



MÉTODO PREDITIVO BASEADO EM RANDOM FOREST PARA ESTIMAR A CONTAGEM DE CÉLULAS SOMÁTICAS DE VACAS JERSEY EM LACTAÇÃO

ROMICY DERMONDES SOUZA; ROSANGELA SILVEIRA BARBOSA; ISABELA SILVA COSTA; SÉRGIO AUGUSTO DE ALBUQUERQUE FERNANDES

RESUMO

A mastite bovina subclínica é a principal causa de perdas econômicas na pecuária leiteira, sendo a Contagem de Células Somáticas (CCS) o indicador chave de inflamação e qualidade do leite. Métodos rápidos e de baixo custo são necessários para o monitoramento eficiente, e a Espectroscopia de Infravermelho por Transformada de Fourier (FTIR), que detecta alterações bioquímicas correlacionadas à CCS, desponta como alternativa promissora. Assim, o objetivo deste trabalho foi avaliar o desempenho de um modelo de regressão baseado no algoritmo Random Forest para a predição da CCS em amostras de leite de vacas da raça Jersey. Amostras de leite foram coletadas na Estação Experimental Terras Baixas da Embrapa Clima Temperado, provenientes de vacas com diferentes graus de mastite, e a coleta seguiu rigorosos procedimentos de higiene. A modelagem preditiva envolveu a seleção de variáveis espectrais utilizando o algoritmo Boruta e uma etapa posterior de filtragem por correlação para reduzir a redundância. O conjunto de dados foi dividido em 80% para treinamento e 20% para teste. A aplicação do Boruta inicialmente identificou 87 números de onda relevantes, mas após a filtragem por correlação, apenas cinco foram retidos: 2681, 1875, 1142, 1130 e 1018 cm^{-1} . O modelo de regressão por Random Forest, treinado com essas cinco variáveis, apresentou desempenho restrito no conjunto de teste, com um coeficiente de determinação de $R^2 = 0,246$ e um RMSE de 728.477 células/mL. O R^2 indica que o modelo explicou apenas cerca de 24,6% da variabilidade da CCS, e o alto RMSE demonstra um erro médio considerável, tornando o modelo inviável para quantificações regulatórias. Além disso, os limites de detecção e quantificação revelaram que o modelo não é sensível o suficiente para níveis típicos de produção leiteira. Conclui-se que, apesar da identificação de bandas espectrais relacionadas à mastite, a redução extrema de variáveis, embora favorecendo a interpretação, comprometeu a capacidade preditiva do modelo. Dessa forma, estudos futuros devem considerar o uso de todo o espectro FTIR ou algoritmos híbridos para aprimorar a acurácia do método.

Palavras-chave: Diagnóstico Rápido; Aprendizado de Máquina; Gado Leiteiro.

1 INTRODUÇÃO

A produção de leite é um dos pilares da economia agropecuária global, com a qualidade do produto final sendo um fator determinante para a rentabilidade da atividade (STANEK; ŻÓŁKIEWSKI; JANUŚ, 2024). Nesse contexto, a mastite bovina, especialmente a forma subclínica, representa a principal causa de perdas econômicas nos rebanhos leiteiros,

resultando na diminuição da produção de leite e gordura, além de afetar a composição geral do produto, como a redução dos teores de lactose (CHENG; HAN, 2020). A Contagem de Células Somáticas (CCS) no leite é o principal indicador de inflamação da glândula mamária e, conseqüentemente, da qualidade sanitária do leite cru, sendo o seu monitoramento essencial para o manejo do rebanho. No Brasil, a legislação estabelece limites máximos de CCS que devem ser rigorosamente atendidos para a comercialização do leite (MAPA, 2018).

Diante da necessidade de métodos de triagem rápidos e de baixo custo, a Espectroscopia de Infravermelho por Transformada de Fourier (FTIR) surge como uma alternativa promissora. Esta técnica avalia a composição química do leite de forma indireta, identificando alterações no perfil bioquímico, como proteínas, lipídios e lactose, que estão correlacionadas com a elevação da CCS causada pela inflamação (CHEN et al., 2014; MOURA et al., 2017). Para converter os dados espectrais complexos em estimativas quantitativas, técnicas avançadas de aprendizado de máquina, como a Random Forest, têm ganhado relevância (JAFARZADEH et al., 2021). O algoritmo Random Forest é um modelo preditivo não linear que oferece simplicidade, flexibilidade e robustez, sendo capaz de lidar com a alta dimensionalidade e colinearidade inerentes aos dados espectrais (BREIMAN, 2001). A aplicação de Random Forest para a predição da CCS a partir de espectros FTIR representa um passo importante na busca por ferramentas rápidas de diagnóstico e manejo da qualidade do leite.

Dessa forma, o presente trabalho teve como objetivo avaliar o desempenho do modelo de regressão baseado no algoritmo Random Forest, utilizando números de onda selecionados por relevância espectral, para a predição da Contagem de Células Somáticas (CCS) em amostras de leite de vacas da raça Jersey em lactação.

2 MATERIAL E MÉTODOS

Amostras de leite foram coletadas na Estação Experimental Terras Baixas (ETB) da Embrapa Clima Temperado, localizada em Capão do Leão, Rio Grande do Sul (31° 49' 27" S e 52° 26' 35" W). O objetivo da coleta foi compor um conjunto de amostras com ampla variação nos teores de células somáticas (CCS), essencial para o desenvolvimento e validação de modelos preditivos. Para isso, foram selecionadas vacas com diferentes graus de mastite, desde animais clinicamente saudáveis até casos subclínicos leves e moderados, garantindo uma representação abrangente do espectro de CCS encontrado na produção leiteira. Todas as vacas incluídas no estudo não haviam recebido antibióticos nos 30 dias anteriores à coleta. Antes da ordenha, os tetos foram higienizados com solução iodada e secos com papel toalha descartável. Os primeiros jatos de leite foram descartados, e as amostras foram coletadas manualmente por um ordenhador previamente treinado e com as mãos higienizadas. Imediatamente após a coleta, as amostras foram acondicionadas em caixas térmicas com temperatura controlada ($5^{\circ}\text{C} \pm 5^{\circ}\text{C}$) e transportadas ao laboratório.

A seleção dos números de onda mais informativos para a modelagem preditiva foi realizada em duas etapas, visando maximizar a relevância preditiva e minimizar a redundância espectral. Inicialmente, todas as variáveis espectrais foram submetidas ao algoritmo Boruta, um método baseado em Random Forest que compara a importância das variáveis reais com a de *shadow features* para identificar, de forma estatisticamente robusta, quais preditores são significativamente mais relevantes do que o acaso. Antes da aplicação do Boruta, os espectros foram pré-processados por meio de centralização na média seguida de escalonamento min-max. Após a identificação das variáveis confirmadas como importantes pelo Boruta, realizou-se uma segunda etapa de filtragem com base na correlação intervável. Calculou-se a matriz de correlação de Pearson entre os números de onda selecionados, e pares de variáveis com coeficiente de correlação absoluta $r \geq 0.8$ foram considerados altamente redundantes e eliminados para reduzir multicolinearidade. O conjunto resultante compõe as variáveis espectrais utilizadas nas etapas subsequentes de modelagem preditiva.

A predição do teor de CCS foi realizada por meio de regressão por Random Forest utilizando as variáveis selecionadas pelo boruta. O conjunto de dados foi dividido aleatoriamente em 80% para treinamento e 20% para teste, mantendo a representatividade da distribuição de concentrações. O modelo Random Forest foi ajustado com 500 árvores, com o número mínimo de observações por nó terminal fixado em 5 e o número de variáveis candidatas a cada divisão otimizado internamente pelo algoritmo. A avaliação do desempenho preditivo foi baseada nas seguintes métricas: coeficiente de determinação (R^2), raiz do erro quadrático médio (RMSE), limite de decisão ($CC\alpha$), capacidade de detecção ($CC\beta$), limite de detecção (LOD) e limite de quantificação (LOQ). A validação do modelo foi realizada no conjunto de teste independente, garantindo uma estimativa realista da capacidade preditiva. Todos os cálculos foram implementados na linguagem R (versão 4.4.3), utilizando o pacote randomForest.

3 RESULTADOS E DISCUSSÃO

Para a seleção dos números de onda com maior capacidade preditiva foi utilizado o algoritmo boruta. A aplicação do algoritmo Boruta identificou 87 números de onda como estatisticamente relevantes para a predição do teor de CCS. Essas variáveis representam regiões espectrais com maior poder discriminante em relação à resposta analítica, superando consistentemente as *shadow features* geradas aleatoriamente (KURSA; RUDNICKI, 2010). Contudo, uma análise de correlação entre essas variáveis revelou alta redundância espectral, comum em dados FTIR devido à natureza contínua e suave dos espectros. Após a remoção de variáveis com $r \geq 0,8$, apenas cinco números de onda foram retidos no conjunto final: 2681, 1875, 1142, 1130 e 1018 cm^{-1} . Essa redução drástica, de 87 para 5 variáveis, demonstra que, embora muitas regiões espectrais contenham informação relevante, grande parte dela é altamente colinear. A retenção de um conjunto mínimo e não redundante não apenas simplifica o modelo, mas também melhora sua interpretabilidade e robustez, mitigando riscos de sobreajuste [referência].

Os números de onda selecionados correspondem a regiões vibracionais frequentemente associadas a ligações funcionais relevantes em matrizes biológicas ou orgânicas. Os números de onda identificados em sua análise FTIR de leite (2681, 1875, 1142, 1130 e 1018 cm^{-1}) correspondem a regiões do espectro infravermelho associadas a diferentes grupos funcionais presentes nos componentes do leite. Em geral, bandas próximas de 2681 cm^{-1} podem estar relacionadas a estiramentos C-H de grupos aldeídos ou a sobretons de ligações C=O, embora essa região não seja típica para os principais constituintes do leite. A região em torno de 1875 cm^{-1} normalmente está associada a sobretons ou combinações de vibrações de carbonilas (C=O), que podem estar presentes em pequenas quantidades em compostos como ácidos orgânicos (TOLEDO-ALVARADO et al., 2021; WANG; HULZEBOSCH; BOVENHUIS, 2016). As bandas em 1142, 1130 e 1018 cm^{-1} são características de estiramentos C-O e C-C, frequentemente atribuídas a carboidratos como a lactose, além de possíveis contribuições de fosfatos e outros ésteres presentes no leite (HOSSEINI et al., 2021; TOLEDO-ALVARADO et al., 2021). Quanto à associação desses números de onda ao CCS, alguns estudos mostram que regiões específicas do espectro FTIR, especialmente a chamada "região de impressão digital" (aproximadamente 925–1582 cm^{-1} , onde estão 1142, 1130 e 1018 cm^{-1}), apresentam forte correlação com o CCS, pois refletem mudanças em componentes como lactose, proteínas e outros constituintes químicos do leite afetados por mastite ou inflamação da glândula mamária (DU et al., 2023). O CCS elevado está associado a alterações na composição do leite, como redução de lactose e caseína, e aumento de certos ácidos e proteínas do soro, o que pode modificar as bandas do FTIR nessas regiões (PEGOLO et al., 2021).

A partir dos números de onda selecionados, foi realizado a análise random forest para a realização da previsão. O modelo de regressão por Random Forest, treinado com os cinco

números de onda selecionados, apresentou um coeficiente de determinação de $R^2 = 0,246$ e um erro quadrático médio raiz de 728.477 células/mL, mostrando um alto grau de dispersão como mostrado na Figura 1. Esses resultados indicam que o modelo possui capacidade limitada de predição, explicando apenas cerca de 24,6% da variabilidade observada na CCS no leite, sugerindo que, embora as bandas espectrais selecionadas contenham informação relevante, elas não capturam toda a complexidade química associada à variação de CCS (PEGOLO et al., 2021). Esse valor relativamente baixo é consistente com a natureza intrínseca da análise de CCS por FTIR: a concentração celular não possui assinatura espectral direta, mas correlaciona-se indiretamente com alterações no perfil bioquímico do leite, como proteínas, lipídios, eletrólitos, causadas pela inflamação mamária (DU, 2024; JULMOHAMMAD et al., 2024).

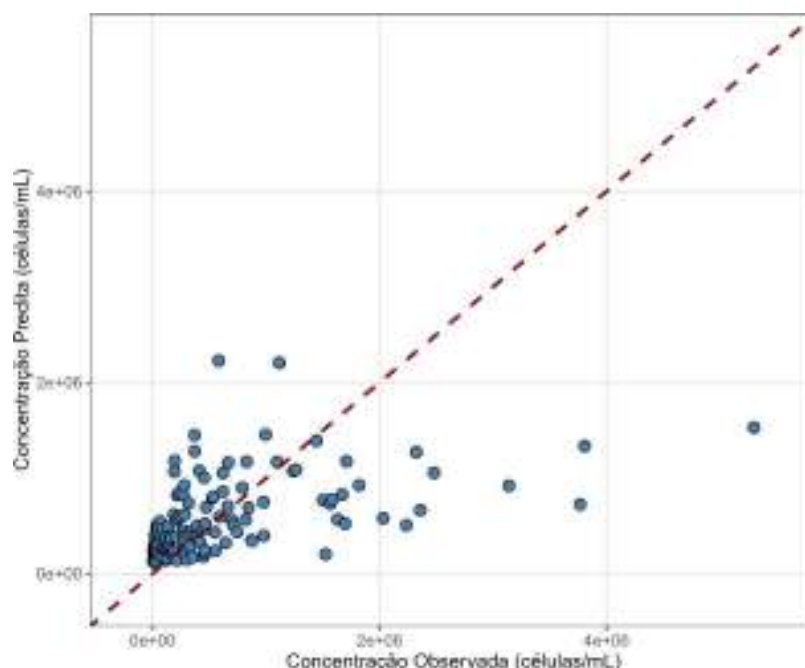


Figura 1. Correlação entre os valores observados e preditos de células somáticas (CCS) por Random Forest.

O RMSE de 728.477 células/mL reflete o erro médio absoluto entre os valores preditos e observados, indicando que, em média, o modelo erra em cerca de 728 mil células/mL, um valor considerável, especialmente quando comparado ao intervalo típico de CCS em leite (de ~50.000 a >2.000.000 células/mL). Este valor é mais alto do que o limite estabelecido pela legislação brasileira, de 500000 células/mL (MAPA, 2018), o que torna inviável para quantificações. Além disso, Os parâmetros de desempenho analítico, $CC\alpha$ (1.200.671 células/mL), $CC\beta$ (2.401.343 células/mL), LOD (10.097.192 células/mL) e LOQ (30.597.551 células/mL), revelam que o modelo não é sensível o suficiente para detecção ou quantificação confiável de CCS em níveis típicos de produção leiteira. Valores de LOD e LOQ superiores a 10^7 e 3×10^7 células/mL, respectivamente, estão muito acima do limite máximo permitido pela regulamentação brasileira.

4 CONCLUSÃO

O presente estudo demonstrou o potencial, mas também as limitações, do uso da espectroscopia FTIR associada ao algoritmo Random Forest para a predição da CCS em leite bovino. Embora tenha sido possível identificar bandas espectrais relacionadas a alterações

químicas características da mastite, o modelo apresentou desempenho restrito, indicando que a relação entre o perfil espectral e a CCS é complexa e não linear. A redução extrema de variáveis, apesar de favorecer a interpretação dos resultados, comprometeu a capacidade preditiva do modelo. Dessa forma, abordagens futuras devem considerar o uso de todo o espectro FTIR, técnicas de pré-processamento mais sofisticadas e algoritmos híbridos ou integrados a dados complementares, como informações químicas e clínicas, a fim de aprimorar a acurácia e a aplicabilidade do método no monitoramento da qualidade do leite.

REFERÊNCIAS

- BREIMAN, Leo. Random Forests. **Machine Learning**, v. 45, n. 1, p. 5–32, 2001.
- CHEN, Yanyan; FURMANN, Agnieszka; MASTALERZ, Maria; SCHIMMELMANN, Arndt. Quantitative analysis of shales by KBr-FTIR and micro-FTIR. **Fuel**, v. 116, p. 538–549, 2014.
- CHENG, Wei Nee; HAN, Sung Gu. Bovine mastitis: risk factors, therapeutic strategies, and alternative treatments — A review. **Asian-Australasian Journal of Animal Sciences**, v. 33, n. 11, p. 1699–1713, 2020.
- DU, Chao et al. Assessing the relationship between somatic cell count and the milk mid-infrared spectrum in Chinese Holstein cows. **Veterinary Record**, v. 193, n. 11, 2023.
- DU, Lijuan. New insights into raw milk adulterated with milk powder identification: ATR-FTIR spectroscopic fingerprints combined with machine learning and feature selection approaches. **Journal of Food Composition and Analysis**, v. 133, p. 106443, 2024.
- HOSSEINI, Elahesadat; GHASEMI, Jahan B.; DARAIEI, Bahram; ASADI, Gholamhassan; ADIB, Nooshin. Application of genetic algorithm and multivariate methods for the detection and measurement of milk-surfactant adulteration by attenuated total reflection and near-infrared spectroscopy. **Journal of the Science of Food and Agriculture**, v. 101, n. 7, p. 2696–2703, 2021.
- JAFARZADEH, Hamid; MAHDIANPARI, Masoud; GILL, Eric; MOHAMMADIMANESH, Fariba; HOMAYOUNI, Saeid. Bagging and boosting ensemble classifiers for classification of multispectral, hyperspectral and polsar data: a comparative evaluation. **Remote Sensing**, v. 13, n. 21, p. 4405, 2021.
- JULMOHAMMAD, Norliza Binti; OSLAN, Siti Nur Hazwani; MD SUHAINI, Iffah Kamilah; TAN, Emeline; AMERAM, Nadiyah; SOLOI, Sabrina; MOHD YAAKUB, Nurul' Azah; JULMOHAMAD, Norziana. Detection of adulterant residues in UHT milk products using ATR-FTIR spectroscopy. **International Journal of Food**, v. 1, n. 1, p. 20–31, 2024.
- KURSA, Miron B.; RUDNICKI, Witold R. Feature selection with the boruta package. **Journal of Statistical Software**, v. 36, n. 11, 2010.
- MAPA. INSTRUÇÃO NORMATIVA Nº 77, DE 26 DE NOVEMBRO DE 2018. Disponível em: https://www.in.gov.br/materia/-/asset_publisher/Kujrw0TZC2Mb/content/id/52750141/do1-2018-11-30-instrucao-normativa-n-77-de-26-de-novembro-de-2018-52749887.
- MOURA, Emmanuella De Oliveira; RANGEL, Adriano Henrique Do Nascimento; BORBA,

Luis Henrique Fernandes; GALVÃO, José Geraldo Bezerra; DA COSTA LIMA, Guilherme Ferreira; DE LIMA, Dorgival Moraes; DOS SANTOS DIFANTE, Gelson; URBANO, Stela Antas; DE AGUIAR, Emerson Moreira. Electrical conductivity and somatic cell count in zebu cow's milk. **Semina: Ciências Agrárias**, v. 38, n. 5, p. 3231–3240, 2017.

PEGOLO, S. et al. Genetic parameters of differential somatic cell count, milk composition, and cheese-making traits measured and predicted using spectral data in Holstein cows. **Journal of Dairy Science**, v. 104, n. 10, p. 10934–10949, 2021.

STANEK, Piotr; ŻÓŁKIEWSKI, Paweł; JANUŚ, Ewa. A Review on Mastitis in Dairy Cows Research: Current Status and Future Perspectives. **Agriculture**, v. 14, n. 8, p. 1292, 2024.

TOLEDO-ALVARADO, H.; PÉREZ-CABAL, M. A.; TEMPELMAN, R. J.; ECCHINATO, A.; BITTANTE, G.; DE LOS CAMPOS, G.; VAZQUEZ, A. I. Association between days open and milk spectral data in dairy cows. **Journal of Dairy Science**, v. 104, n. 3, p. 3665–3675, 2021.

WANG, Qiuyu; HULZEBOSCH, Alex; BOVENHUIS, Henk. Genetic and environmental variation in bovine milk infrared spectra. **Journal of Dairy Science**, v. 99, n. 8, p. 6793–6803, 2016.