Uso de técnicas de Machine Learning na análise da relação da composição do leite bovino com a contagem de células somáticas

João Paulo de Avila

Ciência da Computação – Universidade de Passo Fundo (UPF) – Campus 1
BR 282, Km 292 - Bairro São José - Passo Fundo - RS, 99052-900

178898@upf.br

Abstract. *This work investigates the relationship between some milk components — protein, lactose, and fat — and the somatic cell count (SCC), one of the main indicators of the health of dairy cows. Machine learning techniques were used to perform a statistical analysis, aiming to assess the potential of these components for detecting diseases such as mastitis. The algorithms Simple and Multiple Linear Regression, Random Forest, and XGBoost were trained for predictive modeling, and their performances were evaluated based on metrics such as the coefficient of determination (R^2), mean squared error (MSE), and the relative importance of the predictor variables. The results indicate that lactose was the most influential factor in the prediction of SCC, standing out as a marker for monitoring the udder health of dairy cows. Linear models showed an R^2 performance of 0.14, and non-linear machine learning models showed an R^2 performance of 0.17, reflecting the biological complexity of the studied phenomenon.*

Resumo. *Este trabalho investiga a relação entre alguns componentes do leite — proteína, lactose e gordura — e a contagem de células somáticas (CCS), um dos principais indicadores da saúde de vacas leiteiras. Foram utilizadas técnicas de aprendizado de máquina para realizar uma análise estatística, com o objetivo de verificar o potencial desses componentes para a detecção de doenças, como a mastite. Os algoritmos Regressão Linear Simples e Múltipla, Random Forest e XGBoost foram treinados para modelagem preditiva, e suas performances foram avaliadas com base em métricas como o coeficiente de determinação R^2 e o erro quadrático médio (MSE) e a importância relativa das variáveis preditoras. Os resultados indicam que a lactose foi o fator mais influente na predição do CCS, destacando-se como um marcador para o monitoramento da saúde do úbere de vacas leiteiras. Modelos lineares apresentaram um desempenho R^2 de 0,14 e modelos não lineares de machine learning apresentaram um desempenho R^2 de 0,17. Refletindo a complexidade biológica do fenômeno estudado.*

1. Introdução

A cadeia de produção de leite enfrenta desafios constantes relacionados à manutenção da qualidade do produto, que está ligado diretamente com a saúde dos animais. Diversos fatores influenciam a composição do leite bovino, incluindo estação do ano, estágio da lactação, alimentação, manejo, fatores genéticos e saúde do rebanho. Tanto o volume quanto a composição do leite (gordura, proteína, lactose, sólidos totais e contagem de células somáticas) desempenham um papel importante, pois servem como referência para a estimativa da qualidade e para o preço pago pela matéria-prima (Dürr et al., 2004). Um dos principais desafios da cadeia produtiva do leite é assegurar não

apenas a qualidade industrial, mas também a segurança alimentar, mitigando a presença de contaminantes ou indicadores de doenças, como a mastite (Folchini, 2020).

A contagem de células somáticas (CCS) é amplamente utilizada como indicador laboratorial de infecções na glândula mamária, sendo sua elevação associada à diminuição dos teores de lactose e proteína no leite. Esse cenário reduz a qualidade do produto final e pode resultar em amostras abaixo do padrão legal (Folchini, 2020; Alessio et al., 2016). De 2020 em diante, a aplicação de modelos baseados em aprendizado de máquina tem se destacado como uma alternativa eficiente para previsão de alterações sanitárias e metabólicas do rebanho, superando os métodos estatísticos tradicionais em precisão e capacidade de detecção precoce de doenças (Stygar et al., 2023).

Diante desse contexto, o presente trabalho busca aprofundar a compreensão sobre a relação entre alguns componentes do leite, especialmente proteína, lactose e gordura e a contagem das células somáticas (CCS), reconhecida como um dos principais indicadores da saúde das vacas leiteiras. Para verificar isso, optou-se pela aplicação de técnicas de aprendizado de máquina visando analisar o potencial desses componentes para a detecção de alterações sanitárias do rebanho.

As próximas seções contextualizam os principais fatores que influenciam a composição do leite e sua relação com a saúde do úbere, com destaque para a contagem das células somáticas como indicador central. Também são abordadas aplicações de algoritmos de Machine Learning (ML), especialmente na previsão da CCS com base em dados de rotina, como gordura, proteína e lactose. A seção de materiais e métodos destaca a origem dos dados utilizados, o pré-processamento, os procedimentos adotados para análise estatísticas e a construção de modelos preditivos. Por fim, são apresentados os resultados e discussões, além de sugestões para aprofundamentos futuros que possam fortalecer o uso dessas ferramentas na gestão da qualidade do leite.

2. Revisão bibliográfica

A produção leiteira no Rio Grande do Sul (RS) representa um dos pilares do agronegócio, tanto por seu valor social quanto pelo impacto econômico na renda de milhares de produtores rurais (Folchini et al., 2020). As oscilações na composição do leite, sobretudo nos teores de proteína, lactose, gordura e sólidos não gordurosos, são reflexos de fatores intrínseco do animal, do ambiente e do manejo, tornando desafiador o entendimento contínuo aos padrões de qualidade estabelecidos pela legislação. Esse estudo também identificou que, na região, os níveis de sólidos não gordurosos foram influenciados pela CCS, juntamente com outros fatores como proteína, lactose e estações do ano.

A qualidade do leite é fortemente impactada por alterações na saúde das glândulas mamárias, principalmente a mastite, que eleva a contagem de CCS e interfere na concentração de lactose e proteína. (Folchini 2020). No inverno e outono, são registrados os maiores teores de proteína, lactose e sólidos não gordurosos, enquanto no verão e primavera ocorre a redução desses componentes, além do aumento do CCS, indicando maior desafio sanitário no período mais quente (Bondan et al., 2023).

Diversos estudos têm explorado a aplicação de algoritmos de machine learning para interpretar dados de FTIR e prever componentes específicos do leite e fazer correlações. De Vries, Bliznyuk e Pinedo (2023) destacam que a Inteligência Artificial (IA), e particularmente o ML, oferece oportunidades significativas para analisar dados complexos e heterogêneos gerados em fazendas leiteiras, incluindo dados de composição do leite e espectroscopia. Ribeiro et al. (2023), por exemplo, demonstraram a eficácia da combinação de FTIR com Redes Neurais Convolucionais (CNN) para quantificar lactose residual, glicose e galactose em leite com baixo teor de lactose, alcançando alta acurácia ($R^2 > 80\%$ para quantificação).

A literatura aponta que modelos baseados em técnicas como Random Forest e Redes Neurais Artificiais (RNA) também apresentam boa acurácia na predição de diversos indicadores de qualidade do leite. No estudo de Sunithaman et al. (2024), a integração de sensores e ML permitiu predições da qualidade do leite, demonstrando o potencial dessas tecnologias para rotinas de laboratório e a indústria.

Métodos supervisionados e não supervisionados de ML também já foram aplicados para prever teores de componentes como lactoferrina e proteínas, como demonstrado no estudo de Soyeurt et al. (2020). Neste trabalho, os autores compararam quatro algoritmos de ML para a predição do teor de lactoferrina em leite bovino. Foram avaliados: regressão por mínimos quadrados parciais (PLS, *Partial Least Squares*), *Support Vector Regression* (SVR) linear e polinomial combinados com fatores PLS e RNA. Na comparação destes algoritmos, concluiu-se que a combinação de PLS com RNA apresentou melhor desempenho do teor de lactoferrina.

Aplicações de MKL não se restringem apenas à espectroscopia no contexto da análise do leite. Sunithamani et al. (2024) exploraram a predição da qualidade do leite e classificaram como baixa, média ou alta, utilizando sensores de baixo custo (pH, temperatura e turbidez) integrados a um arduino. Utilizaram dentre os algoritmos de ML Random Forest, que atingiu acurácia próxima a 100%, sugerindo o potencial de sistemas acessíveis baseados em ML.

A aplicação direta de ML para prever CCS ou risco de mastite com base na composição do leite (ou dados espectrais) é uma área que merece maior investigação, especialmente no Brasil e no RS. De acordo com Soyeurt et. al (2020), estudos internacionais indicam a viabilidade, mas a validação e a adaptação desses modelos para as condições locais são fundamentais. A combinação de dados de composição/espectrais com outros dados de sensores (atividade, ruminação) e de manejo, como sugerido por De Vries, Bliznyuk e Pinedo (2023), pode levar a sistemas de monitoramento da saúde animal ainda mais poderosos e preditivos.

2.1. Análise do leite bovino

A análise do leite bovino é essencial para garantir a qualidade, segurança alimentar e a conformidade com a legislação vigente. No Brasil, a obrigatoriedade dessas análises está estabelecida por diversas normativas. O Decreto nº 9.013, de 29 de março de 2017, conhecido como Regulamento de Inspeção Industrial e Sanitária de Produtos de Origem Animal (RIISPOA), determina que a inspeção de leite e derivados abrange desde a sanidade do rebanho até a expedição do produto final (Brasil, 2024).

No norte do RS, segundo o estudo conduzido por Folchini (2020), as amostras de leite são coletadas diretamente de tanques de expansão e imersão em propriedades

rurais que são encaminhadas para o Serviço de Análises de Rebanhos Leiteiros (SARLE) da Universidade de Passo Fundo, laboratório credenciado ao Ministério da Agricultura. A autora descreve que a composição do leite é analisada utilizando a tecnologia de FTIR, método que permite a rápida identificação e quantificação dos constituintes químicos do leite. Já a CCS, é realizada com ciclometria de fluxo, técnica reconhecida sobre sua precisão e confiabilidade para grandes volumes de amostras, como detalha Folchini (2020).

De acordo com Barbano e Clark (1989), as análises físico-químicas modernas, como FTIR são amplamente utilizadas para quantificar componentes do leite, incluindo gordura, proteína, lactose e sólidos totais, permitindo avaliações rápidas e precisas em larga escala. Além de rápido, apresenta boa exatidão e precisão, permitindo a quantificação de componentes como gordura, proteína e lactose no leite cru, com base na absorção de radiação infravermelha em comprimentos de onda específicos (Embrapa, 2014). As moléculas de gordura, proteína e lactose absorvem radiação infravermelho em comprimentos de onda distintos, de acordo com grupos funcionais presentes nas moléculas. Baseado nisso, o software do equipamento calcula a concentração dos componentes usando modelos de calibração multivariadas.

A citometria de fluxo é uma técnica utilizada onde um reagente corante se liga às células e permite a quantificação por fluorescência, possibilitando o cálculo da concentração por um software específico. Este método instrumental, possibilita a análise de um volume conhecido de leite e gera resultados rápidos e precisos, sendo atualmente uma das metodologias de referência para a determinação da contagem de células somáticas (CCS) em laboratórios credenciados (Embrapa, 2014). A citometria de fluxo, junto do FTIR foram as formas de coleta de dados utilizadas nos dados aqui utilizados e apresentados por Folchini (2020), os quais serviram de fonte de dados para este estudo.

2.2 Uso de Machine Learning para análise de amostras de leite

O avanço das tecnologias de IA e ML está revolucionando diversas áreas do agronegócio, incluindo a pecuária leiteira. Ferramentas de desta natureza permitem não apenas automatizar tarefas repetitivas e demoradas, mas gerar modelos preditivos e prescritivos a partir de grandes volumes de dados, potencializando a tomada de decisão do mercado e de diferentes etapas do processo produtivo.

Segundo De Vries, Bliznyuk e Pinedo (2023), métodos de ML já viabilizam aplicações de análise de mensuração da condição corporal, monitoramento de ingestão alimentar, além de prever eventos futuros, como fertilidade, risco de doenças, integrando dados heterogêneos como informações de saúde, genética, comportamento e análise do leite.

Além dessas aplicações, o uso de ML associado a espectros obtidos por FTIR, já demonstrou alto potencial na análise composicional do leite. Em estudo realizado por Ribeiro et al. (2023), a associação da FTIR com CNN permitiu identificar e quantificar os açúcares residuais em leite com elevada precisão.

Mota et al. (2021) comparou diversos métodos tradicionais com técnicas de ML (Random Forest, Gradient Boosting machine, Elastic Net) para predição de fenótipos complexos em vacas leiteiras de raça Holandesa. O estudo concluiu que as abordagens de machine learning superaram o PLS na acurácia da predição de escore corporal,

β-hidroxibutirato no sangue e fração de κ-caseína utilizando espectros FTIR do leite como fonte de dados.

Esses resultados são importantes porque demonstram que métodos baseados em árvore de decisão e redes profundas podem capturar relações não lineares e complexas entre espectros de leite e características biológicas, algo que os métodos lineares convencionais não conseguem com a mesma eficiência. Além disso, os algoritmos de machine learning são capazes de realizar seleção de variáveis relevantes, reduzindo a dimensionalidade e aumentando a robustez preditiva, mesmo diante de conjuntos de dados com muitos ruídos (Mota et al., 2021).

O uso de ML para prever a saúde do úbere e a CCS com base em dados de rotina de composição do leite, como proteína, gordura e lactose, já demonstra resultados promissores para aplicação prática em fazendas leiteiras. Bobbo et al. (2021) analisaram diferentes, entre eles RNA e Random Forest para prever o status de saúde do úbere na ordenha subsequente, classificando como saudável ou mastítico (úbere com mastite).

Outro ponto de destaque do estudo foi o uso de validação cruzada estratificada e validação externa, demonstrando que os modelos mantiveram desempenho estável em diferentes conjuntos de dados. Apesar de pequenas diferenças entre os algoritmos, todos apresentaram acurácia acima de 75%, o que mostra o potencial de sua implementação em sistemas de apoio à decisão na rotina da fazenda. Os métodos de RNA, Random Forest tiveram o melhor desempenho na previsão das classes de saúde do úbere em um determinado dia de teste (saudável ou mastítico de acordo com a CCS abaixo ou acima de um limite predefinido de 200.000 células/mL) com base nas características do leite de vaca registradas no dia de teste anterior" (Bobbo et al., 2021).

Portanto, a aplicação de técnicas de ML na análise dos componentes do leite, tanto para previsão quanto para criar correlações, representam uma ferramenta poderosa e viável para auxiliar o monitoramento da saúde do rebanho e consequentemente melhorar a qualidade do leite.

3. Materiais e métodos

Todas as análises de dados e modelagens deste trabalho realizadas utilizando a linguagem de programação Python (versão 3) no Google Colab, com o apoio das seguintes bibliotecas: Pandas para manipulação de dados, NumPy para operações numéricas, Matplotlib e Seaborn para visualização e análise de dados, Statsmodels para modelos de regressão linear, Scikit-learn para o modelo Random Forest e divisão treino-teste, XGBoost para o modelo Gradient Boosting e SHAP para interpretabilidade de modelos de ML.

3.1. Pré processamento dos dados

Os dados utilizados neste estudo foram originalmente apresentados por Folschini et al. (2020), consistindo em análises de amostras de leite provenientes de rebanhos leiteiros do estado do RS. O dataset continha informações sobre a composição do leite, incluindo percentuais de gordura, proteína, lactose, CCS por mililitro, e informações sazonais como estação do ano da coleta.

Contagens originais de CCS igual ou superior a 1.000 células/mL (indicando limite superior a medição do equipamento) tiveram seu valor fixado para 1.000.000

células/mL. As demais amostras tiveram seu valor original multiplicado por 1.000 para representar a contagem por mililitro.

Linhas que continham valores ausentes, NaN, nas colunas consideradas essenciais para análise (gordura, proteína, lactose e CCS), foram removidas do conjunto de dados para garantir a integridade da análise, dado que as amostras são independentes.

A partir do dataset pré-processado, foi realizada uma análise exploratória dos dados, para compreender as características e distribuições das variáveis estudadas. Foram calculadas médias e medianas, desvio padrão, mínimo e máximo para as variáveis gordura, proteína, lactose e CCS, tanto no conjunto de dados completo quanto segmentado por estação do ano (verão, outono, inverno e primavera).

Histogramas e gráficos de densidade foram gerados para visualizar a forma da distribuição de cada variável numérica. Boxplots foram utilizados para comparar as distribuições, medianas e a presença de valores atípicos entre as estações do ano.

A relação entre as variáveis de composição do leite e a CCS foi investigada por análise de correlação. Foram calculados coeficientes de correlação de Pearson e de Spearman. O coeficiente de Pearson avalia a relação linear entre duas variáveis, sendo sensível a valores extremos (*outliers*) (Khan Academy, 2025). Já o coeficiente de Spearman avalia relações monotônicas e é menos influenciado por valores atípicos (Scipy, 2025).

3.2. Modelagem e Machine Learning

Para aprofundar a compreensão das relações, algumas abordagens de modelagem foram aplicadas. Foi ajustado um modelo de regressão linear simples para quantificar a relação entre lactose e CCS, seguida de regressão linear múltipla, incluindo as variáveis de proteína, lactose e gordura e variáveis dummy para estação.

Modelos de ML foram empregados, como Random Forest e XGBoost, incorporando as mesmas variáveis explicativas. Segundo a IBM (2023), o XGBoost destaca-se por sua eficiência computacional, habilidade em lidar com dados ausentes e suporte a regularização, sendo considerado um dos algoritmos mais eficazes da atualidade para tarefas de regressão e classificação e a interpretabilidade do modelo. XGBoost foi aprimorada por meio da técnica SHAP (*Shapley Additive Explanations*), que, de acordo com o DataCamp (2023), fornecem valores para interpretações localmente consistentes e confiáveis, permitindo entender não apenas quais variáveis são importantes, mas como e quanto elas influenciam o resultado final de cada previsão. Essa abordagem é essencial para evitar que modelos mais complexos se tornem “caixas pretas”. O desempenho dos modelos foram avaliados utilizando métricas como o coeficiente de determinação (R^2) e o erro quadrático médio (MSE), além das análises dos coeficientes *p-values* e importância relativa das variáveis preditoras.

4. Resultados e Discussão

Nesta seção, apresentam-se os resultados obtidos a partir da análise exploratória dos dados, com objetivo de compreender a distribuição estatística das principais variáveis analisadas. Na Tabela 1 apresenta-se as estatísticas descritivas das variáveis

selecionadas para o conjunto de dados, sem distinção por estação de ano ou período. O dataset utilizado é composto por 184.933 amostras válidas após pré-processamento.

Tabela 1. Variáveis consideradas nas estatísticas descritivas gerais.

Variável	Média	Desvio Padrão	Mínimo	Mediana	Máximo
Gordura (%)	3,87	0,42	2,60	3,84	5,14
Proteína (%)	3,23	0,20	2,63	3,23	3,82
Lactose (%)	4,37	0,14	3,94	4,38	4,80
CCS (células/mL)	575.718	284.214	61.000	543.000	1.000.000

Observa-se que as variáveis de gordura, proteína e lactose apresentam distribuição relativamente concentrada em torno da média, com baixos desvios padrão. A lactose, em especial, apresenta maior dispersão (±0,14), variando entre 3,94% e 4,80%. A variável CCS apresentou ampla variação, com valores entre 61.000 e o limite de 1.000.000 de células/mL, valor limite que reflete a capacidade máxima de detecção do equipamento utilizado. O alto desvio padrão (±284.214) evidencia a presença de casos tanto de vacas saudáveis, quanto com inflamação da glândula mamária.

A matriz de correlação de Pearson (Figura 1) destaca a relação negativa moderada entre lactose e CCS (r = -0,37). Tais resultados corroboram a evidência apresentada por Bondan et al. (2023), que destacam a correlação negativa entre os níveis de escore de células somáticas (SCS) e os teores de lactose no leite, evidenciando que a elevação da CCS tende a reduzir a síntese de lactose, em decorrência de alterações inflamatórias na glândula mamária.

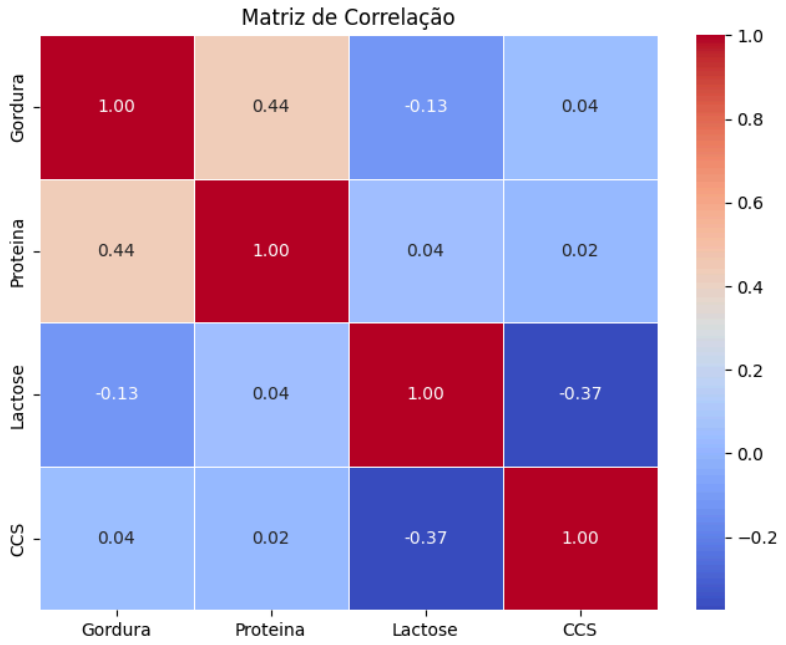

Figura 1. Matriz de correlação de Pearson.

Ao aplicar regressão linear simples realizada, neste caso CCS = $\beta_0 + \beta_1 *$ Lactose + ε) entre o teor de lactose e a CCS demonstrou uma relação negativa estatisticamente significativa entre as variáveis. O coeficiente da lactose (β_1) indica a variação esperada na CCS para cada unidade de aumento da lactose. Como a unidade da lactose é em porcentagem (%), um aumento de 1.0 (por exemplo, de 4.4% para 5.4%) representa um aumento de 1 ponto percentual. Logo, pode-se estimar que, para cada aumento de 1% no teor de lactose, espera-se uma redução média de aproximadamente 758.800 células/mL na CCS. Porém, essa é uma medida estatística linear, e essa relação pode não ser linear em todos os casos.

O coeficiente de determinação ($R^2 = 0,14$) indica que o modelo explica 14% da variabilidade da CCS com base apenas na lactose, o que é considerado uma força de associação fraca a moderada, mas ainda assim significativa em contextos biológicos complexos, nos quais múltiplos fatores interferem na CCS, não deixando de reforçar a lactose como preditor primário da CCS.

A regressão linear múltipla foi aplicada para avaliar as variáveis da composição do leite (neste caso a gordura, proteína e lactose) e das estações do ano sobre a contagem das células somáticas (CCS) (Tabela 2). O modelo teve como resultado $R^2 = 0,151$, indicando que 15% da variabilidade da CCS pode ser explicada por essas variáveis. Embora seja baixo, é compatível com a complexidade do CCS.

A lactose foi a variável que mais influenciou o resultado. O modelo indicou que, quando o teor de lactose no leite aumenta 1%, a CCS tende a cair em cerca de 801 mil células por mililitro. Esse efeito foi considerado estatisticamente confiável, pois o valor de p foi menor que 0,001, o que significa que a chance de esse resultado ter acontecido por acaso é muito pequena. Além disso, o valor de erro padrão foi baixo (9.208), o que mostra que a estimativa do modelo é precisa. A estatística t, usada para medir a força desse efeito, foi de -87 — um número muito alto, reforçando que a relação é forte. O intervalo de confiança de 95% ficou entre -820 mil e -784 mil, indicando que mesmo nos limites, o efeito da lactose continua sendo negativo e relevante. Também foi possível perceber um padrão relacionado à estação do ano: a CCS foi, em média, mais baixa no outono e mais alta na primavera, comparadas ao inverno, que serviu como base para comparação.

Tabela 2. Resultados da Regressão Linear Múltipla.

Variável	Coef,	Erro Padrão	t	p-valor	IC 95%
Gordura (%)	-12.800	2.365	-5,412	< 0,001	[-17.400 ; -8.162]
Proteína (%)	90.940	4.677	19,444	< 0,001	[81.800 ; 100.000]
Lactose (%)	-801.600	9.208	-87,060	< 0,001	[-820.000 ; -784.000]
Estação: Outono	-42.770	1.772	-24,139	< 0,001	[-46.200 ; -39.300]
Estação: Primavera	28.090	2.280	12,320	< 0,001	[23.600 ; 32.600]
Estação: Verão	-13.370	1.740	-7,684	< 0,001	[-16.800 ; -9.960]

Também foi testado um modelo de aprendizado de máquina do tipo Random Forest Regressor, que é um algoritmo baseado em árvore de decisão, conhecido por capturar relações não lineares entre as variáveis. O modelo foi treinado com 70% dos dados totais e testado com os outros 30% restantes. O resultado foi um R^2 de 0,175, o que indica uma leve melhora em relação ao modelo de regressão linear múltipla (que teve R^2 = 0,151), mostrando que o Random Forest conseguiu explicar certa de 17,5% da variação do CCS.

Outro recurso explorado do Random Forest foi a possibilidade de medir a importância relativa de cada variável para o processo de predição. Nessa análise da importância das variáveis mostrou que a lactose foi a variável mais relevante para o modelo, respondendo por 72% da importância total na predição do CCS. Em seguida a gordura com 11% e proteína com 9,9%. As variáveis sazonais, tiveram pesos menores (inferiores a 2% cada), o que sugere que a composição do leite tem mais peso na predição da CCS, ao menos no escopo deste estudo.

Para complementar as análises, foi empregado o algoritmo XGBoost Regressor (Extreme Gradient Boosting), uma técnica baseada em árvores de decisão que combina diversos modelos em um modelo mais robusto. Após o treinamento do modelo com 70% da base de dados para treino e 30% para validação, foi alcançado R^2 de 0,175. Indicando 17,5% da variação na CCS sendo explicada pelas variáveis utilizadas no modelo (proteína, lactose e gordura).

Para interpretar como o XGBoost toma suas decisões, foi aplicada a técnica SHAP, a qual utiliza conceitos da teoria dos jogos para interpretar os modelos, atribuindo a cada variável um valor que representa sua contribuição individual para a predição. O SHAP calcula como a presença (ou ausência) de uma variável altera a previsão, considerando todas as combinações possíveis de atributos (Lundberg e Lee, 2017).

Os valores SHAP médios absolutos revelaram que a variável lactose foi a mais influente na predição da CCS, com um impacto de +91.003. Em seguida vieram gordura com 22.377 e proteína com 19.811. Isso significa que, em média, a lactose contribuiu para alterar a previsão da CCS em cerca de 91 mil células/mL para mais ou para menos, dependendo da amostra, conforme indicado na Figura 2.

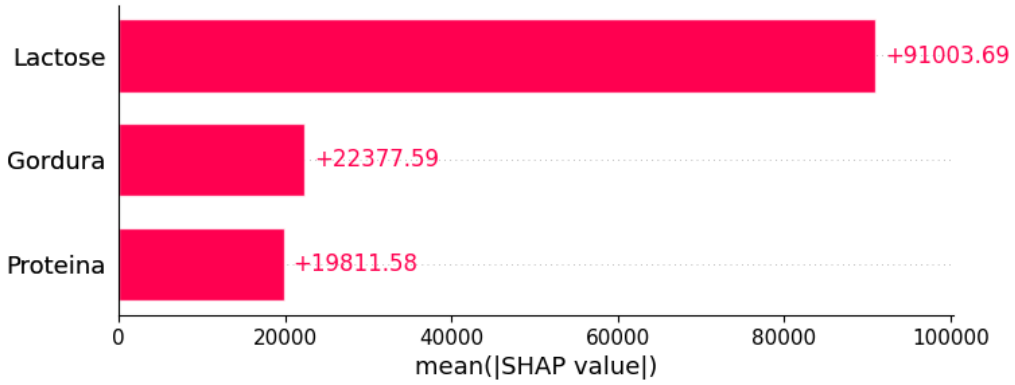

Figura 2. Importância média das variáveis segundo os valores SHAP absolutos. O gráfico apresenta a média dos valores absolutos de SHAP para cada

variável do modelo XGBoost, indicando sua contribuição global para as previsões. Quanto maior o valor, maior a influência da variável na saída do modelo.

Adicionalmente à interpretação do modelo XGBoost, foi gerado um gráfico do tipo *beeswarm* conforme a Figura 3, que mostra a influência individual de cada variável em cada previsão realizada. Nesse gráfico, cada ponto representa uma amostra do dataset utilizado e a posição no eixo horizontal indica se a variável aumentou ou reduziu o valor previsto da CCS.

A variável lactose apresentou a distribuição mais ampla entre todas, com valores SHAP variando aproximadamente de -300.000 a +300.000. Observa-se que valores altos de lactose (em rosa) estão fortemente associados à redução das previsões de CCS, enquanto valores baixos (em azul) puxam as previsões de CCS para cima. As variáveis gordura e proteína apresentaram impacto mais modesto, com variações em ambas as direções e sem um padrão tão marcante.

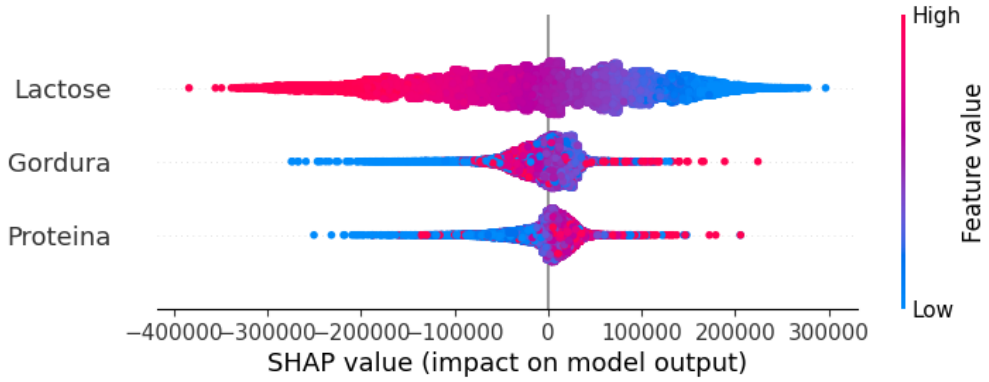

Figura 3. Gráfico beeswarm com valores SHAP para o modelo XGBoost. O eixo horizontal mostra o impacto de cada variável (Lactose, Gordura, Proteína) na saída do modelo. Cada ponto representa uma amostra do conjunto de dados; a cor indica o valor da variável (de baixo [azul] para alto [vermelho]). Variáveis mais importantes têm dispersões maiores de SHAP.

Os resultados obtidos dão uma boa visão da relação entre alguns componentes do leite. A análise exploratória confirma uma ampla variação de CCS no conjunto de dados, com valores baixos até o limite da detecção (1.000.000 células/mL), refletindo a heterogeneidade das características do leite da região.

Nesta primeira constatação já foi demonstrado a correlação negativa da lactose e a CCS, se alinhando perfeitamente com Folchini (2020). A mastite, ao causar danos às células epiteliais secretoras e aumentar a permeabilidade vascular, compromete a criação de lactose, resultando em sua diminuição no leite. Essa correlação negativa em todas as estações do ano, reforça sua importância para o monitoramento da saúde do úbere, corroborando os estudos de Bondan et al. (2023).

A relação da gordura e da proteína com a CCS mostraram-se mais difíceis de serem interpretadas. A correlação foi fraca, mas os resultados obtidos com a regressão linear múltipla mostraram que a proteína teve uma associação positiva e a gordura uma associação negativa. A relação da gordura neste caso pode ser mais investigada, mas também aponta uma complexidade do organismo da vaca durante a mastite.

A influência das estações de ano foi confirmada, com CCS mais alta na primavera e mais baixa no outono/inverno, o que corrobora com as observações de Bondan et al. (2023) sobre os desafios sanitários nos períodos mais quentes e nas variações na composição do leite ao longo do ano no RS. Contudo, nos modelos de ML aqui estudados (Random Forest e XGBoost), a importância das estações é menor, se comparado com a própria composição do leite.

Os modelos aplicados apresentaram um desempenho superior à regressão linear múltipla na explicação da CCS. Embora o R^2 não tenha sido alto, também mostra o nível de complexidade biológica da CCS, que pode ser influenciada por outros fatores além da própria composição do leite. O desempenho alcançado foi similar ao encontrado em outros estudos que aplicaram ML para prever indicadores de saúde em vacas leiteiras, como o trabalho de Bobbo et al. (2021), que obteve acurácia acima de 75% para predição de status de mastite.

A análise da importância das variáveis, tanto no Random Forest quando no XGBoost (utilizando SHAP), destacou a lactose como fator mais determinante dentro do dataset estudado para predição da CCS. O gráfico da Figura 3 demonstrou claramente como níveis altos de lactose diminuem a previsão da CCS e níveis baixos aumentam. Isso reforça o potencial para usar a lactose como entrada para modelos, ganhando ainda mais poder quando combinada com outros componentes do leite, como mencionado por De Vries, Bliznyuk e Pinedo (2023).

5. Conclusão e Trabalhos Futuros

Com base na análise dos dados de composição do leite e CCS de rebanhos de vacas leiteiras do Rio Grande do Sul, este estudo conclui que existe uma relação significativa e consistente entre o teor de lactose no leite e a CCS. A lactose mostrando ser o componente mais fortemente correlacionado com a CCS ($r = -0{,}37$) e o preditor mais importante nos modelos de ML estudados (Random Forest e XGBoost), respondendo pela maior parte de toda a capacidade preditiva estudada.

Gordura e proteína têm relações com a CCS mais complexas, demandando um estudo mais aprofundado para entender a associação dessas duas variáveis com outros componentes do leite.

As estações do ano influenciam os níveis médios de CCS e componentes do leite em geral, mas têm menor importância preditiva, pelo menos, no dataset estudado, quando comparado às próprias variáveis de composição do leite aplicadas aos algorítmos testados.

Modelos de Machine Learning apresentaram um desempenho superior à regressão linear simples e múltipla para explicar a variabilidade da CCS, indicando que capturam algumas relações não lineares. No entanto, a capacidade explicativa geral se mostra moderada, refletindo que a CCS possui uma complexidade biológica alta.

Em síntese, a análise da composição do leite, especialmente do teor da lactose, utilizando ferramentas de machine learning, demonstra ser uma abordagem viável, informativa e bastante auxiliar no monitoramento da saúde de rebanhos leiteiros.

Considerando estes resultados e as próprias limitações do estudo e do conjunto de dados, pode-se sugerir algumas linhas de investigação a serem desenvolvidas no futuro.

Para aprimorar ainda mais o entendimento e a explicabilidade dos resultados, poderia se considerar a realização de testes com métodos como PCA ou agrupamento por KNN, agrupando proteína e gordura em uma única variável, reduzindo a dimensionalidade e possibilitando uma melhor interpretação. Isso poderia aumentar a robustez e a eficiência das previsões.

O presente trabalho utilizou amostras provenientes de tanques de expansão, o que limita a análise da variabilidade individual entre as vacas. A coleta de amostras diretamente de cada animal no momento da ordenha permitiria um monitoramento mais detalhado e preciso. Com isso, seria possível construir um histórico individual da CCS para cada vaca, possibilitando acompanhar alterações ao longo do tempo, como, por exemplo, durante o pós-parto e nas diferentes fases da lactação. Essa abordagem traria vantagens importantes, como o aumento da quantidade e da qualidade dos dados, permitindo não apenas uma análise mais aprofundada das correlações entre a CCS e os componentes do leite (gordura, proteína, lactose etc.), mas também a identificação precoce de casos de mastite subclínica. Além disso, o acompanhamento individual viabiliza um controle mais eficiente do rebanho, auxiliando na tomada de decisões práticas dentro da rotina da fazenda.

Com a coleta individual, também seria possível investigar se há uma relação direta entre o aumento da CCS e a redução da produção de leite de cada vaca. Essa análise traria uma visão mais clara sobre os impactos econômicos da mastite no rebanho, fornecendo dados objetivos que podem fundamentar ações de manejo, descarte seletivo ou intervenções veterinárias.

Por fim, outra possibilidade de expansão seria integrar à análise dados relacionados à alimentação dos animais. O registro detalhado da dieta — incluindo informações sobre a qualidade, quantidade de energia, teor de fibra e presença de minerais — permitiria investigar como diferentes combinações nutricionais influenciam na CCS e nos componentes do leite. Essa linha de pesquisa pode auxiliar produtores e nutricionistas a tomarem decisões mais embasadas sobre a formulação das dietas, promovendo a saúde do úbere e a produtividade do rebanho.

6. Referências Bibliográficas

BARBANO, D. M.; CLARK, J. L. Use of infrared milk analysis for rapid determination of milk composition. Journal of Dairy Science, v. 72, n. 2, p. 321–330, 1989. Disponível em: https://www.journalofdairyscience.org/article/S0022-0302%2889%2979275-4/pdf

BOBBO, T. et al. Comparison of machine learning methods to predict udder health status based on somatic cell counts in dairy cows. Scientific Reports, v. 11, p. 13642, 2021. https://doi.org/10.1038/s41598-021-93056-4.

BONDAN, C. et al. Artificial intelligence-based method using infrared spectral data for predicting the somatic cell score of dairy cows. Journal of Dairy Science, v. 106, n. 5, p. 3391–3404, 2023. Disponível em: https://doi.org/10.3168/jds.2022-22617.

BRASIL. Ministério da Agricultura, Pecuária e Abastecimento. Instrução Normativa nº 77, de 26 de novembro de 2018. Disponível em: https://evomilk.com.br/tudo-sobre-in-77/.

BRASIL. Ministério da Agricultura e Pecuária. Qualidade do leite – PNQL. Disponível em: https://www.gov.br/agricultura/pt-br/assuntos/inspecao/produtos-animal/qualidade-do-leite-pnql. Acesso em: 29 maio 2025.

DATACAMP. An Introduction to SHAP Values for Machine Learning Interpretability. 2023. Disponível em: https://www.datacamp.com/tutorial/introduction-to-shap-values-machine-learning-interpretability. Acesso em: 04 jun. 2025.

DE VRIES, A.; BLIZNYUK, N.; PINEDO, P. Invited review: Examples and opportunities for artificial intelligence (AI) in dairy farms. Applied Animal Science, v. 39, p. 14–22, 2023. https://doi.org/10.15232/aas.2022-02345.

DÜRR, J. W. et al. Milk recording as an indispensable procedure to assure milk quality. Revista Brasileira de Zootecnia, v. 40, p. 76–81, 2011. Disponível em: https://www.alice.cnptia.embrapa.br/alice/bitstream/doc/921024/1/Artigo-meta-2011-RBZ-Claudio-66262.pdf. Acesso em: 04 jun. 2025.

EMBRAPA. Qualidade físico-química, higiênico-sanitária e composicional do leite cru: indicadores e aplicações práticas da Instrução Normativa 62. Porto Velho, RO: Embrapa Rondônia, 2014. (Documentos / Embrapa Rondônia, 158). Disponível em https://www.infoteca.cnptia.embrapa.br/infoteca/bitstream/doc/1018827/1/Doc158leite.pdf. Acesso em: 04 jun. 2025.

FOLCHINI, Jéssica Aneris. Estudo retrospectivo dos sólidos não gordurosos em amostras de leite cru no estado do Rio Grande do Sul. 2020. Dissertação (Mestrado em Agronomia) – Universidade de Passo Fundo.

KHAN ACADEMY. Pearson correlation coefficient intuition. Disponível em: https://pt.khanacademy.org/math/statistics-probability/describing-relationships-quantitative-data/correlation/v/pearson-correlation-coefficient-intuition. Acesso em: 3 jun. 2025.

LUNDBERG, Scott M.; LEE, Su-In. A unified approach to interpreting model predictions. In: ADVANCES IN NEURAL INFORMATION PROCESSING SYSTEMS (NeurIPS), 30., 2017, Long Beach. Proceedings [...]. [S.l.]: Curran Associates, 2017. Disponível em: https://arxiv.org/abs/1705.07874. Acesso em: 14 jun. 2025.

MOTA, L. F. M. et al. Evaluating the performance of machine learning methods and variable selection methods for predicting difficult-to-measure traits in Holstein dairy cattle using milk infrared spectral data. Journal of Dairy Science, v. 104, p. 8107–8121, 2021. https://doi.org/10.3168/jds.2020-19861.

REHAGRO. Análise físico-química do leite: importância para qualidade do leite. Disponível em: https://rehagro.com.br/blog/analise-fisico-quimica-do-leite/. Acesso em: 04 jun 2025.

RIBEIRO, D. C. S. Z. et al. Determination of the lactose content in low-lactose milk using Fourier-transform infrared spectroscopy (FTIR) and convolutional neural network. Heliyon, v. 9, p. e12898, 2023. https://doi.org/10.1016/j.heliyon.2023.e12898.

SCIPY. scipy.stats.spearmanr — Spearman rank-order correlation coefficient. Disponível em: https://docs.scipy.org/doc/scipy/reference/generated/scipy.stats.spearmanr.html. Acesso em: 3 jun. 2025.

SOYEURT, H. et al. Estimation of fatty acid content in bovine milk using mid-infrared spectrometry. Journal of Dairy Science, v. 89, n. 9, 2006. https://doi.org/10.3168/jds.S0022-0302(06)72409-2.

STYGAR, A. H. et al. Measuring dairy cow welfare with real-time sensor-based data and farm records: a concept study. Animal, v. 17, 101023, 2023. Disponível em: https://www.sciencedirect.com/science/article/pii/S1751731123003403. Acesso em: 14 jun. 2025.