https://doi.org/10.22319/rmcp.v16i1.6616

Nota de investigación

Comparación de métodos de aprendizaje automático para predicción de valores de cría genómicos en características de crecimiento en bovinos Suizo Europeo

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. Campus Montecillo. Socioeconomía Estadística e Informática. Texcoco, Estado de México, México.

[c] Livestock Improvement Corporation, Hamilton, New Zealand.

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

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 desempeño inferior. La metodología SVM

destacó como la herramienta con mayor potencial, obteniendo los valores más altos de correlación Pearson entre fenotipos corregidos y valores genéticos predichos para PD. A pesar de un mayor costo computacional, RN tuvo un desempeño razonable, especialmente para PN y PA. La selección del modelo final depende 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.

Recibido: 08/01/2024

Aceptado: 28/11/2024

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, por la complejidad inherente a los datos genómicos, así como su dimensionalidad, plantean obstáculos importantes[1]. La diversidad de la información genómica, que abarca desde secuencias de ADN hasta datos fenotípicos asociados, añade una capa adicional de complejidad. Además, la variabilidad en la calidad y la estructura de los datos genómicos puede dificultar la extracción de conocimientos útiles y significativos. En este contexto, los métodos de aprendizaje automático (AA) emergen como herramientas valiosas para abordar estos desafíos, estas metodologías ofrecen la capacidad de procesar y analizar grandes volúmenes de datos de manera eficiente y precisa[2]. Su capacidad para identificar patrones complejos y relaciones no lineales en datos genómicos y fenotípicos los convierten en una herramienta poderosa para la extracción de conocimientos[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 con 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 por su capacidad para manejar grandes volúmenes de datos, su flexibilidad en el modelar relaciones no lineales, mejorar la precisión predictiva y las continuas innovaciones en algoritmos y técnicas, 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 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.

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). El conjunto de datos utilizado fue analizado previamente por Valerio-Hernández *et al*[14,15] para ajustar otros modelos, por lo que algunos resultados de los aquí obtenidos se comparan en forma directa con los de los autores mencionados. El tratamiento de la información fenotípica para PN, PD y PA se realizó el procedimiento descrito por Valerio-Hernández *et al*[14,15], es decir, 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). Se realizó una selección de marcadores SNP que estuvieran presentes en ambos chips, lo que resultó en un conjunto de 12,835 SNP en común.

La recodificación y control de calidad de la información genotípica realizada por Valerio-Hernández *et al*[15], se basó en lo realizado por Jarquín *et al*[16] recodificando para efectos genéticos aditivos como AA=0, AB=1 y BB=2. Para la imputación de genotipos faltantes en el presente estudio, se utilizó el software FImpute[17] (versión 2.2), para lo cual se obtuvo: 1) el mapa de los marcadores (marcador, cromosoma, posición en pares de base), eliminando marcadores duplicados o marcadores con posiciones desconocidas, 2) el pedigrí de los individuos y su correspondiente sexo. Se eliminaron marcadores monomórficos y aquellos con una frecuencia del alelo menor (MAF) más pequeña que 0.04. Se obtuvieron 9,008 marcadores que se utilizaron para la construcción de la matriz de relaciones genómicas **G**; 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 crecimiento

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

PN= peso al nacimiento, PD= peso al destete, PA= peso al año. $\mathbf{G}^2$ Animales con fenotipos e información genómica.

La matriz de relaciones genómicas $\boldsymbol{G}$ se obtuvo de acuerdo con la metodología descrita por Pérez-Rodríguez *et al*[18], $\boldsymbol{G = WW^t/p}$, donde $\boldsymbol{W}$ es la matriz de marcadores centrada y estandarizada y p el número total de marcadores. Adicionalmente se calculó la matriz de relaciones $\boldsymbol{H}$ que combina información de la matriz $\boldsymbol{G}$ con la información de la matriz de relaciones genéticas aditivas $\boldsymbol{A}$, obtenida para los individuos con pedigrí.

Modelos mixtos lineales (modelos base). Con el objetivo de comparar los resultados del poder predictivo para valores de cría para PN, PD y PA para modelos mixtos lineales *vs* modelos de aprendizaje automático se consideran la secuencia de modelos y de resultados descritos 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:

$$\boldsymbol{y} = \boldsymbol{Xb} + \boldsymbol{Z}_1\boldsymbol{c} + \boldsymbol{Z}_2\boldsymbol{a} + \boldsymbol{e}, \ldots (1)$$

donde $\boldsymbol{y}$ es el vector de fenotipos, $\boldsymbol{X}$ es la matriz de incidencias para efectos fijos, que para este estudio son: sexo del animal, edad de la madre de cada animal y el grupo contemporáneo ya descritas, $\boldsymbol{b}$ es el vector de efectos fijos, $\boldsymbol{Z}_1$ es una matriz de incidencias que conecta los fenotipos con los grupos contemporáneos, cuyos efectos se suponen aleatorios y representan la variabilidad en los fenotipos que se debe a diferencias entre grupos de individuos que están sujetos a las mismas condiciones ambientales y de manejo, $\boldsymbol{c} \sim NM(\boldsymbol{0}, \sigma^2_{gc}\boldsymbol{I})$, con NM que denota la distribución normal multivariada, con media $\boldsymbol{0}$ y σ^2_{gc} parámetro de varianza asociado, $\boldsymbol{I}$ la matriz identidad, $\boldsymbol{Z}_2$ es una matriz es una matriz de incidencias que conecta los fenotipos con efectos genéticos aditivos los cuales se suponen como efectos aleatorios, $\boldsymbol{a} \sim NM(\boldsymbol{0}, \sigma^2_a\boldsymbol{K})$, con $\boldsymbol{K} \in \{\boldsymbol{A}, \boldsymbol{G}, \boldsymbol{H}\}$, $\boldsymbol{e} \sim NM(\boldsymbol{0}, \sigma^2_e\boldsymbol{I})$ representa el vector de errores aleatorios, donde σ^2_e denota la variabilidad asociada con los mismos. Dependiendo de los datos utilizados, el modelo (1) da lugar a tres modelos distintos que se denotan como sigue: 1) BLUP, $\boldsymbol{K} = \boldsymbol{A}$, 2) GBLUP, $\boldsymbol{K} = \boldsymbol{G}$ y 3) ssGBLUP (GBLUP en un solo paso) $\boldsymbol{K} = \boldsymbol{H}$. Los modelos mixtos lineales descritos previamente fueron ajustados por Valerio-Hernández *et al*[15] utilizando el paquete estadístico BGLR[19].

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 $\boldsymbol{H}$, 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 $\boldsymbol{H}$ en los modelos de aprendizaje se realizó la descomposición espectral de la misma, es decir, $\boldsymbol{H} = \boldsymbol{\Gamma\Lambda\Gamma}^t$ a partir de la cual se obtuvo $\boldsymbol{X} = \boldsymbol{\Gamma\Lambda}^{\frac{1}{2}}$ (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[20,21].

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[3,22]. 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[23]. 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; $k = 1, \ldots, S$ (neuronas); $j = 1, \ldots, p$ (predictores); $i = 1, \ldots, n$ (observaciones); y $g(\cdot)$ es la función de activación, de acuerdo con Bai *et al*[24] y Gianola *et al*[20], 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, éste 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 capa oculta es una capa intermedia entre la capa de entrada y la capa de salida, es donde se lleva a cabo la mayor parte del procesamiento y la extracción de características del conjunto de datos; está compuesta por un número específico de neuronas (*S*) es un hiperparámetro del modelo que se ajusta durante el proceso de entrenamiento. Un valor más alto de *S* permite que la red neuronal capture una mayor complejidad en los datos, pero también puede aumentar el riesgo de sobreajuste. La RN ajusta los parámetros $(\beta's, w's)$ 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*[25] donde $g(x) = 2/[1 + \exp(-2x)] - 1$ se conoce como la función de activación tangente hiperbólica (*tanh*). Para el ajuste del modelo de red neuronal se utilizó la función "brnn"[26] incluida en el paquete del mismo nombre (versión 0.9.3) en el paquete estadístico R[27] (versión 4.3.0).

Árboles de regresión. Este modelo se basa en el planteado por Breiman *et al*[28], $y_i = \sum_{j=1}^{J} y_j I(\boldsymbol{x}_i \epsilon\, R_j)$, donde y_i es una variable respuesta (PN, PD y PA), y_j es el valor de regresión que asociado con una "hoja", $\boldsymbol{x}_i$ 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 parada del proceso, como la profundidad máxima del árbol o el número mínimo de casos en una hoja. El ajuste del modelo se realizó con la función "rpart"[29] incluida en la biblioteca de funciones del mismo nombre (versión 4.1.19) en el paquete estadístico R[27] (versión 4.3.0).

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*. La implementación del algoritmo bosques aleatorios se realizó utilizando la función "randomForest" [30] (versión 4.7-1.1) incluida en la biblioteca de funciones del mismo nombre en el paquete estadístico R[27] (versión 4.3.0).

Máquina de soporte vectorial. El modelo de máquina de soporte vectorial (SVM), se utiliza para clasificación y regresión[31]. En el contexto de regresión, dado un conjunto de datos $\{y_1, \boldsymbol{x}_1\}, \ldots, \{y_n, \boldsymbol{x}_n\}$, donde y_i representa el valor de la variable respuesta (continua) para el i-ésimo individuo, $\boldsymbol{x}_i$ el valor de las covariables asociadas, el objetivo es obtener una función $f(\boldsymbol{x})$ 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*[32] la función de regresión se aproxima en términos de funciones base $\{h_m(\boldsymbol{x})\}, m = 1, \ldots, M$ como sigue:

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

donde $\beta = (\beta_0, \beta_1, \ldots, \beta_M)^t$ son coeficientes que se obtienen al minimizar: $Q(\boldsymbol{\beta}) = \sum_{i=1}^{n} L\big(y_i - f(\boldsymbol{x}_i)\big) + \frac{\lambda}{2}\sum \beta_m^2$, 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: $\hat{f}(\boldsymbol{x}) = \sum_{i=1}^{n} \hat{a}_i K(\boldsymbol{x}, \boldsymbol{x}_i)$, 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 $(\boldsymbol{x}_i^t \boldsymbol{x}_j)$, polinomial $(\boldsymbol{x}_i^t \boldsymbol{x}_j + coef_0)^d, d = 2,3, \ldots$, radial $(e^{-\gamma||x_i - x_j||^2})$ y sigmoide $(\tanh(\gamma \boldsymbol{x}_i^t \boldsymbol{x}_j), +coef_0)$, 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. El ajuste del modelo se realizó con el paquete "e1071"[33] (versión 1.7-13), con ayuda de la función svm en el paquete estadístico R[27] (versión 4.3.0). Los códigos para ajuste de los modelos están disponibles previa solicitud al autor para correspondencia.

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[20,34]. 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 los procedimientos realizados 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.

En el Cuadro 2 se muestran los promedios de las cien correlaciones de Pearson (en validación cruzada) entre valores corregidos y predichos para las características PN, PD y PA, utilizando los cuatro algoritmos de AA comparados en el estudio. Para PD, el algoritmo SVM fue con el que se obtuvieron los valores más altos para los coeficientes de correlación de Pearson entre valores corregidos y predichos en los conjuntos de validación (PD= 0.256). 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 mejor desempeño con 0.402 y 0.195, respectivamente.

Con la metodología BA se realizaron pruebas con diferente cantidad de "árboles" dentro de los parámetros del modelo, siendo 150 árboles para PN y PD; y 250 para PA los que obtuvieron valores óptimos de predicción; para las características PD y PA obtuvieron el tercer mejor rendimiento en cuestión de sus predicciones. En relación con la metodología AR mostró menor capacidad predictiva para las características PD y PA de esta investigación.

Con los resultados obtenidos y con la finalidad de probar la significancia de los coeficientes de correlación obtenidos se planteó el siguiente juego de hipótesis: $H_0: \mu_r \leq 0$ *vs* $H_1: \mu_r > 0$, donde μ_r es la media de la distribución del coeficiente de correlación de Pearson y se desea probar si la asociación es o no positiva. El juego de hipótesis planteado fue contrastado utilizando la prueba de *t* para 1 muestra, verificando primero el supuesto de normalidad en cada uno de los casos[35], en todos los casos se concluyó que el supuesto de normalidad es adecuado ($P - valor > 0.05$).

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

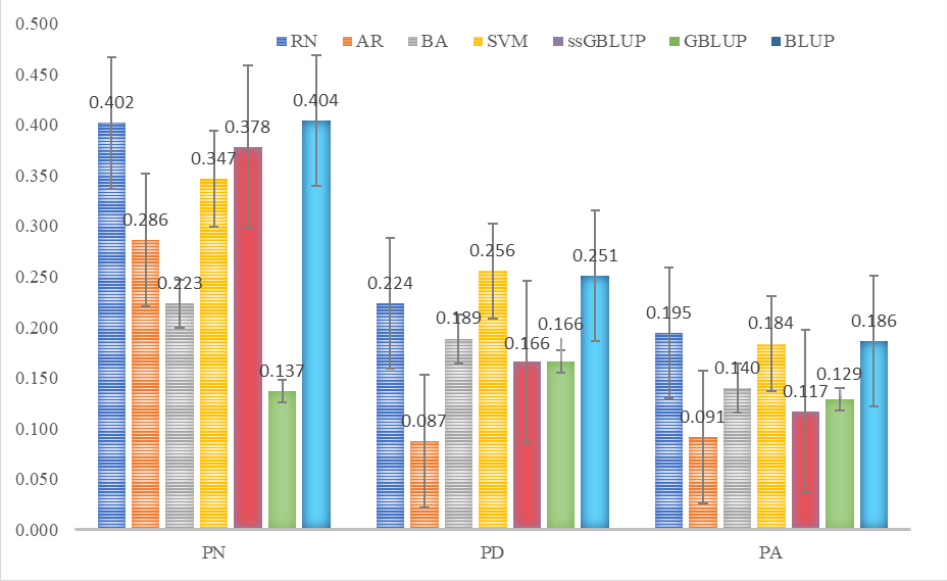
Característica	Algoritmo	Cor	DE
PN	Red neuronal	0.402	0.160
	Árbol de regresión	0.286	0.153
	Bosques aleatorios	0.223	0.163
	Máquina de soporte vectorial	0.347	0.129
PD	Red neuronal	0.224	0.126
	Árbol de regresión	0.087	0.163
	Bosques aleatorios	0.189	0.117
	Máquina de soporte vectorial	0.256	0.144
PA	Red neuronal	0.195	0.152
	Árbol de regresión	0.091	0.178
	Bosques aleatorios	0.140	0.128
	Máquina de soporte vectorial	0.184	0.160

PN= peso al nacimiento, PD= peso al destete, PA= peso al año; Cor = coeficiente de correlación de Pearson; DE= desviación estándar de los 100 estimadores de la correlación para particiones seleccionadas al azar.

Para determinar la capacidad predictiva de los modelos de AA, se compararon los estimadores del coeficiente de correlación Pearson entre los fenotipos corregidos y los valores genéticos predichos con los modelos comparados[36], esto realizado en los conjuntos de prueba para cada característica de la metodología validación cruzada descrita anteriormente, donde no se produjo variación en la información utilizada en los análisis en comparación con otros estudios previos. Esto garantiza la consistencia en las comparaciones realizadas y proporciona una base sólida para evaluar el rendimiento relativo de los métodos tradicionales y los algoritmos de AA. El problema de inferencia de valores genéticos y predicción de fenotipos para caracteres cuantitativos gobernados por formas complejas de interacción entre genes resulta difícil de resolver utilizando los modelos mixtos lineales utilizados de manera rutinaria[37,38], 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[20]. 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 la Figura 1 se muestra que 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[20]. Por su parte, Rodríguez-Alcántar[3] comparó algoritmos de AA utilizando diversos conjuntos de SNP 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.

Figura 1: Comparación de los coeficientes de correlación (promedio de las 100 validaciones) de los fenotipos corregidos y valores genéticos predichos



Valores genéticos 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.

Los resultados indican que el desempeño de los modelos varía según la característica y la cantidad de información[20], 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[39,40], pese a las bajas correlaciones y grandes varianzas de las predicciones, estas pueden atribuirse a una serie de factores genéticos y metodológicos. En consonancia con los hallazgos de Cuyabano *et al*[41], es importante considerar las diferencias genéticas entre las poblaciones de referencia y las poblaciones objetivo al calcular la precisión de las predicciones. Además, sugiere que existe un límite teórico superior para la precisión de estas predicciones, que está determinado por la raíz cuadrada de la heredabilidad. Zhang *et al*[42] mencionan que varios factores pueden influir en la precisión de las predicciones de valores de cría genómicos; la heredabilidad (empleando el modelo descrito como BLUP, Valerio-Hernández *et al*[15] reporta 0.260 para PN; 0.223

para PD y 0.231 para PA), la densidad de marcadores genéticos, la frecuencia del alelo menor (MAF por sus siglas en inglés) utilizado durante el proceso de depuración de datos y el modelo estadístico utilizado son solo algunos factores que pueden afectar la precisión. Esto plantea desafíos significativos en la predicción de rasgos complejos.

Las metodologías de SVM, RN y BA mostraron un desempeño similar en términos de los coeficientes de correlación de Pearson de los fenotipos corregidos y los valores predichos para las tres características de crecimiento utilizadas; comparando los resultados de éstas con valores obtenidos por Valerio-Hernández *et al*[15] utilizando metodologías tradicionales BLUP, GBLUP y ssGBLUP. El costo computacional de RN fue mayor que el de los otros tres algoritmos comparados, se determinó, midiendo el tiempo de ejecución necesario para entrenar y validar cada uno de los algoritmos en los conjuntos de datos de entrenamiento y prueba, registrando el tiempo transcurrido desde el inicio del entrenamiento hasta la finalización del proceso de validación; este resultado es similar al que reportaron Zhao *et al*[43] 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[31]; este algoritmo aporta a las aplicaciones de AA para el análisis de conjuntos de datos provenientes de información genética y genómica[44,45].

Los resultados obtenidos en este estudio demuestran que los algoritmos de AA tienen el potencial de generar predicciones útiles incluso bajo condiciones de información limitada, como el tamaño reducido de muestra y la baja densidad de marcadores genéticos. Este hallazgo resalta su aplicabilidad en escenarios prácticos donde los recursos son restringidos. Sin embargo, se identificaron desafíos importantes, como el costo computacional y la dependencia de una cantidad suficiente de datos de calidad para maximizar la capacidad predictiva. A pesar de estas limitaciones, los algoritmos como RN y SVM mostraron un desempeño consistente, lo que sugiere que pueden ser herramientas valiosas en el análisis genómico. Estos resultados no solo brindan información práctica sobre el uso de los algoritmos de AA, sino que también abren la puerta a investigaciones futuras enfocadas en evaluar su comportamiento con bases de datos más amplias y detalladas, optimizando tanto su implementación como su capacidad predictiva en diferentes contextos.

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.

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 [tesis Doctorado]. Baja California, México: Universidad Autónoma de Baja California; 2019.

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. Jarquín D, Howard R, Graef G, Lorenz A. Response surface analysis of genomic prediction accuracy values using quality control covariates in soybean. Evol Bioinform Online 2019;15:1176934319831307. https://doi.org/10.1177/117693431 9831307.

17. Sargolzaei M, Chesnais JP, Schenkel FS. A new approach for efficient genotype imputation using information from relatives. BMC Genomics 2014;15:478. https://doi.org/10.1186/1471-2164-15-478.

18. 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.

19. Pérez P, de los Campos G. Genome-wide regression and prediction with the BGLR statistical package. Genetics 2014;198:483–95. https://doi.org/10.1534/genetics.114.164442.

20. 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.org/10.1186/1471-2156-12-87.

21. 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.

22. 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.

23. 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.

24. 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.

25. Pérez-Rodríguez P, Gianola D, Weigel KA, Rosa GJM, Crossa J. 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.

26. Pérez-Rodriguez P, Gianola D. Title Bayesian regularization for Feed-Forward Neural Networks. 2022. https://cran.r-project.org/package=brnn.

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. Breiman L, Friedman JH, Olshen RA, Stone CJ. Classification and Regression Trees. Wadsworth. 1984. https://doi.org/10.1201/9781315139470.

29. Therneau T, Atkinson B, Ripley B. Package "rpart". 2022. https://cran.r-project.org/package=rpart.

30. Liaw A, Wiener M. Classification and Regression by random Forest. R News 2002;2(3): 18-22. https://CRAN.R-project.org/doc/Rnews/.

31. Chih-Chung C, Chih-Jen L. LIBSVM: A library for support vector machines. ACM Trans Intell Syst Technol 2011;2(3):1-27. http://www.csie.ntu.edu.tw/~cjlin/libsvm.

32. Hastie T, Tibshirani R, Friedman J. The elements of statistical learning: Data mining, inference, and prediction. Berlin: Springer Science & Business Media. 2008.

33. Meyer D, Dimitriadou E, Hornik K, Weingessel A, Leisch F, Chang C, Lin C. Package "e1071". 2023. https://cran.r-project.org/package=e1071.

34. Berrar D. Cross-Validation. ABC of bioinformatics. Elsevier 2018; 542-545. https://doi.org/10.1016/B978-0-12-809633-8.20349-X.

35. Royston P. An extension of Shapiro and Wilk's W test for normality to large samples. Applied Statistics 1982;(31):115-124. https://doi.org/10.2307/2347973.

36. 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.

37. Gianola D, de los Campos G. Inferring genetic values for quantitative traits non-parametrically. Genet Res (Camb) [Internet]. 2009/01/06. 2008;90(6):525–540.

38. Gianola D, Fernando RL, Stella A. Genomic-Assisted prediction of genetic value with semiparametric procedures. Genetics 2006;173(3):1761–1776. https://doi.org/10.1534/genetics.105.049510.

39. 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. https://doi.org/10.1016/j.csbj.2021.07.021.

40. 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.

41. Cuyabano BCD, Boichard D, Gondro C. Expected values for the accuracy of predicted breeding values accounting for genetic differences between reference and target populations. Genet Sel Evol 2024;56:15. https://doi.org/10.1186/s12711-024-00876-9.

42. Zhang H, Yin L, Wang M, Yuan X, Liu X. Factors affecting the accuracy of genomic selection for agricultural economic traits in maize, cattle, and pig populations. Front Genet 2019;14(10):189. https://doi.org/10.3389/fgene.2019.00189.

43. 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.

44. 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.

45. Libbrecht M, Noble W. Machine learning applications in genetics and genomics. Nat Rev Genet 2015;16:321-332. https://doi.org/10.1038/nrg3920.