Universidade de São Paulo
Escola Superior de Agricultura "Luiz de Queiroz"

Aprendizado de máquina aplicado a dados do microbiota do leite bovino cru para classificação em relação à mastite

Luan Gaspar Clemente

Trabalho de conclusão de curso para obtenção do título de graduação em Engenharia Agronômica.

Piracicaba
2019

Luan Gaspar Clemente
Engenharia Agronômica

Aprendizado de máquina aplicado a dados do microbiota do leite bovino cru para classificação em relação à mastite

Orientador:
Prof. Dr. **LUIZ LEHMANN COUTINHO**

Trabalho de conclusão de curso para obtenção do título de graduação em Engenharia Agronômica.

Piracicaba
2019

Dados Internacionais de Catalogação na Publicação
DIVISÃO DE BIBLIOTECA – DIBD/ESALQ/USP

Clemente, Luan Gaspar

Aprendizado de máquina aplicado a dados do microbiota do leite bovino cru para classificação em relação à mastite

Trabalho de Conclusão de Curso - - USP / Escola Superior de Agricultura “Luiz de Queiroz”.

1. CLEMENTE, L.G. I. Aprendizado de máquina aplicado a dados do microbiota do leite bovino cru para classificação em relação à mastite

SUMÁRIO

RESUMO

Aprendizado de máquina aplicado a dados do microbiota do leite bovino cru para classificação em relação à contagem de células somáticas e para rastreabilidade

A produção de leite é uma atividade cada vez mais competitiva. Somente em 2017, o Brasil produziu cerca de 35,1 bilhões de litros de leite. Um dos principais problemas da pecuária de leite é a mastite, que acarreta em diminuição ou até a perda total da produção do animal doente. O diagnóstico da mastite é realizado por meio de testes que detectam a presença de glóbulos brancos ou de microrganismos em amostras de leite. Uma vez detectada a presença de mastite, é imprescindível o seu controle, podendo ser feito através do manejo do rebanho e pela aplicação de antibióticos específicos ou de amplo espectro. Os algoritmos de aprendizado de máquina aplicado a uma estratégia de reamostragem podem ser utilizados para prever ou classificar amostras de leite, devido à sua eficiência computacional em encontrar padrões de generalização em dados com uma baixa quantidade de amostras. Neste trabalho foram aplicados três algoritmos de aprendizado de máquina, para *labels* de 300 mil CSS, 500 mil CSS, local de coleta de amostras e fornecedores em 213 amostras de leite bovino cru. Como forma de selecionar o grupo de gêneros mais relevante foi aplicado a abordagem de florestas aleatórias não enviesadas com inferência condicional e suas importâncias foram medidas pelo algoritmo de permutação condicional para preditores correlacionados. Os resultados demonstraram que o algoritmo de aprendizado de máquina mais eficiente é o de florestas aleatórias e que os perfis microbiológicos distintos dependem da quantidade de CSS, local de coleta de amostra e fornecedor. Foi encontrado também que os gêneros de microrganismos de maior importância para CSS são aqueles que possuem respaldo na literatura acerca de sua influência em processos inflamatórios, contaminação e degradação do leite. Portanto, pode-se concluir que dados do microbiota presente em amostras do leite possuem uma habilidade preditiva para amostras do leite, em alguns casos, dependendo da qualidade de seus dados e das técnicas utilizadas.

Palavras-chave: 1. Redução de dimensionalidade 2. Leite bovino 3. Mastite 4. Aprendizado de máquina.

ABSTRACT

Machine learning applied to microbiome data of raw bovine milk samples related to somatic cell counts and for traceability

In Brazil, milk production is an increasingly competitive activity. Only in 2017 it produced about 35.1 billion liters of milk. One of the main problems of dairy farming is mastitis, which leads to a decrease or even a total loss of production of the diseased animal. The diagnosis of mastitis is performed by tests that detect the presence of white blood cells or microorganisms in milk samples. Once the presence of mastitis is detected, its control is essential, and can be done through the management of the herd and the application of specific or broad spectrum antibiotics. The machine learning algorithms applied to a resampling strategy can be used to predict or classify milk samples because of their computational efficiency in finding generalization patterns in data with a low amount of samples. In this work three machine learning algorithms were applied, for 300 thousand CSS labels, 500 thousand CSS, sample collection site and storage types in 213 raw bovine milk samples. As a way to select the most relevant group of genera, the approach of non-skewed random forests with conditional inference was applied and their importance was measured by the conditional permutation algorithm for correlated predictors. The results showed that the most efficient machine learning algorithm is that of random forests and that the different microbiological profiles depend on the amount of CSS, sample collection site and supplier. It was also found that the genera of microorganisms of major importance for CSS are those that have support in the literature about their influence on inflammatory processes, contamination and milk degradation. Therefore, it can be concluded that microbiota data present in milk samples have a predictive ability to predict milk samples in some cases, depending on the quality of their data and the techniques used.

Keywords: 1. Dimensionality reduction, 2. Bovine milk, 3. Mastitis, 4. Machine Learning

1. INTRODUÇÃO

No Brasil, a produção de leite é uma atividade cada vez mais competitiva, sendo necessário compreender os fatores que podem influenciar a sua eficiência. De acordo com dados disponibilizados pela Embrapa (2018), o Brasil produziu cerca de 35,1 bilhões de litros de leite no ano de 2017, tendo como a volatilidade de preços e de custos uma das principais preocupações dos produtores.

Um dos grandes problemas da produção de leite brasileiro intrinsicamente relacionada ao custo é a mastite. A mastite acarreta a diminuição ou até a perda total da secreção láctea, além de apresentar modificação em sua composição, alterando suas características organolépticas, físicas, químicas e microbiológicas (VIANNI, 1986; LEITE et al., 1976). Radostis et al. (2007) descreveram a existência de mais de 140 espécies, subespécies e sorotipos de microorganismos envolvidos na mastite.

O diagnóstico da mastite clínica é feito para localizar os animais infectados dentro do rebanho, que devido à transmissibilidade são um risco potencial para a saúde de outras vacas. Os testes de diagnósticos mais comuns, para a mastite clínica, são os sinais clínicos e o teste de caneca, enquanto que para a mastite subclínica é a análise da contagem de células somáticas (CCS) e análise microbiológica do leite, para identificação das espécies. Uma vez detectada a presença de mastite clínica é imprescindível o controle da mesma, que pode ser realizado através do manejo do rebanho e pela aplicação de antibióticos específicos ou de amplo espectro (UFLA, 2012).

Nesse contexto, o uso de metodologias alternativas que são passíveis de prover uma boa predição dos resultados é de grande interesse. Os algoritmos de aprendizado de máquina aplicado a uma estratégia de reamostragem para predizer ou classificar amostras podem ser utilizadas para esse propósito, Como evidenciado por Piles (2019), recentemente, os algoritmos de aprendizado de máquina vem sendo utilizados para analisar dados de sequênciamento de última geração devido à sua eficiência computacional em encontrar padrões de generalização em dados de alta dimensão através de uma quantidade pequena de amostras.

O objetivo deste trabalho é identificar gêneros associados à abundância de células somáticas, tipo de armazanamento e fornecedores, além de verificar a

possibilidade do uso de dados da composição do leite, de mais fácil acesso e baixo custo, para a classificação das amostras de leite bovino cru em diversos *labels,* possibilitando então a criação de bancos de dados que possam ser utilizados para a previsão do conjunto de gêneros mais relevantes à cada contexto de controle do rebanho bovino.

2. MATERIAL E MÉTODOS

2.1. Amostras de leite bovino cru refrigerado

O conjunto de dados utilizados no estudo consistem em 301 amostras de leite bovino cru refrigerado de tanques, caminhões e silos, de produtores das cidades de Três Rios/MG e Araraquara/SP. Os dados gerados pelas amostras contem informações determinadas pela Clínica do Leite/ESALQ acerca da composição do leite como a contagem de estratos sólidos desengordurados, contagem de células somáticas, contagem bacteriana total, gordura, proteínas, lactose e estratos totais.

2.2. Sequênciamento das amostras

As amostras de leite bovino cru refrigerado foram coletadas em conservante bromanata; em seguida extraiu-se o DNA da amostra, via kit de extração da QIAGEN®. Para a preparação da biblioteca metagenômica foi utilizado o protocolo de preparação de bibliotecas 16S Illumina.

O protocolo Illumina consiste em uma amplificação em dupla etapa com purificação, seguida de normalização e *pool* das bibliotecas. Na primeira etapa de amplificação foi utilizada a técnica de PCR para amplificar as amostras de DNA utilizando *primers* específicos para a região V4.

As purificações das PCR foram feitas com *beads* AMPure XP e placas magnéticas. Na segunda etapa de PCR foram conectados os adaptadores Illumina utilizando o kit de indexação Nextera XP. A quantificação foi feita em NanoDrop e após a normalização, uma alíquota de 5uL de cada amostra foi adicionada ao *pool* de bibliotecas.

O sequenciamento das bibliotecas foi realizado pelo sistema MiSeq Illumina, tendo suas informações interpretadas e processadas. Os dados, em formato fasta, foram comparados aos do banco de dados SILVA, versão 132 (QUAST *et al.*, 2012), através da plataforma QIIME2, versão 2018.2 (https://qiime2.org).

2.3. Seleção dos dados

Foram selecionadas as amostras que continham todas as informações acerca da composição do leite, rastreabilidade e perfil microbiológico. O conjunto completo de amostras apresentou o total de 213 amostras.

2.4. Classificação das amostras em relação à contagem de células somáticas

As amostras de leite bovino cru foram rotuladas com base em sua contagem de células somáticas. Foram desenvolvidos dois rótulos independentes para o conjunto de amostras. O primeiro rótulo é referente à contagem de células somáticas tendo como limiar a contagem de 500 mil células somáticas por mililitro de leite bovino cru resfriado, limiar esse instruído pelo Ministério da Agricultura, Pecuária e Abastecimento no dia 30 de novembro de 2018, através das instruções normativas 76 e 77 (BRASIL, 2018). Contagens acima do limiar são classificadas como impróprias e contagens abaixo do limiar são classificadas como próprias. O segundo rótulo é referente aos limiares obtidos por Fonseca e Veiga (2000), onde foi detectado que a contagem de células somáticas de animais sadios, normalmente é inferior a 300 mil células por mililitro.

2.5. Classificação das amostras em relação a dados da rastreabilidade

Além da rotulação por contagem de células somáticas foram desenvolvidos dois rótulos referentes à rastreabilidade das amostras de leite bovino cru. O primeiro rótulo, denominado de tipo, é uma variável categórica com as categorias tanque, caminhão e silo. O segundo *label*, denominado forcedor, é uma variável categórica contendo a identificação de diversos produtores de leite das cidades de Três Rio, no estado de Minas Gerais e Araraquara, no estado de São Paulo.

2.6. Classificação das amostras em relação à presença de patógenos

Foram criados rótulos em função da presença de patógenos nas amostras de leite cru bovino, sendo estes microrganismos pertencentes aos gêneros *Mycobacterium*, *Staphylococcus* e *Streptococcus*. A escolha desses gêneros se deve principalmente pela presença desses patógenos e por sua relevância em termos de saúde pública, sanidade de produção e qualidade do leite.

No que tange a saúde pública, o *Mycobacterium bovis*, causador da tuberculose bovina, é responsável por parte dos casos de tuberculose em humanos. Estimativas indicam que o M. *bovis* seja responsável por 3% de todas as formas de tuberculose humana na América Latina (COSIVI *et al.*, 1998).

Em relação a qualidade do leite e à sanidade do sistema de produção, duas espécies importantes do gênero *Staphylococcus* são a *Staphylococcus aureus* e *Staphylococcus agalatiae*, considerados como os principais agentes contagiosos da mastite bovina, enquanto que espécies do gênero *Streptococcus*, como *Streptococcus uberis* e *Streptococcus bovis* são considerados os principais agentes ambientais da mastite bovina (BRAMLEY, 1984).

2.7. Seleção de variáveis

Um dos objetivos principais do trabalho foi identificar o menor número de gêneros que resultem na classificação de maior qualidade possível. Segundo Piles (2019), existem três abordagens principais no aprendizado de máquina para a realização da redução de dimensiondalidade: *filter*, *wrapper*, e *embedded*. Neste estudo foi escolhida a abordagem por *filter,* em especial por suas características, como a independência de técnicas específicas de aprendizado de máquina, escalabilidade e velocidade computacional (STROBL *et al.*, 2007).

Os gêneros foram selecionados com base em sua medida de importância, calculados através de florestas aleatórias não enviesadas baseadas em inferência condicional (STROBL *et al.*, 2007).

A floresta aleatória não enviesada baseada em inferência condicional é um método não linear e não paramétrico que pode ser aplicado a uma variedade de problemas, inclusive aqueles que apresentem interações complexas não lineares e nos quais o número de dados é fortemente menor do que o número de preditores,

como no estudo em questão. Tais análises foram realizadas na plataforma RStudio com o auxílio do pacote "*party*" através da função "cforest". A função "cforest" realiza uma reamostragem sem reposição que, segundo Strobl e colaboradores (2007), é a única abordagem que garante uma seleção de variáveis confiáveis, produzindo medidas de importância não enviesadas mesmo em situações nas quais as variáveis preditoras possuem diferentes escalas de medidas e diferentes números de categorias.

O hiperparâmetro ajustado neste algoritmo de aprendizado de máquina foi o número de variáveis de entrada que foram amostradas como candidatas em cada nó. O número de árvores foi fixado em 1000.

Para o cálculo da medida de importância das variáveis foi utilizado a abordagem de permutação condicional para preditores correlacionados. Nessa abordagem a importância de cada variável é medida pela permutação interna definida pelas covariações que são associadas (dentro de 1 – p-valor maior que um número de corte) para a variável de interesse. A medida de importância da variável é condicional em relação ao coeficiente beta no modelo de regressão, mas representa o efeito da variável tanto no efeito principal como na interação (STROBL *et al.*, 2008).

Os algoritmos de aprendizado de máquina foram testados utilizando o subconjunto de gêneros de microrganismos com medida de importância positiva e o conjunto completo de gêneros.

2.8. Algoritmos de aprendizado de máquina

A classificação das amostras de leite bovino cru em relação à contagem de células somáticas, CCS, utilizando dados do microbiota e da composição do leite foram realizadas utilizando três métodos de aprendizado de máquina descritos a seguir: máquina de vetores de suporte (SVM), floresta aleatória (RF), e k-vizinhos mais próximos (KNN). Esses algoritmos de aprendizado de máquina foram implementados utilizando o pacote R "caret" (KUHN, 2008), que é uma interface para um grande número de técnicas de classificação e regressão que possibilita que resultados de diferentes técnicas de aprendizado de máquina sejam comparados nas mesmas condições e permite encontrar os hiperparâmetros mais adequados

para cada técnica de aprendizado de máquina automaticamente, garantindo que os resultados sejam confiáveis e não enviesados.

Segundo Piles (2019), a máquina de vetor de suporte é uma técnica de aprendizado de máquina bem conhecida e utilizada em diferentes domínios, com bons resultados em muitos casos; seu objetivo é encontrar o hiperplano que melhor separa as amostras, enquanto maximiza a distância entre as amostras e o hiperplano, encontrando portanto a melhor generalização possível (BULGES, 1998).

A floresta aleatória combina inúmeras árvores de classificação que são ajustadas para subamostras do conjunto original de amostras utilizando seleção randômica de variáveis preditores. Como resultado, uma única predição é obtida com a média (no caso da regressão) ou o voto majoritário (no caso da classificação), da predição de todas as árvores (BREIMAN, 2001). As vantagens da floresta aleatória são: é simples e os resultados são facilmente interpretráveis no caso de poucos preditores, e pode ser aplicado a vários problemas, mesmo aqueles com efeitos de interação ou relações não lineares entre as variáveis (BREIMAN, 2001).

A abordagem de k-vizinhos mais próximos (KNN) foi proposto por Fukunaga e Narendra em 1975. Atualmente é considerado um dos classificadores mais simples de ser implementado, de fácil compreensão e com bons resultados dependendo de sua aplicação. Segundo Fukunaga (1975) a ideia principal do KNN é determinar o rótulo de classificação de uma amostra baseado nas amostras vizinhas advindas de um conjunto de treinamento. Dois pontos chaves que devem ser determinados para aplicação do KNN, sendo eles: a métrica de distância e o valor de k. Existem diversas métricas de distância, e a escolha de qual usar varia de acordo com o problema. A mais utilizada é a distancia eucladiana, porém outras distâncias como a de Minkowsky e Chebyshev também podem ser utilizadas (FUKUNAGA, 1975).

2.9. Reamostragem

A reamostragem foi realizada utilizando a abordagem de *nested resampling.* A *nested resampling* permite a obtenção de estimadores de desempenho confiáveis para os algoritmos de aprendizado de máquina e a quantificação da habilidade de generalização do modelo de classificação utilizado (FERNANDEZ-LOPEZ *et al.*, 2016); consiste em dois *loops* de reamostragem aninhada. No primeiro *loop*, uma

validação cruzada de dez vezes é realizada randomicamente, dividindo o conjunto de dados em 10 subgrupos de tamanhos iguais, levando em consideração a leve desigualdade dos dados. Deste modo, os subcojuntos são estratificados de modo que se mantenham similares as frequências de cada classe no subconjunto. Um grupo foi utilizado como conjunto de validação, enquanto os outros 9 grupos restantes foram utilizados como conjunto de treinamento. O processo é repetido 10 vezes, com um diferente grupo de dados, sendo utilizados como conjunto de validação, resultando então em 10 pares de conjuntos de treinamento/validação. O ajuste de parâmetro foi feito para cada conjunto de treinamento, executando o *loop* interno de reamostragem, que também consistia em uma validação cruzada de dez vezes, resultando em um conjunto de hiperparâmetros selecionados para cada conjunto de treinamento. O aprendiz foi ajustado em cada conjunto de treinamento externo usando os hiperparâmetros selecionados correspondentes e seu desempenho foi avaliado no conjunto de validação correspondente. O pacote “caret” facilita a execução de todas as tarefas necessárias e agrega os resultados obtidos de cada algoritmo de aprendizado de máquina, comparando-os exatamente nas mesmas condições.

2.10. Medidas de desempenho dos classificadores

Inúmeras medidas de desempenho estão disponíveis para a classificação como a taxa média de erro, acurácia e medidas baseadas em análises. ROC. Neste trabalho foi utilizada a acurácia e o coeficiente de concordância Kappa. O coeficiente de concordância kappa é um teste proposto por Jacob (1960) e possui a finalizade de medir o grau de concordância entre proporções derivadas de amostras dependentes (FLEISS, 1981). Essa medida tem como valor máximo o valor unitário, que representa total concordância. Os valores próximos e até mesmo abaixo de zero indicam nenhuma concordância, ou a presença de uma eventual discordância entre as classificações. Como tentativa de interpretação, os autores Landis e Koch (1977) sugerem a seguinte tabela de interpretação do valor de kappa.

Tabela 1. Valores de interpretação do coeficiente Kappa por Landis e Koch (1977).

Valor de Kappa	**Interpretação**
Menor do que zero	Insignificante
Entre 0 e 0,2	Fraca
Entre 0,21 e 0,4	Razoável
Entre 0,41 e 0,6	Moderada
Entre 0,61 e 0,8	Forte
Entre 0,81 e 1	Quase perfeita

Fonte: Landis e Koch (1977)

3. RESULTADOS E DISCUSSÕES

3.1. Conjunto de variáveis relacionadas a gêneros de microrganismos

O número de gêneros selecionados para cada *label* variou, sendo de 46 gêneros com maior importância para classificação de amostras com quantidades acima de 300 mil células somáticas por ml, 52 gêneros para amostras com quantidades acima de 500 mil células somáticas, 86 gêneros para classificação com base em seu local de coleta e 82 gêneros para classificação com base em seus fornecedores.

3.1.1 Seleção de variáveis

3.1.1.1 Labels 500 e 300mil CCS/ml

Os dez principais gêneros para classificação de amostras com quantidades acima de 300 mil células somáticas são *Streptococcus, Anoxybacillus, Lactobacilus, Empedobacter, Enhydrobacter, Leuconostoc, Veillonela, Shewanella, Haemophilus, Singulisphaera*, enquanto que para amostras com quantidades acima de 500 mil células somáticas são *Lactococcus, Empedobacter, Streptococcus, Aeromonas,, Shewanella, Massilia, Anoxybacillus, Allorhizobium, Leuconostoc* e *Solibacilus.*

Os valores de medida de importância apresentaram um intervalo de -0,0005 a 0,0047 para a classificação com base em limites de 300 mil células somáticas e -0,00045 a 0,00385 para classificação com base em limites de 500 mil células somáticas.

É interessante observar que dentre os gêneros que mais auxiliam na classificação das amostras, grande parte deles possuem relação, constatada previamente em outros estudos, com prevalência em amostras, processos de deterioração do leite, infecção em ruminantes e falta de sanidade no processo de produção.

Relacionados à prevalência em amostras, foi constatado como primeiro gênero mais importante para classificação, com limar em 300 mil CSS, o gênero *Anoxybacillus*, previamente observado no estudo de Miller et al. (2015), como sendo o segundo gênero mais prevalente em amostras de leite cru e leite em pó; indicando

portanto que sua abundância pode estar relacionada à altas contagens de células somáticas.

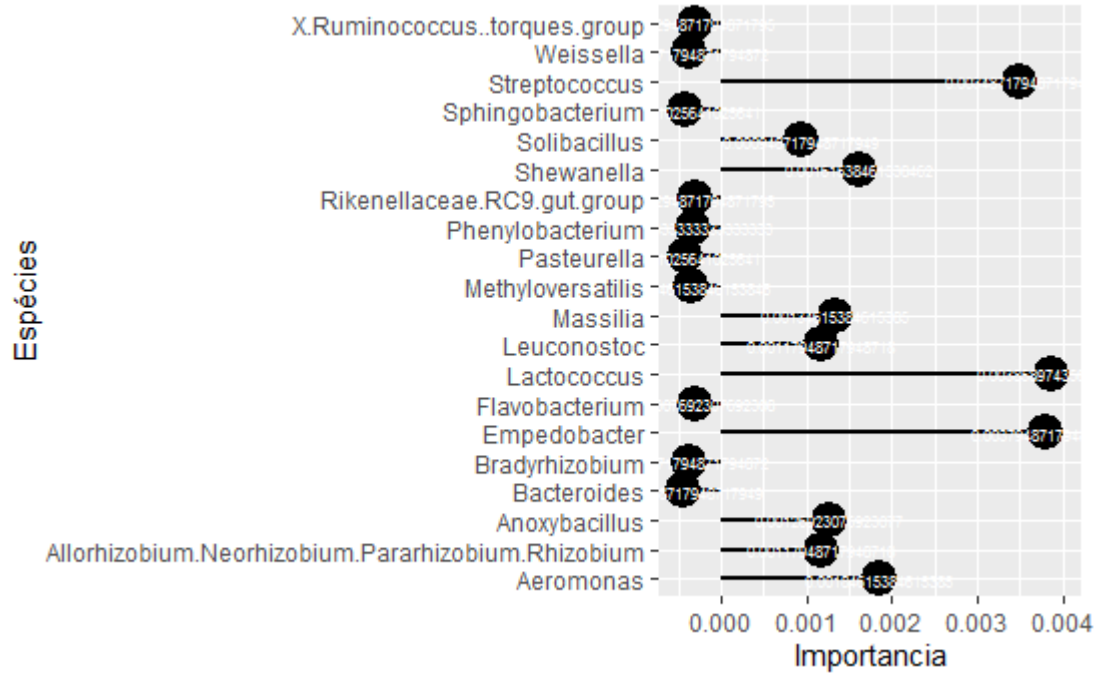
Para a classificação com limiar em 500 mil CSS, os gêneros de maior importância diferem em partes, com a interseção dos gêneros *Streptococcus* e *Leuconostoc*. Os valores de medida de importância apresentaram um intervalo de -0,000102 a 0,00824 para classificação com base no meio de local de coleta e por fim, -0,0002 a 0,00484 para classificação com base nos fornecedores.

Em seu trabalho, Catozzi (2017) encontrou que a presença de mastite está relacionada a mudança relativa dos gêneros *Psychrobatecter*, *SMB53* e *Solibacillus*. As abundâncias relativas decrescem em amostras de leite de bufalo com mastite clínica, padrão semelhante encontrado em nosso estudo, evidenciando portanto a possibilidade do uso de analises discriminantes e aprendizado de máquina para discriminar comunidades microbiotas de animais saudáveis e com mastite clínica com base em seu perfil microbiológico.

Em relação a *Aeromonas*, é conhecido que esteja relacionada à falta de sanidade no processo produtivo com alta importância na saúde pública (KIROV, 1993).

Segue abaixo um gráfico apresentando os 10 gêneros com a maior e menor medida de importância para o label 300 mil células somáticas por ml.

Figura 1. Dez gêneros com a maior e menor medida de importância para o label 300 mil células somáticas por ml.



3.1.1.2 *Labels* local de coleta e fornecedor

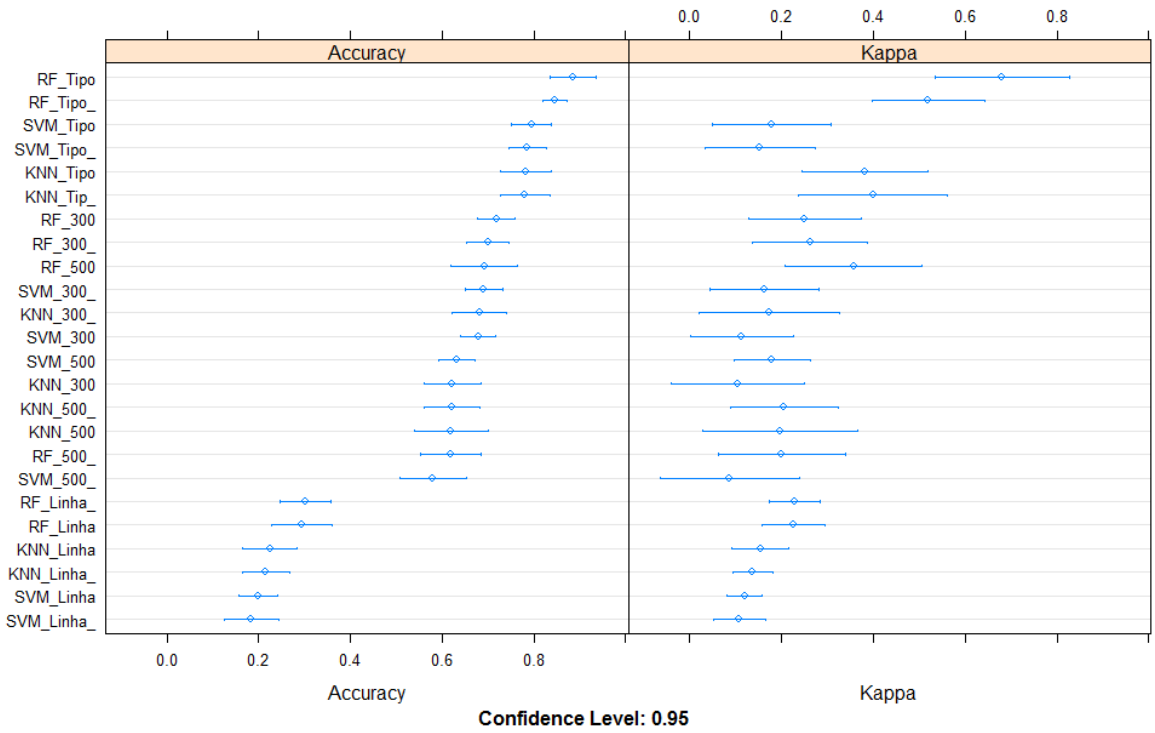
Os resultados de seleção de variáveis para *label* local de coleta indicam que os dez principais gêneros com maior importância são o *Carnobacterium, Lactococcus, Leuconostoc, Anoxybacilus, Enterobacteriacea, Macrococcus, Shewanella, Mycoplasma, Bifidobacterium, Pseudomonas*, enquanto que para a classificação em função dos fornecedores os genêros são *Anoxybacillus, Empedobacter, Carnobacterium, Janibacter, Lactococcus, Macrococcus, Burkholderia, Pantoea, Enhydrobacter e Enterobacter.*

Um gênero relevante nesse cenário é o pertencente à família *Enterobacteriaceae*. A *Enterobacteriacea* apresenta importância não só por poder indicar contaminação fecal, mas também por estar geralmente implicado em processos infecciosos, demonstrando um grau considerável de ineficiência higiênico-sanitária. (HOFFMANN et al., 2004), corroborando portanto com o potencial da técnica de classificação por aprendizado de máquina, uma vez que a mesma qualifica tal gênero como relevante para a classificação com base no diferentes locais de coleta, uma potencial fonte de contaminação do leite.

3.1.2 Classificação, acurácia e medida de desempenho dos algoritmos

A Figura 2 mostra os resultados para a classificação das amostras de leite bovino cru baseado em dados da microbiota do leite bovino cru. O conjunto de dados contém as variáveis referentes as informações da microbiota, expressa pelos gêneros presentes na amostra e suas quantidades aproximadas. A quantificação dos gêneros foi realizada pelo produto entre a porcentagem do gênero presente na amostra e a contagem bacteriana total. Os *labels* utilizados na classificação desse conjunto de dados foram os de 300 mil células somáticas por ml, 500 mil células somáticas por ml, local de coleta, podendo ser por caminhão, silo ou tanque e fornecedores.

Figura 2. Valores de acurácia e do coeficiente Kappa resultantes da aplicação de três modelos de aprendizado de máquina em amostras de leite bovino cru com as *labels* 300, 500, local de coleta e linha/fornecedor de grupos de variáveis selecionadas e conjunto de variáveis totais, representadas por um *underline* terminal.



Para dados da microbiota do leite é possível verificar que o desempenho dos algoritmos de aprendizado de máquina variou em função da técnica escolhida, *label* utilizado e conjunto de variáveis aplicadas.

Para a técnica dos k-vizinhos mais próximos os resultados de acurácia foram de, em média, 0,620 para o *label* 500, 0,622 para o *label* 300, 0,225 para o label linha/fornecedor e 0,783 para o local de coleta empregado, em função ao conjunto de variáveis previamente selecionados pelo método de floresta aleatória não enviesada baseada em inferência condicional, enquanto que para o conjunto completo de variáveis os resultados de acurácia foram de, em média, 0,621 para o *label* 500, 0,681 para o *label* 300, 0,215 para o *label* linha/fornecedor e 0,78 para o local de coleta empregado.

Os valores do coeficiente de kappa são de, em média, 0,197 para o *label* 500, 0,104 para o *label* 300, 0,153 para o label linha/fornecedor e 0,381 para o local de coleta selecionado, em função ao conjunto de variáveis previamente selecionados pelo método de floresta aleatória não enviesada baseada em inferência condicional, enquanto que para o conjunto completo de variáveis os resultados de acurácia foram de, em média, 0,2 para o *label* 500, 0,172 para o *label* 300, 0,137 para o *label* linha/fornecedor e 0,399 para o local de coleta selecionado.

Para a técnica de máquina de vetores de suporte os resultados de acurácia foram de, em média, 0,63 para o *label* 500, 0,67 para o *label* 300, 0,199 para o label linha/fornecedor e 0,795 para o local de coleta selecionado, em função ao conjunto de variáveis previamente selecionados pelo método de floresta aleatória não enviesada baseada em inferência condicional, enquanto que para o conjunto completo de variáveis os resultados de acurácia foram de, em média, 0,58 para o *label* 500, 0,69 para o *label* 300, 0,184 para o *label* linha/fornecedor e 0,786 para o local de coleta selecionado.

Os valores do coeficiente de kappa são de, em média, 0,179 para o *label* 500, 0,113 para o *label* 300, 0,119 para o label linha/fornecedor e 0,179 para o local de coleta selecionado, em função ao conjunto de variáveis previamente selecionados pelo método de floresta aleatória não enviesada baseada em inferência condicional, enquanto que para o conjunto completo de variáveis os resultados de acurácia foram de, em média, 0,086 para o *label* 500, 0,162 para o *label* 300, 0,107 para o *label* linha/fornecedor e 0,152 para o local de coleta selecionado.

Para a técnica de florestas aleatórias os resultados de acurácia foram de, em média, 0,691 para o *label* 500, 0,718 para o *label* 300, 0,293 para o label linha/fornecedor e 0,885 para o local de coleta selecionado, em função ao conjunto de variáveis previamente selecionados pelo método de floresta aleatória não

enviesada baseada em inferência condicional, enquanto que para o conjunto completo de variáveis os resultados de acurácia foram de, em média, 0,618 para o *label* 500, 0,70 para o *label* 300, 0,184 para o *label* linha/fornecedor e 0,845 para o local de coleta selecionado.

Os valores do coeficiente de kappa são de, em média, 0,357 para o *label* 500, 0,250 para o *label* 300, 0,226 para o label linha/fornecedor e 0,68 para o local de coleta selecionado, em função ao conjunto de variáveis previamente selecionados pelo método de floresta aleatória não enviesada baseada em inferência condicional, enquanto que para o conjunto completo de variáveis os resultados de acurácia foram de, em média, 0,19 para o *label* 500, 0,261 para o *label* 300, 0,227 para o *label* linha/fornecedor e 0,5195 para o local de coleta selecionado.

Verifica-se, portanto, que a técnica de seleção de variáveis utilizada não apresentou um ganho de acurácia e aumento do coefiicente Kappa em grande parte das análises, com exceção de três casos; aplicação de máquinas de vetor suporte para o label 500, resultando em aumento de acurácia e do coeficiente kappa, no entanto, ainda considerada como um classificador fraco em relação à interpretação do coeficiente kappa por Landis e Koch (1977); floresta aleatória aplicada ao label 500, com aumento de acurácia e coeficiente kappa, resultando em uma interpretação de classificação moderada; floresta aleatória aplicada ao local de coleta, com aumento de acurácia e coeficiente kappa, mantendo a interpretação da classificação como forte.

De todos os cenários analisados, os que possuiram a melhor aplicabilidade são a técnica de florestas aleatória para detecção do local de coleta e classificação de amostras de leite bovino como limiares de contagem de células somáticas em 500 mil.

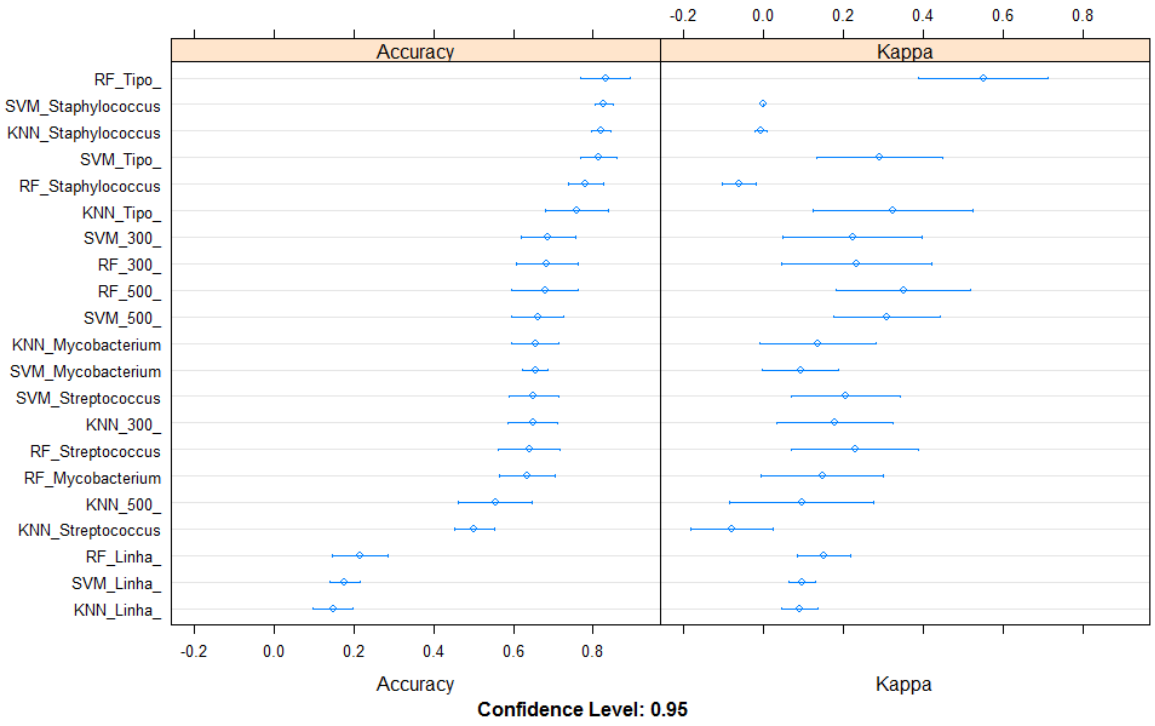
3.2. Conjunto de variáveis relacionadas à composição do leite

O conjunto de variáveis relacionadas à composição do leite possuem informações acerca das variáveis contagem bacteriana total, porcentagem de gordura, proteína, lactose, estratos totais e estratos totais desengodurados.

3.2.1 Classificação, acurácia e medida de desempenho dos algoritmos

A Figura 3 mostra os resultados para a classificação das amostras de leite bovino cru baseado em dados da composição do leite bovino cru. Os *labels* utilizados na classificação desse conjunto de dados foram os de 300 mil células somáticas por ml, 500 mil células somáticas por ml, local de coleta, podendo ser por caminhão, silo ou tanque, linha/fornecedores e presença de gêneros como *Mycobacterium, Staphylococcus* e *Streptococcus*.

Figura 3. Valores de acurácia e do coeficiente Kappa resultantes da aplicação de três modelos de aprendizado de máquina em amostras de leite bovino cru com as *labels* 300, 500, local de coleta e linha/fornecedor, e presença de gêneros como *Mycobacterium, Staphylococcus e Streptococcus.*



Para dados da composição do leite, é possível verificar que o desempenho dos algoritmos de aprendizado de máquina variou, assim como para os dados da microbiota do leite, em função da técnica escolhida e do *label* utilizado.

Para a técnica dos k-vizinhos mais próximos os resultados de acurácia foram de, em média, 0,554 para o *label* 500, 0,648 para o *label* 300, 0,237 para o label

linha/fornecedor , 0,758 para o local de coleta selecionado, 0,655 para presença do gênero Mycobacterium, 0,82 para presença do gênero Staphylococcus e 0,502 para presença do gênero Streptococcus.

Os valores do coeficiente de kappa são de, em média, 0,095 para o *label* 500, 0,178 para o *label* 300, 0,09 para o label linha/fornecedor, 0,32 para o local de coleta selecionado, 0,136 para presença do gênero Mycobacterium, -0,007 para presença do gênero Staphylococcus e -0,079 para presença do gênero Streptococcus.

Para a técnica de máquina de vetor de suporte os resultados de acurácia foram de, em média, 0,660 para o *label* 500, 0,687 para o *label* 300, 0,177 para o label linha/fornecedor , 0,813 para o local de coleta selecionado, 0,655 para presença do gênero Mycobacterium, 0,82 para presença do gênero Staphylococcus e 0,651 para presença do gênero Streptococcus.

Os valores do coeficiente de kappa são de, em média, 0,30 para o *label* 500, 0,22 para o *label* 300, 0,09 para o label linha/fornecedor, 0,289 para o local de coleta selecionado, 0,09 para presença do gênero Mycobacterium, 0,0 para presença do gênero Staphylococcus e 0,206 para presença do gênero Streptococcus.

Para a técnica de árvores aleatórias os resultados de acurácia foram de, em média, 0,679 para o *label* 500, 0,684 para o *label* 300, 0,215 para o label linha/fornecedor , 0,8309 para o local de coleta selecionado, 0,634 para presença do gênero Mycobacterium, 0,781 para presença do gênero Staphylococcus e 0,640 para presença do gênero Streptococcus.

Os valores do coeficiente de kappa são de, em média, 0,30 para o *label* 500, 0,34 para o *label* 300, 0,23 para o label linha/fornecedor , 0,55 para o local de coleta selecionado, 0,14 para presença do gênero Mycobacterium, -0,062 para presença do gênero Staphylococcus e 0,229 para presença do gênero Streptococcus.

O cenário que apresentou o melhor resultado foi a técnica de florestas aleatória para detecção do local de coleta selecionado.

4. CONCLUSÃO E CONSIDERAÇÕES

Os resultados permitiram concluir que o uso de dados do microbiota presente em amostras do leite possui uma habilidade preditiva para classificação do leite em alguns casos, dependendo da qualidade de seus dados e das técnicas utilizadas.

Além disso, foi possível demonstrar um bom desempenho dos algoritmos de aprendizado de máquina para realizer predições em contextos complexos como a ocorrência de mastite e interações com o local de coleta.

Dentre os algoritmos analisados, o que se mostrou com maior potencial foi o de florestas aleatórias, apresentando os melhores resultados nos cenários analisados, tanto em sua acurácia, como seu coeficiente kappa.

Os dados também indicam a possibilidade de uso da técnica de redução de dimensionalidade, a floresta aleatória não enviesada baseada em inferência condicional para a seleção de gêneros mais associados com algum fenômeno de interesse.

Por fim, vale ressaltar que os dados utilizados no estudo são referentes aos tanques de armazenamento de leite, podendo, portanto, sua comunidade microbiota estar enviesada, tendo seu perfil microbiológico distante daquele que seria encontrado em amostra de leite bovino individualizado, limitando a eficiência das técnicas aplicadas. Portanto, é necessário que mais estudos sejam realizados nesse sentido, para uma melhor aferição do potencial da técnica utilizada.

REFERÊNCIAS

Bramley A.J. & Dodd F.H. (1984). Reviews of the progress of dairy science: mastitis control-progress and prospects. **J. Dairy Res**. v.51. p.481-512.

Brasil **(2018).** Critérios e procedimentos para a produção, acondicionamento, conservação, transporte, seleção e recepção do leite cru em estabelecimentos registrados no serviço de inspeção oficial. **Instrução Normativa Nº 77, de 26 de novembro de 2018**.

Breiman L. (2001). Random forests. **Mach Learn.** 45:5.

Burges, C.J.C. (1998) A Tutorial on support vector machines for pattern recognition. **Data Min Knowl Discov**. 2:121–67.

Catozzi, C. et al. (2017). The microbiota of water buffalo milk during mastitis. **PloS One.** 12:9.

Cohen, J (1960). A Coefficient of Agreement for Nominal Scales. **Educational and Psychological Measurement**, v. 20, n. 1, p. 37–46.

Cosivi, O.; Grange, J. M.; Daborn, C. J. (1998) Zoonotic tuberculosis due to Mycobacterium bovis in developing countries. **Emerging Infectious Diseases**, v. 4, p. 59-70.

Fernandez-Lozano, C, Gestal, M, Munteanu, C.R., Dorado, J, Pazos, A. (2016). A methodology for the design of experiments in computational intelligence with multiple regression models. **Peer J**. 4:e2721.

Fleiss, J. L (1981). Statistical methods for rates and proportions. **New York: John Wiley**; p. 212-36.

Fonseca, L. F. L.; Santos, M. V. (2000). Qualidade do leite e controle de mastite. **São Paulo: Lemos Editorial**,. 175 p

Fukunaga, K.; Narendra, P. M. (1975). A branch and bound algorithm for computing k-nearest neighbors. **IEEE Transactions on Computers**, v. 100, n. 7, p. 750–753.

Hoffman, F. L. et al. (2004). Qualidade microbiologia de queijos ralados de diversas marcas comerciais, obtidos do comércio varejista do município de São José do Rio Preto, SP. **Rev. Higiene Alimentar**, v. 18, n.122, p. 62-66.

Kuhn, M. (2008). Building Predictive Models in R Using the caret Package. **Journal of Statistical Software**, 28(5), 1 – 26.

Kirov, S.M., HUP, D.S., Hayward, L.J., (1993). Milk as a potencial source of Aeromonas Gastrointestinal Infection. **Journal of Food Protection.** V.56 n.4 p.306-312.

Landis, J.R., Koch, G.G. (1977). The Measurement of Observer Agreement for Categorical Data**. Biometrics**. v. 33, n.1, p159-74.

Leite, R.C. Brito, J.R.F., Figueiredo, J.B. (1976). Alterações da glândula mamária de vacas tratadas intensivamente via mamária com penicilina em meio aquoso. **Arqv. Esc. Vet. UFMG**. v.28, p.27-31.

Miller, R., et al. (2015). Spore populations among bulk tank raw milk and dairy powders are significantly different. **Journal of Dairy Science**, v. 98, p. 8492-8504.

Piles, et al. (2019). Machine learning applied to transcriptomic data to identify genes associated with feed efficiency in pigs. **Genetics Selection Evolution.** 51:10.

Quast, C. et al. (2012), The SILVA ribossomal RNA gene database project: improved data processing and web-based tools. **Nucleid Acid Reserarch,** v.41, n.1, p590-597.

Radostis, O.T., et al. (2007) Veterinary Medicine: a text book of disease of cattle, horses, sheep, pigs and goats, 10th edition, **Philadelphia: Saunders Elsevier**. p. 173 - 187.

Strobl C, Boulesteix AL, Zeileis A, Hothorn T (2007). Bias in random forest variable importance measures: illustrations, sources and a solution. **BMC Bioinformatics** 8:25.

Strobl C, Boulesteix AL, Kneib T, Augustin T, Zeileis A (2008). Conditional variable importance for random forests. **BMC Bioinformatics** 9:307.

UFLA, (2012). Mastite bovina: controle e prevenção. Boletim técnico nº 93. **Editora UFLA.** 30p.

Vianni, M.C.E., Lázaro, N.S. (2003). Perfil de suscetibilidade a antimicrobianos em amostras de cocos Gram-positivos, catalase negativos, isolados de mastite subclínica bubalina. Pesq-Veterin. Bras. n.23, p47-51.