Universidade Estadual de Campinas
Instituto de Matemática, Estatística e Computação Científica
Departamento de Estatística

Modelo de regressão logística para a predição de estado de saúde animal a partir de variáveis escalares e funcionais

---

Relatório Científico Final do projeto na modalidade Iniciação Científica, fomentado pela Fundação de Amparo à Pesquisa do Estado de São Paulo.

Projeto FAPESP: 2020/01436-0

Bolsista Responsável:
Rodrigo Forti

Americana
2021

Informações Gerais do Projeto

- Título do projeto: Modelo de regressão logística para a predição de estado de saúde animal a partir de variáveis escalares e funcionais
- Nome do pesquisador responsável: Rodrigo Forti
- Instituição sede do projeto: Instituto de Matemática, Estatística e Computação Científica da Universidade Estadual de Campinas
- Equipe de pesquisa: Rodrigo Forti (Bolsista) e Mariana Rodrigues Motta (Pesquisadora)
- Número do projeto de pesquisa: 2020/01436-0
- Período de vigência: 01/09/2020 a 31/08/2021
- Período coberto por este relatório científico: 11/02/2021 a 31/08/2021

Conteúdo

Resumo

Introdução

O período de transição entre o final da gravidez em vacas e o começo período de lactação (também chamado período periparturiente) é certamente o estágio mais importante do ciclo de lactação destes animais, pois maioria das doenças infecciosas e distúrbios metabólicos ocorrem durante esse período. A febre do leite, a cetose, as membranas fetais retidas, a metrite e o abomaso deslocado afetam as vacas diminuindo a produção de leite e prejudicando o seu bem-estar durante o período peripartidário. Há uma maior suscetibilidade para doenças durante o período periparturiente que leva, por exemplo, ao aumento de incidência de mastite nos animais. Assim, a ocorrência de problemas de saúde se concentra no período periparturiente, preocupando produtores já que a produção de leite pode diminuir (Drackley, 1999). Como afirmado por Goff and Horst (1997), a transição do estado gestante, não lactante, para o estado não gestante e lactante, é muitas vezes uma experiência desastrosa para a vaca. Alguns problemas de saúde e reprodutivos podem ser resultado do aumento do estresse que se encontram vacas de alta produção que estão sob o início da lactação (Kaufman et al. 2018).

Além disso, no início da lactação, a ingestão alimentar é incapaz de atender às demandas de alta produção de leite. A vaca, portanto, entra em um período de balanço energético negativo que leva à mobilização de reservas corporais para equilibrar déficit entre consumo de energia dos alimentos e produção de energia do leite (Lago et al. 2001). Lago et al. (2001) também defendem que o processo de mobilização parece afetar o bem-estar da vaca e outras vias biológicas comprometidas à medida que a energia de entrada é direcionada para a produção.

Assim, a saúde do animal durante o período de lactação é fator determinante na lucratividade dos produtores. Limitações nutricionais ou de manejo durante esse período podem impedir a capacidade da vaca de atingir a produção máxima de leite. Desta forma monitorar índices vitais da vaca, assim como do ambiente onde está inserido, é condição essencial para aprimorar o bem-estar animal e a lucratividade no período de alta incidência de doenças em vacas periparturientes

Dados

Os dados a serem utilizados neste projeto fazem parte de uma base de dados do laboratório Dairy Cattle Biology and Management da Universidade de Cornell, EUA. Estes dados foram intermediados pelo pesquisador Guilherme Rosa, da University of Wisconsin, EUA, mediante colaboração no projeto FAPESP 2017/15306-9, sob coordenação da Profa. Dra Nancy Lopes Garcia, e tendo como colaboradora no projeto a Profa. Dra. Mariana Rodrigues Motta.

Os dados foram coletados a partir da análise clínica diária do estado de saúde 500 vacas. Os dados contém informações sobre características individuais das vacas, desempenho na produção de leite e eventos das lactações anteriores e da parição atual. Além disso, os dados foram coletados através de múltiplos sensores adaptados ao animal durante 3 sessões diárias para avaliar aspectos do leite, tais como porcentagem de gordura, de proteína, dados de lactose e condutividade, além de dados sobre total de passos dados, tempo de ruminação e tempo de repouso do animal.

Os dados considerados neste estudo devem ser usados para treinar, validar e testar modelos do Sistema de Monitoramento de Saúde Animal (AHMS) do projeto USDA da Dairy Laboratório de Biologia e Gerenciamento de Gado da Universidade de Cornell. Os dados contém informações sobre os estados de saúde da vaca e respectiva curva de lactação. O conjunto de dados principal corresponde ao exame clínico

diário de animais que são necessários para analisar o período de uma semana antes até três semanas depois do parto em todos os animais. Diariamente, informações de uma vaca sobre um distúrbio de saúde são coletadas.

O estudo de Cornell compreende três fases, Fase I, Fase II e Fase III. O objetivo do estudo da Fase I é caracterizar o padrão de doença, avaliando como os parâmetros mudam com relação ao status do animal. No estudo da Fase II, o foco é desenvolver alertas que sirvam de indicadores aos agricultores quando uma vaca está doente, visando criar uma combinação de vários parâmetros que informam sobre a saúde do animal. A Fase III do estudo compreende a validação dos métodos usados com dados da Fase I, usando ferramentas desenvolvidas em campo em tempo real e, em seguida, avaliando falsos positivos e sensibilidade do método.

A descrição das variáveis do conjunto de dados se encontram na Tabelas 1 e 2.

Tabela 1: Descrição das variáveis do conjunto de dados

Variável	Unidade	Descrição
ID	Número	Identificação da vaca
DiasLac	Dias	Dias em lactação (dia 0: dia do parto)
DiasRegimeFechado	Dias	Dias em regime Fechado
DiasSecos	Dias	Dias sem produzir leite
NumBezerro	Número	Número de bezerros nascidos no parto
ProdLeite	Gramas	Produção de leite
Gordura	%	Gordura no leite
Proteina	%	Proteína no leite
Lactose	%	Lactose no leite
Sangue	%	Sangue no leite
NumCelSomaticas	*1000/ml	Número de células somáticas no leite
TempoRepouso	Minutos	Tempo total de repouso por dia

Tabela 2: Descrição das variáveis do conjunto de dados - continuação

Variável	Unidade	Descrição
NumRepousos	Número	Número de repousos por dia
DuracaoMediaRepouso	Minutos	Duração média do cada repouso
Atividade	Passos/Hora	Total de passos dados no dia dividido por 24
RuminacaoUltimas24h	Minutos	Ruminação total nas últimas 24 horas
AlimentacaoUltimas24h	Gramas	Quantidade de comida ingerida nas últimas 24 horas
Hist_DistDigestivo	0-1	Há histórico de distúrbios digestivos - sim(1) não(0)
Hist_Mastite	0-1	Há histórico de mastite - sim(1) não(0)
Hist_Claudicao	0-1	Há histórico de claudicação - sim(1) não(0)
DiasNaoGravida	Dias	Dias não grávida
NInseminada	Número	Número de vezes que a vaca foi inseminada
DuracaoGestacao	Dias	Duração da gestação
DiasPrimeiraParicao	Dias	Idade quando pariu o primeiro bezerro
DuracaoLacAnterior	Dias	Duração da lactação anterior

Objetivos

Neste estudo, os objetivos estão concentrados na Fase I, que busca caracterizar o padrão de doença, avaliando como os parâmetros associados às variáveis de interesse se comportam com relação ao status do animal. As variáveis utilizadas para modelar a probabilidade do animal estar doente serão consideradas na forma escalar e funcional e a ideia é identificar através do ajuste de modelos uma combinação de vários parâmetros que informem sobre a saúde do animal dentro de um determinado período de tempo. Por exemplo, é de interesse predizer a probabilidade de uma animal ficar doente no período de cinco dias após o parto a partir da curva de lactação do período. Para modelar a probabilidade de um animal ficar doente consideramos um modelo de regressão logística com covariáveis escalares (número de inseminações, histórico de doenças, etc) e funções de covariáveis, como por exemplo uma função dos dias em lactação do animal.

1 Modelos estatísticos

Neste estudo, consideramos o ajuste de dois modelos. O primeiro modelo considera como variável resposta uma variável aleatória binária que vale 1 se a vaca ficou ao menos uma vez doente num determinado período de lactação e 0, caso constrário. Já no segundo modelo, consideramos todas as medidas do estado do animal (doente, não doente) num intervalo do período de lactação. Neste modelo, ajustamos o estado do animal no dia t considerando como covariável o estado do animal nos $t - r$ dias anteriores a t. Assim, por exemplo, para o estado do animal no dia do parto levamos em consideração o estado do animal nos $t - r$ dias anteriores ao dia do parto. Veja que havendo um total de R estados de saúde observados, se $r < R$, teremos que acomodar a correlação entre as medidas restantes do animal (aquelas que não entram como covariáveis), de tal forma que consideramos um efeito aleatório no modelo para acomodar marginalmente essa correlação.

Considerando o primeiro modelo, seja Y_i o estado de saúde da vaca i (1 para pelo menos uma vez doente no período de estudo e 0, caso contrário), e seja $E(Y_i) = p_i$ a probabilidade do animal i estar pelo menos uma vez doente nesse período. Inicialmente ajustamos o modelo de regressão logística (Agresti, 2003) da forma

$$\text{logito}(p_i) = \beta_0 + \sum_{j=1}^{J} \beta_j x_{ij}, \tag{1}$$

onde $\text{logito}(p_i) = log(p_i/(1 - p_i)$, x_{ij} representa a j-ésima variável explicativa da vaca i e β_j, o efeito desta variável no logito. Para essa abordagem, cada vaca possui uma única observação.

Considerando o segundo modelo, acomodamos as medidas repetidas de cada animal tomadas entre o dia do parto e 20 dias após este evento. Consideramos um efeito aleatório $u_i \sim N(0, \sigma^2)$ para acomodar a correlação das observações do mesmo. Suponha que Y_{it} seja o estado de saúde da vaca i (1 para doente e 0, caso contrário) em um determinado dia t, e seja $E(Y_{it}) = p_{it}$ a probabilidade do animal i estar doente no dia t. Consideramos o ajuste de p_{it} a partir de variáveis escalares e funcionais, dado por

$$\text{logito}(p_{it}) = \beta_0 + \sum_{j=1}^{J} \beta_j x_{ij(t-1,t-2,\ldots,t-p)} + \sum_{k=1}^{K} f_k(x_{ik(t-1,t-2,\ldots,t-p)}) + u_i. \tag{2}$$

Aqui, $\text{logito}(p_{it}) = log(p_{it}/(1 - p_{it})$; j indexa a variável explicativa escalar e k, a funcional; J e K são o número de covariáveis escalares e covariáveis funcionais no modelo, respectivamente. β_j é efeito das covariável escalar j no logito; $f_k(x)$ são funções suaves de $x_{ik(t-1,t-2,\ldots,t-p)}$; $t - 1$, $t - 2$ e $t - p$ indexam 1, 2 e p dias anteriores ao dia t, respectivamente. A função $f_k(x)$ será estimada através de

$$\hat{f}_k(x) = \sum_{r=1}^{R} B_r(x)\hat{\theta}_r, \tag{3}$$

onde B_r são as funções bases e $\hat{\theta}_r$ seus respectivos coeficientes.

Utilizamos o mecanismo *stepwise* para seleção de variáveis no modelo tendo como critério de qualidade de ajuste do modelo o AIC (Bozdogan, 1987). O *stepwise* é um processo iterativo para seleção de variáveis explanatórias. Para mais informações sobre o método stepwise, veja Yamashita, Yamashita e Kamimura (2007).

Neste trabalho, consideramos um nível de significância igual a 0,05. Para realizar o ajuste da regressão logística em (1) utilizamos a própria função base *glm* do software R (R Core Team, 2019). O modelo em (2) foi ajustado por meio da função *gamm* do pacote *mgcv* do programa R. As funções suaves $f_k(x)$ (3) foram estimadas através da expansão de bases de splines de regressão cúbica pela própria função *gamm*. Na função *gamm*, o modelo é ajustado através da maximização da quasi-verossimilhança penalizada. Para mais informações sobre ajuste de modelos GLM com variáveis funcionais e efeitos aleatórios veja Wood (2017).

2 Estrutura e manipulação dos dados

O conjunto de dados utilizado para este trabalho apresenta 71 observações para cada uma das 500 vacas estudadas, coletadas 30 dias antes do parto até 40 dias depois, resultando em um total de 35500 observações.

Além disso, o conjunto de dados apresentou diversos dados faltantes nas covariáveis. Para contornar esse problema tomou-se o seguinte procedimento: se o animal apresentou pelo menos um dado faltante para uma variável, o dado faltante foi substituído pela média dos dados não faltantes. Para o animal sem nenhuma informação numa determinada variável, o valor imputado foi a mediana da variável para as vacas com informação.

3 Validação cruzada

Para avaliação da capacidade preditiva dos modelos, utilizamos a técnica de validação cruzada. Para isso, primeiramente, separamos o conjunto de dados entre 10 envelopes, cada um contendo observações de 50 animais escolhidos aleatoriamente. Colocamos a restrição que as observações de um mesmo animal não podem estar em mais de um envelope.

Na sequência, realizamos 10 repartições e, em cada repartição, os modelos foram treinados com as observações de 9 dos 10 envelopes; o envelope faltante foi utilizado para o teste dos modelos. Veja a Figura 1 que representa um diagrama ilustrativo da validação cruzada que foi empregada. Para mais informações sobre validação cruzada veja Kohavi (2007).

A forma de checagem da performance dos modelos se deu através da área sob a curva da ROC (Bradley, 1997; Ballings and Van den Poel, 2013) e da acurácia (Novaković et al. 2017) do ajuste aos dados do conjunto de teste. A área sob a curva da ROC quantifica a capacidade do modelo em discriminar entre aqueles animais classificados como doentes dado que de fato são doentes e aqueles classificados como sadios, dado que são sadios. Um modelo com capacidade de predição tão informativo quanto um classificador aleatório tem uma área igual a 0,5, enquanto que um modelo perfeito (predito igual ao observado) tem uma área sob a curva igual a 1. Por outro lado, a acurácia indica qual é a porcentagem de predições corretas que o modelo é capaz de realizar.

Com o intuito de avaliar a capacidade de predição do estado de saúde de um animal que venha fazer parte do rebanho, consideramos que animais no conjunto de treino não fazem parte do conjunto de dados para fins de teste simultaneamente.

As funções do pacote R utilizadas para realizar a validação cruzada e a checagem da qualidade de predição foram desenvolvidas pelo autor do relatório com o auxílio do pacote *Tidyverse* (Wickham et

al. 2019). As funções escritas em linguagem R pelo autor se encontram disponíveis no seguinte link: https://github.com/rodrigoforti2000/funcoes-uteis.

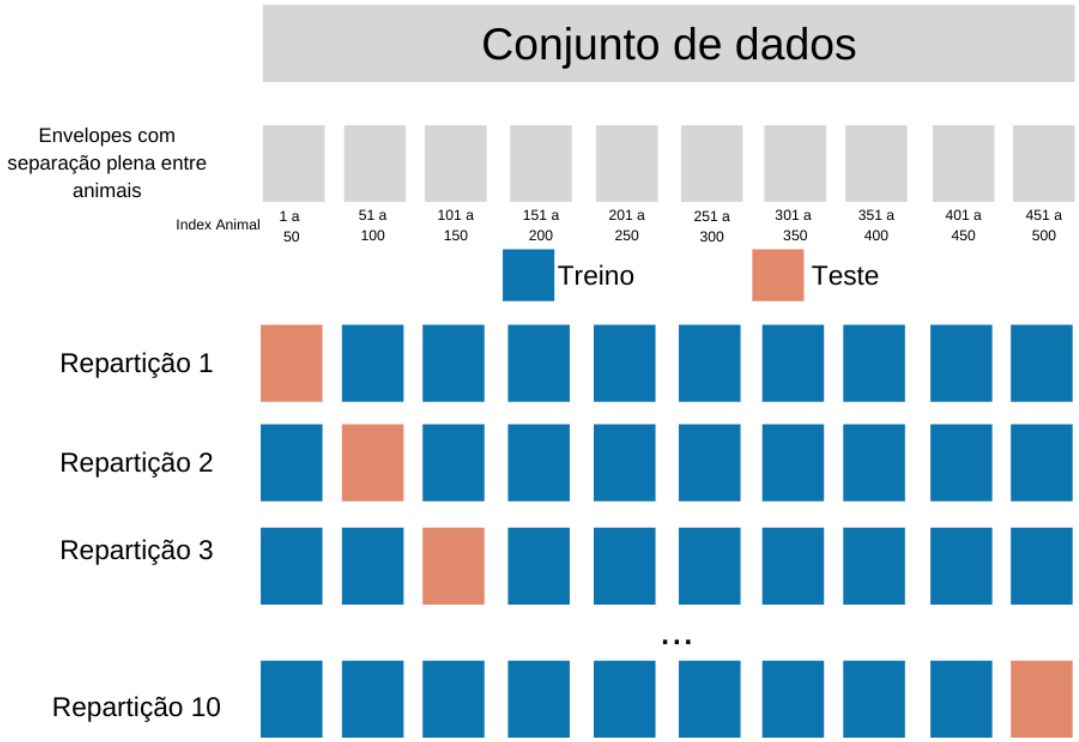

Figura 1: Diagrama ilustrativo do esquema de validação cruzada empregado neste trabalho. O conjunto de dados foi separado em 10 envelopes e em cada envelope há observações de 50 animais escolhidos aleatoriamente. Além disso, as observações de um mesmo animal pertencem à apenas um dos envelopes. Após a separação dos dados em envelopes, ocorreu 10 repartições, onde 9 envelopes foram utilizados para o ajuste e 1 para o teste do modelo.

4 Resultados

Como mostra a Figura 2, as vacas ficaram doentes majoritariamente no período entre o dia do parto (dia 0) e 20 dias após esse evento. Após esse período, menos de 1% das observações são de vacas em estado doentio, o que dificulta a predição, pois o estado doentio se torna um evento raro. Por isso, trabalhamos apenas com os dados entre o parto e o vigésimo dia após o parto.

Foram utilizadas duas abordagens para resolver o problema de classificação do estado de saúde dos animais entre os dias 0 a 20, após o parto. A primeira abordagem faz o ajuste de um modelo a partir de uma única observação do estado de saúde do animal. A segunda abordagem faz o ajuste de um modelo considerando medidas repetidas de cada animal.

4.1 Resultados do ajuste da probabilidade da vaca ter adoecido pelo menos 1 vez no período entre o dia 5 e 20 dias após o parto

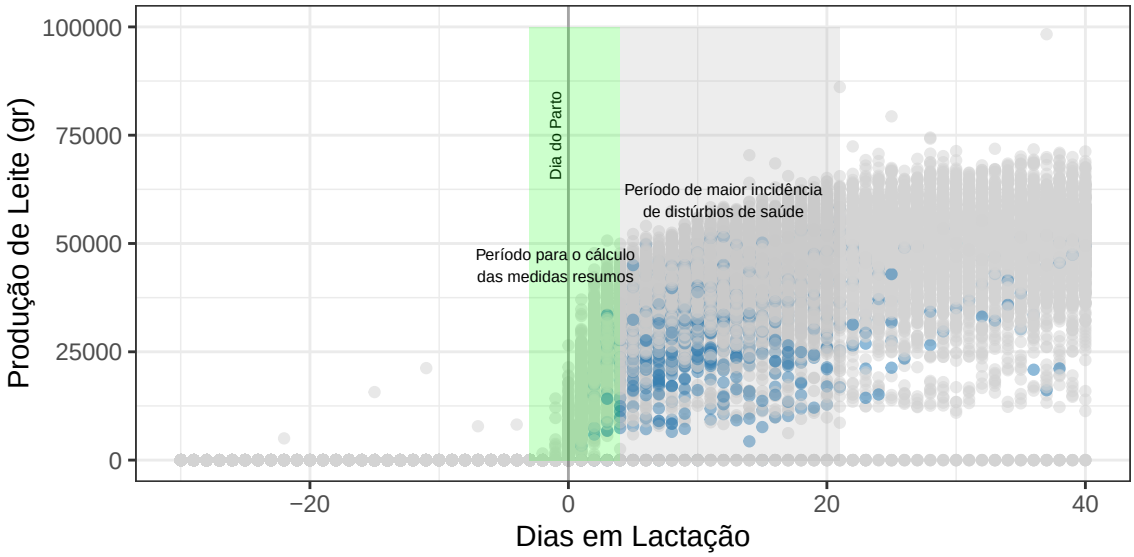

Figura 2: Produção de leite entre os dias em torno do parto. A cor azul indica distúrbio de saúde e o cinza, estado saudável.

Aqui, consideramos como resposta se a vaca ficou doente pelo menos uma vez entre os dias 5 a 20 após o parto. Se o animal ficou doente pelo menos uma vez nesse período, a variável resposta $Y_i = 1$ e caso contrário, $Y_i = 0$. Neste modelo, as covariáveis x_{ij} do modelo em (1) corresponde à média da $j - ésima$ variável no período 3 dias antes e 4 dias depois do parto da vaca i.

Utilizando essa abordagem, aproximadamente um terço das vacas estudadas apresentaram pelo menos um distúrbio de saúde no período entre 5 e 20 dias após o parto.

O modelo em (1) foi ajustado inicialmente considerando todas as covariáveis disponíveis da Tabela 1 e 2 na forma escalar. O método *stepwise* foi empregado para a seleção de covariáveis, pois há uma quantidade considerável de variáveis explanatórias no conjunto de dados, o que dificulta a interpretação dos parâmetros e aumenta o risco de multicolinearidade. A partir do recurso de *stepwise*, o modelo final foi dado pelo modelo com menor valor de AIC, que tem como covariáveis a média entre os dias 3 antes do parto e 4 depois do parto das variáveis referentes aos dias sem produzir leite, à gordura e à proteína no leite, ao tempo e ao número de repousos, à ruminação, aos dias não grávida e à duração da lactação anterior.

Após a definição do modelo final, o conjunto de dados foi separado em 10 envelopes e, feito isso, houve 10 reparticionamentos dos envelopes para realização da validação cruzada com o intuito de testar a capacidade preditiva do modelo. Em cada uma das repatições da validação cruzada, calculou-se a curva ROC e a acurácia, as quais encontram-se na Figura 3. Para calcular a acurácia, utilizamos um valor de corte que define se a observação i será classificada como doente ou não doente baseado na probabilidade p_i de (1). Se a observação i apresentou p_i menor que o valor de corte, esta foi classificada como não doente, e se a observação i apresentou p_i maior que o valor de corte, esta foi classificada como doente.

O gráfico da curva ROC na Figura 3 indica que uma área sob a curva média de 0,735, implicando que há 73,5% de chance que o modelo será capaz de distinguir entre observações de animais doentes e não doentes. O gráfico da acurácia na Figura 3 indica que o modelo tem uma acurácia média entre 62,5% a 75,0% em seu ápice, que é em torno de um corte igual a 0,4.

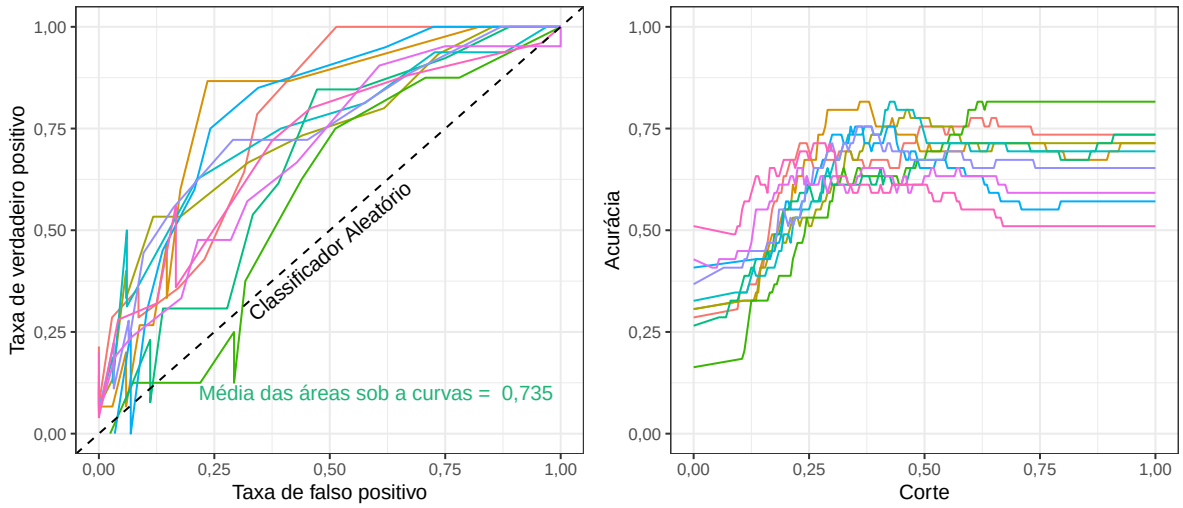

Figura 3: Curvas ROC e acurácias do modelo sem medidas repetidas alcançado por stepwise obtidas por meio de validação cruzada.

As estimativas dos parâmetros para o modelo final encontram-se na Tabela 3. Os resultados mostram que um aumento na porcentagem de gordura no leite e no número de dias que o estado da vaca não está grávida aumentam a probabilidade do animal apresentar um distúrbio de saúde após o parto, pois o efeito estimado das covariáveis é positivo. Já um aumento na quantidade de proteína no leite, no tempo de repouso, no tempo de ruminação e na alimentação diminui a probabilidade do animal ter um problema de saúde, já que o efeito estimado das covariáveis é negativo.

Tabela 3: Resultados do ajuste do modelo sem medidas repetidas obtido por stepwise.

Termo	Estimativa	Erro padrão	Estatística do teste	P-valor
Intercepto	15,344	5,963	2,573	<0,05
DiasSecos.media	-0,044	0,021	-2,098	<0,05
Gordura.media	0,606	0,201	3,013	<0,05
Proteina.media	-0,823	0,284	-2,893	<0,05
TempoRepouso.media	-0,002	0,001	-2,125	<0,05
NumRespousos.media	0,120	0,043	2,802	<0,05
RuminacaoUltimas24h.media	-0,009	0,003	-3,586	<0,05
AlimentacaoUltimas24h.media	-0,007	0,002	-3,653	<0,05
DiasNaoGravida.media	0,052	0,021	2,431	<0,05
DuracaoLacAnterior.media	-0,047	0,021	-2,219	<0,05

4.2 Resultados do ajuste dos dados longitudinais do estado de saúde da vaca no período entre o dia do parto e 20 dias após esse evento

Ajustamos o modelo (2) considerando duas abordagens: uma delas considera somente preditores na forma escalar e a outra considera preditores na forma escalar e na forma funcional. Como o processamento computacional para o ajuste desses modelos é alto, principalmente com o uso de covariáveis na forma funcional, nos guiamos pelos efeitos significativos apresentados na Tabela 3. Consideramos as covariáveis na forma escalar a ruminação, a alimentação e os estados de saúde dos dois dias anteriores.Os outros preditores da Tabela 3 não apresentaram alterações significativas, por isso nós os retiramos do modelo. Além disso, nas duas abordagens um efeito aleatório para cada vaca foi considerado para acomodar a correlação marginal entre as medidas do animal. Chamaremos o modelo da primeira abordagem como modelo misto escalar (MME) e o da segunda de modelo misto escalar-funcional (MMEF).

Para predizer o estado do animal no dia t, do ponto de vista biológico, fez sentido considerar o efeito das covariáveis sobre o estado de saúde do animal que foram medidas no dia $t-1$ e $t-2$. Não ocorreu nenhuma melhora significativa ao utilizar dias anteriores à $t-2$. Sendo assim, para o modelo em (2), consideramos o estado do animal i no dia t como função das covariáveis $x_{i,1,(t-1,t-2)}, \ldots, x_{i,J,(t-1,t-2)}$. Desta forma, para o dia do parto utilizamos informações das variáveis medidas nos dois dias anteriores e assim sucessivamente até o vigésimo dia após esse evento.

Além das covariáveis referentes à ruminação, à alimentação e ao estado de saúde das últimas 24 e 48 horas, adicionamos ao modelo as covariáveis que são a interação entre os efeitos de mesma natureza dos dois dias anteriores. Por exemplo, a interação entre estado de saúde no dia $t-1$ e $t-2$.

As curvas ROC e a acurácias do MME obtidas por meio de validações cruzadas indicam que o poder de classificação do modelo é bastante satisfatório, pois a área sob a curva e a acurácia estão próximas de 0,9, como mostram os gráficos da Figura 4.

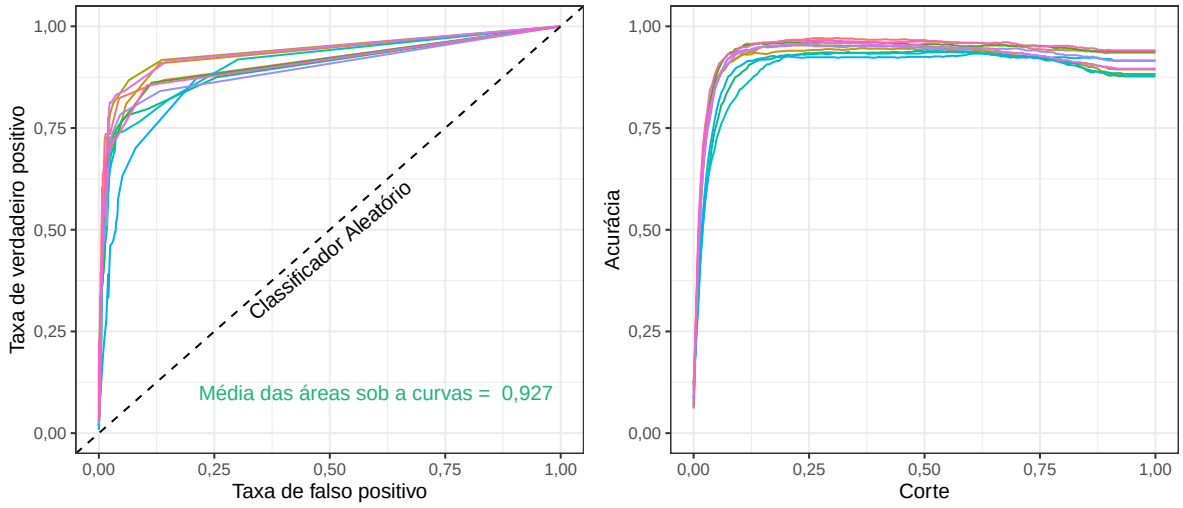

Figura 4: Curvas ROC e acurácias do MME obtidas por meio de validação cruzada.

Para o MMEF, a variável referente aos dias de lactação foi utilizada como funcional. Entre o dia do parto até quinto dia após esse evento, 5,4% das observações das vacas são doentes. Entre o quinto e o

décimo dia, 16,7%. Entre o décimo e o décimo quinto, 11,4%. E entre o décimo quinto e o vigésimo dia após o parto, 4,8%, o que indicia a aparente não linearidade da variável dias de lactação em relação ao estado de saúde dos animais. Logo, nesse caso, uma variável referente aos dias de lactação do tipo funcional se encaixou melhor no modelo. As outras variáveis apresentaram comportamentos lineares e foram mantidas na forma escalar.

As curvas ROC e a acurácia do MMEF, apresentados na Figura 5, apontam que a capacidade de predição é muito boa, pois ambas das métricas foram maiores que 0,9, indicando um bom poder preditivo do modelo, superando os resultados no MME. A média da área sob a curva da ROC e a média de acurácia do MMEF aumentaram 0,007 e 0,0012 em relação ao modelo sem a variável na forma funcional, respectivamente.

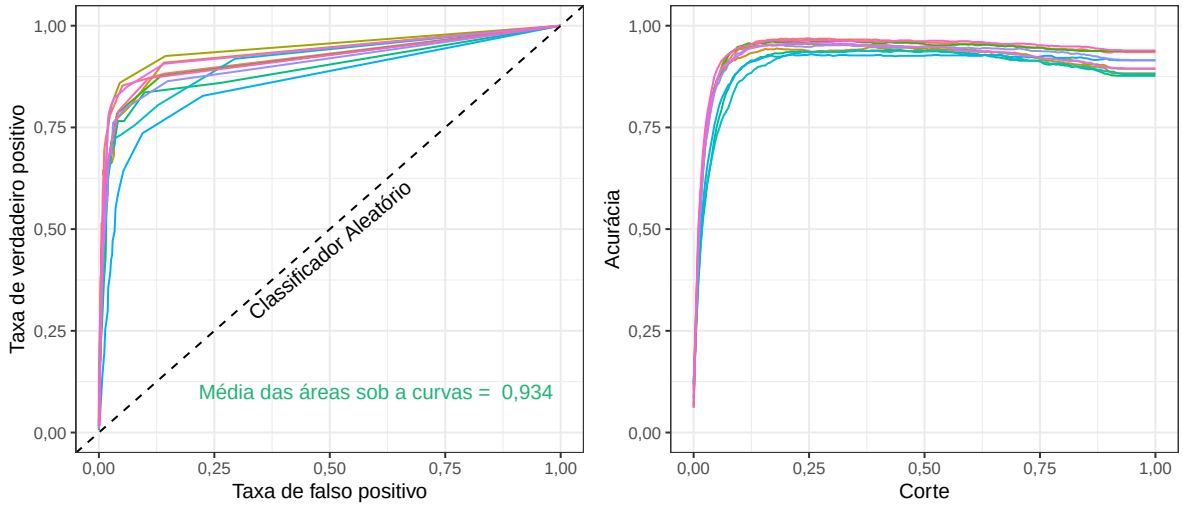

Figura 5: Curvas ROC e acurácias do MMEF obtidas por meio de validação cruzada.

As estimativas dos efeitos escalares do MMEF se encontram na Tabela 4. O efeito da ruminação nas últimas 24 e 48 horas, assim como o estado de saúde do animal nos dois dias anteriores tem efeito positivo na probabilidade do animal adoecer no dia t, indicando que a probabilidade da vaca ficar doente aumenta quando há um aumento na atividade de ruminação ou se a vaca ficou doente nos dois dias anteriores. A alimentação nas últimas 24 horas tem efeito negativo na probabilidade de adoecer, assim a chance de ficar doente diminui quando a vaca está bem alimentada. Além disso, como a estimativa da interação entre o estado de saúde dos dias anteriores tem efeito positivo na probabilidade, isso nos diz que há um aumento na probabilidade de vir a ficar doente quando o animal esteve doente nos dois dias anteriores em comparação à probabilidade de ter estado doente em apenas um dos dias anteriores.

A Figura 6 exibe a curva estimada a variável funcional referente ao dia de lactação. A estimativa dos coeficientes de suavização, $\hat{\theta}_r$, em (3) para a curva de dia de lactação mostrou se estatisticamente significante já que a função estimada $s(DiasLac)$ tem p-valor menor do que 0,001, como descrito na Tabela 5, implicando que não há linearidade na relação entre dia de lactação e probabilidade do animal estar doente.

Tabela 4: Estimativas dos parâmetros do ajuste do modelo misto escalar-funcional. Os sufixos k1 e k2 são referentes à k-1 e k-2, respectivamente.

Termo	Estimativa	Erro padrão	Estatística do teste	P-valor
Intercepto	-3,08e+00	0,440	-7,015	<0,05
RuminacaoUltimas24h_k1	3,60e-03	0,002	2,164	<0,05
RuminacaoUltimas24h_k2	5,99e-03	0,002	3,908	<0,05
AlimentacaoUltimas24h_k1	-8,04e-04	0,002	-0,372	0,710
AlimentacaoUltimas24h_k2	3,55e-03	0,002	1,752	0,080
doente_k1	2,70e+00	0,162	16,619	<0,05
doente_k2	9,41e-01	0,222	4,245	<0,05
RuminacaoUltimas24h_k1:RuminacaoUltimas24h_k2	-3,48e-05	0,001	-5,280	<0,05
AlimentacaoUltimas24h_k1:AlimentacaoUltimas24h_k2	-2,09e-05	0,001	-2,827	<0,05
doente_k1:doente_k2	4,76e-01	0,289	1,650	0,100

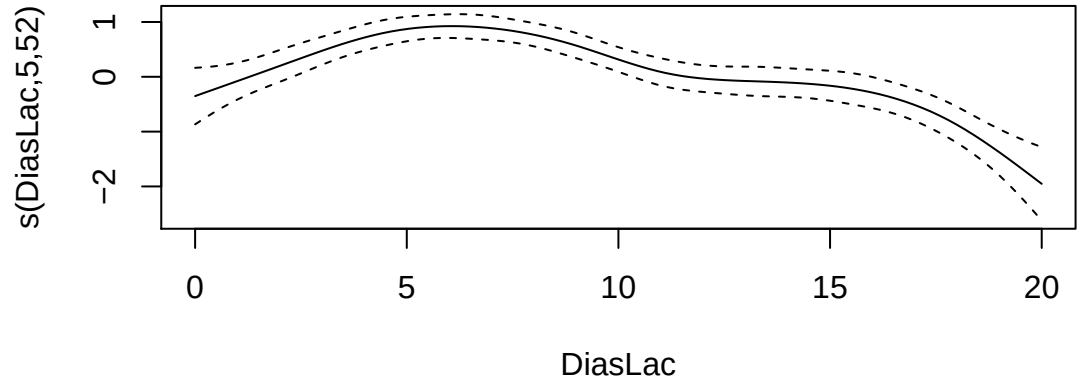

Figura 6: Curva estimada de dia de lactação (DiasLac)

Tabela 5: P-valor do efeito da função de suavização s(DiasLac) no modelo misto escalar-funcional. GLE é referente aos graus de liberdade efetivos.

Termo	GLE	Estatística do teste	P-valor
s(DiasLac)	5,519	19,107	<0,001

A maior parte das observações entre o dia do parto até os vinte dias posteriores a esse evento são de estado de saúde saudável, cerca de 90%. Com isso, um modelo que predissesse todas as observações como saudáveis também teria uma acurácia de 90%. Para investigar mais a fundo como ocorre a classificação do estado de saúde no MMEF, construiu-se o gráfico da Figura 7. Ao invés de analisar a acurácia geral do

modelo, dividiu-se as observações dos envelopes de teste da validação cruzada em intervalos referentes a quantos dias a vaca apresentou ficar doente. Por exemplo, o intervalo 0 contém observações de animais que não ficaram doentes entre o dia do parto e os 20 dias seguintes, já o intervalo de 1 a 5 contém observações de animais que ficaram doentes entre 1 a 5 vezes nesse período, e assim por diante. No gráfico da Figura 7, cada ponto transparente é a acurácia no intervalo em uma repartição da validação cruzada, já os pontos sólidos são a média de acurácia nos intervalos. Podemos ver que o modelo possui uma acurácia perfeita entre as vacas que não ficaram doentes em nenhum dos dias considerados, e cai quando se avança nos intervalos, onde os animais ficaram doentes pelo menos uma vez. Entretanto, a média da acurácia se manteve maior que 75% nesses intervalos. Assim, o modelo ajustado apresenta ser superior a um classificador que prediz todas as observações como doente.

Ainda na Figura 7, há uma linha demarcando a acurácia média do MMEF. A tabela no gráfico mostra a quantidade média de observações dos intervalos de vezes que as vacas ficaram doentes entre o parto e os 20 dias posteriores em cada envelope da validação cruzada. Por exemplo, em um envelope há em média 643 observações de animais, contendo medidas repetidas do mesmo animal, que não ficaram doentes durante esse período e 217 medições de animais que ficaram doentes entre 1 a 5 vezes, e assim por diante.

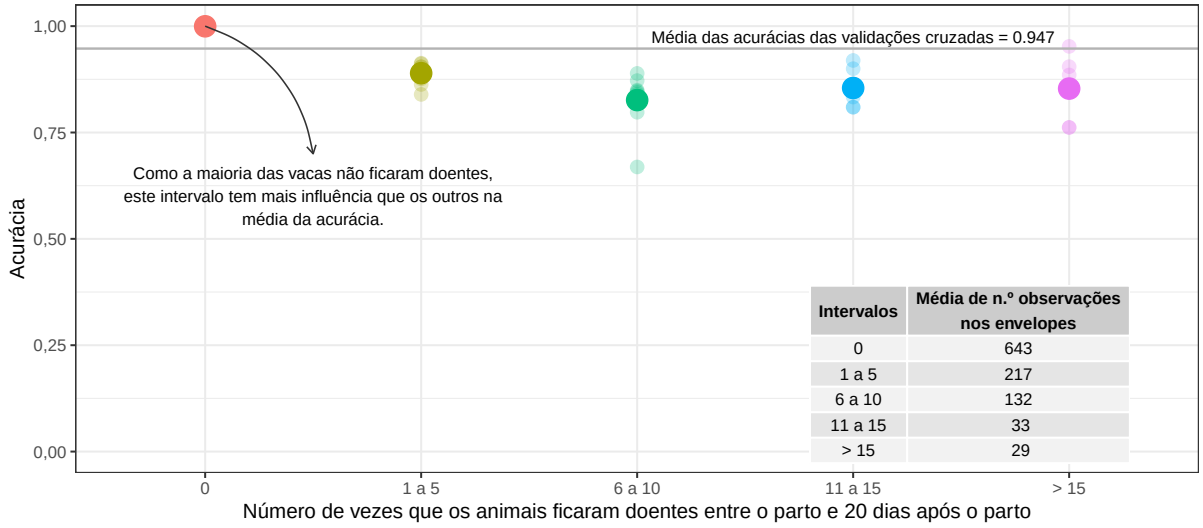

Figura 7: Acurácia em intervalos de quantas vezes as vacas do conjunto de teste ficaram doentes entre o dia do parto e os 20 dias posteriores, obtida pela validação cruzada.

5 Conclusão

Os modelos ajustados neste trabalho auxiliam o pecuarista em relação aos cuidados com as vacas leiteiras, pois apresentam boas capacidades de predição do estado de saúde futuro dos animais.

A incidência de doenças nas vacas leiteiras se concentra entre o dia do parto e os 20 primeiros dias após esse evento. Depois desse período, os distúrbios de saúde tornam-se um evento raro.

Quando consideramos os dados de 3 dias antes do parto até 4 dias depois desse evento para predizer se a vaca ficará doente entre os dias 5 até o 20 após o parto, temos que um aumento na porcentagem de gordura no leite e no número de dias que o estado da vaca não está grávida aumentam a probabilidade do animal apresentar um distúrbio de saúde após o parto. Já um aumento na quantidade de proteína no leite, no tempo de repouso, no tempo de ruminação e na alimentação diminui a probabilidade do animal ter um problema de saúde.

Quando consideramos dados dos dias $t-1$ e $t-2$ para predizer o estado de saúde do animal no dia t, temos que o efeito da ruminação nas últimas 24 e 48 horas, assim como o estado de saúde do animal nos dois dias anteriores tem efeito positivo na probabilidade do animal adoecer, indicando que a probabilidade da vaca ficar doente aumenta quando há um aumento na atividade de ruminação ou se a vaca ficou doente nos dois dias anteriores. A alimentação nas últimas 24 horas tem efeito negativo na probabilidade de adoecer, assim a chance de ficar doente diminui quando a vaca está bem alimentada. Além disso, utilizar a covariável referente aos dias em lactação em forma funcional traz uma melhora na capacidade de predição do estado de saúde.

6 Realizações do período

Durante a segunda quinzena de fevereiro, foi realizado o estudo bibliográfico para validação cruzada, cujas funções foram criadas em março, junto dos gráficos para a checagem da qualidade de predição do modelo. Ainda em março, iniciou-se a escrita do relatório, a qual terminou em junho. Ao longo de todo esse período, ocorreram reuniões semanais com a orientadora, a fim de esclarecer dúvidas gerais sobre o projeto.

Referências bibliográficas

Agresti, A., 2003. Categorical data analysis. John Wiley & Sons.

Ballings, M., Van den Poel, D., 2013. AUC: Threshold independent performance measures for probabilistic classifiers.

Bell, A., Fairbrother, M., Jones, K., 2019. Fixed and random effects models: Making an informed choice. Quality & Quantity 53, 1051–1074. https://doi.org/10.1007/s11135-018-0802-x

Bozdogan, H., 1987. Model selection and Akaike's Information Criterion (AIC): The general theory and its analytical extensions. Psychometrika 52, 345–370.

Bradley, A.P., 1997. The use of the area under the roc curve in the evaluation of machine learning algorithms. Pattern recognition 30, 1145–1159.

Drackley, J.K., 1999. Biology of dairy cows during the transition period: The final frontier? Journal of Dairy Science 82, 2259–2273. https://doi.org/https://doi.org/10.3168/jds.S0022-0302(99)75474-3

Goff, J., Horst, R., 1997. Physiological changes at parturition and their relationship to metabolic disorders1, 2. Journal of dairy science 80, 1260–1268.

Kaufman, E., Asselstine, V., LeBlanc, S., Duffield, T., DeVries, T., 2018. Association of rumination time and health status with milk yield and composition in early-lactation dairy cows. Journal of dairy science 101, 462–471.

Kohavi, R., 1995. A study of cross-validation and bootstrap for accuracy estimation and model selection. IJCAI'95 1137–1143.

Lago, A.V.A.S., Ernani Paulino do AND Pires, 2001. Efeito da condição corporal ao parto sobre alguns parâmetros do metabolismo energético, produção de leite e incidência de doenças no pós-parto de vacas leiteiras. Revista Brasileira de Zootecnia 30, 1544–1549.

Novaković, J.D., Veljović, A., Ilić, S.S., Papić, Ž., Milica, T., 2017. Evaluation of classification models in machine learning. Theory and Applications of Mathematics & Computer Science 7, Pages: 39.

R Core Team, 2020. R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria.

Wickham, H., Averick, M., Bryan, J., Chang, W., McGowan, L.D., François, R., Grolemund, G., Hayes, A., Henry, L., Hester, J., Kuhn, M., Pedersen, T.L., Miller, E., Bache, S.M., Müller, K., Ooms, J., Robinson, D., Seidel, D.P., Spinu, V., Takahashi, K., Vaughan, D., Wilke, C., Woo, K., Yutani, H., 2019. Welcome to the tidyverse. Journal of Open Source Software 4, 1686. https://doi.org/10.21105/joss.01686

Wilcoxon, F., 1945. Individual comparisons by ranking methods. Biometrics Bulletin 1, 80–83.

Wood, S.N., 2017. Generalized additive models: An introduction with r. CRC press.

Yamashita, T., Yamashita, K., Kamimura, R., 2007. A stepwise aic method for variable selection in linear regression. Communications in Statistics - Theory and Methods 36, 2395–2403. https://doi.org/10.1080/03610920701215639