y permite construir modelos sostenibles de ganadería de precisión adaptados a la realidad del trópico colombiano.

Referencias

Aurén Géron. (2023). *Hands - On Machine learning whit Sckit - Learn, Keras & TensorFlow* (3.ª edición). O'Reilly Media, Inc. https://www.oreilly.com/library/view/hands-on-machine/9781098125974/

Bishop Christopher M. (2016). *Pattern Recognition and Machine Learning* (1st ed.). Springer.

Brum Luciano Moraes da Luz, Lampert Vinícius do Nascimento, & Camargo Sandro da Silva. (2019). Business intelligence and data warehouse in agrarian sector: A bibliometric study. *Journal of Agricultural Science*, *11*(2), 353. https://doi.org/10.5539/jas.v11n2p353

Chapman Pete, Clinton Julian, Kerber Randy, Khabaza Thomas, ReinartzThomas, Colin Shearer, & Wirth Rüdiger. (2000). *CRISP-DM 1.0: Step-by-step data mining guide. SPSS Inc*. DaimlerChrysler.

Chaturvedi Shailesh, Yadav Ram Lal, Gupta A K, & Sharma A K. (2013). Life time milk amount prediction in dairy cows using artificial neural networks. *International Journal of Recent Research and Review*, *V*.

Council on Dairy Cattle Breeding. (2025, April 14). *Council on Dairy Cattle Breeding (CDCB)*. https://uscdcb.com/

Dijkinga, F. J. (2023). Uso de aprendizaje supervisado de máquina para la predicción de valores genéticos basada en dos generaciones de ancestros. *Research, Society and Development*, *12*(6), e2812641904. https://doi.org/10.33448/rsd-v12i6.41904

Dongre, V. B., & Gandhi, R. S. (2016). Applications of artificial neural networks for enhanced livestock productivity: A review. *Indian Journal of Animal Sciences*, *86*(11), 1232–1237. https://doi.org/10.56093/ijans.v86i11.62970

Empresa Brasileira de Pesquisa Agropecuária (Embrapa). (2025). *Empresa Brasileira de Pesquisa Agropecuária (Embrapa)*. https://www.embrapa.br/

Estrada Carvaja Verny, Camacho Calvo Marlen, Molina Montero Rafael, & Paniagua Madrigal Wilfrido. (2019). Detección de mastitis subclínica en vacas lecheras por modelos de regresión lineal y algoritmos de inteligencia artificial, San Carlos, Costa Rica. *Revista Agro Innovación En Trópico Húmedo*, *2215–5368*, 18–30. https://doi.org/10.18860/rath.v2i1.4689

Fedegan. (2023). *Balance y perspectivas sector ganadero 2023 - 2024 | Fedegan*. https://www.fedegan.org.co/balance-y-perspectivas-sector-ganadero-2023-2024

Fernández Chuairey, L., Walkiria Guerra Bustillo, C., de Calzadilla Pereyra, J., & Ulises Lim Chang, N. (2017). Desarrollo de la modelación estadístico - matemática en las ciencias agrarias. Retos y perspectivas. *Revista Investigación Operacional*, *38*(5), 462–467. https://revistas.uh.cu/invoperacional/article/view/4118/3622

Fernando Johann, Patiño Hoyos, Liliana Velásquez Carrascal Blanca, Bautista Dewar Rico, & García Díaz Noel. (2023). Impacto transformador de la inteligencia artificial y aprendizaje automático en la producción agropecuaria: un enfoque en la sostenibilidad y eficiencia. *Revista Formación Estratégica*. https://orcid.org/0000-0002-1808-3874

Flores M Julia, Gámez Jose A, Mateo Juan L, & Puerta José M. (2006). Selección genética para la mejora de la raza ovina manchega mediante técnicas de Minería de Datos. *Revista Iberoamericana de Inteligencia Artificial*, *10*, 69–77. http://www.redalyc.org/articulo.oa?id=92502908

Grus Joel. (2019). *Ciencia de datos desde cero Principios básicos con Python 2.ª Edición* (2.ª edición). O'Reilly Media / Anaya Multimedia. https://github.com/joelgrus/data-science-from-scratch

Instituto Colombiano Agropecuario - ICA. (2025). *Buenas Prácticas Ganaderas (BPG)*. https://www.ica.gov.co/areas/agricola-pecuaria/bpa-bpg.aspx

Jeremias Lachman, & Andres lopez. (2018). Innovación, habilidades y nuevas áreas de conocimiento en sectores tecnológicos emergentes: el caso de la Agricultura y Ganadería de Precisión. *Revista Pymes, Innovación y Desarrollo , Vol. 6, No. 3*, 60–85. https://ri.conicet.gov.ar/handle/11336/87911

Kane Frank. (2017). *Hands-On Data Science and Python Machine Learning* (1.ª edición). Packt Publishing. https://www.packtpub.com

Lactanet Canada. (2025). *Lactanet Canada*. https://lactanet.ca/en/

Morales-Cardoso, S., Morales-Morales, M., Andrade-Bazurto, A., & Cevallos-Black, L. (2020a). Analítica de datos puros dentro del ámbito productivo y reproductivo de las ganaderías de leche. *Revista Arbitrada Interdisciplinaria Koinonía*, *5*(9), 287. https://doi.org/10.35381/r.k.v5i9.649

Morales-Cardoso, S., Morales-Morales, M., Andrade-Bazurto, A., & Cevallos-Black, L. (2020b). Analítica de datos puros dentro del ámbito productivo y reproductivo de las ganaderías de leche. *Revista Arbitrada Interdisciplinaria Koinonía*, *5*(9), 287. https://doi.org/10.35381/r.k.v5i9.649

Perdigón Llanes, R., & González Benítez, N. (2022). Redes neuronales artificiales en el pronóstico de la producción de leche bovina. *Revista Colombiana de Computacion*, *23*(1), 20–33. https://doi.org/10.29375/25392115.4209

Perdigon Llanes Rudibel, & Gonzales Benitez Neilys. (2020). Una revisión bibliográfica sobre modelos para predecir las producciones de leche. *Revista Ingeniería Agrícola, Vol. 10, No. 4*(2306). https://doi.org/10.13140/RG.2.2.28326.32325

Raschka Sebastian, & Mirjalili Vahid. (2017). *Python machine learning : machine learning and deep learning with Python, scikit-learn, and TensorFlow* (2.ª edición). Packt Publishing. https://www.packtpub.com

Reglamento (UE) 2016/1012 Del Parlamento Europeo y Del Consejo, de 8 de Junio de 2016, Pub. L. No. Reglamento (UE) 2016/1012 (2016). https://eur-lex.europa.eu/legal-content/ES/LSU/?uri=CELEX:32016R1012

Ríos-Utrera, Á., Calderón-Robles, C., Reyes Galavíz-Rodríguez, J., & Vega-Murillo, V. E. (2012). Análisis genético de la producción láctea de vacas Holstein y Pardo Suizo en pastoreo intensivo en condiciones subtropicales. *Revista Científica, FCV-LUZ, XXII Nº 6*, 545–552. https://www.redalyc.org/pdf/959/95925106008.pdf

Sarkar, U., Bannerjee, G., Das, S., & Ghosh, I. (2018). Artificial Intelligence in Agriculture: A Literature Survey. *International Journal of Scientific Research in Computer Science Applications and Management Studies IJSRCSAMS, 7*(3). https://www.researchgate.net/profile/Gouravmoy-Banerjee/publication/326057794_Artificial_Intelligence_in_Agriculture_A_Literature_Survey/links/5b35ab970f7e9b0df5d83ec6/Artificial-Intelligence-in-Agriculture-A-Literature-Survey.pdf

Sharma Shivangi, Verma Kirti, & Hardaha Palak. (2023). Implementation of Artificial Intelligence in Agriculture. *Journal of Computational and Cognitive Engineering*, *2*(2), 155–162. https://doi.org/10.47852/bonviewJCCE2202174

Spósitto Osvaldo, Blanco Gabriel, Levi Marcelo, Corral Patricio Macías, & Matteo Lorena. (2019). Peso al nacer de terneros Aberdeen Angus mediante algoritmos no supervisados. *Conference: CONAIISI*. https://www.researchgate.net/profile/Lorena-Matteo/publication/337445353_Peso_al_Nacer_de_Terneros_Aberdeen_Angus_mediante_Algoritmos_No_Supervisados/links/5dd7f6afa6fdccdb445a08e0/Peso-al-Nacer-de-Terneros-Aberdeen-Angus-mediante-Algoritmos-No-Supervisados.pdf

Sposito Osvaldo, Blanco Gabriel, & Matteo Lorena. (2020). *TECNICAS DE PREPROCESAMIENTO DE DATOS EN MODELOS NO SUPERVISADOS APLICADOS AL ESTUDIO GENETICO DE LA RAZA ABERDEEN ANGUS*. 2525–1333. http://reddi.unlam.edu.arPág:1Artículooriginal

Zea, A., Bermúdez, D., Jiménez, A., Gómez, G., & Martínez, C. A. (2022). Predicción de la producción diaria de leche en bovinos Gyr a través de métodos de aprendizaje supervisado. *Comunicaciones En Estadística*, *15*(1), 35–47. https://dialnet.unirioja.es/servlet/articulo?codigo=8710101

Apéndices

Apéndice A

Estadísticas de Padres Genéticos en Vacas Tipo Embrión

Padre	N° lactancias	N° Animales	Media Leche día/IEP	desviación	mínimo	máximo	Ranking
FEVER	86	20	12.86	3.88	4.24	35.79	1
BIG POINT	13	13	12.80	4.22	8.58	23.85	2
SAVARD	16	4	12.63	2.42	9.37	17.19	3
AFTERSHOCK	40	11	12.50	2.98	5.34	19.49	4
LAVANGUARD	80	19	12.44	2.40	6.04	18.34	5
SID	43	7	12.23	3.38	3.46	22.62	6
METEORO	18	2	12.21	2.06	7.38	16.7	7
BENFEITOR	22	3	12.13	3.83	2.95	17.83	8
GABOR	31	6	12.12	2.14	7.98	15.98	9
SHADOW	41	8	12.06	3.18	4.96	20.8	10
GILLESPY	106	34	11.68	3.98	5.47	38.91	11
SALOON	51	16	11.62	2.94	5.8	18.94	12
JIVES	34	13	11.28	2.25	5.47	15.69	13
JEDI	49	34	11.23	3.09	7.02	23.1	14
GOLDWIN	15	3	11.13	2.42	6.81	15.38	15
PARADICE	49	13	10.97	3.06	3.51	17.24	16
MOGUL	37	17	10.92	2.87	5.15	18.63	17
DOORMAN	42	15	10.76	4.07	3.12	25.71	18
MONTROSS	136	71	10.46	2.92	4.14	20.77	19

Apéndice B

Estadísticas de Madres Genéticas en Vacas Tipo Embrión

Madre genética	N° Lactancias	N° Animales	Media Leche día/IEP	desviación	mínimo	máximo	Ranking
1137HO	8	1	15.33	2.94	8.88	17.83	1
1533/46GYR	8	4	13.50	4.15	10.08	22.85	2
136/7	13	3	13.06	3.73	6.62	22.62	3
1683/18GYR	5	5	13.06	6.11	9.45	23.85	4
1467/55GYR	16	10	12.46	3.72	8.09	20.73	5
1559HO	9	1	12.30	2.23	7.38	14.55	6
224/7	60	16	12.29	3.38	4.57	25.71	7
194-14	6	3	12.21	5.51	8.36	23.1	8
630HO	9	1	12.12	2.01	9.92	16.7	9
1390/54GYR	14	8	12.12	2.62	8.42	18.63	10
382/8	32	7	12.04	3.25	5.34	19.49	11
933/4	54	12	12.00	3.27	3.46	22.37	12
294/7GY	145	46	11.92	2.57	4.71	20.77	13
214/7	88	29	11.89	3.16	3.51	20.51	14
330/7	61	17	11.74	4.09	5.15	35.79	15
1286/2	78	28	11.73	2.61	6.39	19.36	16
961HO	8	1	11.71	2.26	9.59	15.3	17
1428/15GY	40	21	11.61	5.22	4.14	38.91	18
941/4	30	11	11.53	2.99	5.91	18.65	19
303-16	9	5	11.31	1.11	9.27	13.24	20
1319/13GY	77	31	10.89	3.21	3.12	17.27	21
294-16	14	6	10.88	1.71	7.92	13	22
1282/2GY	29	10	10.87	3.52	6.2	20.8	23
1360/4GY	69	26	10.54	2.68	5.53	16.38	24
306	6	3	9.78	3.85	5.93	16.79	25
1275/41BR	15	3	9.74	2.76	4.96	15.48	26
831HO	6	1	8.41	3.03	2.95	11.28	27

Apéndice C

Agrupamiento de Padres Genéticos Según Clúster K-means

Padre	N° Lactancias	Promedio Leche día/IEP	clúster
FEVER	86	12.857093	0
BIG POINT	13	12.801538	0
SAVARD	16	12.625000	0
AFTERSHOCK	40	12.498250	0
LAVANGUARD	80	12.442250	0
SID	43	12.232791	0
METEORO	18	12.214444	0
BENFEITOR	22	12.125455	0
GABOR	31	12.118387	0
SHADOW	41	12.064634	0
SALOON	51	11.622353	1
JIVES	34	11.281471	1
JEDI	49	11.232041	1
GOLDWIN	15	11.128667	1
PARADICE	49	10.969184	1
MOGUL	37	10.918378	1
DOORMAN	42	10.758095	1
GILLESPY	106	11.676415	2
MONTROSS	136	10.455294	2

Apéndice D

Agrupamiento de Madres Genéticas Según Clúster K-means

Madre genética	N° Lactancias	Promedio Leche día/IEP	Clúster
1137HO	8	15.328750	0
1533/46GYR	8	13.501250	0
136/7	13	13.063846	0
1683/18GYR	5	13.058000	0
1467/55GYR	16	12.457500	0
1559HO	9	12.304444	0
194-14	6	12.211667	0
630HO	9	12.124444	0
1390/54GYR	14	12.115714	0
382/8	32	12.042813	0
961HO	8	11.711250	0
1428/15GY	40	11.610000	0
941/4	30	11.534667	0
303-16	9	11.313333	0
224/7	60	12.290000	1
933/4	54	12.000000	1
294/7GY	145	11.920138	1
214/7	88	11.892727	1
330/7	61	11.742951	1
1286/2	78	11.731410	1
1319/13GY	77	10.885325	1
1360/4GY	69	10.540435	1
294-16	14	10.876429	2
1282/2GY	29	10.866552	2
306	6	9.776667	2
1275/41BR	15	9.738000	2
831HO	6	8.406667	2

Apéndice E

Importancia Relativa de Padres Genéticos en el Modelo Random Forest

Padre	Importancia
GILLESPY	0.043922
FEVER	0.038725
DOORMAN	0.025886
MONTROSS	0.018752
SID	0.012320
MOGUL	0.011920
SALOON	0.011499
LAVANGUARD	0.011444
SAVARD	0.010933
PARADICE	0.010870
GOLDWIN	0.007126
GABOR	0.006636
JIVES	0.005614
BENFEITOR	0.005392
SHADOW	0.004923
JEDI	0.003342
BIG POINT	0.002481
METEORO	0.001423

Apéndice F

Importancia Relativa de Madres Genéticas en el Modelo Random Forest

Madre genética	Importancia
1428/15GY	0.031006
330/7	0.028736
214/7	0.024834
294/7GY	0.024619
1286/2	0.024250
224/7	0.023378
1275/41BR	0.021860
1360/4GY	0.021247
933/4	0.020738
382/8	0.018590
1319/13GY	0.017149
194-14	0.013770
1282/2GY	0.012858
1533/46GYR	0.012469
941/4	0.012393
136/7	0.011577
831HO	0.010317
1467/55GYR	0.009202
1390/54GYR	0.008826
1683/18GYR	0.006439
306	0.004326
961HO	0.002473
303-16	0.002200
1559HO	0.001575
294-16	0.001068
630HO	0.000755