Predição de Indicadores Zootécnicos de Carcaças Bovinas a Partir de Variáveis de Cria

Thales V. Maciel[1], Vinícius do N. Lampert[2], Denizar S. Souza[3], Rodrigo R. da Silva[1]

[1]Instituto Federal de Educação, Ciência e Tecnologia Sul-rio-grandense (IFSUL)
Campus Bagé – Av. Leonel de Moura Brizola, 2501 – 96.418-400 – Bagé – RS – Brasil

[2]Empresa Brasileira de Pesquisa Agropecuária (EMBRAPA)
Unidade Pecuária Sul – Bagé – RS – Brasil

[3]Centro de Ciências Exatas e Aplicadas (CCEA)
Universidade da Região da Campanha (URCAMP) – Bagé, RS – Brasil

thalesmaciel@ifsul.edu.br, vinicius.lampert@embrapa.br,
denizarsouza@urcamp.edu.br, orki2008@gmail.com

Abstract. *This paper describes a method for obtaining decision trees for predicting carcasse zootechnical quality indicators for bovine based on their breeding data. For such, data mining classification tasks were performed after data preprocessing. All numeric attributes were discretized by non-equal frequency binning or by cluster discovery in distinct classification experiments. Obtained results showed that clustering techniques as means for discretization may generate classes in better balancing conjecture when in comparison to the non-equal frequency binning method, allowing the discovery of models that may be applied to real world problems.*

Resumo. *Este artigo descreve uma metodologia para obtenção de árvores de decisão para previsão de indicadores zootécnicos de qualidade de carcaças bovinas com base em variáveis de cria dos animais. Para tal, procedeu-se a tarefas de mineração de dados com classificação após pré-processamento com discretização dos atributos numéricos por particionamento igualitário do intervalo ou por descoberta de agrupamentos em experimentos distintos de classificação. Os resultados obtidos mostraram que a descoberta de agrupamentos como forma de discretização pode gerar classes com balanceamento de melhor qualidade em comparação ao método tradicional, permitindo a indução de modelos utilizáveis em problemas reais.*

1. Introdução

O sistema de produção de gado de corte é o conjunto de tecnologias e práticas de manejo, tipo de animal, propósito de criação, raça e ecorregião onde a atividade é desenvolvida [Euclides Filho 2000]. Compreende uma das principais atividades de exploração econômica no Brasil, onde, há décadas, tem-se afastado o cenário de resistência ao emprego tecnológico, de modo a permitir estudos para o melhoramento dos índices de qualidade na produção de carne, por exemplo, através de computação aplicada [Barbosa 1999].

Em [Costa 2016], foram analisados dados zootécnicos de 401 animais bovinos da raça Hereford com vistas em prever o peso de fazenda e bonificação dos indivíduos. No estudo, foram empregadas redes neurais artificiais como ferramenta para o processo

de descoberta de conhecimento em experimentos distintos para as duas variáveis. Todos os dados envolvidos foram do tipo numérico. Segundo o autor, o trabalho foi concluído com resultados satisfatórios na previsão do peso de fazenda, mas insatisfatórios na previsão da bonificação, atribuindo o não cumprimento do objetivo específico à má qualidade de dados, uma característica não observada no primeiro experimento. O estudo não considerou a praticidade da utilização de redes neurais artificiais pelos produtores pecuários em meio às tarefas cotidianas, tampouco apresentou comparações com outros métodos para descoberta de conhecimento em bancos de dados.

Em [Da Mota et al. 2017], foram empregadas tecnologias de armazém de dados, consultas analíticas e mineração de dados para 1142230 registros de abates bovinos. O objetivo foi o de prever o grau de acabamento e o rendimento das carcaças, em experimentos individuais, que foram conduzidos com algoritmos de classificação e redes neurais artificiais. Os resultados, segundo os autores, foram promissores, devido às acurácias alcançadas nos experimentos, cuja média em acertos de classificação foi de 62%. Embora tenham composto médias de ácurácias da aplicação de diferentes algoritmos nas tarefas preditivas, os autores falharam em apresentar uma comparação das acurácias dos algoritmos utilizados individualmente, bem como avaliações mais aprofundadas dos resultados, que fossem além das acurácias observadas nos experimentos e apresentar os modelos gerados pelos mesmos e que são passíveis desta análise.

Nota-se que trabalhos correlatos publicados recentemente, mesmo que parcialmente eficazes segundo os respectivos autores, não explicam as predições realizadas pelos experimentos que documentam, ou pela impossibilidade disto ser característica do algoritmo empregado (caixa-preta) ou por não apresentar a totalidade dos resultados da classificação nos resultados obtidos nos testes (matrizes de confusão, por exemplo).

O problema de pesquisa abordado no presente estudo é fundamentado em “quais variáveis podem ser coletadas, pelos criadores, sobre os indivíduos de rebanhos bovinos em etapa de desenvolvimento de cria e que explicam a obtenção de indicadores de qualidade zootécnicos das carcaças ótimos após o abate?”

A hipótese trabalhada é que existe uma relação estatística entre o mês de nascimento, o mês de desmame, a idade de desmame e o peso de desmame com o peso e a idade de abate. Também são consideradas a influência das variáveis citadas sobre o ganho médio diário de peso (GMD) dos animais e a bonificação recebida pelas carcaças após o abate.

O objetivo é obter um modelo gráfico, de fácil interpretação, capaz de orientar os criadores bovinos sobre o desempenho de seus rebanhos, ainda em etapa anterior ao desmame, com previsões dos futuros índices de qualidade que serão obtidos após o abate dos animais.

2. Metodologia

Procedeu-se à descoberta de conhecimento em bancos de dados (DCBD), especificamente com as tarefas de mineração de dados descritas nesta seção.

O processo de DCBD pode ser dividido em três etapas [Maciel et al. 2015]: o pré-processamento, onde o conjunto de dados original é preparado para as próximas etapas do processo através de tarefas de filtragem conforme necessário; o processamento, onde algoritmos de mineração de dados são aplicados sobre o conjunto

de dados pré-processado e; o pós-processamento, onde os padrões descobertos no processamento são analisados e transformados em conhecimento útil sobre o domínio estudado.

Para fins de realização das tarefas e experimentos descritos neste estudo, foi empregado o Waikato Environment for Knowledge Analysis (WEKA), um ambiente para análise de conhecimento desenvolvido pela Universidade de Waikato [Hall et al. 2009].

O WEKA é uma coleção de algoritmos que podem ser utilizados em atividades de mineração de dados diversas, como classificação, regressão, associação e clustering, além de diversos métodos de pré-processamento e visualização de resultados através de interface gráfica, linha de comando ou interface de programação [Witten et al. 2017].

O conjunto de dados analisado teve sua apresentação original em 167 instâncias de animais bovinos e 8 atributos, conforme descrição na Tabela 1.

Tabela 1. Descrição do conjunto de dados analisado

#	Nome do Atributo	Significado	Tipo de Dado
1	nascimento_mes	mês de nascimento (01-12)	nominal
2	desmame_mes	mês de desmame (01-12)	nominal
3	desmame_idade	idade de desmame em meses	numérico
4	desmame_peso	peso de desmame em quilogramas	numérico
5	abate_idade	idade de abate em meses	numérico
6	abate_peso	peso de abate em quilogramas	numérico
7	gmd	ganho médio diário de peso	numérico
8	bonificacao	percentual de bonificação da carcaça após abate	numérico

No preprocessamento, os atributos numéricos foram discretizados, conforme a escala de Likert [Likert 1932], de forma a criar segmentos nominais dentre o intervalo numérico com as denominações: muito baixo, baixo, intermediário, alto e muito alto.

Discretização é o particionamento de um intervalo numérico e sucessiva atribuição de um valor categórico como rótulo de cada partição criada [Witten et al. 2017]. No âmbito deste estudo, dois métodos de discretização distintos foram experimentados:

- Discretização dos atributos em 5 segmentos sem balanceamento de peso entre eles, apenas dividindo o intervalo numérico de cada atributo em 5 frações iguais. Cada fração foi tornada em uma classe;
- Descoberta automatizada de 5 agrupamentos unidimensionais (sobre cada atributo numérico) com base na distância de Manhattan, em aplicação do algoritmo Simple k-means [Arthur e Vassilvitskii 2007]. Cada agrupamento descoberto for tornado em uma classe.

As Figuras 2 e 3 respectivamente apresentam os histogramas referentes às distribuições de frequência das instâncias de bovinos nas categorias propostas pela escala de Likert nos atributos discretizados pelos métodos do fracionamento igualitário do intervalo numérico e com a descoberta automatizada dos agrupamentos baseados na distância de Manhattan.

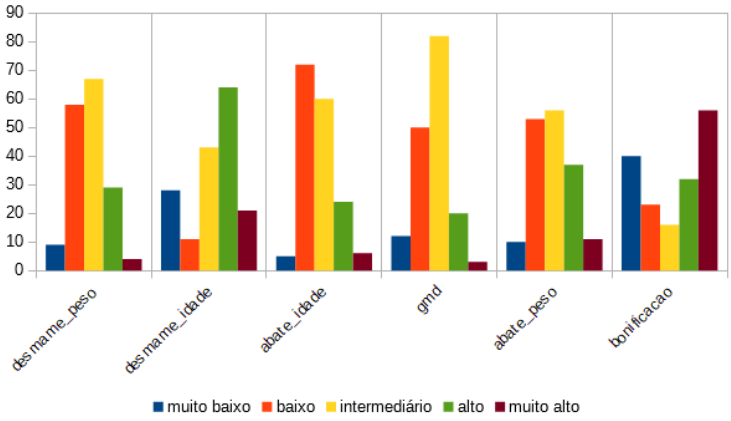

Figura 1. Histogramas referentes aos atributos discretizados por segmentação

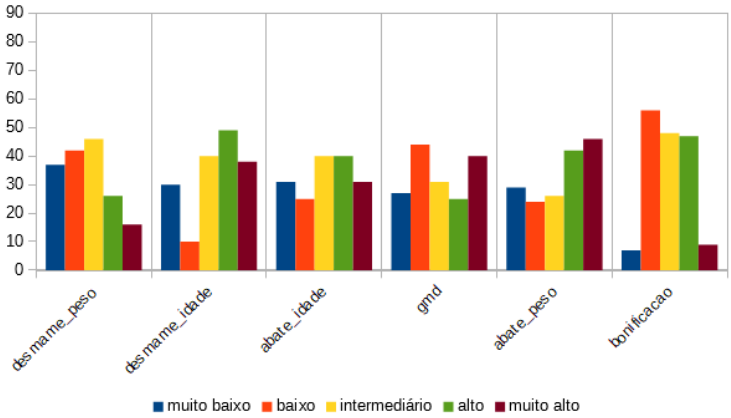

Figura 2. Histogramas referentes aos atributos discretizados por descoberta de agrupamentos.

No total, foram realizados 8 experimentos de predição. Neles, as variáveis de cria (mês de nascimento, mês de desmame, peso de desmame e idade de desmame) foram utilizadas para prever os indicadores zootécnicos de qualidade das carcaças (idade de abate, peso de abate, GMD e bonificação) em experimentos distintos.

Cada indicador zootécnico de qualidade das carcaças foi alvo de predição após discretização dos atributos numéricos através dos dois métodos apresentados. A Tabela 2 elenca as atividades de pré-processamento pelas quais cada atributo foi submetido em cada experimento. Os atributos cujo tipo de dado original é o nominal não passaram pois quaisquer tarefas de pré-processamento (N/A). Os atributos alvo da predição em cada experimento são destacados em sublinhado na Tabela 2.

A mineração de dados é a tarefa de identificação de padrões a partir de dados, de forma automatizada em ambiente computacional, que compreende a etapa de processamento no processo de DCBD [Maciel et al. 2015].

A classificação é um tipo de tarefa da mineração de dados que visa categorizar instâncias supostamente novas com base na análise de dados de instâncias pregressas [Witten et al. 2017]. Há a etapa de treinamento, onde um algoritmo aprende as características inerentes à cada classe e a etapa de teste, onde é verificada a acurária do modelo criado.

Árvores de decisão são um tipo de modelo de dados utilizado como resultado de tarefas de classificação [Quinlan 1993] e apreciado no contexto deste estudo em virtude de sua simplicidade e interpretabilidade.

Tabela 2. Descrição das tarefas de pré-processamento aplicadas a cada atributo do conjunto de dados nos experimentos realizados.

#	nascimento_mes	desmame_mes	desmame_peso	desmame_idade	abate_idade	abate_peso	gmd	bonificacao
1	N/A	N/A	segmentado	segmentado	segmentado	removido	removido	removido
2	N/A	N/A	segmentado	segmentado	removido	segmentado	removido	removido
3	N/A	N/A	segmentado	segmentado	removido	removido	segmentado	removido
4	N/A	N/A	segmentado	segmentado	removido	removido	removido	segmentado
5	N/A	N/A	agrupado	agrupado	agrupado	removido	removido	removido
6	N/A	N/A	agrupado	agrupado	removido	agrupado	removido	removido
7	N/A	N/A	agrupado	agrupado	removido	removido	agrupado	removido
8	N/A	N/A	agrupado	agrupado	removido	removido	removido	agrupado

O J48 [Hall et al. 2009] é um algoritmo de mineração de dados, especificamente para tarefas de classificação, capaz de induzir árvores de decisão, sendo um dos algoritmos mais utilizados em aplicações do tipo no mundo real. Trata-se da implementação em Java da 8ª revisão [Quinlan 1996], do algoritmo C4.5 [Quinlan 1993], originalmente documentado em linguagem C.

A etapa de processamento em todos experimentos foi realizada com o algoritmo J48 configurado para permitir apenas divisões binárias em galhos formados por atributos nominais e desconsiderar limites inferiores de ocorrências de instâncias em folhas para critérios de poda em seu treinamento. Os demais parâmetros do algoritmo foram mantidos em conformação padrão.

Os testes dos modelos descobertos foram realizados sobre os mesmos conjuntos de dados de entrada para as respectivas etapas de treinamento. Embora este não seja apreciado como o mais adequado método de testes em aplicações no mundo real [Witten et al. 2017], é possível observar nas Figuras 1 e 2 que muitas classes apresentam representatividade baixa dentre as 167 instâncias analisadas. Por tal motivo, é inviabilizada a realização de testes por validação cruzada com 5 frações, por exemplo.

3. Resultados Obtidos

Os resultados obtidos nas tarefas de classificação descritas na Seção 2 foram apresentados na forma de árvores de decisão, matrizes de confusão e acurácias dos respectivos modelos. Também é discutida a praticidade dos modelos descobertos.

As acurácias alcançadas em todos experimentos realizados neste estudo estão dispostas na Tabela 3.

Tabela 3. Acurácias resultantes dos experimentos realizados

Experimento	#1 (%)	#2 (%)	#3 (%)	#4 (%)	#5 (%)	#6 (%)	#7 (%)	#8 (%)
Acurácia	63,47	49,70	61,68	53,29	53,29	51,50	53,89	55,09

Foi feita comparação dos resultados obtidos na classificação com os conjuntos de dados cujos atributos numéricos foram discretizados por segmentação igualitária dos intervalos numéricos os conjuntos de dados cujos atributos numéricos foram discretizados pela descoberta automatizada de agrupamentos por aplicação do algoritmo Simple k-means. Foram analisadas as acurácias e matrizes de confusão resultantes dos

experimentos, onde foram evidenciadas falhas cruciais em alguns dos modelos gerados. A Tabela 4 apresenta as matrizes de confusão encontradas nos testes.

Matrizes de confusão são representadas em forma de tabela, onde as linhas representam as classes verdadeiras para cada instância e as colunas representam as classes onde cada instância foi classificada pelo modelo [Witten et al. 2017]. Desta forma, é possível visualizar a quantificação consolidada dos acertos e erros para cada classe nos experimentos.

Tabela 4. Matrizes de confusão resultantes dos experimentos realizados

	abate_idade					abate_peso					gmd					bonificação				
classe verdadeira \ prevista	mb	b	i	a	ma	mb	b	i	a	ma	mb	b	i	a	ma	mb	b	i	a	ma
muito baixo (mb)	0	5	0	0	0	0	3	6	1	0	1	7	4	0	0	15	1	1	11	12
baixo (b)	0	57	11	4	0	0	25	22	6	0	0	28	22	0	0	3	3	2	4	11
intermediário (i)	0	20	32	8	0	0	9	43	4	0	0	13	62	7	0	1	0	5	4	6
alto (a)	0	5	2	17	0	0	6	16	15	0	0	2	6	12	0	5	0	0	16	11
muito alto (ma)	0	5	0	1	0	0	1	6	4	0	0	0	2	1	0	3	0	0	3	50
	Experimento #1					Experimento #2					Experimento #3					Experimento #4				
classe verdadeira \ prevista	mb	b	i	a	ma	mb	b	i	a	ma	mb	b	i	a	ma	mb	b	i	a	ma
muito baixo (mb)	24	1	3	0	3	16	1	5	2	5	30	1	4	3	2	34	0	14	0	8
baixo (b)	11	16	4	2	7	2	6	3	1	14	8	11	8	4	0	3	0	1	0	3
intermediário (i)	6	1	15	3	6	3	3	20	2	14	1	0	16	0	0	6	0	35	1	5
alto (a)	5	1	1	11	7	1	0	6	7	10	7	2	27	17	1	2	0	4	2	1
muito alto (ma)	7	3	2	5	23	1	3	4	1	37	7	0	8	4	6	15	0	12	0	21
	Experimento #5					Experimento #6					Experimento #7					Experimento #8				

Para previsão da idade de abate, foram realizados dois experimentos, #1 e #5, cujos modelos gerados apresentaram acurácias de 63,47% e 53,29% respectivamente, com diferença de 10,18% em favor do primeiro. Contudo, a matriz de confusão referente ao experimento #1 evidencia a incapacidade do modelo em classificar as instâncias nas idades de abate muito baixa e muito alta, o que sobremaneira impede que o mesmo tenha proveito prático em alinhamento com os objetivos do presente trabalho. Esta problemática não foi presente nos resultados obtidos no experimento #5, cuja árvore de decisão resultante é apresentada na Figura 3.

Os experimentos #2 e #6, referentes à previsão do peso de abate, apresentaram acurárias de 49,70% e 51,50% respectivamente, com diferença de 1,80% em favor do segundo. Observou-se que o experimento #2, além de não ter logrado melhor acurácia em comparação com o experimento #6, expõe a mesma problemática apresentada pelos resultados do experimento #1. À exemplo deste, o experimento #2 foi incapaz de classificar as instâncias de animais bovinos nas categorias muito baixo e muito alto, neste caso acerca do peso de abate. A árvore de decisão resultante do experimento #6 é apresentada na Figura 4.

A previsão do ganho médio diário de peso (GMD) foi abordada nos experimentos #3 e #7. Seus resultados apresentaram as acurácias de 61,68% e 53,89%,

respectivamente, com diferença de 7,79% em favor do primeiro. O experimento #3, bem como os experimentos #1 e #2, resultou em um modelo incapaz de classificar bovinos na categoria muito alto para GMD. A classificação dos bovinos em GMD muito baixo ocorreu de forma semelhante, com a diferença de que o modelo foi capaz de classificar 1 instância nesta categoria e corretamente. O experimento #7 resultou num modelo livre desta problemática, conforme apresentado na Figura 5.

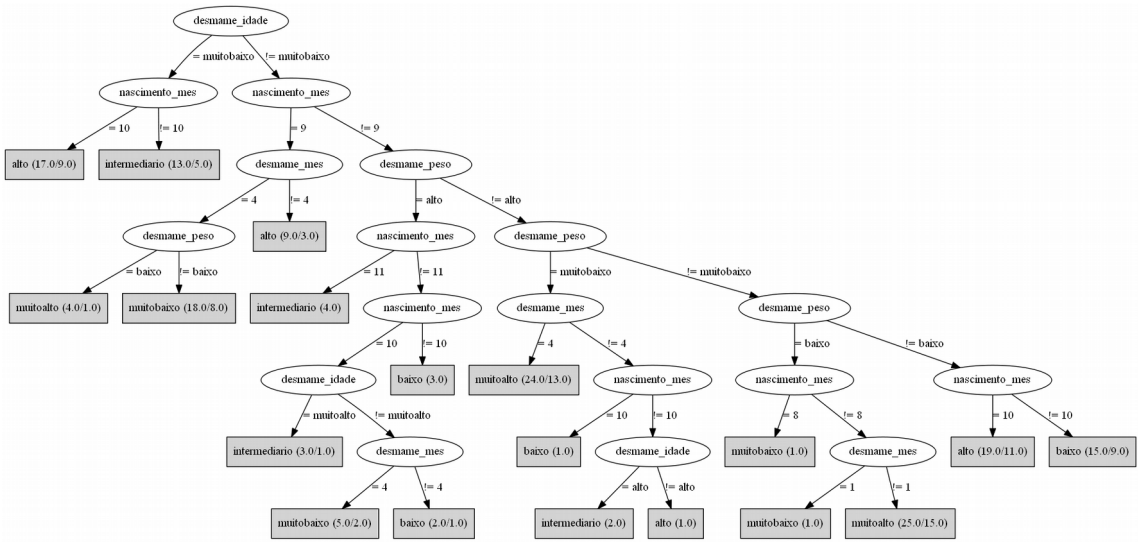

Figura 3. Árvore de decisão para predição da idade de abate

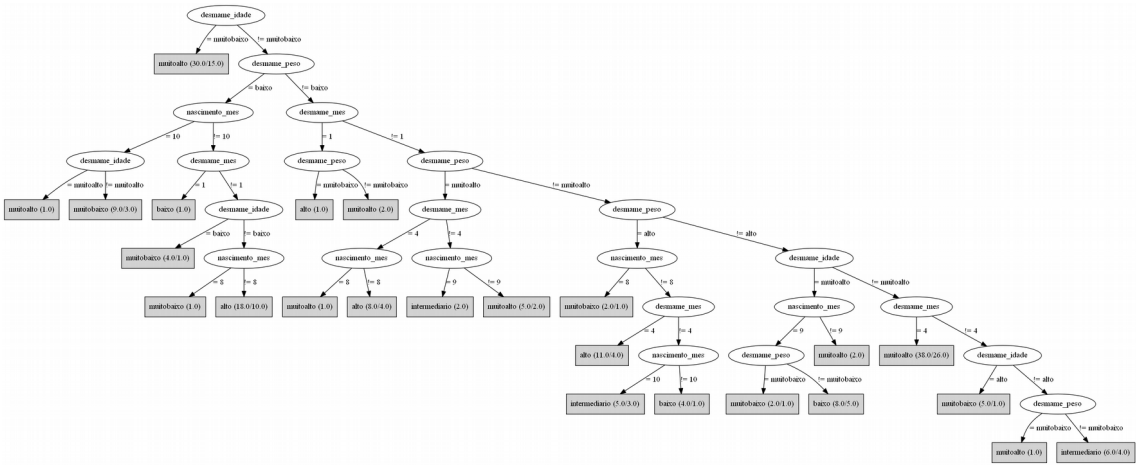

Figura 4. Árvore de decisão para predição do peso de abate

Os experimentos #4 e #8 foram realizados para predição da bonificação e apresentaram os resultados de 53,29% e 55,09%, respectivamente, em acurácias, tendo a diferença de 1,80% em favor do segundo. Contudo, o modelo gerado pelo experimento #8 apresentou incapacidade de classificar instâncias com a bonificação na categoria baixo, tornando o modelo impraticável. O modelo gerado pelo experimento #4 foi isento de tal problemática e sua respectiva árvore de decisão é apresentada na Figura 6.

Em sumário, nos testes realizados, os conjuntos de dados cujos atributos numéricos foram discretizados pela descoberta automatizada dos 5 agrupamentos com aplicações do algoritmo Simple k-means foram as entradas mais adequadas para processamento com tarefas de classificação com o algoritmo J48 sobre as variáveis idade de abate, peso de abate e GMD, ao passo que o conjunto de dados cuja

discretização dos atributos numéricos foi realizada pela segmentação igualitária do intervalo numérico em 5 frações foi a entrada mais adequada para previsão da bonificação.

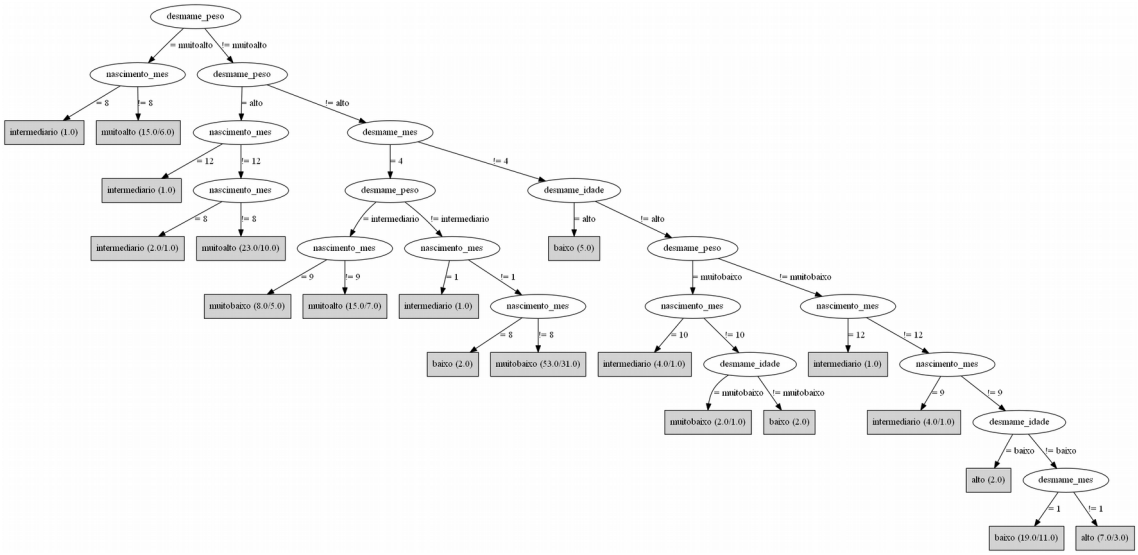

Figura 5. Árvore de decisão para predição do GMD

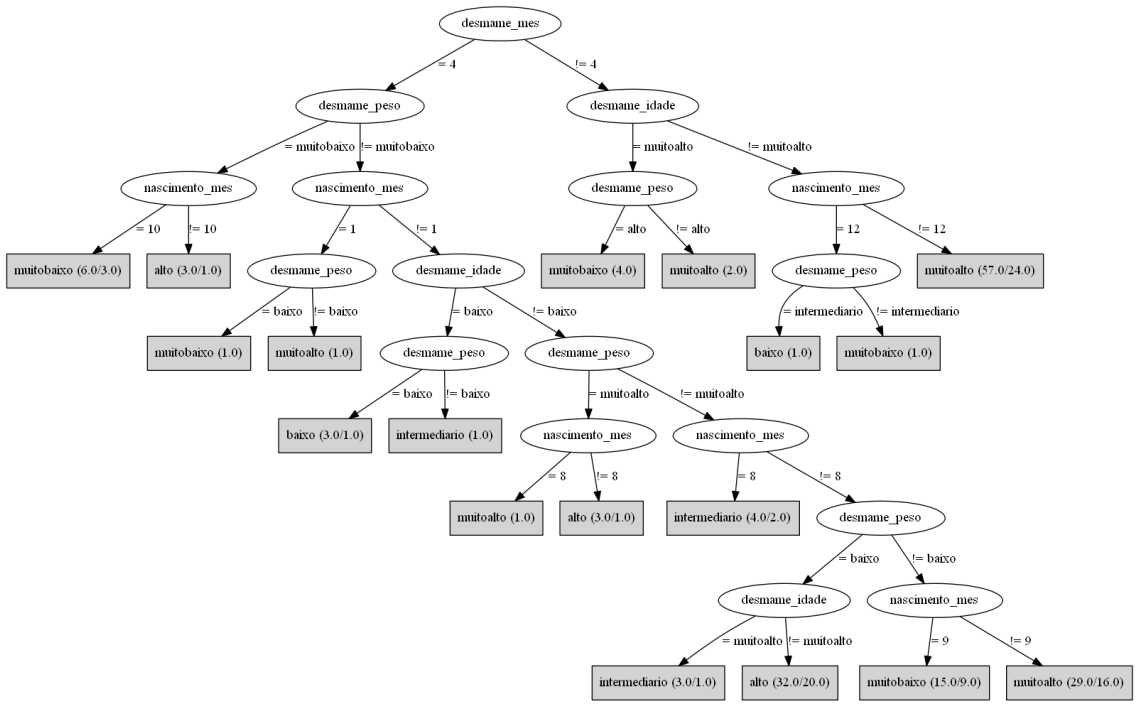

Figura 6. Árvore de decisão para predição da bonificação

Isto ocorreu em virtude do desbalanceamento entre as frequências das categorias após as tarefas de discretização, que podem ser observadas nos histogramas apresentados nas Figuras 1 e 2. Estas figuras evidenciam as diferenças de balanceamento das classes e permitem a comparação dos resultados da discretização para os dois métodos de discretização utilizados sobre os atributos originalmente numéricos. Entende-se que o desbalanceamento observado em alguns histogramas são a

causa da impossibilidade do algoritmo em gerar modelos que não negligenciam quaisquer classes a partir dos respectivos conjuntos de dados.

4. Conclusão

O presente trabalho buscou um método para descoberta de árvores de decisão capazes de auxilizar os produtores de gado de corte na previsão de indicadores zootécnicos da qualidade das carcaças com base nas respectivas variáveis de cria, ou seja, dados que podem ser coletados entre o nascimento e o desmame dos animais.

Foram realizados experimentos de classificação com o algoritmo J48 após pré-processamento do conjunto de dados para discretização dos atributos numéricos. O método tradicional de discretização, pelo particionamento igualitário do intervalo numérico, se mostrou problemático ao produzir categorias com frequências desbalanceadas para atributos do conjunto de dados apresentado. Diante desta situação, foi proposto que as tarefas de discretização fossem realizadas através da descoberta automatizada das categorias por medida da distância de Manhattan, através de aplicação do algoritmo Simple k-means.

Esta abordagem proveu maior qualidade de discretização para 75% dos experimentos realizados, provendo melhor balanceamento entre as categorias criadas em relação à primeira, que foi capaz de produzir categorias sem problemática de desbalaceamento em 25% dos experimentos realizados.

Finalmente, foi possível descobrir árvores de decisão capazes de explicar a influência das variáveis de cria mês de nascimento, mês de desmame, peso de desmame e idade de desmame nos indicadores zootécnicos de qualidade de carcaças, como idade de abate, peso de abate, ganho médio diário de peso e bonificação, cumprindo de forma satisfatória o objetivo do estudo.

Trabalhos futuros envolvem tarefas adicionais de coleta de dados, com vistas no melhoramento da representatividade das classes descobertas por discretização, de modo a possibilitar maior adequação do método de teste dos modelos descobertos. Também será objetivada a otimização da acurácia das árvores de decisão, o que pode ser abordado por diferentes métodos de discretização dos atributos numéricos, diferentes métodos de descoberta de agrupamentos para aplicação na discretização, experimentações com outros algoritmos de indução de árvores de decisão, empilhamento de classificadores e aprendizado sensível à custo.

Referências

Arthur, D. and Vassilvitskii S. (2007) k-means++: the advantages of carefull seeding. In: Proceedings of the eighteenth annual ACM-SIAM symposium on Discrete algorithms, 1027-1035.

Barbosa, P. (1999) Raças e estratégias de cruzamento para produção de novilhos precoces. Embrapa Pecuária Sudeste. In: Simpósio de Produção de Gado de Corte, 1. Viçosa, Brasil.

Costa, C. L. (2016). Utilização de caracteríticas zootécnicas e de manejo na pecuária para previsão do peso final e bonificação de bovinos empregando redes neurais artificiais. Tabalho de conclusão de curso, Universidade Federal do Pampa.

Euclides Filho, K. (2000) Produção de bovinos de corte e o trinômio genótipo-ambiente-mercado. Embrapa Gade de Corte - Documentos (Infoteca-E).

Hall, M., Frank, E., Holmes, G., Pfahringer, B., Reutemann, P. and Witten, I. (2009) The WEKA Data Mining Software: An Update. SIGKDD Explorations, Volume 11, Issue 1.

Likert, R. (1932) A Technique for the Measurement of Attitudes. Archives of Psychology. 140: 1–55.

Maciel, T., Seus, V., Machado, K. and Borges, E. (2015). Mineração de dados em triagem de risco de saúde. Revista Brasileira de Computação Aplicada, 7(2), 26-40.

Mota, F., Souza, K., Ishii, R. and Gomes, R. (2017) BovReveals: uma plataforma OLAP e data mining para tomada de decisão na pecuária de corte. In:: Congresso Brasileiro de Agroinformática, 11. Campinas, Brasil.

Quinlan, R. (1993) C4.5: Programs for Machine Learning. Morgan Kaufmann Publishers, San Mateo, CA.

Quinlan, J. R. (1996). Improved use of continuous attributes in C4. 5. Journal of artificial intelligence research, 4, 77-90. Chicago, IL.

Witten, I., Frank, E., Hall, M. and Pal, C. (2017) Data mining: practical machine learning tools and techniques. Morgan Kaufmann.