Research, Society and Development, v. 12, n. 6, e2812641904, 2023
(CC BY 4.0) | ISSN 2525-3409 | DOI: http://dx.doi.org/10.33448/rsd-v12i6.41904

Utilização de aprendizagem supervisionada de máquina para predição de valores genéticos com base em duas gerações de ascendentes

Use of supervised machine learning for prediction of genetic values based on two generations of ancestors

Uso de aprendizaje supervisado de máquina para la predicción de valores genéticos basada en dos generaciones de ancestros

Recebido: 16/05/2023 | Revisado: 27/05/2023 | Aceitado: 29/05/2023 | Publicado: 03/06/2023

Fernando Jean Dijkinga
ORCID: https://orcid.org/0000-0002-6851-5014
GenMate Genética LTDA, Brasil
E-mail: fernando.dijkinga@gmail.com

Resumo
Uma eficiente utilização de ferramentas que visam acelerar o melhoramento genético como o acasalamento direcionado requer informações de valores genéticos tanto de reprodutores quanto de mães, porém a baixa acurácia ou a carência desta pode comprometer programas de melhoramento genético, dessa forma, estimar valores genéticos de fêmeas é desafiador. Metodologias que utilizem apenas informações de ascendentes machos para o cálculo de valores genéticos de fêmeas já foram propostos, porém, com o surgimento de tecnologias de aprendizado de máquina supervisionado possibilitaram relevantes ao melhoramento genético. Portanto este trabalho tem como objetivo avaliar metodologia baseada em um algoritmo supervisionado de regressão linear em função do método consolidado em literatura, podendo ser empregado em programas de melhoramento genético. Os resultados de análise de correlação de Pearson foram todos significativos ($p<0,0001$) entre valores preditos de ambos os modelos e valores reais, demonstrando que ambos os modelos podem ser utilizados para estimar valores genéticos. Porém ao utilizar o modelo baseado em aprendizagem de máquina foi possível constatar erros com menores desvios-padrão para as características de leite, gordura, proteína e vida produtiva, e iguais para as demais características analisadas, indicando que modelos utilizando tecnologias derivadas de aprendizagem de máquina possuem aplicações promissoras no melhoramento genético.
Palavras-chave: Inteligência artificial; Machine learning; Regressão linear.

Abstract
Efficient use of tools aimed at accelerating genetic improvement such as targeted mating requires genetic value information from both breeders and mothers, however, low accuracy or lack thereof can compromise genetic improvement programs, thus estimating genetic values of females is challenging. Methodologies that use only information from male ancestors to calculate genetic values of females have already been proposed, however, the emergence of supervised machine learning technologies have made relevant contributions to genetic improvement. Therefore, this study aims to evaluate a methodology based on a supervised linear regression algorithm in function of the consolidated method in literature, which can be employed in genetic improvement programs. The results of Pearson's correlation analysis were all significant ($p<0.0001$) between predicted values from both models and actual values, showing that both models can be used to estimate genetic values. However, by using the model based on machine learning it was possible to observe errors with smaller standard deviations for the characteristics of milk, fat, protein and productive life, and equal for the other characteristics analyzed, indicating that models using technologies derived from machine learning have promising applications in genetic improvement.
Keywords: Artificial intelligence; Machine learning; Linear regression.

Resumen
El uso eficiente de herramientas destinadas a acelerar la mejora genética, como el apareamiento dirigido, requiere información sobre los valores genéticos tanto de los criadores como de las madres, sin embargo, la baja precisión o la falta de la misma puede comprometer los programas de mejora genética, por lo tanto, estimar los valores genéticos de las hembras es un desafío. Ya se han propuesto metodologías que utilizan solo información de los antepasados masculinos para calcular los valores genéticos de las hembras, sin embargo, la aparición de tecnologías de aprendizaje automático supervisado ha hecho contribuciones relevantes a la mejora genética. Por lo tanto, este estudio tiene como objetivo evaluar una metodología basada en un algoritmo de regresión lineal supervisado en función del método consolidado en la literatura, que puede emplearse en programas de mejora genética. Los resultados del análisis de

correlación de Pearson fueron todos significativos ($p<0.0001$) entre los valores predichos de ambos modelos y los valores reales, lo que demuestra que ambos modelos pueden utilizarse para estimar valores genéticos. Sin embargo, al utilizar el modelo basado en aprendizaje automático fue posible observar errores con desviaciones estándar menores para las características de leche, grasa, proteína y vida productiva, e iguales para las demás características analizadas, lo que indica que los modelos que utilizan tecnologías derivadas del aprendizaje automático tienen aplicaciones prometedoras en la mejora genética.
Palabras clave: Inteligencia artificial; Aprendizaje automático; Regresión lineal.

1. Introdução

O melhoramento genético na raça Holandesa nas últimas décadas proporcionou significativos avanços na produção leiteira, em um período de 4 décadas, onde animais produziam no ano de 1964 em média 6.309kg de leite em período de 305 dias, saltando para 11.324kg de leite em média em 2004 (Ma et al., 2019). Como o melhoramento genético consiste em gerar animais superiores a cada geração é fundamental que existam decisões entorno da seleção e acasalamento (Bourdon, 2014; Wellmann, 2019). Estratégias de acasalamento utilizando informações de pais e mães superiores de forma racional são denominados como acasalamento direcional, e podem ser divididos em dois métodos. O primeiro, definido como acasalamento direcionado positivo, busca por combinar animais com predições genéticas similares com o objetivo de gerar mais variação genética na próxima geração de progênies. O segundo método estipulado como acasalamento direcionado negativo busca a combinação de animais com predições genéticas não-similares com o intuito de criar uniformidade e corrigir defeitos ou extremos na próxima geração (Groen & Van der Waaij, 1999; Neves et al., 2009; Bourdon, 2014). Sendo ambas as estratégias eficientes para ampliar a proporção de animais geneticamente superiores (Neves et al., 2009). Independentemente da estratégia adotada, para que exista mudanças genéticas significativas em uma população é necessário acurácia nas predições de valores genéticos (Bourdon, 2014).

Quando um animal não possui registro, ou seja, não possui valores genéticos calculados, pode ser utilizado avaliações de seus antecedentes, como os pais, onde cada parente contribui com metade de seus genes para a progênie (Mrode, 2005). Porém, informações de fêmeas são limitadas e de baixa acurácia por conta de possuírem poucas observações e fenótipos, agravado pelo fato de produzirem poucas progênies (Jenko et al., 2013). Por outro lado, touros possuem maior capacidade de produção de progênies devido a inseminação artificial, o que proporciona possuírem estimativas de valores genéticos mais precisas mesmo para características de baixa herdabilidade, em comparação a fêmeas, o que torna relevante estimar o valor genético de fêmeas por meio de ancestrais do sexo masculino (Jenko et al., 2013).

Dessa forma, modelos matemáticos utilizando informações de pai, avô materno e até bisavô materno podem ser convenientes, visto que podem partilhar 50%, 25% e 12,5% de seus genes com progênies, respectivamente (Forabosco et al., 2009; Bourdon, 2014). Segundo Mrode (2005), desvio do valor genético animal pode ser observado em relação a média parental, sendo esse fenômeno conhecido como segregação mendeliana, pois a progênie recebe metade do material genético de seus pais de forma aleatória.

O aprendizado de máquina sendo uma área da inteligência artificial voltada a produção de sistemas de predição e aproximação de funções a partir da experiência acumulada baseadas em aprendizado indutivo, sendo ideal para resoluções de problemas de classificação e regressão (Monard & Baranauskas, 2003; Faceli et al., 2011), como a predição de valores genéticos em sua essência se baseia em modelos matemáticos de regressão (Mrode, 2005; Bourdon, 2014), torna-se relevante a aplicação de algoritmos de aprendizado de maquina no melhoramento genético animal.

2. Metodologia

Banco de dados

Foi empregado um banco de dados contendo registros de 37.618 touros únicos, para cada animal havia registro de pai

(Sire) e avô materno (MGS), além de valores genéticos para características de leite, gordura, proteína, células somáticas, vida produtiva, taxa de sobrevivência, taxa de prenhez e resistência a mastite, assim como os valores genéticos para as mesmas características para cada ascendente.

Os touros foram submetidos a análise descritiva para obtenção de valores de média, desvio-padrão, mínimo e máximo para suporte da explicação do objeto-alvo deste trabalho por meio da função "*.describe()*" da biblioteca Pandas para a linguagem de programação Python.

Modelo Sire-MGS

Para calcular o valor genético para fins comparativos em análise posterior, será empregado a adaptação do modelo proposto por Mrode (2005), conforme segue:

$$BV_p = \frac{1}{2}BV_s + \frac{1}{4}BV_{mgs} + \frac{1}{4}BV_{mean}$$

Onde BV_p representa o valor genético da progênie, BV_s e BV_{mgs} representam o valor genético do pai e avô materno, respectivamente, e BV_{mean} representa o valor médio da característica.

Para a predição do valor genético empregando a metodologia proposta foi empregado o modelo matemático entre colunas no banco de dados, adicionando-se ao final o valor de predição por este modelo em novas colunas.

Algoritmo de aprendizado de máquina supervisionado

O algoritmo de aprendizado de máquina baseado em regressão linear busca o melhor modelo de predição ajustando coeficientes para minimizar a soma de quadrados residual entre os alvos observados e alvos preditos pela aproximação linear, conforme modelo:

$$min_w||X_w - y||_2^2$$

Onde X é vetor contendo as variáveis independentes, y representa a variável dependente e w representa os coeficientes do modelo linear (scikit-learn, 2022).

Comparativo entre métodos

Um novo banco de dados foi formado contendo os valores genéticos reais como também os valores de predição para cada modelo para todas as características. A fim de comparar o grau de associação entre os valores preditos e os reais foi empregado a análise de correlação de Pearson do software *Prism Graph Pad 8* (Silva et al., 2014; Pour Hamidi et al 2017; Ghotbaldini et al., 2019).

Os modelos foram avaliados quanto a sua performance de predição de valores genéticos subtraindo do valor genético os valores genéticos preditos, para obter o erro de predição, em seguida, foram analisados os erros para todas as características de ambos os modelos por meio de análise descritiva.

3. Resultados

Na Tabela 1 está apresentada a análise descritiva das características contidas no banco de dados.

Tabela 1 – Análise descritiva do banco de touros da raça holandesa.

Característica	n	Média	Desvio Padrão	Mínimo	Máximo
Leite	37.618	19,59	831,00	-3238,00	3397,00
Gordura	37.618	6,83	45,66	-116,00	152,00
Proteína	37.618	3,66	28,82	-99,00	101,00
Células somáticas	37.618	3,00	0,20	2,38	4,01
Vida produtiva	37.618	-0,17	3,26	-9,80	8,90
Taxa de sobrevivência	37.618	-0,71	2,35	-10,00	7,10
Taxa de prenhez	37.618	-0,33	1,75	-8,00	7,20

n = número de indivíduos que apresentam a características. Fonte: Elaborado pelos autores.

Para todas as características e para ambos os modelos de regressão a correlação de Pearson foi significativa, para os coeficientes de determinação dos modelos pode-se notar que para todas as características os coeficientes foram maiores para o modelo de aprendizado de máquina em relação ao modelo tradicional com exceção dos coeficientes para a característica de taxa de prenhez que foram iguais, o mesmo acontece com o coeficiente de correlação de Pearson, conforme Tabela 2.

Para todas as características e para ambos os modelos de regressão a correlação de Pearson foi significativa, para os coeficientes de determinação dos modelos pode-se notar que para todas as características os coeficientes foram maiores para o modelo de aprendizado de máquina em relação ao modelo tradicional com exceção dos coeficientes para a característica de taxa de prenhez que foram iguais, o mesmo acontece com o coeficiente de correlação de Pearson, conforme Tabela 2.

Tabela 2 - Análise de correlação de pearson entre os modelos de predição propostos.

Característica	Modelo	r	R^2	Valor de P
Leite	Trad	0,7978	0,6365	<0,0001
	ML	0,8007	0,6411	<0,0001
Gord	Trad	0,9130	0,8336	<0,0001
	ML	0,9137	0,8348	<0,0001
Prot	Trad	0,9069	0,8224	<0,0001
	ML	0,9086	0,8255	<0,0001
CCS	Trad	0,7401	0,5478	<0,0001
	ML	0,7412	0,5493	<0,0001
VP	Trad	0,8963	0,8033	<0,0001
	ML	0,8971	0,8048	<0,0001
LIV	Trad	0,7956	0,6329	<0,0001
	ML	0,7965	0,6344	<0,0001
DPR	Trad	0,6473	0,4189	<0,0001
	ML	0,6473	0,4189	<0,0001
MAST	Trad	0,7878	0,6206	<0,0001
	ML	0,7882	0,6212	<0,0001

Gord = Gordura; Prot = Proteína; CCS = Células somáticas; VP = Vida produtiva; LIV = Taxa de sobrevivência; DPR = Taxa de prenhez; MAST = Resistência a mastite; Trad = Modelo de regressão tradicional; ML = Modelo baseado em aprendizado de máquina; r = Coeficiente de correlação de Pearson; R^2 = Coeficiente de determinação do modelo; Valor de P = Valor de p do modelo. Fonte: Elaborado pelos autores.

Quanto a análise descritiva dos erros obtidos a partir dos dois modelos de predição (Tabela 3) é possível notar que os valores de média para o modelo de aprendizagem de máquina ficam próximos a zero. Também é possível notar que para as características de leite, gordura, proteína e vida produtiva o valor de desvio padrão é menor para o modelo de aprendizado de máquina em relação ao modelo tradicional, para as características de células somáticas, taxa de sobrevivência, taxa de prenhez e resistência a mastite o valor de desvio padrão é o mesmo para ambos os modelos de predição.

Tabela 3 – Análise descritiva dos erros de predição entre os modelos propostos.

Característica	Modelo	Média	Desvio padrão	Mínimo	Máximo
Leite	Trad	-60,0	517,00	-2256,0	2636,0
	ML	0,0	489,00	-2327,0	2630,0
Gord	Trad	1,0	22,00	-90,0	106,0
	ML	0,0	19,00	-82,0	92,0
Prot	Trad	-0,7	14,00	-65,0	67,0
	ML	0,0	12,00	-56,0	57,0
CCS	Trad	-0,01	0,13	-0,5	0,7
	ML	0,0	0,13	-0,5	0,73
VP	Trad	-0,00	1,60	-6,8	6,2
	ML	0,0	1,40	-7,5	6,6
LIV	Trad	0,08	1,40	-7,9	7,7
	ML	0,0	1,40	-8,2	7,8
DPR	Trad	0,06	1,30	-7,2	6,1
	ML	0,0	1,30	-7,2	6,1
MAST	Trad	0,2	1,10	-8,0	5,2
	ML	0,0	1,10	-8,0	5,1

Gord = Gordura; Prot = Proteína; CCS = Células somáticas; VP = Vida produtiva; LIV = Taxa de sobrevivência; DPR = Taxa de prenhez; MAST = Resistência a mastite; Trad = Modelo de regressão tradicional; ML = Modelo baseado em aprendizado de máquina. Fonte: Elaborado pelos autores.

4. Discussão

Na avaliação dos valores genéticos das características de interesse contidas no banco de dados nota-se que o valor médio se aproxima de zero, pois este valor se refere ao valor médio da população de referência na qual os valores genéticos dos touros se baseiam, sendo valores genéticos negativos representam performance abaixo da média da população e valores genéticos positivos representam performance acima da média populacional (Bourdon, 2014). Percebe-se que para o método tradicional as médias são diferentes de zero, e para o método baseado em aprendizagem de máquina (ML) a média é igual a zero, além de que, juntamente com o fato de apresentarem menores desvios-padrão, pode indicar que este último método consegue prever e tomar em consideração a força média de seleção sobre a característica. Ou seja, se o método ML considera como média zero, dentro das constantes computadas pelo modelo o valor é possível que seja computado o efeito médio da força de seleção, uma vez que variâncias genéticas podem ser alteradas devido a seleção (Schenkel & Schaeffer, 2000). E o fato do valor para os resultados com o método baseado em aprendizagem estarem centrados em zero, significa quando o erro médio tende a zero, estamos nos referindo ao fato de que, à medida que a quantidade de dados aumenta, a capacidade do modelo de aprender a função subjacente melhora, o que pode reduzir o erro médio, tendo este ponderamento na constante atribuída a função (Molas, 2022).

Outro motivo que pode levar a menores desvios-padrão pode ser o peso atribuído a cada um dos parâmetros. Quando

onde na literatura é encontrado que uma progênie apresenta 50% de similaridade genética com seu pai, e uma fatoração desse valor é predito para cada geração para com os outros ascendentes (Wright, 1922), no método ML a similaridade genética entre progênie a ascendentes é ajustado por meio do processo de minimização dos quadrados por meio dos coeficientes. A desproporção pode ser também causada por outros fatores como a consideração de efeitos como de desiquilíbrio gamético, mas, afirmações sai difíceis de se confirmarem devido a natureza dos dados (Hedrick, 1987; Lewontin, 1988; White et al., 2008).

A aplicação de tecnologias baseadas aprendizagem de máquina no melhoramento genético animais ainda é recente, havendo mais casos em que modelos mais simples como regressões lineares é superado por métodos tradicionais como o BLUP (Melhor predição linear não-viesada), sendo necessário então realizar estudos comparativos entre métodos e modelos (Nayeri et al., 2019).

Porém, para todas as características analisadas neste estudo por ambos os modelos de predição resultados significativos foram encontrados. Em relação aos erros obtidos nas predições, para o modelo ML utilizado neste estudo, foi menor ou apresentou igual performance em relação ao método tradicional ao apresentar desvio padrão relacionados ao erro de predição menores.

Na particular aplicação de algoritmos de aprendizagem supervisionados para a realização da predição de valores genéticos baseados em valores genéticos de ascendentes carecem de fontes e publicações, demonstrando que este domínio de aplicação ainda necessita desenvolvimento e mais estudos.

5. Conclusão

Tecnologias de aprendizagem de máquina apresentam grande relevância para o campo de melhoramento genético animal, visto que os modelos atuais se baseiam em modelos geralmente de cunho linear, logo, mais pesquisas devem ser realizadas com o objetivo de avaliar a performance e aplicação desta. Com resultados como este, mostram que novas abordagens podem auxiliar no progresso genético da raça Holandesa.

Referências

Bourdon, R. (2014). *Understanding Animal Breeding*. (2a ed.), Pearson. p. 560.

Faceli, K., Lorena, A. C., Gama, J. & Carvalho, A. C. P. L. F. (2011). *Artificial intelligence: a machine learning approach*. Ed. LTC.

Forabosco, F., Jakobsen, J., & Fikse, W. (2009). International genetic evaluation for direct longevity in dairy bulls. *Journal of Dairy Science*, 92(5), 2338–2347. https://doi.org/10.3168/jds.2008-1214.

Ghotbaldini, H., Mohammadabadi, M., Nezamabadi-pour, H., Babenko, O. I., Bushtruk, M. V., & Tkachenko, S. V. (2019). Predicting breeding value of body weight at 6-month age using Artificial Neural Networks in Kermani sheep breed. Acta Scientiarum. *Animal Sciences*, 41(1), 45282. https://doi.org/10.4025/actascianimsci.v41i1.45282.

Groen, A. F., & der Waaij, L. V. (1999). Some basics about mating schemes | *Interbull Bulletin.* https://journal.interbull.org/index.php/ib/article/view/562.

Hedrick, P. W. (1987). Gametic Disequilibrium Measures: Proceed With Caution. *Genetics*, 117(2), 331–341. https://doi.org/10.1093/genetics/117.2.331.

Jenko, J., Gorjanc, G., Kovač, M., & Ducrocq, V. (2013). Comparison between sire-maternal grandsire and animal models for genetic evaluation of longevity in a dairy cattle population with small herds. *Journal of Dairy Science*, 96(12), 8002–8013. https://doi.org/10.3168/jds.2013-6830.

Lewontin, R. C. (1988). On measures of gametic disequilibrium. *Genetics*, 120(3), 849–852. https://doi.org/10.1093/genetics/120.3.849.

Ma, L., Sonstegard, T. S., Cole, J. B., VanTassell, C. P., Wiggans, G. R., Crooker, B. A., Tan, C., Prakapenka, D., Liu, G., Da, Y. (2019). Genome changes due to artificial selection in U.S. Holstein cattle. *BMC Genomics*, 20(1). https://doi.org/10.1186/s12864-019-5459-x.

Molas, A. (2022). Why do we minimize the mean squared error? Acesso em dez 2022. *Towards Data Science*. https://towardsdatascience.com/why-do-we-minimize-the-mean-squared-error-3b97391f54c.

Monard, & Baranauskas. (2003). *Conceitos sobre aprendizado de máquina*. Sistemas Inteligentes Fundamentos e Aplicações. Manole Ltda.

Mrode. (2005). *Linear models for the prediction of animal breeding values* (3rd ed.). CABU.

Nayeri, S., Sargolzaei, M., & Tulpan, D. (2019). A review of traditional and machine learning methods applied to animal breeding. *Animal Health Research Reviews*, 20(1), 31–46. https://doi.org/10.1017/s1466252319000148.

Neves, H. H., Carvalheiro, R., Cardoso, V., Fries, L. A., & de Queiroz, S. A. (2009). Acasalamento dirigido para aumentar a produção de animais geneticamente superiores e reduzir a variabilidade da progênie em bovinos. *Revista Brasileira de zootecnia*, 38(7). https://doi.org/10.1590/S1516-35982009000700006.

Pour Hamidi, S., Mohammadabadi, M. R., Asadi Foozi, M., & Nezamabadi-pour, H. (2017). Prediction of breeding values for the milk production trait in Iranian Holstein cows applying artificial neural networks. *Journal of Livestock Science and Technologies*, 5(2), 53-61. 10.22103/jlst.2017.10043.1188.

Schenkel, F., & Schaeffer, L. (2008). Effects of nonrandom parental selection on estimation of variance components. Journal of Animal Breeding and *Genetics*, 117(4), 225–239. https://doi.org/10.1111/j.1439-0388.2000.00262.x

Silva, G. N., Tomaz, R. S., Sant'Anna, I. D. C., Nascimento, M., Bhering, L. L., & Cruz, C. D. (2014). Neural networks for predicting breeding values and genetic gains. *Scientia Agricola*, 71(6), 494–498. https://doi.org/10.1590/0103-9016-2014-0057.

sklearn.linear_model.LinearRegression. (2022). Acesso em dez 2022. https://scikit-learn/stable/modules/generated/sklearn.linear_model.LinearRegression.html.

Wellmann, R. (2019). Optimum contribution selection for animal breeding and conservation: the R package optiSel. *BMC Bioinformatics*, 20(1). https://doi.org/10.1186/s12859-018-2450-5.

White, D. J., Wolff, J. N., Pierson, M., & Gemmel, N. J. (2008). Revealing the hidden complexities of mtDNA inheritance. *Molecular Ecology*, 17(23), 4925–4942. https://doi.org/10.1111/j.1365-294x.2008.03982.x.

Wright, S. (1922). Coefficients of Inbreeding and Relationship. *The American Naturalist*, 56(645), 330–338. https://doi.org/10.1086/279872.