



UNIVERSIDAD AUTÓNOMA CHAPINGO
DEPARTAMENTO DE ENSEÑANZA, INVESTIGACIÓN
Y SERVICIO EN ZOOTECNIA
POSGRADO EN PRODUCCIÓN ANIMAL

COMPARACIÓN DE MÉTODOS DE APRENDIZAJE AUTOMÁTICO
PARA PREDECIR VALORES GENÓMICOS EN CARACTERÍSTICAS
DE CRECIMIENTO EN BOVINOS

TESIS

Que como requisito parcial
para obtener el grado de:

MAESTRO EN CIENCIAS EN INNOVACIÓN GANADERA

Presenta:

JOSÉ LUIS VELEZ LABRADA

Bajo la supervisión de: **Agustín Ruíz Flores, Ph.D.**
Paulino Pérez Rodríguez, Dr.



Chapingo, Estado de México, enero de 2024

**COMPARACIÓN DE MÉTODOS DE APRENDIZAJE
AUTOMÁTICO PARA PREDECIR VALORES GENÓMICOS EN
CARACTERÍSTICAS DE CRECIMIENTO EN BOVINOS**

Tesis realizada por **JOSÉ LUIS VELEZ LABRADA** bajo la supervisión del Comité Asesor indicado, aprobada por el mismo y aceptada como requisito parcial para obtener el grado de:

MAESTRO EN CIENCIAS EN INNOVACIÓN GANADERA

DIRECTOR:



Ph. D. Agustín Ruíz Flores

CO-DIRECTOR:



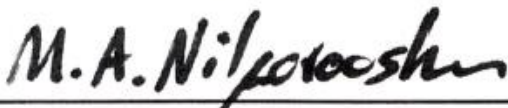
Dr. Paulino Pérez Rodríguez

ASESOR:



Dr. Luis Alberto Miranda Romero

ASESOR:



Ph. D. Mohammad Ali Nilforooshan

CONTENIDO

LISTA DE CUADROS	v
LISTA DE FIGURAS	vi
DEDICATORIAS	vii
AGRADECIMIENTOS	viii
DATOS BIOGRÁFICOS	ix
1. INTRODUCCIÓN GENERAL	1
2. REVISIÓN DE LITERATURA	3
2.1 Diversidad genética	3
2.1.1 Genotipo animal	3
2.1.2 Fenotipo animal	4
2.2 Selección genómica	5
2.2.1 Antecedentes	6
2.2.2 Ventajas de la selección genómica	6
2.2.3 Desventajas de la selección genómica	8
2.3 Aplicación de la genómica en el ganado	8
2.3.1 Polimorfismos de un solo nucleótido (SNP)	9
2.3.2 Selección asistida en bovinos a través de marcadores moleculares ..	10

2.4 Características de crecimiento en bovinos.....	12
2.5 Métodos de aprendizaje automático	14
2.5.1 Árboles de clasificación.....	16
2.5.2 Árboles de regresión	18
2.5.3 Bosques aleatorios	20
2.5.4 Redes neuronales	22
2.5.5 Redes neuronales de una capa	23
2.5.6 Redes neuronales multicapa.....	24
2.5.7 Máquina de soporte vectorial	27
2.6 Entrenamiento de los modelos.....	29
2.7 Validación de los modelos	30
2.8 Aplicaciones de algoritmos de aprendizaje automático en la ganadería..	31
2.9 Literatura citada	32
3. COMPARACIÓN DE MÉTODOS DE APRENDIZAJE AUTOMÁTICO PARA PREDICCIÓN DE VALORES DE CRÍA GENÓMICOS EN CARACTERÍSTICAS DE CRECIMIENTO EN BOVINOS.....	50
3.1 Resumen.....	50
3.2 Abstract.....	51
3.3 Introducción	51

3.4 Materiales y Métodos	52
3.4.1 Información fenotípica y genotípica	53
3.4.2 Modelos mixtos lineales (modelos base)	53
3.4.3 Modelos de aprendizaje automático	55
3.4.4 Red neuronal artificial	56
3.4.5 Árboles de regresión	56
3.4.6 Bosques <i>aleatorios</i>	57
3.4.7 Máquina de soporte vectorial	58
3.4.8 Validación <i>cruzada</i>	58
3.4.9 Software y lenguaje de programación.....	60
3.4.10 Ajuste de <i>los</i> modelos.....	60
3.5 Resultados	61
3.6 Discusión	63
3.7 Conclusiones	63
3.8 Literatura citada	65

LISTA DE CUADROS

Número	Título	Página
1	Número de animales genotipados y fenotipados de una población de bovinos Suizo Europeo para tres características de crecimiento	55
2	Promedios de los estimadores de la correlación Pearson y desviación estándar entre fenotipos corregidos y valores genéticos predichos, para las tres características de crecimiento y los algoritmos comparados	62

LISTA DE FIGURAS

Número	Título	Página
1	Diagrama de flujo de la información en un árbol de clasificación	17
2	Esquema de un árbol de regresión, dentro de los algoritmos de aprendizaje automático	19
3	Diagrama representativo del flujo de información en el algoritmo bosques aleatorios	21
4	Esquema de una red neuronal de una capa oculta con s-número de neuronas y una salida	24
5	Esquema de una red neuronal de 2 capas ocultas y s-número de neuronas	26
6	Representación en el plano de algunas funciones de la máquina de soporte vectorial en la división de clases	28
7	Comparación de los coeficientes de correlación (promedio de las 100 validaciones) de los fenotipos corregidos y valores genéticos predichos obtenidos con los métodos de aprendizaje automático, redes neuronales artificiales (RN), arboles de regresión (AR), bosques aleatorios (BA) y máquina soporte vectorial (SVM) con las metodologías realizadas por Valerio-Hernández et al.(15), mejor predictor lineal insesgado (BLUP), BLUP genómico (GBLUP) y GBLUP de un solo paso (ssGBLUP) para las características peso al nacer (PN), peso al destete (PD) y peso al año (PA) de una población de bovinos suizo europeo.	71

DEDICATORIAS

Al Gran Arquitecto Del Universo quien me ha guiado en todas las decisiones que me han llevado hasta este momento.

Con gratitud hacia quienes iluminaron mi camino académico, este logro es un tributo a la perseverancia, al conocimiento compartido y al inquebrantable apoyo. A mi esposa, familia, mentores y seres queridos, gracias por ser los pilares de este reto.

A quienes ya no están con nosotros, y que me hubiera encantado que vieran la culminación de esta etapa.

“Lo que conocemos es una gota, lo que no conocemos es un océano”

Isaac Newton (1642 – 1727)

AGRADECIMIENTOS

Quiero expresar mi profunda gratitud a todas las personas que han contribuido de manera significativa a la realización de este trabajo. Esta etapa académica ha sido una experiencia enriquecedora, y no habría sido posible sin el apoyo y la colaboración de diversas personas.

En primer lugar, quiero agradecer al Dr. Agustín Ruíz, al Dr. Paulino Pérez y al Dr. Mohammad por su orientación, paciencia y dedicación. Sus valiosos consejos y comentarios han sido fundamentales para dar forma a este proyecto y para mi desarrollo académico, en verdad agradezco su amistad y guía.

Agradezco sinceramente a todos mis profesores que compartieron sus conocimientos y experiencias, y quienes fueron la razón del regresar a la Universidad; Dr. Miranda, Dr. Maximino, Dr. Rafael, Dr. Rodolfo, Dr. Altamirano, M.C. Carlos proporcionaron una base sólida para la investigación y mi desarrollo como profesional.

Quiero reconocer el apoyo financiero brindado por Consejo Nacional de Humanidades, Ciencias y Tecnologías para realizar mis estudios de maestría. Así como a la Asociación Mexicana de Criadores de Ganado Suizo de Registro, por brindar la información base de este estudio. Su respaldo ha sido fundamental para la realización de esta investigación.

Agradezco al Dr. Jonathan por su apoyo y tiempo para resolver dudas, a mis compañeros del posgrado y colegas por su colaboración, discusiones estimulantes y amistad. Sus perspectivas han contribuido significativamente al desarrollo de mis ideas.

A todas las personas que, de una forma u otra, han sido parte de este viaje, les expreso mi agradecimiento.

Gracias.

DATOS BIOGRÁFICOS



Datos personales

Nombre:	José Luis Velez Labrada
Fecha de nacimiento:	04 de diciembre de 1992
Lugar de nacimiento:	México, Distrito Federal
CURP:	VELL921204HDFLBS00
Profesión:	Ingeniero Agrónomo especialista en Zootecnia
Cédula profesional:	12582272

Desarrollo académico

Maestría en Ciencias (2022-2023)	DEIS en Zootecnia Posgrado en Producción Animal Universidad Autónoma Chapingo
Licenciatura Ingeniero Agrónomo Especialista en Zootecnia (2014-2018)	DEIS en Zootecnia Universidad Autónoma Chapingo

1. INTRODUCCIÓN GENERAL

Uno de los usos de la genómica en producción animal es predecir el valor genético de los individuos mediante marcadores moleculares. Esta técnica es una herramienta ampliamente utilizada recientemente en programas de mejoramiento genético, que, en combinación con los modelos tradicionales de evaluación genética, busca incrementar la tasa de avance genético de manera eficaz mediante la selección de animales reproductores. La modelación con dichos marcadores ha permitido mejorar eficientemente y en menor tiempo características de importancia económica y cuya medición es compleja, cara o no repetible.

En las unidades de producción bovina se procura llevar registros de caracteres importantes desde el comienzo de la vida del animal, como pesos al nacimiento, destete y a la venta; ocasionalmente de rendimiento en canal, de la calidad y composición de la carne. Estas características están influenciadas por la interacción de diversos genes, que en combinación con un ambiente favorable brindarán individuos genéticamente superiores (Henderson, 1984).

En la producción de carne bovina, las características de crecimiento como pesos al nacimiento y al destete son relevantes, la primera incide directamente en la sobrevivencia del animal y la segunda muestra el potencial de crecimiento que éste tendrá. Ambas características son importantes económicamente y son fundamentales en la selección de los reproductores de la siguiente generación para este objetivo de producción. En México, se han realizado estudios acerca del tema. Sifuentes et al. (2020) mencionan que, debido a su impacto económico en la producción de ganado, así como la facilidad de registro, estas características se han estudiado con el enfoque de la genética tradicional y actualmente con el genómico. Jahuey-Martínez et al. (2020) evaluaron la precisión de los valores genómicos obtenidos mediante distintos métodos de regresión para características de crecimiento en ganado Charolais, denotando limitaciones en la estructura poblacional y número de animales utilizados en el estudio, así como su relación con la precisión y confianza en estudios genómicos.

Las características de crecimiento están controladas por varios loci (Lu et al., 2013; Martínez et al., 2017), los cuales interactúan entre ellos. A través de marcadores genéticos, como los SNP (polimorfismos de un solo nucleótido), se generan cientos de miles de datos por individuo, lo que permite buscar asociaciones de dichos marcadores con las características de interés. Buzanskas et al. (2014) enfatizaron la importancia de localizar regiones del genoma que deben estudiarse para identificar qué genes actúan en la expresión de determinado fenotipo.

La cantidad de información obtenida mediante herramientas genómicas vuelve complejo su análisis. Rodríguez-Alcántar (2019) estudió herramientas de aprendizaje automático, como árboles de decisión y redes neuronales artificiales, las cuales permiten analizar con más efectividad la información del genoma bovino, implementando herramientas predictivas para características de importancia económica en ganado bovino lechero.

Con los métodos de aprendizaje automático se pueden construir modelos genotipo-fenotipo a través de patrones genéticos, que también proporcionarán predicciones fenotípicas (Okser et al., 2014). El objetivo del presente estudio fue comparar métodos de aprendizaje automático para la predicción de valores de cría genómicos de peso al nacer, al destete y al año en bovinos Suizo Europeo, utilizando registros fenotípicos e información genómica de los animales.

Este documento de tesis está compuesto por tres capítulos. En el Capítulo 1 se presenta brevemente el tema general de la tesis. El Capítulo 2 contiene la Revisión de Literatura general sobre el tema destacándose diversos aspectos relacionados con Genómica y el Aprendizaje Automático. En el Capítulo 3 se presenta un manuscrito derivado de la investigación desarrollada para el programa de Maestría en Ciencias del autor, enviado para su evaluación y posible publicación a la Revista Mexicana de Ciencias Pecuarias.

2. REVISIÓN DE LITERATURA

2.1. Diversidad genética

El concepto de diversidad genética en mejoramiento animal es importante, permite a los ganaderos desarrollar o mejorar características en respuesta a cambios en el ambiente, por enfermedades o condiciones de mercado (Maudet et al., 2002). Los estudios de diversidad genética permiten identificar caracteres de una población, similitudes o diferencias entre especies y variedades; ha permitido seleccionar, clasificar, resguardar y utilizar individuos sobresalientes que aporten una mejora en la siguiente generación (Cortés, 2008).

La diversidad genética es la base para que los productores establezcan programas de mejoramiento en sus animales, los cuales pueden contar con características deseables de adaptación para cumplir con sus objetivos productivos (FAO, 2016).

Es fundamental tener poblaciones bovinas en las que se mantenga diversidad genética en cuestión de adaptación a cambios o situaciones ambientales particulares. En razas especializadas para objetivos productivos, se han encontrado tendencias hacia la disminución del tamaño efectivo de población y acumulación de consanguinidad mediante el análisis genómico (Lozada-Soto et al., 2022). Por tanto, es una prioridad desarrollar planes de respaldo y preservación de recursos genéticos. Los polimorfismos de un solo nucleótido (SNP) son los marcadores elegidos para muchos estudios con estos objetivos (Ramos et al., 2011).

2.1.1. Genotipo animal

El concepto fue introducido por Johannsen (1911), quien propuso que el genotipo de un animal es la combinación de los genes que determina la expresión de sus características. Estas características son la apariencia y el comportamiento del animal, así como su susceptibilidad. Además, los genes se heredan de los padres y se transmiten a las generaciones sucesivas.

La combinación de genes se modifica a través de las generaciones mediante selección artificial. Esto implica manipulación selectiva de los individuos que se reproducen para obtener mejores resultados a través de la diferenciación de los animales. Esta selección ha permitido a los animales obtener mayor resistencia a enfermedades, tolerancia a altas temperaturas y rendimiento en el trabajo (Bickhart & Liu, 2014).

La identificación de los genes de un individuo para ciertas características se realiza a través de una combinación de pruebas como la secuenciación de ADN (Mielczarek et al., 2017), la reacción en cadena de polimerasa (PCR, por sus siglas en inglés) y la genotipificación. Estas pruebas permiten conocer los genes presentes, cómo están organizados y relacionados, esto proporciona información de la estructura genética de características del animal.

Los resultados de pruebas sobre el genotipo animal pueden ser útiles a los criadores al seleccionar los mejores animales para la producción. Un aspecto importante de la mejora genética es la respuesta a la selección, que depende linealmente de la precisión con la que se realiza. La precisión de la selección tradicional depende de la cantidad de información fenotípica del animal y sus parientes, así como del desempeño del propio animal y de la heredabilidad del rasgo considerado (Pszczola et al., 2012).

2.1.2. Fenotipo animal

El fenotipo es el conjunto de características observables de un individuo, resultado de la interacción de sus genes y el ambiente. Estas características pueden incluir tamaño, forma, color, comportamiento y otras características físicas (Núñez-Torres & Almeida-Secaira, 2022). En Biología el término se usa para describir el modo en que los genes y el ambiente interactúan para producir una característica definida (Dorado et al., 2021).

La selección natural es uno de los principales factores que influyen en los cambios de los fenotipos animales; los individuos con mejores fenotipos tienen

más probabilidades de sobrevivir y reproducirse con éxito, transmitiendo sus genes a la próxima generación. Esta selección natural contribuye a los cambios en los fenotipos animales a lo largo del tiempo (Sarmiento, 2009).

2.2. Selección genómica

El avance en el conocimiento del genoma de las especies domésticas, al identificar genes o regiones cromosómicas asociadas con caracteres de importancia económica, evolucionó a la par de la genómica humana, creciendo exponencialmente a partir de la inclusión de marcadores moleculares (Liao et al., 2013). Esto permitió avanzar en el desarrollo de mapas genómicos comparativos de alta resolución (Band et al., 2000).

El proyecto de secuenciación en el ganado bovino se realizó en la raza Hereford (Burt, 2009). En 2009 se publicaron los resultados del análisis genómico de sementales dirigidos a la producción de leche, esto se considera el avance más importante después del congelamiento del semen.

La selección genómica es una técnica que permite identificar y con ello seleccionar organismos con base en la secuenciación y análisis del genoma, que como se mencionó permite la identificación de genes y variantes de estos relacionados con características de interés. Baizabal y Sánchez (2010) comentan que la investigación genómica tiene como uno de sus objetivos, impulsar el cambio genético en los animales, disminuyendo el intervalo generacional, para mantener y desarrollar una ganadería competitiva a través del uso de animales jóvenes probados generando descendencia con caracteres altamente productivos.

La genotipificación es la técnica molecular que permite determinar los diferentes alelos presentes en los loci de un individuo. Posteriormente puede determinarse cuales de estos alelos se relacionan con las diferentes características de interés económico como peso al nacimiento, peso al destete, producción de leche, ganancia de peso, que tienen propiedades que se reflejarán en el fenotipo del

individuo (Sourdis, 2008), generando así información complementaria para los criadores y productores de ganado. La evaluación genética permite identificar a los animales sobresalientes portadores de alelos o genotipos deseables (Goddard & Hayes, 2009) y a partir de esto ubicar reemplazos que aumenten la productividad en los hatos (Goddard, 2012).

2.2.1. Antecedentes

La selección genómica tiene su origen en la selección asistida por marcadores. Esta última asocia información de marcadores y fenotipos. Entre los primeros intentos que se realizaron fue identificar loci asociados con características cuantitativas (QTL, por sus siglas en inglés). Sin embargo, esta asociación no siempre rindió los mejores resultados, principalmente por la interacción QTL-ambiente (Montesinos et al., 2022). Posteriormente, se desarrolló el mapeo de asociación con el objetivo de superar las deficiencias de poder del análisis de ligamiento. Esto, sin embargo, tampoco pudo superar las dificultades que significaron el detectar variantes raras asociadas con características de importancia económica. Finalmente, esto permitió el desarrollo de la GS en dos etapas (Meuwissen et al., 2001): 1) el entrenamiento de un modelo en un conjunto de datos utilizando la información de todos los marcadores moleculares disponibles; y 2) el uso de este modelo para predecir los valores genéticos de animales genotipados, no incluidos en el conjunto de entrenamiento del modelo.

Un desarrollo posterior permitió evaluar en una sola etapa una población sin disponer los genotipos de todos los individuos. Esta metodología se denomina evaluación genómica en una sola etapa (Miszta et al., 2009). Este método permite la inclusión de animales genotipados o no, usando la información genotípica y de pedigrí simultáneamente (Valerio-Hernández et al., 2023).

2.2.2. Ventajas de la selección genómica

La selección genómica es una herramienta cada vez más utilizada puesto que el avance de los conocimientos científicos ha llevado a una mayor comprensión de

la genética y de la información que puede obtenerse del ADN. Ahora el ganadero puede seleccionar los individuos con los mejores genes, para aumentar la probabilidad de que la progenie de ellos sea más productiva, saludable y resistente a enfermedades.

Esta técnica tiene múltiples ventajas:

- Permite una selección precisa de los individuos con la mejor aptitud en cuanto a calidad y rendimiento del producto objeto de producción.
- Reduce la tasa de consanguinidad por generación al permitir la identificación precisa de los mejores individuos en lugar de las mejores familias (Daetwyler et al., 2007).
- Reduce costos de producción, al no invertir tiempo y recursos en la cría de animales improductivos o que no puedan brindar un producto con las características deseadas por el mercado.
- Mejora la resistencia a enfermedades y plagas, se pueden seleccionar individuos con mayor resistencia a enfermedades o que sean portadores de genes que puedan expresar características no deseables que conduzcan a anomalías estructurales y funcionales (May et al., 2019).
- Reduce el tiempo del ciclo de cría, esto permite obtener resultados más rápido; esto se ve como mejora genética (de Albuquerque et al., 2017) en las unidades de producción animal.

Es así como, con el objetivo de satisfacer la demanda mundial de carne bovina, estrategias como la selección genómica, son una herramienta para identificar animales que se adapten a condiciones ambientales difíciles y que tengan resistencia a enfermedades, para esto, es necesario generar iniciativas para rediseñar esquemas de cría (Fernandes-Júnior et al., 2022). Esta estrategia brinda una mayor precisión de selección para individuos jóvenes, permitiendo así reducir drásticamente los intervalos de generación, acelerando el progreso genético, siendo esto relevante para rasgos de importancia económica que no se registran rutinariamente (Meuwissen et al., 2013).

2.2.3. Desventajas de la selección genómica

A pesar de los múltiples beneficios de la selección genómica, hay desventajas que deben considerarse. Una de las principales es el impacto sobre la diversidad genética, puesto que seleccionar a los individuos con los mejores genes, se corre el riesgo de reducir las formas en las que se presentan los alelos. Por esta razón los animales seleccionados pueden verse afectados en otras características como fertilidad, tolerancia a enfermedades, entre otras, lo que podría afectar negativamente a toda la población (Bjelland et al., 2013).

La selección genómica puede ser una práctica costosa, ya que requiere de equipos sofisticados, tiempo de investigación, personal capacitado y el costo de los análisis del ADN. Además, se debe abundar en el estudio de algunos rasgos que pudieran no ser hereditarios y de alto valor para realizar los análisis de la información genética de los individuos (Calus & Veerkamp, 2011). Se requiere tener la mayor cantidad de registros de los animales, puesto que a la par de las evaluaciones tradicionales debe considerarse toda la información disponible para reducir el sesgo causado por la preselección (Meuwissen et al., 2016).

2.3. Aplicación de la genómica en el ganado

La genómica aplicada al ganado ha permitido un mejor conocimiento de la variabilidad genética entre los diferentes animales. Esto ha permitido desarrollar herramientas para seleccionar animales con mayor probabilidad de tener caracteres adecuados para la producción.

La selección genómica ha resultado en índices rápidos de ganancias sobre todo en ganado bovino lechero, con la peculiaridad de que un mayor número de toros activos para la aplicación de inseminación artificial son jóvenes y fueron genómicamente probados en los EE. UU. (Hutchison et al., 2014). Es así como los sistemas de evaluación genética proporcionan una base sólida por el éxito de la selección genómica generando a la par estructuras de mejoramiento bien desarrolladas, particularmente para empresas de cría (Mrode et al., 2019).

La genómica es útil para seleccionar características de bajo índice de herencia (Van Eenennaam, 2006), y en las que solamente hay una medición, como las involucradas en la calidad de la canal y rendimiento en cortes (Goddard, 2009). Esta metodología de evaluación tendrá precisión y validez cuando se acompañe de los registros productivos de las características de importancia económica (Page et al., 2004).

Wiggans y Carrillo (2022) mencionan que la selección genómica revolucionó la ganadería bovina lechera, duplicando la tasa de ganancia genética, principalmente al reducir sustancialmente el intervalo generacional. El Consejo de Cría de Ganado Lechero (2022), reporta que el número de genotipos en la colección aumentó rápidamente y alcanzó 6.6 millones en agosto de 2022 respecto a 1.1 millones existentes en 2021.

La industria bovina lechera es un claro ejemplo del poder y la escala de la mejora genética mediante la evaluación genómica (Scheifers & Weigel, 2012; Meuwissen et al., 2016), ya que se produce más de 80 % de leche con 65 % menos de vacas (Mueller & Van Eenennaam, 2022). La producción de leche es cuatro veces mayor a la de mediados del siglo XX, pasando de 2,000 a 10,000 kg vaca⁻¹ lactancia⁻¹ (Capper & Cady 2019). Esta mejora genética puede aplicarse a características de difícil medición como alimentación, eficiencia, longevidad y fertilidad, teniendo el potencial reducir la huella ambiental por unidad de producto (Fennessy et al. 2019).

2.3.1. Polimorfismos de un solo nucleótido (SNP)

Los SNP son variaciones del código genético ubicados en un solo nucleótido, se consideran una de las formas más comunes de modificaciones puesto que se sustituye por otro en una determinada posición en la secuencia de ADN, pueden estar relacionados con genes asociados con la respuesta cualitativa y cuantitativa del fenotipo (Vallejos-Vidal et al., 2020).

La variación de origen genético en las poblaciones se presenta en dos formas, aunque existen SNP multialélicos, los del primer tipo o bialélicos requieren una frecuencia mayor que 1 % en la población (Burrow et al., 2001). Estos pueden contribuir a cambios en la secuencia genómica, ya sea en la región codificante (exones), intergénica o no codificante (intrones) (Ahmad et al., 2018).

El cambio suele ocurrir en los gametos y puede heredarse a uno o más descendientes en las siguientes generaciones (Yu et al., 2005). Los SNP que se heredan en grupo se les denomina haplotipos, siendo una combinación de distintos loci de un cromosoma, locus, varios loci o un cromosoma completo (Sunnucks, 2001).

Pszczola et al. (2012) mencionan que los SNP están en desequilibrio de ligamiento (LD) con el gen o el loci de carácter cuantitativo (QTL) en el genoma. Por tanto, el uso de marcadores SNP permite que todos los QTL en el genoma se identifiquen indirectamente a través de SNP adyacentes.

Srinivasan et al. (2016) consideran los marcadores SNP como los más útiles para identificar caracteres de interés, como enfermedades, debido a su frecuencia común, facilidad de análisis, bajo costo de genotipado y la posibilidad de realizar estudios de asociación basados en herramientas estadísticas y bioinformáticas.

2.3.2. Selección asistida en bovinos a través de marcadores moleculares

La selección asistida por marcadores (MAS) se basa en marcadores moleculares en equilibrio de enlace con loci de rasgos cuantitativos (QTL) (LE-MAS), marcadores moleculares en desequilibrio de enlace con QTL (LD-MAS), o en la selección de las mutaciones reales que causan el efecto QTL (Gen-MAS) (Hayes et al., 2007). Los LD-MAS son los que actualmente se utilizan de mayor manera en la industria ganadera (Pszczola et al., 2012).

La implementación de la MAS implica estimar los efectos de los marcadores como los SNP en una población de referencia, con registros fenotípicos y genotipos, seguido por la predicción de valores genómicos (GEBV) para

candidatos de selección sin fenotipos propios (Meuwissen et al., 2001). Gracias a la secuenciación del genoma de los animales domésticos y al desarrollo e investigación de marcadores tipo SNP, es posible realizar estudios en los animales con marcadores que resultan significativos en la asociación con QTL (Montesinos et al., 2022).

El incluir en análisis las relaciones genómicas permite capturar el muestreo mendeliano y revela vínculos entre animales aparentemente desconocidos o no emparentados vía pedigrí (Yang et al., 2010).

VanRaden (2008) menciona que a través de la inclusión de la matriz de relaciones genómicas (Matriz **G**), se puede incluir en la metodología BLUP (mejor predictor lineal insesgado), donde dicha matriz reemplaza a la de relaciones genéticas aditivas (Matriz **A**), generando así el modelo GBLUP. Otra manera de realizar el procedimiento es usar simultáneamente las matrices **A** y **G**, utilizando la información fenotípica y genómica (Miztal et al., 2009), para la evaluación genética, obteniendo los valores de cría estimados (EBV por sus siglas en inglés) y valores de cría estimados mediante genómica (GEBV por sus siglas en inglés) en un mismo procedimiento (Gengler et al., 2007).

La precisión de la MAS está determinada por parámetros como el nivel de desequilibrio de ligamiento entre marcadores o haplotipos de marcadores y la cantidad de registros fenotípicos por alelo de marcador o por haplotipo (Graves et al., 2004). Además, la mejora genética depende de la precisión de la selección, en los métodos tradicionales esta depende de la disponibilidad de información fenotípica tanto en los parientes como en el propio animal, así como la heredabilidad del carácter. Goddard (2009) menciona que la exactitud usando genómica depende de la heredabilidad, pero también factores como el tamaño de la población de referencia y su tamaño efectivo (N_e).

La MAS busca medir el efecto de los genes relacionados con características de producción directamente en la composición genética del animal, no estimarlo a partir del fenotipo, a la par integrar los dos métodos de selección, es decir,

métodos tradicionales con métodos de genética molecular que generen mejor respuesta de selección (Wakchaure & Ganguly, 2015).

2.4. Características de crecimiento en bovinos

El valor fenotípico de un individuo para una característica específica se explica como el resultado del genotipo del animal, el ambiente donde se desarrolla y la interacción entre éstos, de tal manera que los criadores pueden cambiar las propiedades genéticas de sus animales en dos formas: 1) seleccionar los animales que serán su pie de cría, y 2) decidir cómo aparearlos (Sifuentes et al., 2020).

El objetivo de los ganaderos es lograr criar animales sanos y productivos, mediante la selección adecuada de los animales para características que influyen en el objetivo de producción. Weik et al. (2022) comentan que la rentabilidad de la industria de la carne de res depende del desempeño tanto del ganado en la engorda como del hato de cría de donde provienen los animales.

Las características de crecimiento en bovinos se refieren a los atributos físicos que afectan el crecimiento y el desarrollo de los animales. Estas características incluyen el tamaño, propiedades de la canal, el metabolismo y la habilidad para alimentarse. Es así como se consideran impulsores clave para un sistema de engorda y finalización de carne (Wolcott et al., 2013).

El peso al destete es una característica importante en los sistemas de producción de carne bovina, influye directamente en la eficiencia económica de la unidad de producción por lo que puede utilizarse como criterio de selección (Montes et al., 2008). Un componente importante para evaluar la rentabilidad en la producción de carne bovina es el crecimiento de las crías, lo que implica que el óptimo desarrollo en las fases pre- y postdestete, arrojará la eficiencia productiva de la vaca. Por otro lado, las características de crecimiento por lo general presentan índices de herencia o heredabilidad considerados como valores medios y altos, lo que sugiere que pueden establecerse programas de mejoramiento genético en

poblaciones de bovinos de carne basados en este parámetro (Martínez et al., 2007).

La heredabilidad de los rasgos de crecimiento es alta, se tenía la idea contraria puesto que la recopilación de datos y el número limitado de años en los que se registraban no permitió la estimación de los efectos maternos en el rendimiento de la progenie para estas características (Koots et al., 1994), esto es relevante para peso al año, donde el impacto materno es considerablemente mayor. Wolcott et al. (2013) indican que una gran proporción del fenotipo para los rasgos de crecimiento se atribuye a diferencias en el genotipo, por lo que es probable que la selección de animales con mayor mérito genético para estos rasgos mejore el crecimiento en la próxima generación.

Khorshidi et al. (2020) mencionan que para crear una raza de carne sobresaliente en crecimiento y en uso eficientemente del alimento con el objetivo de que el ganado llegara al mercado en peso a edades más tempranas, se seleccionaron toros con registros fenotípicos de pesos al destete y al año realizando una evaluación del potencial de crecimiento. Por ello los registros de peso se recopilan suponiendo que el crecimiento es lineal a lo largo de la curva de crecimiento y se estandariza a 205 y 365 d de edad para peso al destete y al año, respectivamente (Beef Improvement Federation, 2018).

Se debe considerar que los efectos de sexo, raza, edad de la madre y edad al nacer afectan la expresión de las características de crecimiento (Dadi et al., 2002; Goyache et al., 2003; Krupa et al., 2005). Además, la eficiencia alimenticia, entendida como la relación entre el alimento consumido y el aumento de peso corporal observado (conversión alimenticia), afectan la selección para mayores tasas de crecimiento en el ganado para carne (Seabury et al., 2017), pero utilizarla para la selección y mejorar el peso al destete o al año puede conducir a mayores costos asociados con el mantenimiento demandando más nutrientes, costos de alimentación, dificultad de parto debido a pesos al nacer altos, etc. (Archer et al., 1999).

2.5. Métodos de aprendizaje automático

El aprendizaje automático (Machine Learning) es un campo de estudio de la Inteligencia Artificial en el que se diseñan algoritmos para que una computadora aprenda a partir de datos (Nichols et al., 2019). Esto implica que una computadora puede desarrollar reglas y patrones a partir de los datos para resolver problemas, sin que se le enseñe explícitamente (Staniewicz & Midgley, 2015).

Hakem et al. (2021) mencionan que las técnicas de aprendizaje automático han surgido recientemente como enfoque efectivo para el análisis orientado a la predicción. Las aplicaciones son amplias y pueden proporcionar resultados adecuados con base en los datos que se proporcionen. El mundo está creciendo a un ritmo exponencial y también los datos recopilados están siendo cada vez más significativos y relevantes, abriendo caminos para el aprendizaje automático, en particular para el aprendizaje profundo y la inteligencia artificial (Kersting, 2018).

Existen una variedad de métodos de aprendizaje automático que se pueden utilizar para abordar diferentes problemas, estos incluyen Aprendizaje Supervisado, Aprendizaje No Supervisado, Aprendizaje Profundo y Aprendizaje por Refuerzo (Rashidi et al., 2023).

Kersting (2018) comenta que el Aprendizaje Supervisado se usa para predecir resultados basados en datos de entrenamiento, implica que estos contienen etiquetas de respuesta conocidas, esto facilita a la computadora aprender el patrón entre las entradas y las etiquetas de salida.

El Aprendizaje No Supervisado, por otro lado, no se basa en etiquetas de salida. En lugar de eso, busca descubrir patrones en los datos de entrada y descubrir reglas que describan dichos datos (Harrison et al., 2021). Esto es útil para descubrir relaciones entre diferentes conjuntos de datos que de otra manera no se detectarían.

El Aprendizaje Profundo es una forma avanzada de aprendizaje automático que consiste en la creación de una red neuronal profunda para realizar tareas de aprendizaje (Ching et al., 2018). Estas redes se componen de capas de neuronas interconectadas que procesan los datos de entrada para producir una salida (Glorot et al., 2011), lo que permite que la computadora aprenda a partir de datos más complejos, rindiendo así resultados más precisos.

Finalmente, el Aprendizaje por Refuerzo se usa para programar una computadora para tomar acciones basadas en un entorno dado e implica que la computadora aprende de sus acciones para maximizar una recompensa, bien sea positiva o negativa (Moerland et al., 2022). Esto se usa para problemas que requieren toma de decisiones complejas.

Dentro del aprendizaje automático se consideran diferentes tipos de problemas, donde cada uno de ellos se resuelve por un algoritmo:

- Problema de clasificación: el objetivo de este problema es intentar predecir sobre un conjunto de clases fijas. En años recientes se ha tenido buen desempeño de los algoritmos para esta función (Tan, 2021).
- Problema de regresión: es un problema cuyo principal objetivo es predecir una variable continua. Esta es la principal diferencia con respecto al problema de clasificación, pues el valor a predecir es una magnitud numérica continua, mientras que en el problema de clasificación es una magnitud discreta. El objetivo que se persigue en el problema de regresión según Black et al. (2023) es pronosticar ese valor de modo que se minimice el error o diferencia entre el valor predicho y el real.
- Problema de agrupamiento: consiste en obtener grupos a partir de los datos. La diferencia entre los problemas de clasificación y de agrupamiento radica en que este último analiza los datos para posteriormente crear unos determinados grupos donde incluirlos (Liu et al., 2021), mientras que en el problema de clasificación los grupos o clases están predefinidos.

El crecimiento de la inteligencia artificial y del aprendizaje automático ha sido exponencial en años recientes (Jiang et al., 2023). Las aplicaciones de la predicción usando datos genómicos masivos (GWP por sus siglas en inglés) se han desarrollado enormemente, con aplicaciones variadas en agricultura y ganadería (Reinoso et al., 2021). Por tanto, se debe tener como una opción para obtener soluciones a cuestionamientos que involucran gran cantidad de información, el cual es el caso con la información de los genotipos en términos de SNP.

2.5.1. Árboles de clasificación

El árbol de clasificación es una técnica de aprendizaje automático supervisado que se utiliza para predecir el valor de una variable objetivo. Esta técnica se aplica a los datos que contienen variables predictoras (Ozcan & Peker, 2023). El árbol de clasificación se crea a partir de un conjunto de reglas de decisión que se utilizan para segmentar los datos y que se construyen usando los valores de los atributos presentes en los datos. Se compone de una raíz, ramas y nodos hoja (Ghiasi & Zendehboudi, 2019). La raíz del árbol representa la variable de respuesta, mientras que las ramas representan los atributos predictores y los nodos hoja representan el valor de la variable objetivo. Batra y Agrawal (2018) infieren que el árbol de decisión es uno de los algoritmos de aprendizaje automático más populares debido a su facilidad de interpretación, su alto grado de precisión y su capacidad de manejar problemas complejos.

En la Figura 1 se observa la división de un conjunto de datos en clases o categorías. La raíz se considera el nodo principal del árbol, se encuentra en la parte superior y ésta representa la característica o atributo más importante en el conjunto de datos que se utilizará para tomar la primera decisión de clasificación (D'Onofrio et al., 2023; Ozcan & Peker, 2023). Los nodos son puntos en el árbol donde se toman decisiones basadas en características particulares, se dividen en nodos internos y nodos hoja. Los nodos internos representan puntos de decisión donde se evalúa una característica y se decide qué rama tomar en función de ciertas condiciones. Los nodos hoja, también se conocen como nodos

terminales y representan las clases o categorías finales a las que se llega después de seguir un camino de la raíz a estos puntos (Maray et al., 2022).

Las ramas son las conexiones entre los nodos y representan el resultado de una decisión basada en una característica. Cada rama sale de un nodo y conduce a otro, ya sea interno o a una hoja. Las ramas se utilizan para guiar el flujo de los datos a través del árbol, desde la raíz hasta un nodo hoja. Las hojas son los nodos finales del árbol y representan las clases de salida o categorías en las que se clasifican los datos (Brown et al., 1996). Cada hoja tiene asignada una etiqueta que indica la clase a la que pertenecen los datos que llegan hasta ese punto en el árbol. Cuando un dato sigue un camino desde la raíz hasta una hoja, se clasifica en la categoría representada por esa hoja.

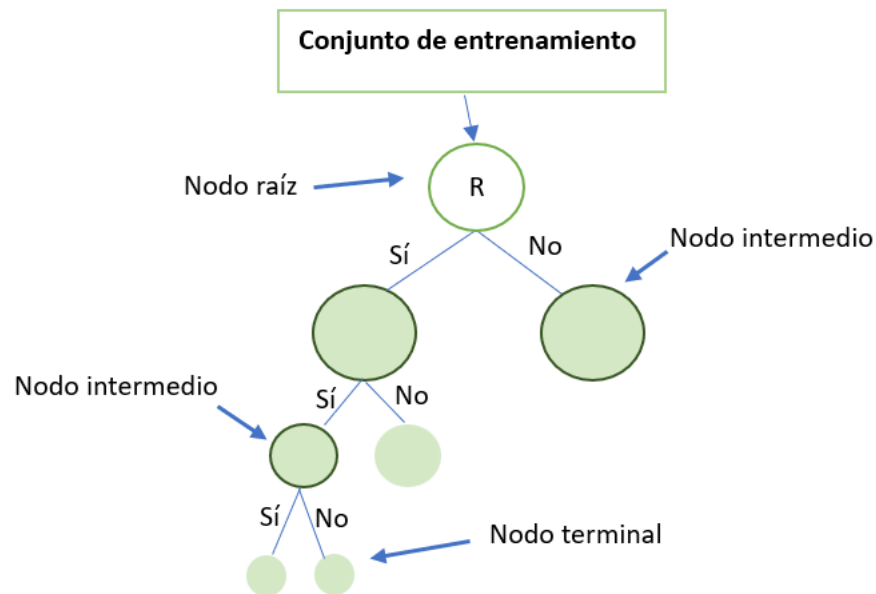


Figura 1. Diagrama de flujo de la información en un árbol de clasificación.

Esta metodología se considera eficiente en la solución de problemas de clasificación, como su nombre lo indica consiste en un árbol que no cuenta con arcos de entrada en la raíz, pero que en los nodos internos y externos existe una salida (Maimon & Rokach, 2014). Las muestras alojadas en nodos internos se separan en subconjuntos con base en un atributo, repitiéndose en cada

subconjunto, finalizando cuando el subconjunto de una muestra en un nodo tiene el mismo valor objetivo, cuando la división no mejora la clasificación o cuando la separación es imposible debido a restricciones impuestas por el usuario (Xu et al., 2005).

Una de las principales ventajas de usar árboles de clasificación es su habilidad para manejar variables medidas en una escala de razón o proporción (continuas) y categóricas simultáneamente. Además, los árboles de clasificación proporcionan una representación visual intuitiva de los datos (Xie et al., 2019). Esto hace que sea fácil comprender el patrón de los datos y los resultados de la clasificación. Los árboles de clasificación también son una excelente herramienta para manejar problemas de clasificación con muchos atributos (Song & Lu, 2015).

Sin embargo, los árboles de clasificación tienen algunos inconvenientes, por ejemplo, una tendencia a sobreajustarse a los datos de entrenamiento (Bramer, 2002). Lo que significa que pueden tener buen rendimiento en el conjunto de datos de entrenamiento, pero bajo rendimiento en los conjuntos de datos de prueba.

Kim (2016) menciona que otra manera de superar estos problemas, es utilizar otros clasificadores como la regresión logística, la máquina soporte vectorial y las redes neuronales en combinación con un árbol de decisión.

2.5.2. Árboles de regresión

Los árboles de regresión son una técnica de aprendizaje automático utilizado para predecir variables continuas. Esta técnica permite construir un modelo de predicción basado en la estructura de árbol de decisión (Loh, 2014). El árbol de regresión es una herramienta útil tanto para la regresión lineal como para la no lineal, la ventaja principal de esta técnica es que construye un árbol de decisión que hace que sea más fácil para los usuarios comprender la relación entre la variable dependiente y la independiente (Blanquero et al., 2022).

En la Figura 2 se presenta un diagrama parecido al de un árbol de clasificación, pero en lugar de predecir clases o categorías, se utiliza para predecir valores en una escala continua, sus componentes son similares. La raíz en un árbol de regresión representa lo que mejor divide el conjunto de datos en función de algún criterio (por ejemplo, la reducción de la varianza). Los nodos son puntos en el árbol donde se toman decisiones basadas en características particulares (Loh, 2014). Los nodos internos representan puntos de decisión donde se evalúa una característica y se decide qué rama tomará en función de ciertas condiciones. Los nodos hoja representan valores numéricos de regresión que se predicen (Wu et al., 2016).

Las ramas son las conexiones entre nodos y representan el resultado de una decisión basada en una característica, al igual que en un árbol de clasificación. Cada rama sale de un nodo y conduce a otro, ya sea interno u hoja, y está etiquetada con la condición de que debe cumplirse para seguir esa ruta. Las hojas en un árbol de regresión representan valores numéricos que corresponden a la predicción.

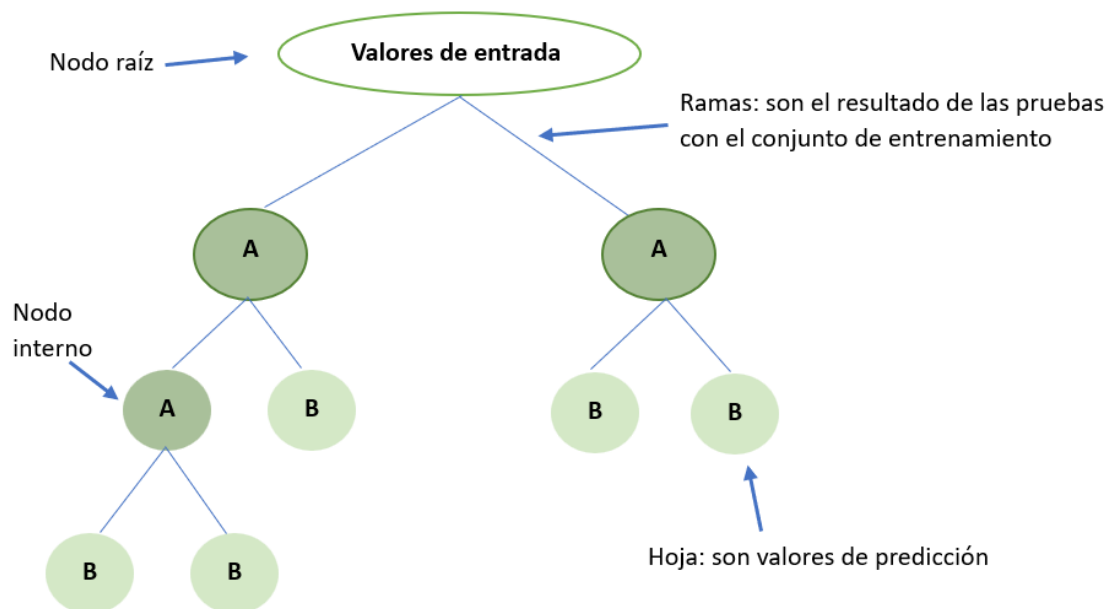


Figura 2. Esquema de un árbol de regresión, dentro de los algoritmos de aprendizaje automático.

El árbol de regresión puede usarse para optimizar los parámetros de un modelo de regresión y mejorar su desempeño, por lo que puede ayudar a encontrar los mejores parámetros para un modelo de regresión (Hu et al., 2022).

2.5.3. Bosques aleatorios

Los bosques aleatorios (random forest o ensemble learning en inglés) son una técnica de aprendizaje automático no supervisada que puede utilizarse para la clasificación, regresión y otros problemas de aprendizaje. Esta metodología combina varios árboles de decisión para crear un modelo predictivo más preciso que un único árbol de decisión (Schonlau & Zou 2020). Breiman (2001) menciona que este algoritmo estima con mayor precisión la tasa de error comparada con los árboles de decisión, puesto que se ha demostrado matemáticamente que la tasa de error siempre converge a medida que aumenta el número de árboles.

Los bosques aleatorios tienen ciertas ventajas sobre otros algoritmos de aprendizaje automático. Por ejemplo, son más robustos a problemas de sobreajuste a los datos de entrenamiento (Witten et al., 2016). Además, pueden entrenarse de forma paralela, lo que permite que el entrenamiento de un bosque aleatorio puede ser bastante rápido, generando un costo computacional menor.

Como se observa en la Figura 3, un Bosque Aleatorio consta de un conjunto de árboles de clasificación o regresión, donde como se describió previamente, cada árbol se construye utilizando un subconjunto aleatorio de los datos de entrenamiento (Sekhar et al., 2016). Cada árbol se entrena de manera independiente, se elige una característica en cada nodo basándose en ciertos criterios como la ganancia de información, la impureza de Gini (clasificación), o la reducción de la varianza en el caso de regresión (Nembrini et al., 2018).

Los árboles se expanden hasta alcanzar una profundidad máxima o hasta que se cumple un criterio de parada, como un número mínimo de instancias en un nodo hoja. Una vez que se han construido todos los árboles, para realizar una predicción, cada árbol en el bosque emite una (clasificación o regresión) para

cada entrada (Schonlau et al., 2020; Nyman & Ormerod, 2017). En el caso de la clasificación, el Bosque Aleatorio realiza una suma entre los árboles para determinar la clase más común. En regresión, promedia las predicciones de todos los árboles (Borup et al., 2023).

Las predicciones individuales de los árboles se combinan para obtener una predicción final robusta y menos propensa al sobreajuste. La mayoría de las veces, la combinación se hace mediante votación (resultado más frecuente) en problemas de clasificación o promediando las predicciones para caso de variables respuesta continuas, aunque existen otros métodos.

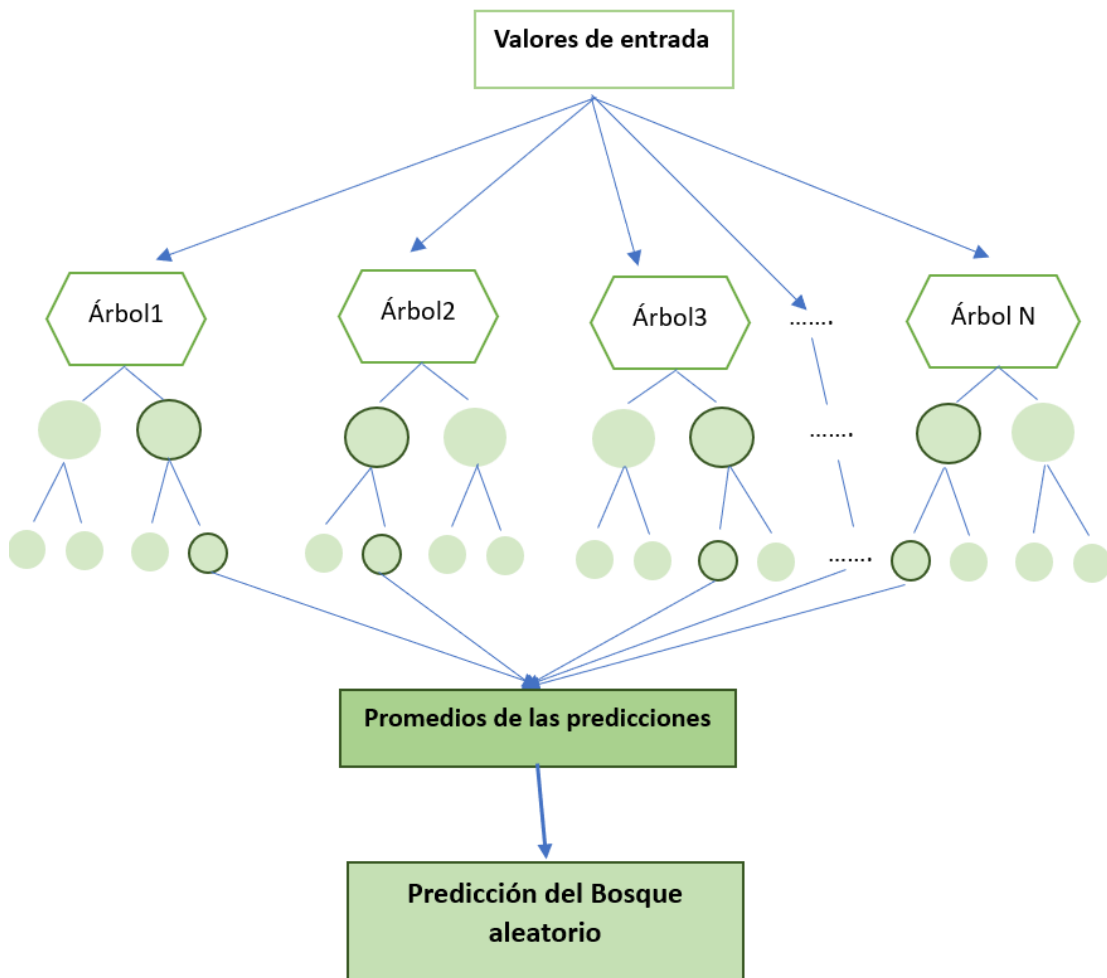


Figura 3. Diagrama representativo del flujo de información en el algoritmo bosques aleatorios.

Uno de los principales inconvenientes de los bosques aleatorios es que pueden ser difíciles de interpretar, esto se debe a que el modelo está compuesto por muchos árboles de decisión, y no se puede determinar fácilmente qué características del modelo son importantes para la predicción (Menze et al., 2009). Para solucionar este problema, se han desarrollado herramientas para ayudar a interpretar los resultados de predicción usando bosques aleatorios, como el análisis de importancia de los atributos (Lin & Chen, 2012).

Se debe considerar que, en el algoritmo de bosque aleatorio, el tamaño del subconjunto de variables predictoras es crucial para controlar la profundidad final de los árboles (Scott & Su-In, 2017). Por tanto, es un parámetro que debe ajustarse durante la selección del modelo.

Para obtener una calificación del modelo, se calcula la suma de la mejora en la función objetivo proporcionada en el criterio de división en todos los nodos internos de un árbol y en todos los que conforman el bosque, esto se realiza por separado para cada variable predictora (He et al., 2018). La puntuación se normaliza dividiendo todas entre el valor máximo.

2.5.4. Redes neuronales

Una red neuronal artificial (RNA) es un algoritmo de aprendizaje automático que en sus inicios trató de imitar el comportamiento de las neuronas biológicas en el cerebro humano (Arbib, 2003; Selverston, 2013). Estos algoritmos se usan para resolver eficientemente problemas complejos, como la clasificación de imágenes, la predicción, la detección de patrones y la optimización (Sahin et al., 2023).

La metodología de redes neuronales inicialmente trató de imitar el funcionamiento del cerebro y la interacción con las neuronas en el procesamiento de información. Su desarrollo data de 1960 pero en ese entonces no se contaba con la cantidad y calidad de datos, así como la potencia de las computadoras para hacer eficiente esta metodología (Kim, 2016). El principio se basa en alimentar la red con los valores iniciales realizando ajustes necesarios para

modelar las relaciones entre la entrada y la salida, esto es útil si se utilizan varias capas en la zona intermedia.

Los algoritmos de redes neuronales artificiales generalmente se componen de varias etapas, en la primera se recopila la información de entrada, esta información se compila en una matriz de datos (Walczak & Cerpa, 2003). En la segunda, los datos se pasan a través de una serie de nodos conectados entre sí, esto se conoce como proceso de propagación. En la tercera, los resultados de la propagación se utilizan para entrenar la red neuronal para que sea capaz de realizar tareas específicas. Por último, en la cuarta etapa, la red neuronal se evalúa para verificar su exactitud y fiabilidad (Lee & Almond, 2003).

2.5.5. Redes neuronales de una capa

Una Red Neuronal de una Capa es una estructura de aprendizaje automático que se basa en la teoría de redes neuronales artificiales. Básicamente, se trata de una red de neuronas artificiales conectadas entre sí, cuyo objetivo es aprender patrones de entrada a través de la propagación de señales a lo largo de las conexiones entre las neuronas (Goel et al., 2023). Esta técnica se aplica comúnmente en problemas de aprendizaje supervisado, en los que hay un conjunto de datos de entrenamiento ya etiquetados y se desea aprender un modelo predictivo para los datos sin etiquetar (Teodoro et al., 2022).

La Red Neuronal de una Capa es un modelo de aprendizaje sencillo con una sola capa oculta. Esta capa oculta contiene un conjunto de neuronas conectadas entre sí y con una capa de entrada (Sarkar et al., 2021). Estas neuronas reciben señales de entrada desde la capa inicial y luego las procesan para producir señales de salida (Figura 4). Estas señales se comparan con los datos de entrenamiento para estimar el error y se usan para ajustar los pesos de las conexiones entre las neuronas. Los pesos actualizados se usan para producir una salida más precisa.

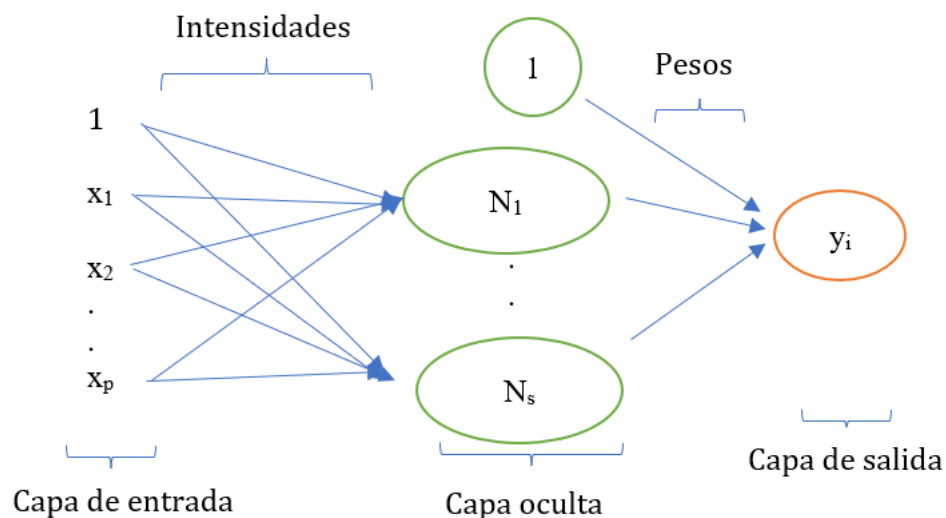


Figura 4. Esquema de una red neuronal con una capa de entrada, una capa oculta con s neuronas y una capa de salida.

Es un modelo útil para problemas de aprendizaje supervisado, como reconocimiento de patrones, clasificación, regresión y otros problemas relacionados. Esta técnica se puede combinar con otros algoritmos de aprendizaje para abordar problemas más complejos (Babatunde et al., 2020). Este método a menudo se utiliza para resolver problemas con gran número de variables de entrada, ya que no requiere la manipulación manual de los datos de entrada (Yang & Maiti, 2020). El procedimiento es fácil de entender, ya que el aprendizaje puede rastrearse y comprenderse a través de la propagación de señales a lo largo de las conexiones de la red (Feng & Simon, 2017).

2.5.6. Redes neuronales multicapa

Una red neuronal multicapa es un modelo de aprendizaje automático, capaz de aprender a partir de datos, esto se logra con la creación de una red de nodos conocidos como neuronas conectadas entre sí (Gardner & Dorling, 1998). Se utilizan para tareas como clasificación, reconocimiento de patrones, detección de errores y predicción. Estas redes se usan con éxito en diversas aplicaciones, como el reconocimiento de imágenes, de voz, procesamiento del lenguaje natural y sistemas de recomendación (Feng et al., 2022; Maggiori et al., 2017).

Uno de los beneficios principales de utilizar esta metodología es que las redes son capaces de aprender a partir de los datos, sin tener que programar explícitamente. Se basa en la estructura de una red y puede ser más robusta a los cambios en los datos (Zhang et al., 2019); si los datos cambian, la red puede ajustarse para considerar estos cambios. Esto les permite ser más flexibles que otros modelos de aprendizaje automático como los árboles de decisión (Curram & Mingers, 1994).

La estructura de una red neuronal multicapa se muestra en la Figura 5, donde se observa que la capa de entrada es la primera de la red y está compuesta por neuronas que representan los valores de entrada, no se realizan cálculos en esta capa, solo transmiten los valores de entrada a las capas ocultas (Madhiarasan & Louzazni, 2022). Las capas ocultas son intermedias entre la capa de entrada y la de salida, cada capa oculta está compuesta por un conjunto de neuronas que procesan la información y realizan cálculos mediante operaciones lineales como multiplicación matricial y funciones de activación no lineales. El número de capas ocultas y la cantidad de neuronas en cada capa son hiperparámetros que se ajustan durante el entrenamiento y pueden variar según el problema (Elansari et al., 2023; Kadhim et al., 2022).

Cada neurona en una capa está conectada con todas las neuronas de la capa anterior y la siguiente. Estas conexiones entre neuronas tienen un peso asociado que se ajusta durante el proceso de entrenamiento, estos determinan la influencia relativa de las entradas en la neurona de destino. Los pesos son parámetros de la red que se optimizan para minimizar la diferencia valores observados y predichos durante el entrenamiento (De Almeida et al., 2016; Gardner & Dorling, 1998; Teodoro et al., 2022).

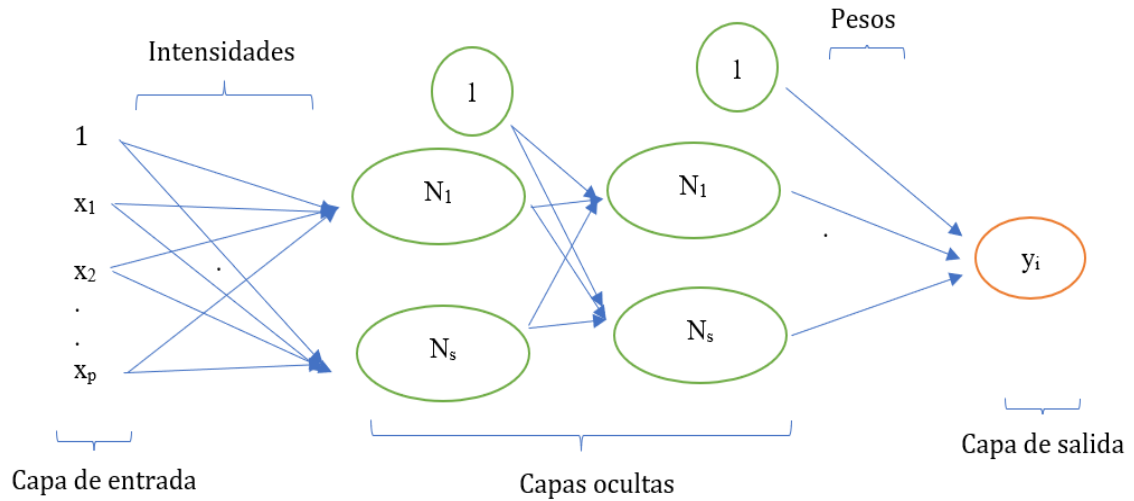


Figura 5. Diagrama de una red neuronal de 2 capas ocultas y s neuronas en cada capa oculta.

Las funciones de activación introducen no linealidad a la red, permitiendo aprender relaciones y patrones complejos en los datos (Goebbels, 2020). Algunas de las funciones de activación comunes incluyen la función ReLU (Rectified Linear Unit) la cual convierte los valores negativos a cero y mantiene valores positivos; función sigmoide, la cual transforma cualquier valor real en un rango de 0 a 1; y la función tangente hiperbólica ($\tanh$), la cual mapea valores reales a un rango de -1 a 1 (similar a la función sigmoide pero centrada en cero) (Salam et al., 2021).

La capa de salida genera las predicciones finales de la red, esta capa depende del tipo de problema que se esté abordando (Aitkin & Foxall, 2003). En problemas de clasificación binaria, puede tener una neurona con una función de activación sigmoide, mientras que en problemas de clasificación multiclase, puede tener múltiples neuronas con una función de activación softmax (Montesinos et al., 2022). Las predicciones de la capa de salida se comparan con los valores reales durante el entrenamiento para calcular el error y ajustar los pesos.

2.5.7. Máquina de soporte vectorial

El método de máquina de soporte vectorial (SVM) es una forma de aprendizaje automático supervisado usada para resolver problemas de clasificación y regresión (Zhao et al., 2020). El objetivo principal de SVM es construir un modelo que prediga la clase de una nueva muestra dada a partir de un conjunto de entrenamiento.

La principal ventaja de esta metodología es que es una herramienta poderosa para trabajar con datos de alta dimensión (Ghaddar & Naoum-Sawaya, 2018; Purnami et al., 2015). Lo cual permite trabajar con datos de muchas variables que podrían ser difíciles de manejar con otros algoritmos como redes neuronales y árboles de decisión. El SVM generalmente puede producir mejores resultados con tiempo de entrenamiento corto (Kurani et al., 2023).

El SVM se considera como un método no paramétrico, por lo que no requiere que se especifiquen parámetros adicionales para ajustar la función de decisión. Esto es útil cuando hay datos limitados o cuando se desconoce la distribución de estos (Gholami & Fakhari, 2017). Sin embargo, también tiene algunas desventajas. Por ejemplo, es un método computacionalmente costoso, es decir, puede ser mucho más lento que otros algoritmos de clasificación cuando se trabaja con grandes conjuntos de datos; además, no es adecuado para datos de clasificación múltiple o multiclase, ya que es un método de clasificación binario (Sunitha & Raju, 2021).

El objetivo de este algoritmo es encontrar una superficie de decisión (hiperplano) que separe de la mejor manera los datos en dos o más clases (Figura 6) o que mejor ajuste los datos (regresión). Fayed y Atiya (2021) describen el problema como cuyo objetivo es encontrar un hiperplano que maximice la distancia entre las clases, se busca el margen máximo, siendo la distancia perpendicular más corta entre el hiperplano y los puntos de datos más cercanos de las clases.

Los puntos de datos más cercanos al hiperplano se denominan vectores de soporte, los cuales son críticos para definir el hiperplano de separación y la

distancia de margen. Cuando los datos no son linealmente separables en su espacio original se utiliza un kernel para mapear los datos a un espacio de mayor dimensionalidad donde sí sean linealmente separables (Savas & Dovic, 2019). Los kernel lineal, polinómico y radial permiten transformar los datos de manera eficiente sin necesidad de calcular las coordenadas en el espacio de alta dimensión (Ben-Hur et al., 2008).

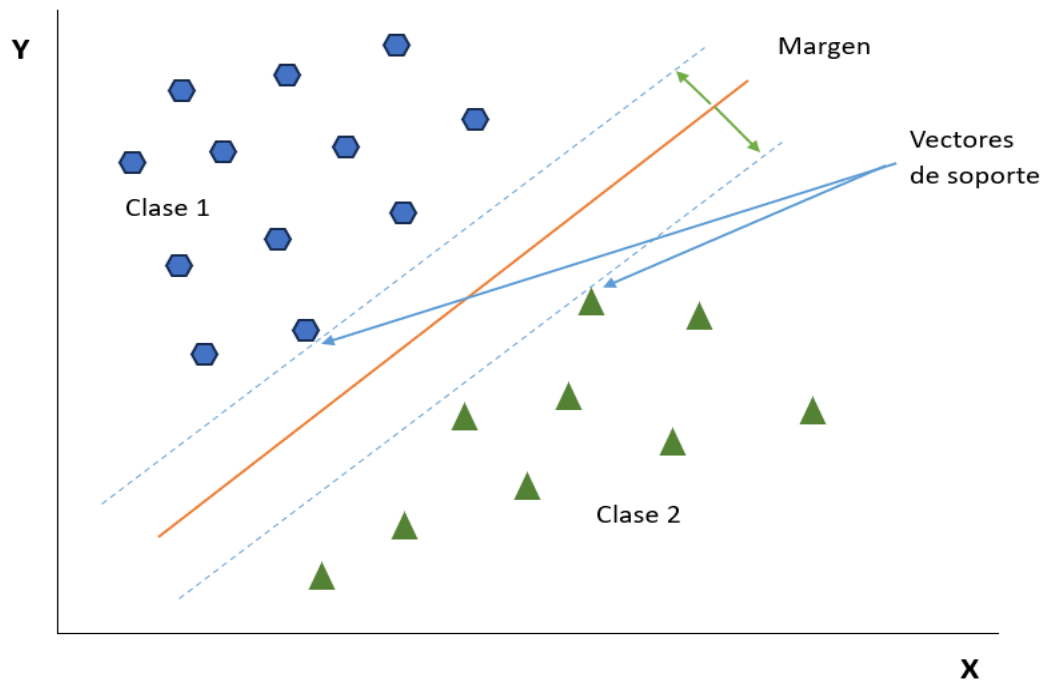


Figura 6. Representación en el plano de algunas funciones de la máquina de soporte vectorial en la división de clases.

Después de encontrar el hiperplano óptimo, se utiliza para clasificar nuevos datos. Los datos se asignan a una clase según en qué lado del hiperplano se encuentren. Cuando, los datos no son completamente separables, se introduce el concepto de un margen suave (soft margin) (Ben-Hur et al., 2008; Gholami & Fakhari, 2017; Marín et al., 2022).

El parámetro C controla la penalización por puntos mal clasificados; un valor más alto de C permite clasificar más puntos correctamente, pero puede llevar a un margen más estrecho (Liu et al., 2014).

2.6. Entrenamiento de los modelos

Este proceso en los modelos de aprendizaje automático se refiere a enseñar a un modelo a realizar tareas específicas a partir de datos de entrenamiento (Straub, 2021). El entrenamiento comienza al seleccionar el algoritmo o metodología, como regresión lineal, redes neuronales, máquinas de soporte vectorial o árboles de decisión, se inicializan los parámetros del modelo, los cuales son valores que se ajustan durante el entrenamiento. Ciampiconi et al. (2023) comentan que es conveniente definir una función de pérdida que mida las diferencias entre las predicciones del modelo y los valores reales en los datos de entrenamiento. El objetivo es minimizar esta función de pérdida durante esta fase del entrenamiento.

La siguiente fase es la de optimización (Yi et al., 2020), la cual utiliza un algoritmo de optimización, como lo es el descenso de gradiente estocástico (Adam), para ajustar gradualmente los parámetros y minimizar la función de pérdida. El optimizador ajusta los parámetros en la dirección que reduce la pérdida (Bottou et al., 2018).

El proceso de entrenamiento por ejemplo para el caso de redes neuronales con el algoritmo de propagación hacia atrás se realiza en ciclos. En cada ciclo se realiza lo siguiente:

- Paso hacia adelante (Forward Pass): se alimenta el conjunto de entrenamiento a través del modelo, lo que genera predicciones.
- Cálculo de la pérdida: se calcula el error utilizando la función de pérdida definida anteriormente, comparando las predicciones con las etiquetas reales.
- Paso hacia atrás (Backpropagation): se calculan los gradientes de la función de pérdida con respecto a los parámetros del modelo (Wunderlich y Pehle, 2021). Esto se hace mediante el cálculo del gradiente de la función de pérdida y la retroalimentación de los gradientes.

- Actualización de parámetros: se utilizan los gradientes calculados para ajustar los parámetros del modelo en la dirección que minimice la pérdida (la tasa de aprendizaje controla la magnitud de los ajustes).
- Convergencia: el entrenamiento continúa durante un número fijo de épocas o hasta que se alcance un criterio de convergencia, como cuando la pérdida deja de disminuir significativamente en el conjunto de entrenamiento o en un conjunto de validación (Carmona & Laurière, 2021).

2.7. Validación de los modelos

La validación cruzada es una técnica importante en el entrenamiento y evaluación de los modelos de aprendizaje automático, tiene como objetivo estimar el rendimiento del modelo de manera robusta y certera (Tougui et al., 2021).

El proceso implica dividir el conjunto de datos en subconjuntos, posteriormente se entrena y evalúa el modelo varias veces, a través de combinaciones de estas divisiones (Ferdinandy et al., 2020). Los subconjuntos pueden variar, esto según el tamaño del conjunto total de datos y recursos disponibles. Generalmente se utiliza de 70 a 80 % de datos para entrenamiento y de 20 a 30 % para la prueba.

El proceso de validación se realiza k veces. En cada iteración, se selecciona uno de los subconjuntos como conjunto de prueba y los otros subconjuntos se utilizan como conjunto de entrenamiento. El modelo se entrena y se evalúa en cada iteración (Chang et al., 2023; Carmona & Laurière, 2021; Kaliappan et al., 2023). Se registran las métricas de evaluación, como la precisión, el error cuadrático medio, la F1-score, etc.

Después de completar todas las iteraciones, se calcula un promedio de las métricas de evaluación obtenidas en cada iteración. Así se obtiene una estimación más robusta de la capacidad de predicción del modelo en comparación con una sola división de datos en entrenamiento y prueba.

Coelho-Ribeiro et al. (2021) comentan que esta metodología proporciona una estimación confiable de la capacidad de predicción del modelo, ya que utiliza

múltiples subconjuntos de datos para entrenar y evaluar, busca evitar problemas de sobre y subajuste al evaluar el modelo en diferentes conjuntos de datos.

Existen varios tipos de validación cruzada, a continuación, se mencionan algunos que destacan (Kaliappan et al., 2023):

- Validación K-subconjuntos: es el más común, los datos se dividen en k -subconjuntos, y se realizan k -iteraciones.
- Validación dejar uno fuera (Leave-One-Out , LOOCV): cada instancia se usa como conjunto de prueba una vez, mientras que las demás se utilizan como conjunto de entrenamiento, esto se repite n veces, donde n es el número de registros en los datos.
- Validación Estratificada: similar a la primera, pero se asegura de que las proporciones de clases en los subconjuntos sean representativas de las proporciones en el conjunto original. Esta técnica es de utilidad cuando se trabaja con datos desequilibrados.
- Validación Aleatoria: en esta técnica en lugar de dividir los datos en subconjuntos secuenciales, se realizan divisiones aleatorias en cada iteración. Esto es útil cuando la distribución de los datos es heterogénea.

La elección de la técnica de validación dependerá de la naturaleza de los datos y del problema que se está trabajando (Bottou et al., 2018; Straub, 2021; Wunderlich & Pehle, 2021).

2.8. Aplicaciones de algoritmos de aprendizaje automático en la ganadería

Las tecnologías avanzadas de aprendizaje automático y computación se han aplicado en la gestión ganadera de precisión, incluida la detección de enfermedades críticas, vacunación, gestión de la producción, seguimiento y control de la salud (Hossain et al., 2022).

Hay interés creciente en el uso del aprendizaje profundo en la investigación genómica. Los métodos clásicos como regresión lineal se han utilizado comúnmente en el análisis de esta información para detectar efectos lineales, el

aprendizaje profundo puede aprender características complejas de la información, lo que lo convierte en un poderoso método para considerar los efectos no lineales y de interacción (Shen et al., 2022). Algunas aplicaciones de estas metodologías son:

1. Predicción del peso de la cría: los algoritmos de aprendizaje automático pueden ayudar a predecir el peso de la cría en función de factores como edad y tamaño de la madre y sexo de la cría (Ruchay et al., 2022).

2. Detección de enfermedades: pueden emplearse para determinar si un animal está enfermo o no (Kavlak et al., 2023). Esto mediante el uso de datos recopilados por dispositivos móviles, incluyendo datos de temperatura, frecuencia cardíaca y respiratoria.

3. Optimización de la alimentación y el cuidado: puede incluir aspectos como la cantidad de alimento que necesitan los animales o la cantidad de tiempo que se necesita para mantenerlos limpios (Neethirajan, 2020; Sokullu et al., 2022; Wijeyakulasuriya et al., 2020).

2.9. Literatura citada

Ahmad, T., Valentovic, M. A., & Rankin, G. O. (2018). Effects of cytochrome P450 single nucleotide polymorphisms on methadone metabolism and pharmacodynamics. *Biochemical Pharmacology*, 153, 196–204. <https://doi.org/10.1016/j.bcp.2018.02.020>

Aitkin, M., & Foxall, R. (2003). Statistical modelling of artificial neural networks using the multi-layer perceptron. *Statistics and Computing*, 13(3), 227–239. <https://doi.org/10.1023/A:1024218716736>

Archer, J. A., Richardson, E. C., Herd, R. M., & Arthur, P. F. (1999). Potential for selection to improve efficiency of feed use in beef cattle: A review. *Australian Journal of Agricultural Research*, 50: 147–62. <https://doi.org/10.1071/A98075>

Arbib, M. (2003). The handbook of brain theory and neural network. *Computers in Biology and Medicine*, 26. ISBN: 0262011972

Babatunde, D., Anozie, A., & Omoleye, J. (2020). Artificial neural network and its applications in the energy sector – An overview. *International Journal of*

Energy Economics and Policy, 10, 250–264.
<https://doi.org/10.32479/ijeep.8691>

- Baizabal, E., & Sánchez, H. (2010). La genómica mejorando el programa genético. Primer foro sobre la ganadería lechera de la zona alta de Veracruz. Veracruz, México. Disponible en: https://www.uv.mx/apps/agronomia/foro_lechero/Bienvenida_files/LAGENOMICA_MEJORANDOELPROGRAMAGENETICO.pdf. Consultado el 13 de septiembre de 2022.
- Band, M. R., Larson, J. H., Rebeiz, M. et al. (2000). An ordered comparative map of the cattle and human genomes. *Genome Research*, 10: 1359–1368.
- Batra, M., & Agrawal, R. (2018). Comparative analysis of decision tree algorithms. *Advances in Intelligent Systems and Computing*, 652, 31–36, https://doi.org/10.1007/978-981-10-6747-1_4
- Beef Improvement Federation. (2018). Guidelines for Uniform Beef Improvement Programs. 9th ed. https://beefimprovement.org/wp-content/uploads/2018/03/BIFGuidelinesFinal_updated0318.pdf Consultado el 04 de agosto de 2022.
- Ben-Hur, A., Ong, C. S., Sonnenburg, S., Schölkopf, B., & Rätsch, G. (2008). Support vector machines and kernels for computational biology. *PLoS Computational Biology*, 4(10). <https://doi.org/10.1371/journal.pcbi.1000173>
- Bickhart, D., & Liu, G. (2014). The challenges and importance of structural variation detection in livestock. *Frontiers in Genetics*, 5. <https://doi.org/10.3389/fgene.2014.00037>
- Bjelland, D. W., Weigel, K. A., Vukasinovic, N., & Nkrumah, J. D. (2013). Evaluation of inbreeding depression in Holstein cattle using whole-genome SNP markers and alternative measures of genomic inbreeding. *Journal of Dairy Science*, 96, 4697–4706. <https://doi.org/10.3168/jds.2012-6435>.
- Black, J. E., Kueper, J. K., & Williamson, T. S. (2023). An introduction to machine learning for classification and prediction. *Family Practice*, 40(1), 200–204. <https://doi.org/10.1093/fampra/cmac104>
- Blanquero, R., Carrizosa, E., Molero-Río, C., & Morales, D. R. (2022). On sparse optimal regression trees. *European Journal of Operational Research*, 299(3), 1045–1054. <https://doi.org/10.1016/j.ejor.2021.12.022>
- Bramer, M. (2002). Using J-Pruning to Reduce Overfitting in Classification Trees. In M. Bramer, F. Coenen, & A. Preece (Eds.), *Research and Development in Intelligent Systems XVIII* (pp. 25–38). Springer London.

- Breiman, L. (2001). Random Forests. *Machine Learning*, 45, 5–32. <https://doi.org/10.1023/A:1010933404324>
- Brown, D. E., Pittard, C. L., & Park, H. (1996). Classification trees with optimal multivariate decision nodes. *Pattern Recognition Letters*, 17(7), 699–703. [https://doi.org/10.1016/0167-8655\(96\)00033-5](https://doi.org/10.1016/0167-8655(96)00033-5)
- Borup, D., Christensen, B. J., Mühlbach, N. S., & Nielsen, M. S. (2023). Targeting predictors in random forest regression. *International Journal of Forecasting*, 39(2), 841–868. <https://doi.org/10.1016/j.ijforecast.2022.02.010>
- Bottou, L., Curtis, F. E., & Nocedal, J. (2018). Optimization methods for large-scale machine learning. *SIAM Review*, 60(2), 223–311. <https://doi.org/10.1137/16M1080173>
- Burrow, H. M., Moore, S. S., Johnston, D. J., Barendse, W. & Bindon, B. M. (2001). Quantitative and molecular genetic influences on properties of beef: a review. *Australian Journal of Experimental Agriculture*, 41: 893. <https://doi.org/10.1071/EA00015>
- Burt D. W. (2009). The cattle genome reveals its secrets. *Journal of Biology*, 8(4), 36. <https://doi.org/10.1186/jbiol137>
- Buzanskas, M. E., Grossi, D. A., Ventura, R. V., Schenkel, F. S., Sargolzaei, M., Meirelles, S. L. C., Mokry, F. B., Higa, R. H., Mudadu, M. A., Da Silva, M. V. G. B., Niciura, S. C. M., Torres, R. A. A., Alencar, M. M., Regitano, L. C. A., & Munari, D. P. (2014). Genome-wide association for growth traits in canchim beef cattle. *PLoS ONE*, 9(4). <https://doi.org/10.1371/journal.pone.0094802>
- Calus, M. P. L., & Veerkamp, R. F. (2011). Accuracy of multi-trait genomic selection using different methods. *Genetics Selection Evolution*, 43(1). <https://doi.org/10.1186/1297-9686-43-26>
- Capper, J. L., & Cady, R. A. (2019). The effects of improved performance in the US dairy cattle industry on environmental impacts between 2007 and 2017. *Journal of Animal Science*, 98: 1. <https://doi.org/10.1093/jas/skz291>
- Carmona, R., & Laurière, M. (2021). Convergence analysis of machine learning algorithms for the numerical solution of mean field control and games I: The ergodic case. *SIAM Journal on Numerical Analysis*, 59(3), 1455–1485. <https://doi.org/10.1137/19M1274377>
- Chang, C. C., Chen, C. H., Hsieh, J. G., & Jeng, J. H. (2023). Iterated cross validation method for prediction of survival in diffuse large B-cell lymphoma for small size dataset. *Scientific Reports*, 13(1), 1438. <https://doi.org/10.1038/s41598-023-28394-6>

- Ching, T., Himmelstein, D. S., Beaulieu-Jones, B. K., Kalinin, A. A., Do, B. T., Way, G. P., et al. (2018). Opportunities and obstacles for deep learning in Biology and Medicine. *Journal of The Royal Society Interface*, 15, 20170387. <https://doi.org/10.1098/rsif.2017.0387>
- Ciampiconi, L., Elwood, A., Leonardi, M., Mohamed, A., & Rozza, A. (2023). A survey and taxonomy of loss functions in machine learning. *arXiv Computer Science; Machine Learning* <https://doi.org/10.48550/arXiv.2301.05579>
- Coelho-Ribeiro, L. A., Bresolin, T., Rosa, G. J. de M., Rume Casagrande, D., Danes, M. de A. C., & Dórea, J. R. R. (2021). Disentangling data dependency using cross-validation strategies to evaluate prediction quality of cattle grazing activities using machine learning algorithms and wearable sensor data. *Journal of Animal Science*, 99(9). <https://doi.org/10.1093/jas/skab206>
- Cortés, O. (2008). Análisis de la variabilidad genética en la raza bovina de Lidia utilizando información molecular. (Tesis doctoral, Universidad Complutense de Madrid, Madrid, España). Disponible en: <https://dialnet.unirioja.es/servlet/tesis?codigo=196352>. Consultada el 13 de febrero de 2022.
- Council on Dairy Cattle Breeding. (2022). Genotype counts by chip type, breed code, and sex code in database as of 2022-06-27. Available at https://queries.uscdcb.com/Genotype/cur_freq.html. Consultada el 24 de abril de 2023.
- Curram, S. P., & Mingers, J. (1994). Neural networks, decision tree induction and discriminant analysis: An empirical comparison. *The Journal of the Operational Research Society*, 45(4), 440–450. <https://doi.org/10.2307/2584215>
- Dadi, H., Jordaan, G. F., Schoeman, S. J., and van der Westhuizen, J. (2002). The effect of Charolais and Hereford sires straight bred and crossbred dams on pre- weaning growth of calves. *South African Journal of Animal Science*, 32, 38–43.
- Daetwyler, H. D., Villanueva, B., Bijma, P., and Woolliams, J. A. (2007). Inbreeding in genome-wide selection. *Journal of Animal Breeding and Genetics*, 124:369–376. <https://doi.org/10.1111/j.1439-0388.2007.00693.x>
- de Albuquerque, L. G., Fernandes Júnior, G. A., & Carvalheiro, R. (2017). Beef cattle genomic selection in tropical environments. In: Proceedings of the Association for the Advancement of Animal Breeding and Genetics, pp. 255–263. Disponible en: <https://www.researchgate.net/publication/323029639>. Consultado el 13 de septiembre de 2022.

- De Almeida, F., Ribeiro, A., Ordonez, E., & Estombelo-Montesco, C. (2016). Performance evaluation of an artificial neural network multilayer perceptron with limited weights for detecting denial of service attack on internet of things. Conference: AICT 2016: The Twelfth Advanced International Conference on Telecommunications. Disponible en: <https://api.semanticscholar.org/CorpusID:54963163>. Consultada el 15 de agosto de 2023.
- D'Onofrio, F., Grani, G., Monaci, M., & Palagi, L. (2023). Margin optimal classification trees. *Computers & Operations Research*, 106441. <https://doi.org/10.1016/j.cor.2023.106441>
- Dorado, G., Luque, F., Pascual, P., Jiménez, I., Javier, F., Sánchez-Cañete, S., Raya, P., Sáiz, J., Sánchez, A., Rosales, T. E., Vásquez, V. F., & Hernández, P. (2021). Biología molecular para inferir fenotipos de restos forenses y antiguos en bioarqueología-Revisión. *ARCHAEOBIOS* 15(1), ISSN 1996-5214. Disponible en: https://www.arqueobios.org/en/archaeobios-journal/file/cat_view/16-archaeobios-2021.html?limit=10&limitstart=0&order=date&dir=DESC.org. Consultada el 13 de septiembre de 2022.
- Elansari, T., Ouanan, M., & Bourray, H. (2023). Modeling of multilayer perceptron neural network hyperparameter optimization and training. Preprint *Research Square*. <https://doi.org/10.21203/rs.3.rs-2570112/v1>
- FAO – DADis. (2016). Sistema de Información sobre la Diversidad de los Animales Domésticos hospedado en la FAO. Disponible en: <http://dad.fao.org>. Consultado el 15 de junio de 2022.
- Fayed, H. A., & Atiya, A. F. (2021). Decision boundary clustering for efficient local SVM. *Applied Soft Computing*, 110, 107628. <https://doi.org/https://doi.org/10.1016/j.asoc.2021.107628>
- Feng, L., Liao, W., Pang, J., Hu, R., & Feng, L. (2022). Artificial neural network in classification of multisource remote sensing images. *International Transactions on Electrical Energy Systems*, 2022, 5439941. <https://doi.org/10.1155/2022/5439941>
- Feng, J., & Simon, N. (2017). Sparse-input neural networks for high-dimensional nonparametric regression and classification. arXiv preprint. <https://doi.org/10.48550/arXiv.1711.07592>
- Fennessy, P., Byrne, T., Proctor, L., & Amer, P. (2019). The potential impact of breeding strategies to reduce methane output from beef cattle. *Animal Production Science*, 59(9), 1598–610.
- Ferdinandy, B., Gerencsér, L., Corrieri, L., Perez, P., Újváry, D., Csizmadia, G., & Miklósi, Á. (2020). Challenges of machine learning model validation

using correlated behaviour data: Evaluation of cross-validation strategies and accuracy measures. *PLoS ONE*, 15(7). <https://doi.org/10.1371/journal.pone.0236092>

Fernandes Júnior, G. A., Peripolli, E., Schmidt, P. I., Campos, G. S., Mota, L. F. M., Mercadante, M. E. Z., Baldi, F., Carneiro, R., & de Albuquerque, L. G. (2022). Current applications and perspectives of genomic selection in *Bos indicus* (Nellore) cattle. *Livestock Science*, 263, 105001. <https://doi.org/10.1016/j.livsci.2022.105001>

Gardner, M. W., & Dorling, S. R. (1998). Artificial neural networks (the multilayer perceptron)—a review of applications in the atmospheric sciences. *Atmospheric Environment*, 32(14), 2627–2636. [https://doi.org/10.1016/S1352-2310\(97\)00447-0](https://doi.org/10.1016/S1352-2310(97)00447-0)

Gengler, N., Mayeres, P., & Szydlowski, M. (2007). A simple method to approximate gene content in large pedigree populations: Application to the myostatin gene in dual-purpose Belgian Blue cattle. *Animal*, 1, 21–28. <https://doi.org/10.1017/S1751731107392628>

Ghaddar, B., & Naoum-Sawaya, J. (2018). High dimensional data classification and feature selection using support vector machines. *European Journal of Operational Research*, 265(3), 993–1004. <https://doi.org/10.1016/j.ejor.2017.08.040>

Ghiasi, M. M., & Zendejboudi, S. (2019). Decision tree-based methodology to select a proper approach for wart treatment. *Computers in Biology and Medicine*, 108, 400–409. <https://doi.org/10.1016/j.compbiomed.2019.04.001>

Gholami, R., & Fakhari, N. (2017). Chapter 27 - Support Vector Machine: Principles, Parameters, and Applications. In P. Samui, S. Sekhar, & V. E. Balas (Eds.), *Handbook of Neural Computation* (pp. 515–535). Academic Press. <https://doi.org/10.1016/B978-0-12-811318-9.00027-2>

Glorot, X., Bordes, A., & Bengio, Y. (2011). “Deep sparse rectifier neural networks”. In *Proceedings of the Fourteenth International Conference on Artificial Intelligence and Statistics*, 315–323. Disponible en: <https://api.semanticscholar.org/CorpusID:2239473>. Consultada el 08 de agosto de 2023.

Goel, A., Goel, A. K., & Kumar, A. (2023). Performance analysis of multiple input single layer neural network hardware chip. *Multimedia Tools and Applications*, 82(18), 28213–28234. <https://doi.org/10.1007/s11042-023-14627-3>

Goddard, M. E. (2009). How can we best use DNA data in the selection of cattle? In: *Proceedings of the Beef Improvement Federation 41st Annual*

Research Symposium April 30 – May 3, 2009, Sacramento, California, EE. UU. Disponible en: https://www.bifconference.com/bif2009/proceedings/g1_pro_goddard.pdf. Consultada el 15 de julio de 2022.

- Goddard, M. E., & Hayes, B. J. (2009). Mapping genes for complex traits in domestic animals and their use in breeding programmes. *Nature Reviews Genetics*, 10 (6): 381–391. <https://doi.org/10.1038/nrg2575>
- Goddard, M. E. (2012). Uses of genomics in livestock agriculture. *Animal Production Science*, 52, 73-77. <https://doi.org/10.1071/AN11180>
- Goebbels, S. (2020). ReLU activated multi-layer neural networks trained with mixed integer linear programs. *ArXiv*. /abs/2008.08386. <https://doi.org/10.48550/arXiv.2008.08386>
- Goyache, F., Fernández, I., Royo, L. J., Álvarez, I., & Gutiérrez, J. P. (2003). Factors affecting actual weaning weight, preweaning average daily gain and relative growth rate in Asturiana de los Valles beef cattle. *Archiv fur Tierzucht Dummerstorf*, 46, 235–243. <https://doi.org/10.5194/aab-46-235-2003>
- Grapes, L., Dekkers, J. C., Rothschild, M. F., & Fernando, R. L. (2004). Comparing linkage disequilibrium-based methods for fine mapping quantitative trait loci. *Genetics*, 166, 1561–1570. <https://doi.org/10.1534/genetics.166.3.1561>
- Hakem, M., Boulouard, Z., & Kissi, M. (2021). Classification of body weight in beef cattle via machine learning methods: A review. *Procedia Computer Science*, 198, 263–268. <https://doi.org/10.1016/j.procs.2021.12.238>
- Harrison J. H., Gilbertson J. R., Hanna, M. G., Olson, N. H., Seheult, J. N., Sorace, J. M., et al. (2021). Introduction to artificial intelligence and machine learning for pathology. *Archives Pathology & Laboratory Medicine* 145(10), 1228–54. <https://doi.org/10.5858/arpa.2020-0541-CP>
- Hayes, B. J., Chamberlain, A. J., McPartlan, H., MacLeod, I., Sethuraman, L., & Goddard, M. E. (2007). Accuracy of marker-assisted selection with single markers and marker haplotypes in cattle. *Genetical Research*, 89(4): 215–220. <https://doi.org/10.1017/S0016672307008865>
- He, L., Levine, R. A., & Beemer, J. (2018). Random forest as a predictive analytics alternative to regression in institutional research. *Practical Assessment, Research & Evaluation*, 23(1). <https://doi.org/10.7275/1wpr-m024>
- Henderson, C. R. (1984). Best Linear Unbiased Prediction of performance and breeding value. Cornell University and the University of Illinois.

- Hossain, M. E., Kabir, M. A., Zheng, L., Swain, D. L., McGrath, S., & Medway, J. (2022). A systematic review of machine learning techniques for cattle identification: Datasets, methods and future directions. *Artificial Intelligence in Agriculture*, 6, 138–155. <https://doi.org/10.1016/j.aiia.2022.09.002>
- Hu, J., Ou, X., Liang, P. et al. (2022). Applying particle swarm optimization-based decision tree classifier for wart treatment selection. *Complex & Intelligent Systems*, 8, 163–177. <https://doi.org/10.1007/s40747-021-00348-3>
- Hutchison, J. L., Cole, J. B., & Bickhart, D. M. (2014). Short communication: Use of young bulls in the United States. *Journal of Dairy Science*, 97(5), 3213–3220. <https://doi.org/10.3168/jds.2013-7525>
- Jahuey-Martínez, F. J., Parra-Bracamonte, G. M., Garrick, D. J., López-Villalobos, N., Martínez-González, J. C., Sifuentes-Rincón, A. M., & López-Bustamante, L. A. (2020). Accuracies of direct genomic breeding values for birth and weaning weights of registered Charolais cattle in Mexico. *Animal Production Science*, 60(6), 772–779. <https://doi.org/10.1071/AN18363>
- Jiang, V. S., Pavlovic, Z. J., & Hariton, E. (2023). The Role of Artificial Intelligence and Machine Learning in Assisted Reproductive Technologies. *Obstetrics and Gynecology Clinics of North America*. <https://doi.org/10.1016/j.ogc.2023.09.003>
- Johannsen, W. L. (1911). The genotype conception of heredity. *The American Naturalist*, 45, 129–159.
- Kadhim, Z., Abdullah, H., & Ghathwan, K. (2022). Artificial neural network hyperparameters optimization: A survey. *International Journal of Online and Biomedical Engineering*, 18, 59–87. <https://doi.org/10.3991/ijoe.v18i15.34399>
- Kaliappan, J., Bagepalli, A. R., Almal, S., Mishra, R., Hu, Y. C., & Srinivasan, K. (2023). Impact of cross-validation on machine learning models for early detection of intrauterine fetal demise. *Diagnostics*, 13(10). <https://doi.org/10.3390/diagnostics13101692>
- Kavlak, A. T., Pastell, M., & Uimari, P. (2023). Disease detection in pigs based on feeding behaviour traits using machine learning. *Biosystems Engineering*, 226, 132–143. <https://doi.org/10.1016/j.biosystemseng.2023.01.004>
- Kersting, K. (2018). Machine learning and artificial intelligence: Two fellow travelers on the quest for intelligent behavior in machines. *Frontiers in Big Data*, 1. <https://doi.org/10.3389/fdata.2018.00006>.

- Khorshidi, R., Macneil, M., Hays, D. P., Abo-Ismael, M., Crowley, J. J., Akanno, E., Wang, Z., & Plastow, G. (2020). Estimation of genetic parameters and trends for growth traits in Hays Converter cattle using multiple-trait and random regression models. *Livestock Science*, 241, 104245. <https://doi.org/10.1016/j.livsci.2020.104245>
- Kim, K. (2016). A hybrid classification algorithm by subspace partitioning through semi-supervised decision tree. *Pattern Recognit*, 60, 157–163. <https://doi.org/10.1016/j.patcog.2016.04.016>
- Koots, K. R., Gibson, J. P., Smith, C., & Wilton, J.W. (1994). Analyses of published genetic parameter estimates for beef production traits. 1. Heritability. *Archives Animal Breeding Abstracts*, 62, 309–338
- Krupa, E., Oravcová, M., Polák, P., Huba, J., & Krupová, Z. (2005). Factors affecting growth traits of beef cattle breeds raised in Slovakia. *Czech Journal of Animal Science*, 50(1), 14–21. <https://doi.org/10.17221/3990-CJAS>
- Kurani, A., Doshi, P., Vakharia, A., & Shah, M. (2023). A comprehensive comparative study of artificial neural network (ANN) and support vector machines (SVM) on stock forecasting. *Annals of Data Science*, 10(1), 183–208. <https://doi.org/10.1007/s40745-021-00344-x>
- Lee, J. A., & Almond, D. P. (2003). 21 - A neural-network approach to fatigue-life prediction. In B. Harris (Ed.), *Fatigue in Composites* (pp. 569–589). Woodhead Publishing. <https://doi.org/https://doi.org/10.1533/9781855738577.4.569>
- Liao, X., Peng, F., Forni, S., McLaren, D., Plastow, G., & Stothard, P. (2013). Whole genome sequencing of Gir cattle for identifying polymorphisms and loci under selection. *Genome*, 56, 592–598. <https://doi.org/10.1139/gen-2013-0082>
- Lin, S. W., & Chen, S. C. (2012). Parameter determination and feature selection for C4.5 algorithm using scatter search approach. *Soft Computing*, 16, 63–75. <https://doi.org/10.1007/s00500-011-0734-z>
- Liu, X., Fang, C., Wu, C., Yu, J., & Zhao, Q. (2021). DRG grouping by machine learning: from expert-oriented to data-based method. *BMC Medical Informatics and Decision Making*, 21(1), 312. <https://doi.org/10.1186/s12911-021-01676-7>
- Liu, Y., Lian, J., Bartolacci, M. R., & Zeng, Q. A. (2014). Density-based penalty parameter optimization on C-SVM. *The Scientific World Journal*, 2014, 851814. <https://doi.org/10.1155/2014/851814>

- Loh, W. Y. (2014). Fifty years of classification and regression trees. *International Statistical Review*, 82, 329-348. <https://doi.org/10.1111/insr.12016>
- Lu, D., Miller, S., Sargolzaei, M., Kelly, M., VanderVoort, G., Caldwell, T., Wang, Z., Plastow, G., & Moore, S. (2013). Genome-wide association analyses for growth and feed efficiency traits in beef cattle. *Journal of Animal Science*, 91(8), 3612-33. <https://doi.org/10.2527/jas.2012-5716>
- Maggiori, E., Tarabalka, Y., Charpiat, G., & Alliez, P. (2017). Convolutional neural networks for large-scale remote-sensing image classification. *IEEE Transactions on Geoscience and Remote Sensing*, 55(2), 645–657. <https://doi.org/10.1109/TGRS.2016.2612821>
- Madhiarasan, M., & Louzazni, M. (2022). Analysis of artificial neural network: Architecture, types, and forecasting applications. *Journal of Electrical and Computer Engineering*, 2022, 5416722. <https://doi.org/10.1155/2022/5416722>
- Maimon, O. Z., & Rokach, L. (2014). Data mining with decision trees: theory and applications. *World Scientific*, 81. https://doi.org/10.1142/9789812771728_0001
- Maray, M., Albraikan, A. A., Alotaibi, S. S., Alabdan, R., Duhayyim, M. Al, Al-Azzawi, W. K., & alkhayyat, A. (2022). Artificial intelligence-enabled coconut tree disease detection and classification model for smart agriculture. *Computers and Electrical Engineering*, 104. <https://doi.org/10.1016/j.compeleceng.2022.108399>
- Marín, A., Martínez-Merino, L. I., Puerto, J., & Rodríguez-Chía, A. M. (2022). The soft-margin Support Vector Machine with ordered weighted average. *Knowledge-Based Systems*, 237, 107705. <https://doi.org/10.1016/j.knosys.2021.107705>
- Martínez G., J. C., Castillo-Rodriguez, S. P., Lucero-Magana, F. A., & Ortega-Rivas, E. (2007). Influencias ambientales y heredabilidad para características de crecimiento en ganado Sardo Negro en México. *Zootecnia Tropical*, 25(1), 1-7. Disponible en: http://ve.scielo.org/scielo.php?script=sci_arttext&pid=S0798-72692007000100002&lng=es&tlng=es. Consultado el 21 de junio de 2022.
- Martínez, R., Bejarano, D., Gómez, Y., Romain, D., Jiménez, A., Even, G., Sölkner, J., & Mészáros, G. (2017). Genome-wide association study for birth, weaning and yearling weight in Colombian Brahman cattle. *Genetics and Molecular Biology*, 40(2), 453-459. <https://doi.org/10.1590/1678-4685-GMB-2016-0017>
- Maudet, C., Luikart, G. & Taberlet, P. (2002). Genetic diversity and assignment tests among seven French cattle breeds based on microsatellite DNA

- analysis. *Journal of Animal Science*, 80, 942–950. <https://doi.org/10.2527/2002.804942x>
- May, K., Scheper, C., Brügemann, K., Yin, T., Strube, C., Korkuć, P., Brockmann, G. A., & König, S. (2019). Genome-wide associations and functional gene analyses for endoparasite resistance in an endangered population of native German Black Pied cattle. *BMC Genomics*, 20(1), 277. <https://doi.org/10.1186/s12864-019-5659-4>
- Menze, B. H., Kelm, B. M., Masuch, R., Himmelreich, U., Bachert, P., Petrich, W., et al. (2009). A comparison of random forest and its Gini importance with standard chemometric methods for the feature selection and classification of spectral data. *BMC Bioinformatics*, 10:213. <https://doi.org/10.1186/1471-2105-10-213>
- Meuwissen, T. H., Hayes, B. J., & Goddard, M. E. (2001). Prediction of total genetic value using genome-wide dense marker maps. *Genetics*, 157, 1819–1829. <https://doi.org/10.1093/genetics/157.4.1819>
- Meuwissen, T., Hayes, B., & Goddard, M. (2013). Accelerating improvement of livestock with genomic selection. *Annual Review of Animal Biosciences*, 1, 221–237. <https://doi.org/10.1146/annurev-animal-031412-103705>.
- Meuwissen, T., Hayes, B., & Goddard, M. (2016). Genomic selection: a paradigm shift in animal breeding. *Animal Frontiers*, 6, 6–14. <https://doi.org/10.2527/af.2016-0002>.
- Mielczarek, M., Frąszczak, M., Giannico, R., Minozzi, G., Williams, J. L., Wojdak-Maksymiec, K., & Szyda, J. (2017). Analysis of copy number variations in Holstein-Friesian cow genomes based on whole-genome sequence data. *Journal of Dairy Science*, 100(7), 5515–5525. <https://doi.org/10.3168/jds.2016-11987>
- Misztal, I., Legarra, A., & Aguilar, I. (2009). Computing procedures for genetic evaluation including phenotypic, full pedigree, and genomic information. *Journal of Dairy Science*, 92, 4648–4655. <https://doi.org/10.3168/jds.2009-2064>.
- Moerland, T. M., Broekens, J., Plaat, A., & Jonker, C. M. (2022). A unifying framework for reinforcement learning and planning. *Frontiers in Artificial Intelligence*, 5. <https://doi.org/10.3389/frai.2022.908353>
- Montesinos-López, O. A., Montesinos-López, A., & Crossa, J. (2022). Fundamentals of Artificial Neural Networks and Deep Learning. In: Multivariate Statistical Machine Learning Methods for Genomic Prediction. Springer, Cham. https://doi.org/10.1007/978-3-030-89010-0_10

- Mrode, R., Ojango, J. M. K., Okeyo, A. M., & Mwacharo, J. M. (2019). Genomic selection and use of molecular tools in breeding programs for indigenous and crossbred cattle in developing countries: Current status and future prospects. *Frontiers in Genetics*, 10, <https://doi.org/10.3389/fgene.2018.00694>
- Mueller, M. L., & Van Eenennaam, A. L. (2022). Synergistic power of genomic selection assisted reproductive technologies, and gene editing to drive genetic improvement of cattle. *CABI Agriculture and Bioscience*, 3(1). <https://doi.org/10.1186/s43170-022-00080-z>
- Neethirajan, S. (2020). The role of sensors, big data and machine learning in modern animal farming. *Sensing and Bio-Sensing Research*, 29, 100367. <https://doi.org/10.1016/j.sbsr.2020.100367>
- Nembrini, S., König, I. R., & Wright, M. N. (2018). The revival of the Gini importance? *Bioinformatics*, 34(21), 3711–3718. <https://doi.org/10.1093/bioinformatics/bty373>
- Nichols, J. A., Herbert Chan, H. W., & Baker, M. A. B. (2019). Machine learning: applications of artificial intelligence to imaging and diagnosis. *Biophysical Reviews*, 11(1), 111–118. <https://doi.org/10.1007/s12551-018-0449-9>
- Nuñez-Torres, O. P., & Almeida-Secaira, R. I. (2022). Quantitative genetics: principles of farming in livestock production. *Journal of the Selva Andina Animal Science*, 9(1):23-36. <https://doi.org/10.36610/j.jsaas.2022.090100023>
- Nyman R., & Ormerod, P. (2017). Predicting economic recessions using machine learning algorithms. *ArXiv Working Paper No. arXiv:1701.01428*. <https://doi.org/10.48550/arXiv.1701.01428>
- Okser, S., Pahikkala, T., Airola, A., Salakoski, T., Ripatti, S., et al. (2014). Regularized machine learning in the genetic prediction of complex traits. *PLoS Genetics*, 10(11): e1004754. <https://doi.org/10.1371/journal.pgen.1004754>
- Ozcan, M., & Peker, S. (2023). A classification and regression tree algorithm for heart disease modeling and prediction. *Healthcare Analytics*, 3. <https://doi.org/10.1016/j.health.2022.100130>
- Page, B. T., Casas, E., Quaas, R. L., Thallman, R. M., Wheeler, T. L., Shackelford, S. D., Koohmaraie, M., et al. (2004). Association of markers in the bovine CAPN1 gene with meat tenderness in large crossbred populations that simple influential industry sires. *Journal of Animal Science*, 82, 3474-3481. <https://doi.org/10.2527/2004.82123474x>

- Pszczola, M., Strabel, T., van Arendonk, J. A. M., & Calus, M. P. L. (2012). The impact of genotyping different groups of animals on accuracy when moving from traditional to genomic selection. *Journal of Dairy Science*, 95(9), 5412–5421. <https://doi.org/10.3168/jds.2012-5550>
- Purnami, S. W., Andari, S., & Pertiwi, Y. D. (2015). High-dimensional data classification based on smooth support vector machines. *Procedia Computer Science*, 72, 477–484. <https://doi.org/10.1016/j.procs.2015.12.129>
- Ramos, A. M., Megens, H. J., Crooijmans, R. P. M. A., Schook, L. B., & Groenen, M. A. M. (2011). Identification of high utility SNPs for population assignment and traceability purposes in the pig using high-throughput sequencing. *Stichting International Foundation for Animal Genetics*, 42, 613–620. <https://doi.org/10.1111/j.1365-2052.2011.02198.x>
- Rashidi, H. H., Albahra, S., Robertson, S., Tran, N. K., & Hu, B. (2023). Common statistical concepts in the supervised Machine Learning arena. *Frontiers in Oncology*, 13. <https://doi.org/10.3389/fonc.2023.1130229>
- Reinoso-Peláez, E. L., Gianola, D. & González-Recio, O. (2021). Aprendizaje automático aplicado a la predicción de caracteres complejos. Instituto Nacional de Investigación y Tecnología Agraria y Alimentaria. XIX Jornadas sobre Producción Animal, AIDA (195). Disponible en: https://www.aidaitea.org/aidaitea/files/jornadas/2021/comunicaciones/2021_Gen_44.pdf. Consultado el 13 de julio de 2022.
- Rodríguez-Alcántar, E. (2019). Aplicación de algoritmos de aprendizaje automático para la clasificación de ganado lechero utilizando SNP de genoma completo. (Tesis de Doctorado, Instituto de Ingeniería, Universidad Autónoma de Baja California. México). Disponible en: <https://repositorioinstitucional.uabc.mx/bitstream/20.500.12930/1834/1/MXL122258.pdf>. Consultada el 10 de octubre de 2023.
- Ruchay, A., Kober, V., Dorofeev, K., Kolpakov, V., Dzhulamanov, K., Kalschikov, V., & Guo, H. (2022). Comparative analysis of machine learning algorithms for predicting live weight of Hereford cows. *Computers and Electronics in Agriculture*, 195, 106837. <https://doi.org/10.1016/j.compag.2022.106837>
- Salam, A., Hibaoui, A. El, & Saif, A. (2021). A comparison of activation functions in multilayer neural network for predicting the production and consumption of electricity power. *International Journal of Electrical and Computer Engineering*, 11(1), 163–170. <https://doi.org/10.11591/ijece.v11i1.pp163-170>
- Sarmiento, C. (2009). Beyond natural selection. *Acta Biológica Colombiana*, 14, 187–198. ISSN 0120-548X.

- Sarkar, K., Bonnerjee, D., Srivastava, R., & Bagh, S. (2021). A single layer artificial neural network type architecture with molecular engineered bacteria for reversible and irreversible computing. *Chemical Science*, 12. <https://doi.org/10.1039/D1SC01505B>
- Savas, C., & Dosis, F. (2019). The impact of different kernel functions on the performance of scintillation detection based on Support Vector Machines. *Sensors*, 19(23), 5219. <https://doi.org/10.3390/s19235219>
- Scheifers, J. M., & Weigel, K.A. (2012). Genomic selection in dairy cattle: Integration of DNA testing into breeding programs. *Animal Frontiers*, 2(1): 4-9. <https://doi.org/10.2527/af.2011-0032>
- Schonlau, M., & Zou, R. Y. (2020). The random forest algorithm for statistical learning. *The Stata Journal*, 20(1), 3–29. <https://doi.org/10.1177/1536867X20909688>
- Scott, M. L., & Su-In, L. (2017). A unified approach to interpreting model predictions. *Advances in neural information processing systems*, 30. Disponible en: <https://proceedings.neurips.cc/paper/2017/file/8a20a8621978632d76c43dfd28b67767-Paper.pdf>. Consultado el 04 de febrero de 2022
- Seabury, C. M., Oldeschulte, D. L., Saatchi, M., et al. (2017). Genome-wide association study for feed efficiency and growth traits in U.S. beef cattle. *BMC Genomics*, 18, 386. <https://doi.org/10.1186/s12864-017-3754-y>
- Sekhar, C. R., Minal, & Madhu, E. (2016). Mode choice analysis using random forest decision trees. *Transportation Research Procedia*, 17, 644–652. <https://doi.org/10.1016/j.trpro.2016.11.119>
- Silverston, A. (2013). *Model neural networks and behavior*. Springer Science & Business Media. <https://doi.org/10.1007/978-1-4757-5858-0>
- Sahin, G., Isik, G., & van Sark, W. G. J. H. M. (2023). Predictive modeling of PV solar power plant efficiency considering weather conditions: A comparative analysis of artificial neural networks and multiple linear regression. *Energy Reports*, 10, 2837–2849. <https://doi.org/https://doi.org/10.1016/j.egyr.2023.09.097>
- Shen, X., Jiang, C., Wen, Y., Li, C., & Lu, Q. (2022). A brief review on deep learning applications in genomic studies. *Frontiers in Systems Biology*, 2. <https://doi.org/10.3389/fsysb.2022.877717>
- Sifuentes-Rincón, A. M., Parra-Bracamonte, G. M., Arellano-Vera, W., Ambriz-Morales, P., Cantú-Covarrubias, A., & Moreno-Medina, V. R. (2020). Frecuencia de SNP en genes candidatos para crecimiento y su efecto en caracteres de peso vivo en ganado para carne de Tamaulipas. *Revista*

Mexicana de Ciencias Pecuarias, 11(1), 283-293.
<https://doi.org/10.22319/rmcp.v11i1.4684>

- Sokullu, R., Tanriverdi, B. Y., & Goleva, R. (2022). Iot Based Livestock Precision Feeding System Using Machine Learning. 8th International Conference on Energy Efficiency and Agricultural Engineering 2022 (EE&AE), 1–4. <https://doi.org/10.1109/EEAE53789.2022.9831244>
- Song Y. Y., & Lu Y. (2015). Decision tree methods: applications for classification and prediction. *Shanghai Archives of Psychiatry*, 25;27(2):130-5. doi: 10.11919/j.issn.1002-0829.215044. PMID: 26120265; PMCID: PMC4466856
- Sourdis, A. (2008). Ganadería en Colombia cinco siglos construyendo país. Federación Nacional de Ganaderos, FEDEGAN. Bogotá. pp. 381. Disponible en: https://books.google.com.co/books/about/Ganader%C3%ADa_en_Colombia.html?id=0l1CQwAACAAJ. Consultado el 15 de marzo de 2022.
- Srinivasan, K., Friedman, B. A., Larson, J. L., Lauffer, B. E., Goldstein, L. D., Appling, L. L., Borneo, J., Poon, C., Ho, T., Cai, F., Steiner, P., van der Brug, M. P., Modrusan, Z., Kaminker, J. S., & Hansen, D. V. (2016). Untangling the brain's neuroinflammatory and neurodegenerative transcriptional responses. *Nature Communications*, 7, 11295. <https://doi.org/10.1038/ncomms11295>
- Staniewicz, L., and Midgley, P. A. (2015). Machine learning as a tool for classifying electron tomographic reconstructions. *Advanced Structural and Chemical Imaging*, 1(1), 9. <https://doi.org/10.1186/s40679-015-0010-x>
- Straub, J. (2021). Machine learning performance validation and training using a 'perfect' expert system. *MethodsX*, 8, 101477. <https://doi.org/10.1016/j.mex.2021.101477>
- Sunitha, L., & Raju, M. B. (2021). Multi-class classification for large datasets with optimized SVM by non-linear kernel function. *Journal of Physics: Conference Series*, 2089(1). <https://doi.org/10.1088/1742-6596/2089/1/012015>
- Sunnucks, P. (2001). Efficient genetic markers for population biology. *Tree*, 15, 199 -203. [https://doi.org/10.1016/S0169-5347\(00\)01825-5](https://doi.org/10.1016/S0169-5347(00)01825-5)
- Tan, H. (2021). Machine learning algorithm for classification. *Journal of Physics: Conference Series*, 1994(1). <https://doi.org/10.1088/1742->
- Teodoro, A. A. M., Gomes, O. S. M., Saadi, M., Silva, B. A., Rosa, R. L., & Rodríguez, D. Z. (2022). An FPGA-based performance evaluation of artificial neural network architecture algorithm for IoT. *Wireless Personal*

Communications, 127(2), 1085–1116. <https://doi.org/10.1007/s11277-021-08566-1>

- Tougui, I., Jilbab, A., & Mhamdi, J. E. (2021). Impact of the choice of cross-validation techniques on the results of machine learning-based diagnostic applications. *Healthcare Informatics Research*, 27(3), 189–199. <https://doi.org/10.4258/hir.2021.27.3.189>
- Valerio-Hernández, J. E., Ruíz-Flores, A., Nilforooshan, M. A., & Pérez-Rodríguez, P. (2023). Single-step genomic evaluation for growth traits in a Mexican Braunvieh cattle population. *Animal Bioscience*, 36(7), 1003–1009. <https://doi.org/10.5713/ab.22.0158>
- Vallejos-Vidal, E., Reyes-Cerpa, S., Rivas-Pardo, J. A., Maisey, K., Yáñez, J. M., Valenzuela, H., Cea, P. A., Castro-Fernandez, V., Tort, L., Sandino, A. M., Imarai, M., & Reyes-López, F. E. (2020). Single-nucleotide polymorphisms (SNP) mining and their effect on the tridimensional protein structure prediction in a set of immunity-related expressed sequence tags (EST) in Atlantic salmon (*Salmo salar*). *Frontiers in Genetics*, 10. <https://doi.org/10.3389/fgene.2019.01406>
- Van-Eenenname, A. (2006). DNA-based technologies. Beef sire selection manual. *National beef cattle evaluation consortium* (Eds.), USA. pp. 66-73.
- VanRaden, P. M. (2008). Efficient methods to compute genomic predictions. *Journal of Dairy Science*, 91, 4414–4423. <https://doi.org/10.3168/jds.2007-0980>.
- Walczak, S., & Cerpa, N. (2003). Artificial Neural Networks. In R. A. Meyers (Ed.), *Encyclopedia of Physical Science and Technology* (Third Edition) (pp. 631–645). *Academic Press*. <https://doi.org/https://doi.org/10.1016/B0-12-227410-5/00837-1>
- Wakchaure, R., & Ganguly, S. (2015). Marker assisted selection (MAS) in animal breeding: A review. *Journal of Drug Metabolism & Toxicology*, 06(05). <https://doi.org/10.4172/2157-7609.1000e127>
- Weik, F., Hickson, R. E., Morris, S. T., Garrick, D. J., & Archer, J. A. (2022). Genetic parameters for growth, ultrasound and carcass traits in new zealand beef cattle and their correlations with maternal performance. *Animals*, 12(1). <https://doi.org/10.3390/ani12010025>
- Wiggans, G. R., & Carrillo, J. A. (2022). Genomic selection in United States dairy cattle. *Frontiers in Genetics*, 13. <https://doi.org/10.3389/fgene.2022.994466>

- Wijeyakulasuriya, D. A., Eisenhauer, E. W., Shaby, B. A., & Hanks, E. M. (2020). Machine learning for modeling animal movement. *PLoS ONE*, 15. <https://doi.org/10.1371/journal.pone.0235750>
- Witten, I. H., Frank, E., Hall, M. A., & Pal, C. J. (2016). The WEKA workbench online appendix. In *Data Mining: Practical Machine Learning Tools and Techniques*, 4th ed. Burlington, MA: Morgan Kaufmann.
- Wolcott, M. L., Johnston, D., Barwick, S., Corbet, N., & Burrow, H. M. (2013). Genetic relationships between steer performance and female reproduction and possible impacts on whole herd productivity in two tropical beef genotypes. *Animal Production Science*, 54, 85–96. <https://doi.org/10.1071/AN13141>
- Wu, C.-C., Chen, Y.-L., Liu, Y.-H., & Yang, X.-Y. (2016). Decision tree induction with a constrained number of leaf nodes. *Applied Intelligence*, 45(3), 673–685. <https://doi.org/10.1007/s10489-016-0785-z>
- Wunderlich, T. C., & Pehle, C. (2021). Event-based backpropagation can compute exact gradients for spiking neural networks. *Scientific Reports*, 11(1). <https://doi.org/10.1038/s41598-021-91786-z>
- Xie, T., Li, R., Zhang, X., Zhou, B., & Wang, Z. (2019). Research on heartbeat classification algorithm based on cart decision tree. In: 8th International Symposium on Next Generation Electronics, ISNE 2019, pp. 27–29, <http://dx.doi.org/10.1109/ISNE.2019.8896650>.
- Xu, M., Watanachaturaporn, P., Varshney, P. K., & Arora, M. K. (2005). Decision tree regression for soft classification of remote sensing data. *Remote Sensing of Environment*, 97(3), 322–336. <https://doi.org/10.1016/j.rse.2005.05.008>
- Yang, K., & Maiti, T. (2020). Statistical aspects of high-dimensional sparse artificial neural network models. *Machine Learning and Knowledge Extraction*, 2(1), 1–19. <https://doi.org/10.3390/make2010001>
- Yang, J., Benyamin, B., McEvoy, B. P., Gordon, S., Henders, A. K., Nyholt, D. R., Madden, P. A., Heath, A. C., Martin, N. G., Montgomery, G. W., Goddard, M. E., & Visscher, P. M. (2010). Common SNPs explain a large proportion of the heritability for human height. *Nature Genetics*, 42: 565–569. <https://doi.org/10.1038/ng.608>
- Yi, D., Ahn, J., & Ji, S. (2020). An effective optimization method for machine learning based on ADAM. *Applied Sciences (Switzerland)*, 10(3). <https://doi.org/10.3390/app10031073>
- Yu, S., Kim, J., Chung, H., Jung, K., Lee, Y., Yoon, D., Sang, B. (2005). Molecular cloning and characterization of bovine PRKAG3 gene: structure,

expression and single nucleotide polymorphism detection. *Journal of Animal Breeding and Genetics*, 122(5), 294-301.
<https://doi.org/10.1111/j.1439-0388.2005.00545.x>

Zhao, W., Lai, X., Liu, D., Zhang, Z., Ma, P., Wang, Q., Zhang, Z., & Pan, Y. (2020). Applications of support vector machine in genomic prediction in pig and maize populations. *Frontiers in Genetics*, 11.
<https://doi.org/10.3389/fgene.2020.598318>

Zhang, Y., Fitch, P., Vilas, M. P., & Thorburn, P. J. (2019). Applying multi-layer artificial neural network and mutual information to the prediction of trends in dissolved oxygen. *Frontiers in Environmental Science*, 7.
<https://doi.org/10.3389/fenvs.2019.00046>

Este capítulo se envió a la Revista Mexicana de Ciencias Pecuarias en enero de 2024

3. COMPARACIÓN DE MÉTODOS DE APRENDIZAJE AUTOMÁTICO PARA PREDICCIÓN DE VALORES DE CRÍA GENÓMICOS EN CARACTERÍSTICAS DE CRECIMIENTO EN BOVINOS

José Luis Velez-Labrada^a, Paulino Pérez-Rodríguez^{b*}, Mohammad Ali Nilforooshan^c, Agustín Ruíz-Flores^{a*}

^a Universidad Autónoma Chapingo. Posgrado en Producción Animal. Carretera Federal México-Texcoco Km 38.5, 56230, Texcoco, Estado de México, México.

^b Colegio de Postgraduados. Socio Economía Estadística e Informática. Carretera Federal México-Texcoco Km 36.5, 56227, Texcoco, Estado de México.

^c Livestock Improvement Corporation, Private Bag 3016, Hamilton 3240, New Zealand

*Autores para correspondencia: perpdgo@colpos.mx y aruizf@chapingo.mx

3.1. Resumen

Los algoritmos de Aprendizaje Automático (AA) han demostrado ventaja al abordar desafíos asociados con la cantidad y la complejidad de la información, permiten descubrir patrones, realizar análisis eficientes y servir como herramienta para la toma de decisiones. Este estudio, tuvo como objetivo comparar cuatro métodos de AA: redes neuronales artificiales (RN), árboles de regresión (AR), bosques aleatorios (BA) y máquina de soporte vectorial (SVM) para predecir el valor genómico en bovinos Suizo Europeo utilizando registros fenotípicos de pesos al nacimiento (PN), destete (PD) y al año (PA), así como información genómica. Los resultados indican que la capacidad predictiva de los modelos varía según la característica y la cantidad de información disponible. En general, RN, BA

y SVM mostraron un desempeño similar, mientras que AR tuvo un rendimiento inferior. La metodología SVM destacó como una herramienta prometedora, obteniendo los valores más altos de correlación Pearson entre fenotipos corregidos y valores genéticos predichos para PD y PA. A pesar de un mayor costo computacional RN, tuvo buen desempeño, especialmente para PN y PA. La selección del modelo final dependerá de las necesidades particulares de la aplicación, así como de factores prácticos como la disponibilidad de datos, recursos computacionales y la interpretabilidad; pero en general, RN y SVM surgieron como opciones sólidas en varias categorías.

Palabras clave: Árboles de regresión, bosques aleatorios, redes neuronales, capacidad predictiva.

COMPARISON OF MACHINE LEARNING METHODS TO PREDICT GENOMIC VALUES IN GROWTH TRAITS IN CATTLE

3.2. Abstract

Machine learning algorithms have proven advantageous to address challenges associated with the volume and complexity of information. They enable the discovery of patterns, perform efficient analysis, and serve as a tool in decision-making. This study compared four machine learning methods: artificial neural networks (ANN), regression trees (RT), random forests (RF), and support vectorial machines (SVM) to predict genomic breeding values in Braunvieh cattle. The phenotypic records of birth weight (BW), weaning weight (WW), and yearling weight (YW), along with genomic information were used. The results indicate that the predictive ability of the models varies depending on the trait and the amount of available information. Overall, ANN, RF, and SVM performed similarly, while RT had a lower performance. The SVM methodology resulted a promising tool. It reached

the highest Pearson's correlation of corrected phenotypes and predicted genetic values estimate for WW and YW. Despite the higher computational costs, ANN showed a good performance, especially for BW and YW. The selection of the final model will depend on the specific needs of the application, as well as practical factors such as data availability, computational resources, and interpretability; but overall, ANN and SVM emerged as robust options in several categories.

Key words: Regression tree, random forest, artificial neural networks, predictive ability.

3.3. Introducción

La genómica ha evolucionado en años recientes gracias a los avances en la tecnología de secuenciación de ADN. Estos avances han permitido la generación de grandes cantidades de datos a una velocidad sin precedentes. Sin embargo, el análisis y la interpretación de esta información masiva presentan desafíos significativos debido a su complejidad y dimensionalidad⁽¹⁾. En este contexto, los métodos de aprendizaje automático (AA) son herramientas valiosas que permiten extraer patrones y conocimientos de la información genómica y fenotípica de manera eficiente y precisa^(2,3).

La aplicación de técnicas de AA permite abordar tareas como la identificación de genes relevantes para características específicas, predicción de funciones génicas, detección de variantes genéticas asociadas a características particulares y clasificación de especies con base en información genómica^(4,5,6). Recientemente, el AA se ha vuelto atractivo en la predicción genómica, pero es necesario investigar cómo se compara en la predicción de valores genéticos con los métodos GBLUP convencionales⁽⁷⁾. Por lo que la combinación de datos genómicos con algoritmos de AA permitiría la creación de modelos predictivos

y descriptivos confiables, que a su vez tendría implicaciones en la cría selectiva, conservación de especies y la comprensión de la evolución^(8,9).

Entre los métodos de AA más utilizados están las redes neuronales, máquinas de soporte vectorial, árboles de decisión, regresión lineal y métodos de agrupación^(3,8-11). La diversidad de enfoques disponibles refleja la versatilidad de estos métodos en la resolución de desafíos con información genómica, como la clasificación de secuencias de ADN y la predicción de la estructura de proteínas⁽¹²⁾. El éxito de la aplicación de estos métodos en la genómica animal depende en gran medida de la disponibilidad de información⁽¹³⁾. Además, de elegir el método de AA óptimo, ya que se han propuesto una serie de métodos, cada uno con características propias y capacidades de predicción específicas con diferentes conjuntos de datos y características^(3,7).

Por lo anteriormente expuesto el presente estudio tuvo como objetivo comparar los métodos de AA: redes neuronales (RN), árboles de regresión (AR), bosques aleatorios (BA), y máquina de soporte vectorial (SVM) para predecir valores genómicos de cría utilizando registros fenotípicos de pesos al nacimiento, destete y al año, así como la información genómica de una población de bovinos Suizo Europeo en México.

3.4. Materiales y Métodos

3.4.1 Información fenotípica y genotípica

La información utilizada provino de la base de datos de la Asociación Mexicana de Criadores de Ganado Suizo de Registro (AMCGSR), la cual contiene registros fenotípicos e identificación de los animales, rancho de origen o propietario, genealogía y de caracteres de importancia económica como pesos al nacimiento (PN), al destete (PD) y al año (PA). Para implementar las metodologías se utilizaron estas características, realizando la

limpieza de la información como lo describen Valerio-Hernández *et al.*⁽¹⁴⁾, se omitieron individuos con información faltante en edad de la madre, manejo, hato de procedencia, así como individuos no relacionados genéticamente. Los grupos contemporáneos (GC) se definieron combinando los efectos de hato, año y época de nacimiento. Para PD los GC se conformaron considerando el manejo alimenticio que se da al hato, así como ajuste a días específicos para el destete. Se descartaron GC con menos de tres individuos o con varianza cero, de acuerdo con la metodología citada⁽¹⁴⁾.

La información genómica se obtuvo a través del análisis de muestras de pelo recolectadas de 300 animales de ranchos pertenecientes a la AMCGSR ubicados en Colima, Jalisco y Veracruz. El genotipado lo realizó la empresa GeneSeek (Lincoln, NE, USA), mediante el chip Genomic Profile Bovine LDv.4, el cual se ha utilizado para genotipar diversas razas *Bos indicus* y *Bos taurus*. El genotipado de 150 animales se hizo con un chip de 30,000, otros 150 animales se genotiparon con 50,000 marcadores SNP (Single Nucleotide Polymorphism, por sus siglas en inglés). Se filtraron los marcadores SNP en común para los dos chips, se obtuvieron 12,835 SNPs en común.

La recodificación y control de calidad de la información genotípica fue realizada siguiendo la metodología descrita por Valerio-Hernández *et al.*⁽¹⁵⁾ recodificando para efectos genéticos aditivos como AA=0, AB=1 y BB=2. Se obtuvo una matriz con 300 filas correspondientes a los individuos y 12,835 columnas a los marcadores, los valores faltantes se imputaron con $2\hat{\theta}_j$ donde $\hat{\theta}_j$ es la frecuencia estimada del alelo codificado como 1 para el j -ésimo marcador. Se omitieron marcadores monomórficos o con frecuencia de alelos menores (MAF)<0.04. Se obtuvieron 11,646 de los 12,835 SNP en

común que se utilizaron en análisis posteriores; en el Cuadro 1, se muestra el número de animales que se incorporaron en el estudio por cada característica después del filtrado.

Cuadro 1. Número de animales genotipados y fenotipados de una población de bovinos Suizo Europeo para tres características de crecimiento¹.

Grupo/Variable	PN ¹	PD ¹	PA ¹
Genotipado	300	300	300
Genotipado y fenotipado	330	267	232
Fenotipado en $\mathbf{G}^2$	232	218	191

¹ PN=peso al nacimiento, PD=peso al destete, PA=peso al año. ² Animales con fenotipos e información genómica.

La matriz de relaciones genómicas $\mathbf{G}$ se obtuvo de acuerdo con la metodología de Pérez-Rodríguez *et al.*⁽¹⁶⁾, $\mathbf{G} = \mathbf{W}\mathbf{W}^t/p$, donde $\mathbf{W}$ es la matriz de marcadores centrada y estandarizada y p el número total de marcadores. Adicionalmente se calculó la matriz de relaciones $\mathbf{H}$ que combina información de la matriz $\mathbf{G}$ con la información de la matriz de relaciones genéticas aditivas $\mathbf{A}$, obtenida para los individuos con pedigrí.

3.4.2. Modelos mixtos lineales (modelos base)

Con la finalidad de comparar los resultados del poder predictivo para valores de cría para peso al nacimiento, peso al destete y peso al año se consideran la secuencia de modelos analizados por Valerio-Hernández *et al.*⁽¹⁵⁾. Con la finalidad de presentar toda la información pertinente se describe a continuación en modelo mixto lineal utilizado por el mencionado autor, el cual está dado por:

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{Z}_1\mathbf{c} + \mathbf{Z}_2\mathbf{a} + \mathbf{e}, \dots (1)$$

donde $\mathbf{y}$ es el vector de fenotipos, $\mathbf{X}$ es la matriz de incidencias para efectos fijos, $\boldsymbol{\beta}$ es el vector de efectos fijos, $\mathbf{Z}_1$ es una matriz de incidencias que conecta los fenotipos con los grupos contemporaneos, $\mathbf{Z}_2$ es una matriz es una matriz de incidencias que conecta los fenotipos con efectos genéticos aditivos, $\mathbf{c} \sim NM(\mathbf{0}, \sigma_{gc}^2 \mathbf{I})$, con NM que denota la distribución normal multivariada, con media $\mathbf{0}$ y σ_{gc}^2 parámetro de varianza asociado, $\mathbf{I}$ la matriz identidad, $\mathbf{a} \sim NM(\mathbf{0}, \sigma_a^2 \mathbf{K})$, con $\mathbf{K} \in \{\mathbf{A}, \mathbf{G}, \mathbf{H}\}$. Dependiendo de los datos utilizados, el modelo (1) da lugar a tres modelos distintos que se denotan como sigue: 1) BLUP, $\mathbf{K} = \mathbf{A}$, 2) GBLUP, $\mathbf{K} = \mathbf{G}$ y 3) ssGBLUP (GBLUP en un solo paso) $\mathbf{K} = \mathbf{H}$.

3.4.3. Modelos de aprendizaje automático

Las variables de entrada para los algoritmos de AA que se utilizaron fueron la matriz de relaciones genéticas que combina la información genómica e información de pedigrí denominada $\mathbf{H}$, así como los efectos de edad de la madre para cada animal, variables indicadoras de sexo, grupo contemporáneo descritas previamente. Para poder incluir la información de la matriz $\mathbf{H}$ en los modelos de aprendizaje se realizó la descomposición espectral de la misma, es decir, $\mathbf{H} = \mathbf{\Gamma} \mathbf{\Lambda} \mathbf{\Gamma}^t$ a partir de la cual se obtuvo $\mathbf{X} = \mathbf{\Gamma} \mathbf{\Lambda}^{\frac{1}{2}}$ (componentes principales) mismos que se utilizaron como covariables (variables explicatorias) en los modelos, esta estrategia computacional y otras relacionadas han sido utilizadas por otros autores anteriormente^(17,18).

3.4.4. Red neuronal artificial

Las redes neuronales (RN) son modelos que en sus inicios pretendían emular el funcionamiento del sistema nervioso, donde a través de operadores matemáticos procesan información de entrada, generando valores de salida o el resultado final^(3,19). Las variables de entrada afectan el desempeño del modelo, puede generar sobreajuste si la cantidad de

información es grande por lo que es importante optimizar dichas variables⁽²⁰⁾. Una de las ventajas de las redes neuronales es su capacidad de aprender patrones no lineales⁽³⁾. El modelo de una RN con una capa de entrada con p predictores, una capa oculta con S neuronas y una capa de salida con una respuesta continua puede expresarse como:

$$y_i = \beta_0 + \sum_{k=1}^S w_k g(\beta_0^{(k)} + \sum_{j=1}^p \beta_j^{(k)} x_{ij}) + e_i,$$

donde $e_i \sim NIID(0, \sigma_e^2)$, con *NIID* denotando variables aleatorias normales, independientes e idénticamente distribuidas; $k = 1, \dots, S$ (neuronas); $j = 1, \dots, p$ (predictores); $i = 1, \dots, n$ (observaciones); y $g(\cdot)$ es la función de activación, de acuerdo con Bai *et al.*⁽²¹⁾ y Gianola *et al.*⁽¹⁷⁾, donde, y_i es la variable respuesta para el i -ésimo individuo, para este caso los pesos de características de crecimiento en bovinos Suizo Europeo (PN, PD, PA) que la red predice en función de las entradas; β_0 es el término de sesgo o el intercepto, este puede representar el valor predicho cuando las entradas son igual a cero y w_k son pesos asociados a cada una de las neuronas y determinan la contribución de cada una de ellas en la predicción final. La RN ajusta los parámetros (β 's, w 's) durante el proceso de entrenamiento para minimizar el error de predicción. La función de activación, $g(\cdot)$, mapea las entradas de la línea real al intervalo abierto acotado $(-1, 1)$, ejemplificado por Pérez-Rodríguez *et al.*⁽²²⁾ donde $g(x) = 2/[1 + \exp(-2x)] - 1$ se conoce como la función de activación tangente hiperbólica (*tanh*).

3.4.5. Árboles de regresión

Este modelo se basa en el planteado por Breiman *et al.*⁽²³⁾, $y_i = \sum_{j=1}^J y_j I(\mathbf{x}_i \in R_j)$, donde y_i es una variable respuesta (PN, PD y PA), y_j es el valor de regresión que asociado con una “hoja”, $\mathbf{x}_i$ es el conjunto de características de la observación, R_j es la región asociada a la “hoja j ” definida por características y valores de corte en el camino desde la “raíz”

hasta la “hoja”. $I(\cdot)$ es una función indicadora que toma el valor 1 si la observación i pertenece a la región R_j . El árbol busca encontrar las divisiones que minimicen el error en cada región y dividir recursivamente hasta alcanzar un criterio de alto del proceso, como la profundidad máxima del árbol o el número mínimo de ejemplos en una hoja.

3.4.6. Bosques aleatorios

Este modelo combina múltiples AR donde las predicciones de cada uno se promedian para obtener una predicción final optimizada, $y_i = \frac{1}{N} \sum_{j=1}^N y_{ij}$, donde N es el número de árboles del bosque aleatorio, y_i es una variable aleatoria observada (PN, PD y PA) y y_{ij} es la predicción del j -ésimo AR para la observación i .

3.4.7. Máquina de soporte vectorial

El modelo de máquina de soporte vectorial (SVM), se utiliza para clasificación y regresión⁽²⁴⁾. En el contexto de regresión, dado un conjunto de datos $\{y_1, \mathbf{x}_1\}, \dots, \{y_n, \mathbf{x}_n\}$, donde y_i representa el valor de la variable respuesta (continua) para el i -ésimo individuo, $\mathbf{x}_i$ el valor de las covariables asociadas, el objetivo es obtener una función $f(\mathbf{x})$ de tal forma que la distancia con y no sea más grande que ε para cada uno de los puntos de entrenamiento. De acuerdo con Hastie *et al.*⁽²⁵⁾ la función de regresión se aproxima en términos de funciones base $\{h_m(\mathbf{x})\}, m = 1, \dots, M$ como sigue:

$$f(\mathbf{x}) = \beta_0 + \sum_{m=1}^M \beta_m h_m(\mathbf{x}),$$

donde $\beta = (\beta_0, \beta_1, \dots, \beta_M)^t$ son coeficientes que se obtienen al minimizar:

$$Q(\beta) = \sum_{i=1}^n L(y_i - f(\mathbf{x}_i)) + \frac{\lambda}{2} \sum \beta_m^2,$$

en la que $L(\cdot)$ se denomina función de pérdida (por ejemplo, cuadrática o valor absoluto) y λ es un parámetro de regularización positivo. Para cualquier selección de $L(\cdot)$, la solución tiene la forma:

$$\hat{f}(\mathbf{x}) = \sum_{i=1}^n \hat{a}_i K(\mathbf{x}, \mathbf{x}_i),$$

con $K(\cdot, \cdot)$ conocida como función kernel. Los kernel se consideran en el modelo como elementos fundamentales, sirven como funciones que permiten transformar los datos y generar un espacio de mayor dimensión, ayudan a modelar relaciones complejas en los datos. Los kernel más comunes son el lineal $(\mathbf{x}_i^t \mathbf{x}_j)$, polinomial $(\mathbf{x}_i^t \mathbf{x}_j + coef_0)^d, d = 2, 3, \dots$, radial $(e^{-\gamma \|\mathbf{x}_i - \mathbf{x}_j\|^2})$ y sigmoide $(\tanh(\gamma \mathbf{x}_i^t \mathbf{x}_j) + coef_0)$, donde γ se conoce como ancho de banda que se ajusta en el proceso de entrenamiento mediante validación cruzada, $coef_0$ es una constante que se puede ajustar durante el proceso de entrenamiento del modelo, aunque usualmente se fija a 1.

3.4.8. Validación cruzada

La validación cruzada es un método de re-muestreo de datos muy utilizado para estimar el verdadero error de predicción de los modelos y ajustar los parámetros del modelo^(17,26). Por tanto, para obtener la capacidad de predicción de los modelos RN, AR, BA y SVM, y con ello hacer la comparación, se realizó la validación cruzada teniendo como referencia lo realizado por Valerio-Hernández *et al.*⁽¹⁵⁾. Estos autores dividieron los datos aleatoriamente en 80% para el conjunto de entrenamiento y 20% para el de validación y el proceso se repitió 100 veces. Se ajustaron los modelos de AA y se obtuvieron las correlaciones entre los valores observados vs predichos, donde los valores observados se consideraron los valores de la variable respuesta corregidos por efectos fijos y otros efectos aleatorios. Se obtuvo el coeficiente de correlación de Pearson de los fenotipos corregidos y valores genéticos predichos para cada una de las particiones y se obtuvieron los promedios para cada uno de los modelos.

3.4.9. Software y lenguaje de programación

Se utilizaron un equipo de cómputo con sistema operativo Windows 11 Home Single Language de 64 bits, con procesador Intel(R) Core (TM) i5-10210U CPU @ 1.60GHz, 2.11 GHz, 8 GB RAM. El lenguaje de programación utilizado fue R⁽²⁷⁾ (v. 4.3.0), y usando RStudio como entorno de desarrollo integrado para el lenguaje (v. 2023.06.1+524). Se decidió utilizar R ya que es un lenguaje de uso libre y distribución gratuita y además, incluye bibliotecas de funciones para el desarrollo y utilización de algoritmos de métodos de AA.

3.4.10. Ajuste de los modelos

Los códigos de los ajustes de cada uno de los modelos en algoritmos de AA se realizaron utilizando las bibliotecas de funciones para el software R, que a continuación se describen. Para el modelo de RN se utilizó la función “brnn”⁽²⁸⁾, incluida en la biblioteca de funciones del mismo nombre (versión 0.9.3), el cual ajusta la RN descrita anteriormente e implementa el algoritmo de Gauss-Newton para la estimación de parámetros. Con relación al algoritmo árbol de regresión la función utilizada fue “rpart”⁽²⁹⁾ incluida en la biblioteca de funciones del mismo nombre (versión 4.1.19), el software implementa el modelo descrito por Breiman *et al.*⁽²³⁾ donde los árboles de regresión dividen repetidamente los datos en grupos pequeños de acuerdo con características específicas, para hacer predicciones y tomar el promedio de los valores en el grupo final.

La implementación del algoritmo bosques aleatorios se realizó utilizando la función “randomForest” incluida en el paquete del mismo nombre (versión 4.7-1.1) la cual implementa los algoritmos descritos por Breiman⁽³⁰⁾, donde usa una combinación de árboles que dependen de valores de un vector aleatorio que se muestrea aleatoriamente y

con distribución igual para la totalidad de los árboles del bosque. En lo referente al modelo máquina de soporte vectorial, el ajuste se realizó con el paquete “e1071”⁽³¹⁾ (versión 1.7-13), con ayuda de la función svm. Los códigos están disponibles previa solicitud al autor para correspondencia.

3.5. Resultados

En el Cuadro 2 se muestran los promedios de las cien estimaciones de validación cruzada de las correlaciones Pearson entre valores corregidos y predichos para las características PN, PD y PA, utilizando los cuatro algoritmos de AA comparados en el estudio. Para PD y PA, el algoritmo SVM fue con el que se obtuvieron los valores más altos para las correlaciones entre valores corregidos y predichos en los conjuntos de validación (PD=0.350 y PA=0.208). En este método para las tres características el mejor ajuste se obtuvo utilizando el “Kernel Radial” optimizando los hiperparámetros γ (gamma) y costo (PN: 0.045 y 0.05; PD y PA: 0.05 y 0.01). Con la metodología de RN Artificial se realizaron varias pruebas en relación con el número de neuronas en la capa oculta del modelo, siendo 3 neuronas para PN y 2 para PD y PA cuando se obtuvieron estimadores de los parámetros adecuados de ponderadores generando un modelo parsimonioso. Para PN y PA, este algoritmo obtuvo el primer y segundo mejor rendimiento con 0.425 y 0.183, respectivamente.

Con la metodología BA se realizaron pruebas con diferente cantidad de “árboles” dentro de los parámetros del modelo, siendo 150 árboles para PN, 200 para PD y 250 para PA los que obtuvieron valores óptimos de predicción; para las características PN y PA obtuvieron el tercer mejor rendimiento en cuestión de sus predicciones; en la característica PD se comportó de manera similar a RN obteniendo correlaciones de 0.234 y 0.232. En

relación con la metodología AR mostró menor capacidad predictiva para las tres características de esta investigación.

Cuadro 2. Promedios de los estimadores de la correlación Pearson y desviación estándar entre fenotipos corregidos y valores genéticos predichos, para las 100 validaciones cruzadas para las tres características de crecimiento y los algoritmos comparados.

Característica	Algoritmo	Cor	SD
PN ¹	Red neuronal	0.425	0.159
	Árbol de regresión	0.162	0.152
	Bosques aleatorios	0.218	0.148
	Máquina de soporte vectorial	0.350	0.133
PD ¹	Red neuronal	0.232	0.111
	Árbol de regresión	0.137	0.157
	Bosques aleatorios	0.234	0.121
	Máquina de soporte vectorial	0.278	0.125
PA ¹	Red neuronal	0.183	0.157
	Árbol de regresión	0.024	0.176
	Bosques aleatorios	0.101	0.120

¹ PN=peso al nacimiento, PD=peso al destete, PA=peso al año; Cor = correlación de Pearson; SD= desviación estándar de los 100 estimadores de la correlación seleccionados al azar.

3.6. Discusión

Para determinar la capacidad predictiva de los modelos de AA, se compararon los estimadores de la correlación de Pearson entre los fenotipos corregidos y los valores genéticos predichos con los modelos comparados⁽³²⁾. El utilizar estas metodologías para la obtención de los valores predichos responde a que la estructura de la información no permite realizar un modelo estadístico paramétrico^(33,34), por lo que el uso de algoritmos de AA son una alternativa para modelar funciones complejas identificando relaciones no lineales entre las covariables y la variable respuesta⁽¹⁷⁾. Las correlaciones entre fenotipos corregidos y valores predichos con las metodologías utilizadas permiten evaluar los algoritmos de aprendizaje automático RN, AR, BA y SVM, para las características de crecimiento en bovinos PN, PD y PA. En general los algoritmos RN, BA y SVM mostraron un desempeño predictivo similar a las metodologías evaluadas por Valerio-Hernández *et al.*⁽¹⁵⁾, quienes trabajaron con las mismas variables. En un estudio donde se comparó la capacidad predictiva de redes neurales no lineales (RNNL) con modelos lineales, se encontró que éstas pueden ser útiles en la predicción para características complejas utilizando información genómica, situación en la que ordinariamente el número de parámetros a estimar supera el tamaño de muestra⁽¹⁷⁾. Por su parte, Rodríguez-Alcántar⁽³⁾ compararon algoritmos de AA utilizando diversos conjuntos de SNPs generados a partir de cromosomas con alto número de QTL asociados con alta producción lechera. Este autor encontró que la precisión de la clasificación varió de 90.9 a 94.5% con

árboles de decisión, y de 79.0 a 87.3% con redes neuronales. El autor concluye que tanto el método de redes neuronales para clasificación binaria, como los árboles de decisión son herramientas eficientes para la identificación temprana de vacas lecheras altas productoras.

Los resultados indican que el desempeño de los modelos varía según la característica y la cantidad de información⁽¹⁷⁾, entre otros factores. Lo anterior sugiere que pueden obtenerse mejores resultados con estos modelos al incluir más información de variables y covariables para ajustar el modelo en entrenamiento^(35,36).

Las metodologías de SVM, RN y BA mostraron buen desempeño en términos de los coeficientes de correlación de Pearson de los fenotipos corregidos y los valores predichos para las tres características de crecimiento utilizadas. El costo computacional de RN fue mayor que el de los otros tres algoritmos comparados, este resultado es similar al que reportaron Zhao *et al.*⁽³⁷⁾ quienes mencionan que el ajuste de la RN es más complicado y requiere más tiempo. El algoritmo SVM destacó como una herramienta prometedora para la predicción utilizando información genómica, considerando la cantidad de información y los parámetros utilizados con esta metodología, así como el Kernel⁽²⁵⁾; este algoritmo aporta a las aplicaciones de AA para el análisis de conjuntos de datos provenientes de información genética y genómica^(38,39).

3.7. Conclusiones

La capacidad predictiva de los modelos de AA depende en gran medida de la cantidad de información disponible. Una mayor cantidad de información permite entrenar mejor los modelos y obtener predicciones más confiables. Debe tenerse en consideración el costo computacional al utilizar estas metodologías. Los métodos redes neuronales artificiales y

máquina de soporte vectorial tienden a tener un desempeño sólido en varios indicadores. Sin embargo, la elección final del modelo depende de las necesidades específicas de la aplicación y de los aspectos de evaluación.

Agradecimientos

Al Consejo Nacional de Humanidades, Ciencias y Tecnologías, México, por el financiamiento para el primer autor durante sus estudios de maestría. A la Asociación Mexicana de Criadores de Ganado Suizo de Registro por permitir el uso de su información.

Conflictos de interés

Los autores declaran que no existen conflictos de interés.

3.8. Literatura citada

1. Pérez-Enciso M, Steibel JP. Phenomes: the current frontier in animal breeding. *Genet Sel Evol* 2021;53(1):22. doi: 10.1186/s12711-021-00618-1. PMID: 33673800; PMCID: PMC7934239.
2. Song H, Dong T, Yan X, Wang W, Tian Z, Hu H. Using Bayesian threshold model and machine learning method to improve the accuracy of genomic prediction for ordered categorical traits in fish. *Agric Comm* 2023;1(1):100005. <https://doi.org/10.1016/j.agrcom.2023.100005>
3. Rodríguez-Alcántar E. Aplicación de algoritmos de aprendizaje automático para la clasificación de ganado lechero utilizando SNP de genoma completo. 2019. (Tesis de Doctorado, Instituto de Ingeniería, Universidad Autónoma de Baja California. México). Disponible en:

<https://repositorioinstitucional.uabc.mx/bitstream/20.500.12930/1834/1/MXL122258.pdf>. Consultado el 18 de septiembre de 2023.

4. Campos TL, Korhonen PK, Hofmann A, Gasser RB, Young ND. Harnessing model organism genomics to underpin the machine learning-based prediction of essential genes in eukaryotes – Biotechnological implications. *Biotechnol Adv* 2022;54, 107822. <https://doi.org/10.1016/J.BIOTECHADV.2021.107822>
5. Zhao T, Wu H, Wang X, Zhao Y, Wang L, Pan J, *et al.* Integration of eQTL and machine learning to dissect causal genes with pleiotropic effects in genetic regulation networks of seed cotton yield. *Cell Rep* 2023;42(9). <https://doi.org/10.1016/j.celrep.2023.113111>
6. Guo T, Li X. Machine learning for predicting phenotype from genotype and environment. *Curr Opin Biotechnol* 2023;79,102853. <https://doi.org/10.1016/J.COPBIO.2022.102853>
7. Wang X, Shi S, Wang G, Luo W, Wei X, Qiu A, *et al.* Using machine learning to improve the accuracy of genomic prediction of reproduction traits in pigs. *J Anim Sci Biotechnol* 2022;13(60): <https://doi.org/10.1186/s40104-022-00708-0>
8. Long N, Gianola D, Rosa GJM, Weigel KA. Application of support vector regression to genome-assisted prediction of quantitative traits. *Theor Appl Genet* 2011;123(7):1065-1074. <https://doi.org/10.1007/s00122-011-1648-y>
9. González-Camacho JM, Ornella L, Pérez-Rodríguez P, Gianola D, Dreisigacker S, Crossa J. Applications of machine learning methods to genomic selection in breeding wheat for rust resistance. *Plant Genome* 2018;11(2):170104. <https://doi.org/10.3835/plantgenome2017.11.0104>

10. Müller AC, Guido S. Introduction to machine learning with Python: a guide for data scientists. O'Reilly Media, Inc. 2016.
11. Azodi CB, Bolger, E, McCarren, A, Roantree, M, de los Campos, G, Shiu, SH. Benchmarking parametric and machine learning models for genomic prediction of complex traits. *G3: Genes, Genomes, Genetics* 2019;9(11):3691-3702. <https://doi.org/10.1534/g3.119.400498>
12. Fa R, Cozzetto D, Wan C, Jones DT. Predicting human protein function with multitask deep neural networks. *PLoS ONE* 2018;13(6). <https://doi.org/10.1371/journal.pone.0198216>
13. VanRaden, PM. Efficient methods to compute genomic predictions. *J Dairy Sci* 2008;91(11):4414-4423. <https://doi.org/10.3168/jds.2007-0980>
14. Valerio-Hernández JE, Pérez-Rodríguez P, Ruíz-Flores A. Quantile regression for prediction of complex traits in Braunvieh cattle using SNP markers and pedigree. *Rev Mex Cienc Pecu* 2023;14(1):172–189. <https://doi.org/10.22319/rmcp.v14i1.6182>
15. Valerio-Hernández JE, Ruíz-Flores A, Nilforooshan MA, Pérez-Rodríguez P. Single-step genomic evaluation for growth traits in a Mexican Braunvieh cattle population. *Anim Biosci* 2023;36(7):1003-1009. <https://doi.org/10.5713/ab.22.0158>
16. Pérez-Rodríguez P, Crossa J, Rutkoski J, Poland J, Singh R, Legarra A, *et al.* Single-step genomic and Pedigree Genotype × Environment interaction models for predicting wheat lines in international environments. *Plant Genome* 2017;10(2). <https://doi.org/10.3835/plantgenome2016.09.0089>
17. Gianola D, Okut H, Weigel KA, Rosa GJM. Predicting complex quantitative traits with Bayesian neural networks: a case study with Jersey cows and wheat. *BMC Genetics* 2011;12:87. doi:10.1186/1471-2156-12-87

18. Pérez-Rodríguez P, Flores-Galarza S, Vaquera-Huerta H, del Valle-Paniagua DH, Montesinos-López OA, Crossa J. Genome-based prediction of Bayesian linear and non-linear regression models for ordinal data. *Plant Genome* 2020;13(2). <https://doi.org/10.1002/tpg2.20021>
19. Peng J, Yan G, Druzhinin Z. Applying an artificial neural network- Developed collective animal behavior algorithm for seismic reliability evaluation of structure. *Measurement* 2023;220:113355. Disponible en: <https://www.sciencedirect.com/science/article/pii/S0263224123009193>
20. Wang C, Xu S, Liu J, Yang J, Liu C. Building an improved artificial neural network model based on deeply optimizing the input variables to enhance rutting prediction. *Constr Build Mater* 2022;348:128658. Disponible en: <https://www.sciencedirect.com/science/article/pii/S0950061822023157>
21. Bai S, Kolter JZ, Koltun V. An empirical evaluation of generic convolutional and recurrent networks for sequence modeling. 2018. <http://arxiv.org/abs/1803.01271>
22. Pérez-Rodríguez P, Gianola D, Weigel KA, Rosa GJM, Crossa J. Technical Note: An R package for fitting Bayesian regularized neural networks with applications in animal breeding. *J Anim Sci* 2013;91(8):3522–3531. <https://doi.org/10.2527/jas.2012-6162>
23. Breiman L, Friedman JH, Olshen RA, Stone CJ. *Classification and Regression Trees*. Wadsworth. 1984. <https://doi.org/10.1201/9781315139470>
24. Chih-Chung C, Chih-Jen L. LIBSVM: A library for support vector machines. *ACM Trans Intell Syst Technol* 2011;2(3):1-27. Software available at <http://www.csie.ntu.edu.tw/~cjlin/libsvm>

25. Hastie T, Tibshirani R, Friedman J. The Elements of Statistical Learning: Data Mining, Inference, and Prediction. Berlin: Springer Science & Business Media. 2008. Disponible en: <https://books.google.com.mx/books?id=VRzITwgNV2UC>
26. Berrar D. Cross-Validation. ABC of Bioinformatics. Elsevier 2018; 542-545. <https://doi.org/10.1016/B978-0-12-809633-8.20349-X>
27. R Core Team. R: A language and environment for statistical computing. R Foundation for statistical computing 2021. Vienna, Austria. <https://www.R-project.org/>.
28. Pérez-Rodríguez P, Gianola D. Title Bayesian Regularization for Feed-Forward Neural Networks. 2022. Disponible en: <https://cran.r-project.org/package=brnn>
29. Therneau T, Atkinson B, Ripley B. Package “rpart”. 2022. Disponible en: <https://cran.r-project.org/package=rpart>
30. Breiman L. Random forests. Machine Learning 2001;45(1):5–32. <https://doi.org/10.1023/A:1010933404324>
31. Meyer D, Dimitriadou E, Hornik K, Weingessel A, Leisch F, Chang C, Lin C. Package “e1071”. 2023. Disponible en: <https://cran.r-project.org/package=e1071>
32. González-Recio O, Forni S. Genome-wide prediction of discrete traits using bayesian regressions and machine learning. Genet Sel 2011;43(1): <https://doi.org/10.1186/1297-9686-43-7>
33. Gianola D, de los Campos G. Inferring genetic values for quantitative traits non-parametrically. Genet Res (Camb) [Internet]. 2009/01/06. 2008;90(6):525–40. Disponible en: <https://www.cambridge.org/core/article/inferring-genetic-values-for-quantitative-traits-nonparametrically/F60174E4AF179F871DB0ECB9904FDC46>

34. Gianola D, Fernando RL, Stella A. Genomic-Assisted prediction of genetic value with semiparametric procedures. *Genetics*. 2006;173(3):1761–76. <https://doi.org/10.1534/genetics.105.049510>
35. Monaco A, Pantaleo E, Amoroso N, Lacalamita A, Lo Giudice C, Fonzino A, *et al.* A primer on machine learning techniques for genomic applications. *Comput Struct Biotechnol J* 2021;19:4345-4359. Elsevier B.V. <https://doi.org/10.1016/j.csbj.2021.07.021>
36. Alves AAC, da Costa RM, Bresolin T, Fernandes Júnior GA, Espigolan R, Ribeiro AMF, *et al.* Genome-wide prediction for complex traits under the presence of dominance effects in simulated populations using GBLUP and machine learning methods. *J Anim Sci* 2020;98(6): 1–11. <https://doi.org/10.1093/jas/skaa179>
37. Zhao W, Lai X, Liu D, Zhang Z, Ma P, Wang Q, *et al.* Applications of support vector machine in genomic prediction in pig and maize populations. *Front Genet* 2020;11:598318. <https://doi.org/10.3389/fgene.2020.598318>
38. González-Camacho JM, Ornella L, Pérez-Rodríguez P, Gianola D, Dreisigacker S, Crossa J. Applications of machine learning methods to genomic selection in breeding wheat for rust resistance. *Plant Genome* 2018;11(2). <https://doi.org/10.3835/plantgenome2017.11.0104>
39. Libbrecht M, Noble W. Machine learning applications in genetics and genomics. *Nat Rev Genet* 2015;16:321-332. <https://doi.org/10.1038/nrg3920>

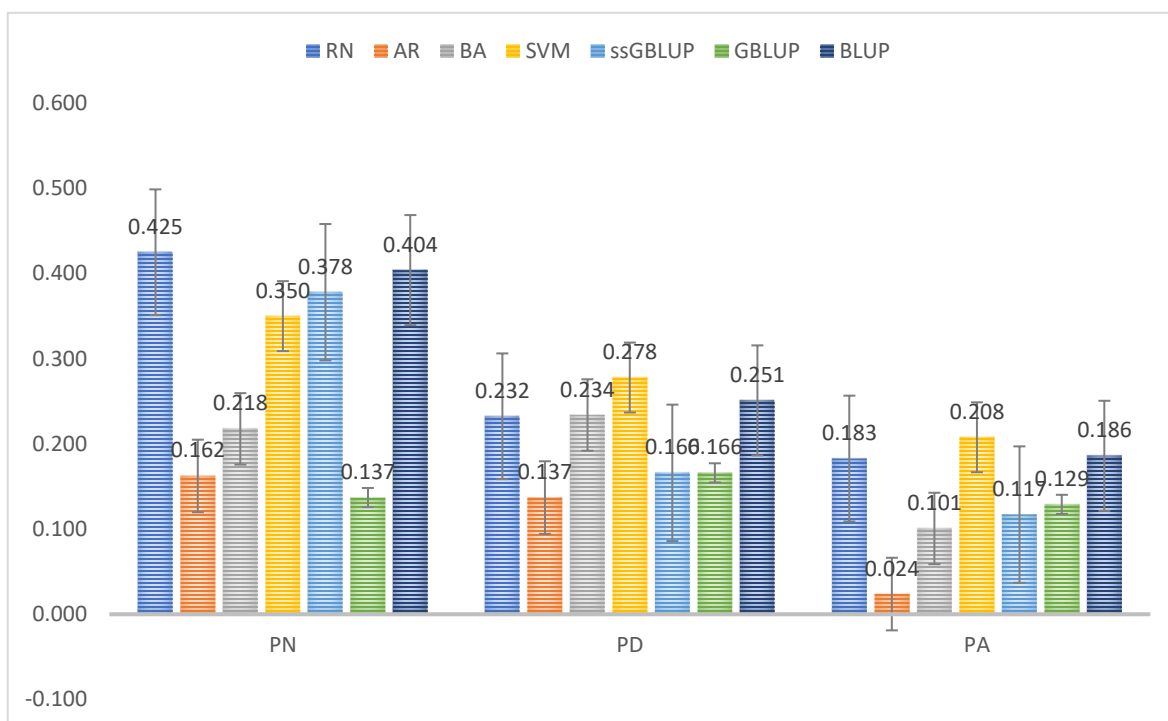


Figura 1. Comparación de los coeficientes de correlación (promedio de las 100 validaciones) de los fenotipos corregidos y valores genéticos predichos obtenidos con los métodos de aprendizaje automático, redes neuronales artificiales (RN), árboles de regresión (AR), bosques aleatorios (BA) y máquina soporte vectorial (SVM) con las metodologías realizadas por Valerio-Hernández *et al.*⁽¹⁵⁾, mejor predictor lineal insesgado (BLUP), BLUP genómico (GBLUP) y GBLUP de un solo paso (ssGBLUP) para las características peso al nacer (PN), peso al destete (PD) y peso al año (PA) de una población de bovinos suizo europeo.