APLICAÇÃO DO APRENDIZADO DE MÁQUINA SOBRE A QUALIDADE DA PRODUÇÃO AGROECOLÓGICA DE LEITE

JULIA ELISABETT KLOCOSKI BOLSONELLO [1], ADEMIR ROBERTO FREDDO[2], ANDRÉ LAZARIN GALLINA [3], FABIANA ELIAS [4], KARINA RAMIREZ STARIKOFF[5]

1 Introdução/Justificativa

A bovinocultura de leite no país é caracterizada por pequenas propriedades de base familiar, que carecem de assistência e informações técnicas, o que reflete diretamente em um leite de baixa qualidade (LORDÃO et al., 2013). Desse modo, ferramentas que auxiliem na interpretação dos dados, podem favorecer os produtores a produzirem leite de qualidade, de forma sustentável, agroecológica e segura para os consumidores.

O aprendizado de máquina é um subcampo da ciência da computação que trabalha com reconhecimento de padrões e teoria da aprendizagem computacional em inteligência artificial na análise de dados. Na mineração de dados são utilizadas técnicas de aprendizado de máquina para descoberta de padrões e extração de conhecimento (HAN; KAMBER; PEI, 2012) e tem como objetivo ensinar ao computador a aprender a partir das próprias experiências (MITCHELL, 1997). A partir de um conjunto de dados, aplica-se alguma técnica de aprendizado de máquina, entre elas, a técnica de árvores de decisão (BERRY; LINOFF, 1997). Por meio dessas técnicas, espera-se desenvolver um software para auxiliar os produtores na produção agroecológica e de melhor qualidade do leite.

2 Objetivo

Empregar técnicas de aprendizado de máquina, mais especificamente, a técnica de mineração de dados denominada de árvores de decisão, para prever a qualidade do leite no que se refere

1 Acadêmica de Medicina Veterinária, Universidade Federal da Fronteira Sul, *campus* Realeza, contato: bolsonellojulia@gmail.com.
2 Professor Adjunto, Doutor em Engenharia Elétrica e Informática Industrial, Universidade Federal da Fronteira Sul, *campus* Realeza.
3 Professor Adjunto, Doutor, Químico, Universidade Federal da Fronteira Sul, *campus* Realeza.
4 Professora Adjunta, Doutora, Médica veterinária, Universidade Federal da Fronteira Sul, *campus* Realeza.
5 Professora Adjunta, Doutora, Médica veterinária, Universidade Federal da Fronteira Sul, *campus* Realeza., **Orientadora**.

a contagem de células somáticas (CCS) e contagem bacteriana total (CBT).

3 Material e Métodos/Metodologia

A pesquisa foi aprovada pelo protocolo número 88506018.7.0000.5564 emitido pelo Comitê de Ética em Pesquisa (CEP) da Universidade Federal da Fronteira Sul (UFFS).

Na atividade de mineração de dados, utilizou-se o método KDD (Knowledge Discovery in Databases - KDD) para extrair conhecimento a partir dos dados, que possui as seguintes etapas: **Seleção de dados:** Os dados foram obtidos de duas origens: de questionários aplicados aos produtores de leite do oeste e sudoeste do Paraná e dos resultados das análises físico-químicas do leite fornecidas por um laticínio do sudoeste paranaense. Selecionou-se 27 questões do questionário sobre características das propriedades: número e raça dos animais, tipo de alimentação, origem da água, manejo e limpeza nos procedimentos de ordenha, bem como, instalações e tipo de resfriador. Os dados obtidos das análises do leite foram: contagem de células somáticas (CCS), contagem bacteriana total (CBT), proteína total, sólidos totais, lactose e gordura, de um período de quatro meses (dezembro/2018 a março/2019). Assim, foi criada uma base de dados composta por 743 registros. **Pré-processamento:** Nesta etapa ocorreu a identificação e retirada de dados duplicados, incorretos, faltantes, valores atípicos e discrepantes. **Transformação e Preparação dos dados:** Foi realizada a categorização de valores numéricos para os itens seguindo os parâmetros de referência, classificando o leite em A, B (conforme a legislação vigente IN 76/2018 do Ministério da Agricultura, Pecuária e Abastecimento) e C (daqueles com valores acima dos máximos exigidos na normativa). Considerando que neste trabalho foi utilizado o ambiente R-Studio4 (interface funcional para o ambiente R), o formato exigido pela ferramenta para entrada de dados é o CSV (Comma Separated Values). **Mineração de Dados:** Compreendeu a aplicação do algoritmo RPART que implementa a árvore de decisão utilizando a ferramenta R e o método CART (Classification and Regression Trees) com processos de poda para diminuir as taxas de erro. Nesta fase, foram criadas duas árvores de decisão para a base, uma para CCS e outra para CBT. Optou-se por utilizar 70% das bases de dados para treinamento e 30% para testes: (744 registros, 520 foram utilizados para o treinamento e 223 para testar). A escolha dos registros de testes e treinamento é feita aleatoriamente pelas ferramentas computacionais na base de dados, com

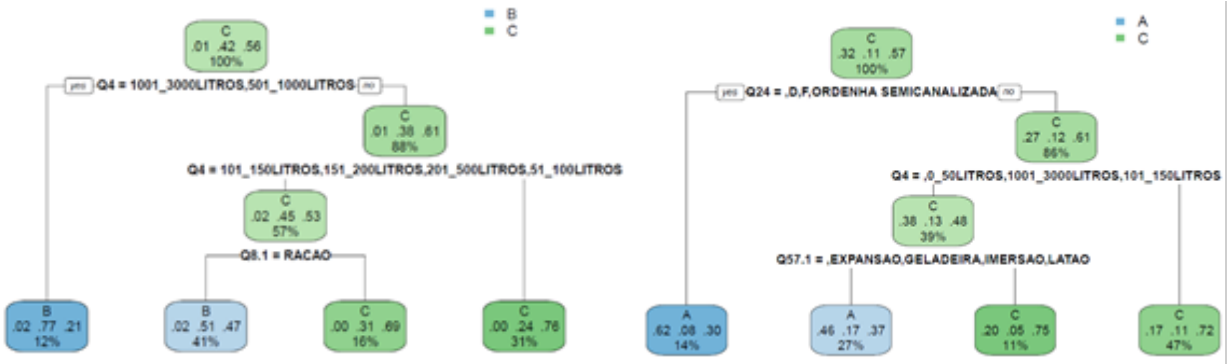
intuito de diminuir os erros. **Extração do conhecimento:** Foram criados os modelos ou padrões que são interpretados nesta etapa para a geração ou extração de conhecimento. Nesta etapa, verifica-se quais são as regras criadas pela árvore de decisão e também quais são os registros (neste caso questões) que mais influenciam sobre os valores de CCS e CBT, que por sua vez determinam a qualidade do leite.

4 Resultados e Discussão

A Figura 1A ilustra uma parte da árvore (profundidade 3) ou modelo criado para análise do CBT, e é possível observar que os atributos mais importantes (dentre os 27 utilizados) para determinação do CBT foram o Q4 (quantidade da produção) e Q8.1 (uso de ração na alimentação do rebanho). Na Figura 1B é ilustrado o modelo criado para determinar o CCS. Percebe-se na árvore que não há informações relevantes para determinar o parâmetro B do CCS, porém, os atributos Q24 (tipo de ordenha), Q4 (quantidade da produção) e Q57.1 (armazenamento do leite) foram os mais importantes para a classificação A ou C do CCS.

A eficiência do modelo para CCS foi de 59,82%, ou seja, a árvore criada foi testada para verificar sua eficácia ou eficiência na classificação do CCS a partir de novos dados que o computador ainda não conhece. A qualidade de leite dos produtores para CCS variou entre A, B e C. Já o modelo ou árvore obtida para CBT a precisão ou eficiência do modelo foi de 72,2%. Isto significa, que sem conhecer o valor do CBT a árvore consegue determinar a qualidade do leite apenas com base nos atributos existentes no modelo. Para o CBT a classificação da qualidade do leite foi apenas B e C.

Figura 1. Parte dos modelos de árvores de decisão gerados para Contagem Bacteriana Total (1A) e Contagem Células Somáticas (2B).



Fonte da imagem: Elaborado pelo autor.

ISSN 2526-205x

5 Conclusão

As eficiências dos testes foram baixas devido às poucas entradas de dados e talvez devido a escolha dos atributos. Por esses motivos, para se melhorar os resultados em trabalhos futuros deve-se ter um maior número de dados (questionários) para cruzar com os resultados das análises do leite e, assim, avaliar a qualidade do leite e os fatores que a influenciam. Também deve-se avaliar a relevância dos atributos, bem como realizar novos testes dentro da mineração de dados.

Palavras-chave: bacia leiteira; software; mineração de dados; agricultura familiar.

Financiamento

Bolsa concedida pela UFFS na modalidade Iniciação em Desenvolvimento Tecnológico e de Inovação (PIBIT).

Referências

BERRY, M. J. A.; LINOFF, G. Data Mining Techniques: For Marketing, Sales, and Customer Support. New York: **Wiley Computer Publishing**, 1997.

HAN, J.; KAMBER, M.; PEI,J. **Data mining concepts and techniques**. USA: Elsevier, 2012.

LORDÃO, A. C. et al. Implantação de medidas de higiene na ordenha para melhoria da qualidade do leite no município de Paty do Alferes/RJ, Brasil. **Archives of Veterinary Science**, v. 18, n. 4, 2013.

MITCHELL TM. **Machine Learning**: McGraw–Hill Science/Engineering/Math; 1997.