Classificação do Desempenho dos Rebanhos de Seleção Nelore por meio de Aprendizado de Máquina

Urbano G. P. Abreu[1], Patrícia Tholon[2], Helano P. de Lima[3]

[1]Embrapa Pantanal (CPAP)l
Caixa Postal 109 – 79320-900 – Corumbá – MS – Brasil

[2]Embrapa Pecuária Sudeste (CPPSE)
São Carlos, SP – Brasil.

[3]Embrapa Agricultura Digital (CNPTIA)
Campinas – SP – Brasil.

{urbano.abreu,patrícia.tholon,helano.lima}@embrapa.br

Abstract. *The aims of this work was, through data mining techniques, to classify the animals of two Embrapa herds, according to the genomic DEP rules (DEPg), to identify the main attributes (characteristics) that direct the understanding of the different objectives of selection in both herds. Eight attributes were selected, with greater importance for the analysis of animal classification, in relation to the selection herds. To perform the classification of animals according to the herd, three supervised algorithms were used, seeking to verify which would present the best performance: decision tree (J48), logistic model trees (LMT) and random forest (RF). The most accurate algorithm was the RF, which modeled the data with greater fit and accuracy.*

Resumo. *O objetivo deste trabalho foi, por meio de técnicas de mineração de dados, classificar os animais de dois rebanhos da Embrapa, em função das réguas de DEP genômicas (DEPg), para identificar os principais atributos (características) que direcionam o entendimento dos diferentes objetivos de seleção nos dois rebanhos. Selecionaram-se oito atributos, para análise de classificação dos animais. Para realizar a classificação dos animais em função do rebanho foram utilizados três algoritmos supervisionados, buscando verificar qual apresentaria o melhor desempenho: árvore de decisão (J48), árvores de modelo logístico (LMT) e floresta randômica (Random Forest - RF). O algoritmo mais acurado foi o Random Forest, que modelou os dados com maior ajuste e acurácia.*

1. Introdução

Buscando auxiliar o objetivo de difundir o melhoramento genético da raça Nelore para os rebanhos de produção comercial, diferentes unidades da Embrapa possuem rebanhos sob seleção, em diferentes biomas e sistemas de produção. Os núcleos possuem como objetivo fornecer subsídio para definição de objetivos e critérios de seleção considerando os principais sistemas de produção de carne no Brasil. Além de, em parceria com produtores, realizar testes com linhagens e tourinhos diferenciados para diferentes biomas do país.

O melhoramento genético, de modo geral, visa obter níveis de produção mais eficientes, produtividade e/ou qualidade do produto direcionado ao sistema de produção e as exigências do mercado. Para o alcance deste objetivo, várias características precisam ser acompanhadas: adaptabilidade, eficiência reprodutiva, viabilidade, pesos corporais, taxas de crescimento, qualidade da carcaça e da carne são alguns exemplos (Rosa et al., 2013). O melhoramento animal é um processo contínuo de criação, seleção e reprodução dos animais domésticos, tendo como objetivo básico modificar em direção ao desejado pelo homem as características dos animais. O aumento dos índices de produção animal pode ser obtido, então, por melhor condição do ambiente, por meio de mudanças nos manejos nutricionais, sanitários e reprodutivos e pelo melhoramento genético (seleção, sistemas de acasalamento e cruzamento).

A implantação de um programa de melhoramento genético nos rebanhos de seleção da Embrapa foi fundamental para direcionar os trabalhos de pesquisa em gado de corte, levando ao desenvolvimento de melhores práticas de gestão dos rebanhos. Seu estabelecido foi de fundamental importância para alcançar visão mais abrangente sobre os vários rebanhos que compõem uma raça. Para que tais programas sejam funcionais e eficientes, é necessário que sejam devidamente estruturados do ponto de vista de avaliação genética (Nobre, et al., 2013). Os resultados das avaliações genéticas apresentadas pelo Programa Geneplus são apresentadas na forma de Diferença Esperada na Progênie (DEP) associada à sua respectiva precisão (acurácia). A DEP deve, portanto, ser utilizada como critério de seleção ou de descarte dos indivíduos. A DEP, que é uma medida relativa (depende da base genética da avaliação), deve ser o elemento de decisão de utilização de um ou outro indivíduo (Torres, Jr, et al., 2013).

O objetivo deste trabalho foi, por meio de técnicas de mineração de dados, classificar os animais de dois rebanhos da Embrapa, em função das réguas de DEP genômicas (DEPg), para identificar os principais atributos (características) que direcionam o entendimento dos diferentes objetivos de seleção nos dois rebanhos.

2. Material e Métodos

Foram analisados 1.596 dados (instâncias) de animais dos rebanhos da Embrapa, sendo da Embrapa Pecuária Sudeste - Fazenda Canchim (1.065 animais) e da Embrapa Pantanal - Fazenda Nhumirim (531 animais). Utilizando a metodologia de análise genômica desenvolvida pelo Programa Geneplus (Nobre et al., 2013). Foram estimadas DEPg de 16 características (atributos): peso ao nascer (PN), peso à desmama (PD), total maternal (TMD), peso ao sobreano (PS), ganho pós-desmama (GPD), conformação frigorífica à desmama (CFD), conformação frigorífica ao sobreano (CFS), habilidade de permanência (HP), perímetro escrotal ao sobreano (PES), idade ao primeiro parto (IPP), relação de desmama (RD), área de olho de lombo (AOL), espessura de gordura

subcutânea (EGS), marmoreio (MAR), consumo alimentar residual (CAR) e índice de qualificação genética (IQG).

A área da computação denominada aprendizado de máquina (ML, do inglês, *machine learning*) fornece ferramentas pelas quais grandes volumes de dados podem ser analisados automaticamente, sendo fundamental para tal, a identificação e seleção correta dos principais atributos para o treinamento dos modelos. Neste trabalho foi utilizada a metodologia descrita por Hall (2000), que seleciona os atributos para classificação supervisionada por meio das correlações entre os atributos. O processo de seleção de atributos pode ser benéfico para uma variedade de algoritmos com viés semelhante na área de aprendizado de máquina. Desta maneira são retiradas das análises atributos irrelevantes e redundantes resultando, em muitos casos, na melhora do desempenho dos algoritmos.

A utilização dos algoritmos de ML permite diferentes abordagens de aprendizado, os tipos supervisionado e não supervisionado são os mais comuns. No aprendizado supervisionado, o conjunto de dados usado para treinar o algoritmo é composto de exemplos (instâncias), onde cada um tem a forma (X,y), onde 'X' representa o vetor de atributos e 'y' representa o resultado esperado de 'X'. A partir do conjunto de instâncias, o aprendizado supervisionado induz modelo capaz de avaliar outros bancos de dados não analisados. Quando este resultado esperado ('y') é uma variável discreta, ou seja, uma 'classe', chamamos esses modelos de classificadores.

Para realizar a classificação dos animais em função do rebanho foram utilizados três algoritmos supervisionados, buscando verificar qual apresentaria o melhor desempenho: árvore de decisão (J48), árvores de modelo logístico (LMT) e floresta randômica (*Random Forest* - RF).

O J48 possui o viés de dividir para conquistar, com a vantagem de possibilitar a descoberta de conhecimento, e a tomada de decisão a partir da construção de uma árvore com base em uma massa de dados. Algumas das vantagens da aplicação deste algoritmo na tomada de decisão é que ele é capaz de abordar problemas envolvendo variáveis qualitativas contínuas, e discretas presentes nas bases de dados é tolerante a valores ausentes e tem uma seleção de atributos embutida, apresentando em suas ramificações os atributos de maior relevância no topo da árvore. Este método usa a abordagem de dividir um problema complexo em partes menores onde proporciona a aplicação de uma estratégia recursiva para cada problema, com isto o mesmo divide o espaço definido pelos atributos em subespaços, associando se a eles uma classe, usando como medida de separação a taxa de ganho de entropia(Witten, et al., 2011). O método LMT emprega o algoritmo LogitBoost (Friedman et al., 2000) que constrói funções de regressão logística nos nós de uma árvore. O número ótimo de iterações é determinado por validação cruzada. Esse método produz modelos finais que contêm consideravelmente menos parâmetros do que os gerados pelo máximo padrão regressão logística de probabilidade, sem diminuir significativamente a precisão e, às vezes, aumentando-o significativamente. *Random Forest* (RF) é um algoritmo classificador que faz uso do método de árvores de decisão criada por Breiman (2001). Esta técnica possui uma ideia diferente dos algoritmos de árvores de decisão, a qual pertence. Enquanto uma árvore possui o objetivo de construção total de uma estrutura a partir de uma base de dados o RF tem o objetivo de efetuar a criação de várias árvores de decisão (floresta) usando um subconjunto de atributos selecionados aleatoriamente a partir do conjunto original, contendo todos os atributos. Este tipo de amostragem (com reposição) é chamado de *bootstrap*. Para realizar a inferência nessa floresta, é atribuído a cada

subconjunto um voto sobre qual classe o atributo chave deve pertencer, este voto possui um "peso" onde o mesmo é afetado pela igualdade entre as árvores, sendo que quanto menor a similaridade entre duas árvores melhor, e pela força que cada árvore tem individualmente, ou seja, quanto mais precisa uma árvore for, melhor será sua nota (Witten et al., 2011).

Com a quebra das massas de dados e construção de vários submodelos (árvores), o RF é capaz de apresentar alta robustez (capacidade de generalização) aliado a alta performance também. Para analisar a qualidade da modelagem e ajuste foram utilizadas os seguintes estimativas:

1) O Coeficiente Kappa, que pode ser definido como medida de associação usada para descrever e testar o grau de concordância (confiabilidade e precisão) na classificação. Apesar de largamente utilizado para o estudo de confiabilidade, este método estatístico apresenta limitações na medida em que não fornece informações a respeito da estrutura de concordância e discordância, muitas vezes, não considerando aspectos importantes presentes nos dados;
2) Cálculo da área abaixo da curva ROC (*Receiving Operating Characteristic*), conhecida como *AUC* (*Area Under the Curve*). Como a *AUC* é uma porção da probabilidade, seu valor será sempre entre 0 e 1. A área sob a curva de um classificador tem uma importante propriedade estatística: ela é equivalente à probabilidade de o modelo ranquear uma observação positiva escolhida ao acaso, com um valor superior a uma observação negativa escolhida ao acaso. Segundo Bissacot et al. (2016), a *ROC* foi desenvolvida para a teoria da detecção de sinais/visualização e análise do comportamento de sistemas de diagnósticos; e

   3) A raiz do erro médio quadrático (RMSC) é um dos principais indicadores de desempenho para um modelo de previsão de regressão. Ela mede a diferença média entre os valores previstos por um modelo de previsão e os valores reais.

3. Resultados

O uso do método de Hall (2000), selecionou oito atributos (DepCFD, DepCFS, DepIPP, DepRD, DepAOL, DepEGS, DepMar e DepCar), com maior importância para análise de classificação dos animais, em relação aos rebanhos de seleção. Na Tabela 1 se observa as médias e desvios padrões das DEPg dos atributos selecionados, em função dos rebanhos.

Os atributos selecionados são ligados aos aspectos de Carcaça (AOL, EGS e Mar), Conformação (CFD e CFS), Reprodução (IPP e RD) e Eficiência Alimentar (CAR). Ou seja, compõem aspectos importantes para o aumento da eficiência da cadeia produtiva da pecuária de corte, do nascimento dos animais até o abate. A combinação destas características compondo índice para os diferentes rebanhos poderá ter aspecto importante na condução futura do projeto, inclusive com a inserção dos pesos econômicos das características para estruturar um índice de seleção próprio para os rebanhos, adequados ao sistema e ambiente no qual os rebanhos são manejados (Portes, et al., 2021). Para realizar o aprendizado com os métodos J48, RF e LMT foram utilizado 80% da base de dados para 'ensinar' o algoritmo (1357 animais). Os 20% restantes (319 animais), foram utilizados na validação dos modelos construídos. Os resultados da validação dos três modelos são apresentados na Tabela 2 e Gráfico 1.

Foram utilizadas para verificar a qualidade da modelagem as estimativas do: Coeficiente Kappa, Área Abaixo da Curva (*AUC*) e raiz do erro médio quadrático (*RMSC*). Pode-se observar que os três algoritmos apresentaram desempenho satisfatório, sendo que os métodos J48 e LMT tiveram resultados de qualidade de ajuste semelhantes.

Tabela1 – Estatísticas descritivas das DEPg dos animais dos rebanhos da Fazendas Canchim e Nhumirim.

Campo Experimental		DepCFD	DepCFS	DepIPP	DepRD	DepAOL	DepEGS	DepMAR	DepCAR
Canchim 1065 animais	Média	1,49	1,96	-5,64	0,29	0,92	0,31	0,37	0,00
	Erro padrão	0,04	0,06	0,30	0,03	0,03	0,02	0,03	0,00
	Mediana	1,57	1,96	-7,07	0,27	0,91	0,30	0,42	0,00
	Desvio padrão	1,46	1,88	9,78	0,88	1,07	0,77	0,84	0,07
	Mínimo	-2,74	-3,59	-36,95	-3,59	-3,57	-2,66	-3,53	-0,36
	Máximo	8,10	9,17	65,77	3,04	7,97	3,98	3,81	0,22
Nhumirim 531 animais	Média	1,60	2,56	-10,60	0,21	0,84	0,47	0,01	-0,01
	Erro padrão	0,04	0,05	0,26	0,03	0,03	0,02	0,03	0,00
	Mediana	1,49	2,60	-10,57	0,14	0,82	0,40	-0,01	-0,01
	Desvio padrão	0,93	1,09	6,01	0,66	0,64	0,58	0,64	0,03
	Mínimo	-0,50	-0,41	-25,93	-1,51	-1,41	-1,09	-1,86	-0,11
	Máximo	4,82	6,39	9,17	2,48	3,21	2,73	2,32	0,07

Observa-se na matriz de confusão, onde verificamos os falsos positivos e falsos negativos, que como o esperado, o número e percentual de classificações erradas no algoritmo RF foi também menor, caracterizando maior acurácia.

O algoritmo *Random Forest* obteve maior eficiência em relação aos outros algoritmos utilizados, os valores observados foram Acurácia de 92,16% e Coeficiente Kappa de 0,812, além da AUC e RMSC estimados de 0,96 e 0,26.

4. Discussão

Segundo Dresch et al., (2022) a recente revolução tecnológica nos trouxe uma imensa quantidade de dados e a necessidade de processar, armazenar, analisar e compreendê-los em grandes volumes e diferentes aplicações. Ganham força as áreas de estudo como a Inteligência Artificial (IA), o Aprendizado de Máquina (ML) e a Ciência de Dados (Data Science). A seleção artificial é eficaz em alterar o desempenho dos sistemas de produção animal. A predição genômica com utilização de ML para aumentar a eficiência dos programas de melhoramento genético de gado de corte e os ganhos genéticos, por meio do treinamento de algoritmos e análises genômicas devem ser realizadas e ferramentas de predição genômica devem ser disponibilizadas para

criadores e outras partes interessadas (Garrick, 2011). O uso de algoritmos treinados de ML, como RF, possibilitará nas grandes bases de dados de seleção genômica, a exploração de dados, buscando padrões, correlações e desvios com base em técnicas descritivas, fase que proporcionará os *insights* necessários que antecedem a modelagem, construção de modelo, apresentação dos resultados e automatização da análise.

Tabela 2 – Matriz de confusão (diagonal principal representa a quantidade de instâncias classificadas corretamente), estimativa de acurácia e índice Kappa. dos animais das Fazendas Canchim e Nhumirim, em relação a classificação dos algoritmos.

J48		C	N	Total
Predito	C	**197**	23	220
	N	25	**74**	99
	Total	222	97	**319**
	Acurácia	88,73%	76,29%	85,00%
	Kappa		0,65	
LMT		C	N	Total
Predito	C	**194**	26	220
	N	23	**76**	99
	Total	217	102	**319**
	Acurácia	89,40%	74,50%	84,63%
	Kappa		0,641	
Random Forest		C	N	
Predito	C	**212**	8	220
	N	17	**82**	99
	Total	229	90	**319**
	Acurácia	92,57%	91,11%	92,16%
	Kappa		0,812	

C – Canchim, N –Nhumirim.

O estudo comparativo entre os algoritmos de mineração de dados pertencentes aos métodos de classificação chamados, *Random Forest*, J48 e LMT buscou contribuir na descoberta de qual das técnicas tem melhor desempenho, exatidão, agilidade entre outras características em sua implementação, juntamente com o desenvolvimento de uma aplicação onde fosse possível apresentar as informações da mineração para o usuário de uma forma de fácil identificação e com uma interface amigável.

Os resultados das avaliações, dentro do rebanho ou entre rebanhos, constituirão instrumentos de decisão. De posse destas informações, cabe ao criador, identificar quais indivíduos e de quais rebanhos serão utilizados no processo reprodutivo. A avaliação genética é uma ferramenta de trabalho na condução de um programa de melhoramento genético, que pode ser fortalecida com o uso de técnicas de ML, em função da flexibilidade e capacidade de extrair padrões em conjunto de muitos dados. Além disso, o sistema de análise e modelagem deve ser sensível à avaliação continua, com os ajustes

ao longo do tempo entre os objetivos definidos e os resultados obtidos (Chafai, et al., 2023).

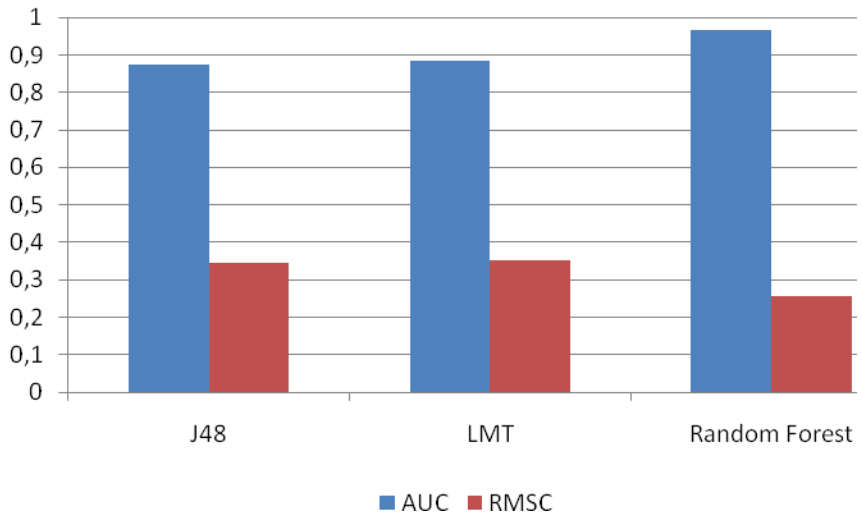

Figura 1 – Estimativas dos indicadores de qualidade de ajuste Área Abaixo da Curva (*AUC*) e raiz do erro médio quadrático (*RMSC*).

5. Conclusão

A indução por diferentes algoritmos de árvores de decisão permitiu a classificação de animais pertencentes a dois rebanhos de seleção Nelore. A descoberta de padrões e conhecimento proporcionados pelas árvores de decisão continua sendo uma ferramenta valiosa na área de ML. O algoritmo mais acurado foi o *Random Forest*, que modelou os dados com maior ajuste a modelagem e classificou as instâncias com acurácia, mas, por ser um algoritmo caracterizado por não se ter conhecimento por acesso direto, sendo identificado meio da verificação da relação existente entre os dados (tipo caixa-preta), foi utilizado como referência para avaliar o desempenho adequado dos outros modelos.

6. Referências

Bissacot, A. C. G., Salgado, S. A. B., Balestrassi, P. P., Paiva, A. P., Zambroni Souza, A. C., Wazen, R. (2016). Comparison of neural networks and logistic regression in assessing the occurrence of failures in steel structures of transmission lines. *The Open Electrical & Electronic Engineering Journal*, 10, p. 11-26.

Breiman, L. (2001) Random forests. *Machine learning*, v. 45, p. 5-32.

Chafai N., Hayah, I., Houaga, I., Badaoui, B. A. (2023) A review of machine learning models applied to genomic prediction in animal breeding. *Frontiers in Genetics*. Sep 6;14:1150596. doi: 10.3389/fgene.2023.1150596.

Dresch, L. de O., Figueiredo, A. M. R., Fagundes, M. B. B. (2022) A digitalização do campo e a democratização da ciência de dados: perspectivas para aplicação por

produtores agropecuários. *COLÓQUIO - Revista do Desenvolvimento Regional***,** 19, p. 310-328.

Friedman, J., Hastie, T., Tibshirani, R. (2000) Special invited paper. additive logistic regression: A statistical view of boosting. *Annals of Statistics*, p. 337-374.

Hall, M. (2000). Correlation-based feature selection for discrete and numeric class machine learning. In P. Langley (Ed.), *Proceedings of the Seventeenth International Conference on Machine Learning* (pp. 359–366). Stanford, CA. San Francisco: Morgan Kaufmann..

Garrick, D. (2011) The nature, scope and impact of genomic prediction in beef cattle in the United States.*Genetic Selection Evolution*. p. 43:17. doi: 10.1186/1297-9686-43-17

Landwehr, N., Hall, M., Frank, E. (2005). Logistic Model Trees. *Machine Learning*. 95, p. 161-205.

Nobre, P. R. C., Silva, L. O. C. da, Rosa, A. do N, Menezes, G. R. de O. (2013) Programa Embrapa de melhoramento de gado de corte - GENEPLUS.). In: Rosa, A. do N., Martins. E. N., Menezes, G. R. de O., Silva, L. O. C. da (Ed.). *Melhoramento genético aplicado em gado de corte: Programa Geneplus-Embrapa***.** Brasília, DF: Embrapa; Campo Grande, MS: Embrapa Gado de Corte. Capítulo 19. p. 224-35

Portes, J. V., Menezes, G. R. O., Silva, L. O. C., MacNeil, M. D., Abreu, U. G. P., Lacerda, V. V., Braccini Neto, J. (2021). Selection indexes for Nellore production system in the Brazilian Pantanal. *Revista Brasileira de Zootecnia* 2021. 50:e20200264.

Rosa, A do N F, Menezes, G. R. de O., Egito, A. A. do.(2013). Recursos genéticos e estratégias de melhoramento.(2013). In: Rosa, A. do N., Martins. E. N., Menezes, G. R. de O., Silva, L. O. C. da (Ed.). *Melhoramento genético aplicado em gado de corte: Programa Geneplus-Embrapa***.** Brasília, DF: Embrapa; Campo Grande, MS: Embrapa Gado de Corte. Capítulo 2, p. 11-26.

Torres, Jr. R. A. de A., Silva, L. O. C. da, Menezes, G. R. de O., Nobre, P. R. C., (2013) Melhoramento animal na era das DEPS.(2013) In: Rosa, A. do N.; Martins. E. N.; Menezes, G. R. de O.; Silva, L. O. C. da (Ed.). *Melhoramento genético aplicado em gado de corte: Programa Geneplus-Embrapa***.** Brasília, DF: Embrapa; Campo Grande, MS: Embrapa Gado de Corte. Capítulo 13. p. 149-166

Witten, I. H., Frank, E., Hall, M. A. (2011) *Data mining: practical machine learning tools and techniques.* 3ed. San Francisco: Morgan Kaufmann.