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

Pesquisador 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 e Mariana Rodrigues Motta
- 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 período de lactação precoce (também chamado período periparturiente) é certamente o estágio mais importante do ciclo de lactação destes animais, pois maioria das doenças infecciosas e distúrbios metabólicos ocorre 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 peri-parturiente que leva, por exemplo, ao aumento de incidência de mastite no animais. Assim, a ocorrência de problemas de saúde se concentra proporcionalmente 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 de que vacas de alta produção estão sob no 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 processo de mobilização parece afetar o bem-estar da vaca e outras vias biológicas comprometidos à medida que a energia de entrada é direcionada para 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 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 em produção de leite e eventos das lactações anteriores e da parição atual. Além disso, os dados foram coleatdos 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 conductividade, 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 vacas e leite coletado. O conjunto de dados principal corresponde ao exame clínico diário dos animais 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 Tabela 1 e 2.

Objetivos

Neste estudo, os objetivos estão concentrados na Fase I, que busca caracterizar o padrão de doença, avaliando como os parâmetros 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 de função 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.

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

1 Modelos estatísticos

Seja Y_i o estado de saúde da vaca i (1 para doente e 0, caso contrário) em um determinado período, e seja $E(Y_i) = p_i$ a probabilidade do animal i estar 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.

Posteriormente, ajustamos o modelo que acomoda 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 animal, devido a repetição de medições das vacas em tempos diferentes. Suponha que Y_{ik} seja o estado de saúde da vaca i (1 para doente e 0, caso contrário) em um determinado dia k, e seja $E(Y_{ik}) = p_{ik}$ a probabilidade do animal i estar doente no dia k. Consideramos o ajuste de

$$\text{logito}(p_{ik}) = \beta_0 + \sum_{j=1}^{J} f_j(x_{ij(k-1,k-2)}) + u_i. \tag{2}$$

Aqui, $\text{logito}(p_{ik}) = log(p_{ik}/(1-p_{ik})$, $f_j(x)$ são funções suaves de $x_{ij(k-1,k-2)}$, $k-1$ e $k-2$ indexam 1 e 2 dias anteriores ao dia k, respectivamente, e j indexa a variável explicativa. A função $f_j(x)$ será estimada através de

$$\hat{f}_j(x) = \sum_{p=1}^{P} B_p(x)\hat{\theta}_p, \tag{3}$$

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

Utilizamos o mecanismo *stepwise* para seleção de variáveis no modelo tendo como critério 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 $f_j(x)$ (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 substituido 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, foi realizada uma validação cruzada com 10 reparticionamentos e com 10 envelopes contendo observações de 50 vacas cada, pois há 500 animais no banco de dados. Em cada repartição, os envelopes foram separados em duas partes: 9 envelopes foram utilizados para o ajuste, i.e. treino, e um envelope foi usado para o verificação de qualidade de predição do modelo, i.e. teste . Veja a Figura 1 que representa um diagrama simplificado da validação cruzada que foi empregado. Para mais informações sobre validação cruzada veja Kohavi (2007).

A forma de checagem da performance do modelo 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) no 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. Um modelo com capacidade de predição tão informativo quanto um classificador aleatório tem uma área igual a 0,5, enquanto que um modelo perfeito 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 modelo para animais que futuramente venham a fazer parte do local onde foram coletados os dados, simulamos esta situação considerando 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.

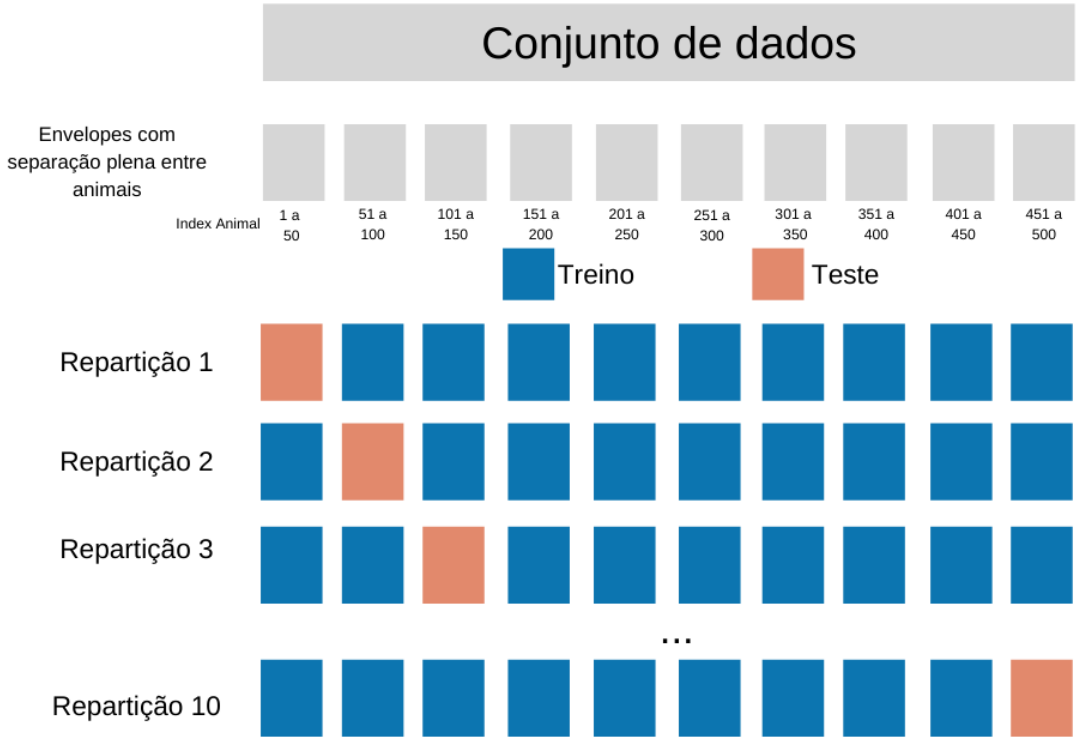


Figura 1: Diagrama simplificado da validação cruzada empregado neste trabalho. O conjunto de dados foi separado em 10 envelopes e em cada envelope há observações de 50 animais. Além disso, não há observações de um mesmo animal em envelopes diferentes. Após isso ocorreu 10 repartições, onde 9 envelopes eram utilizados para treino 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 até 20 dias após o parto. 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 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 do modelo sem medidas repetidas de saúde da vaca

Para a primeira parte ajustamos o modelo em (1). Caso a vaca tenha ficado doente pelo menos uma vez entre os dias 5 a 20 após o parto, a variável resposta $Y_i = 1$ e caso contrário, $Y_i = 0$. Neste modelo,

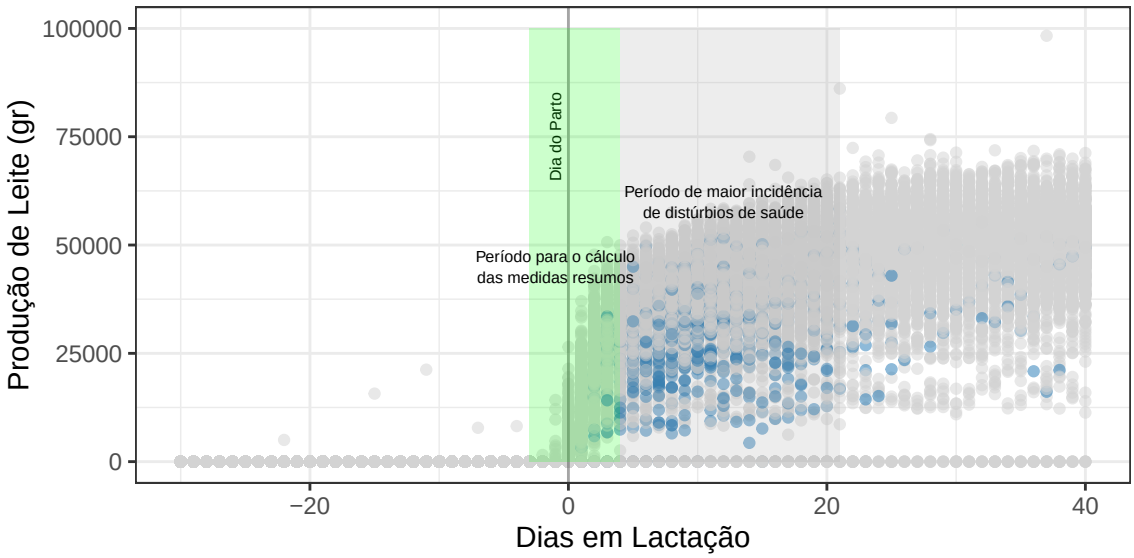


Figura 2: Produção de leite entre os dias em torno do parto. A cor azul indica distúrbio de saúde e o cinza, estado saudável.

para todo j a variável x_{ij} do modelo em (1) corresponde à média da variável j no período 3 dias antes e 4 dias depois do parto.

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 de *stepwise* tendo como critério o AIC, o modelo com menor AIC foi escolhido como melhor opção para essa Seção 3.1, pois apresentou acurácia e área sob a curva da ROC maiores que o modelo com todas as covariáveis.

Após a seleção das variáveis por meio de stepwise, o conjunto de dados foi repartido 10 vezes para realização da validação cruzada com o intuito de testar a capacidade preditiva do modelo. Cada uma das iterações da validação cruzada, calculo-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 classificado 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 classificado como não doente, e se a obsevação i apresentou p_i maior que o valor de corte, esta foi classificada como doente.

Podemos ver que pela Figura 3 que o modelo tem uma acurácia em média de 62,5% a 75,0% em seu ápice, em torno de um corte igual a 0,4. Já as curvas ROC possui uma área sob a curva 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.

A partir do ajuste do modelo final (modelo com menor AIC), os resultados na Tabela 3 mostram que um aumento média de porcentagem de gordura no leite e o número médio de dias que o estado da vaca

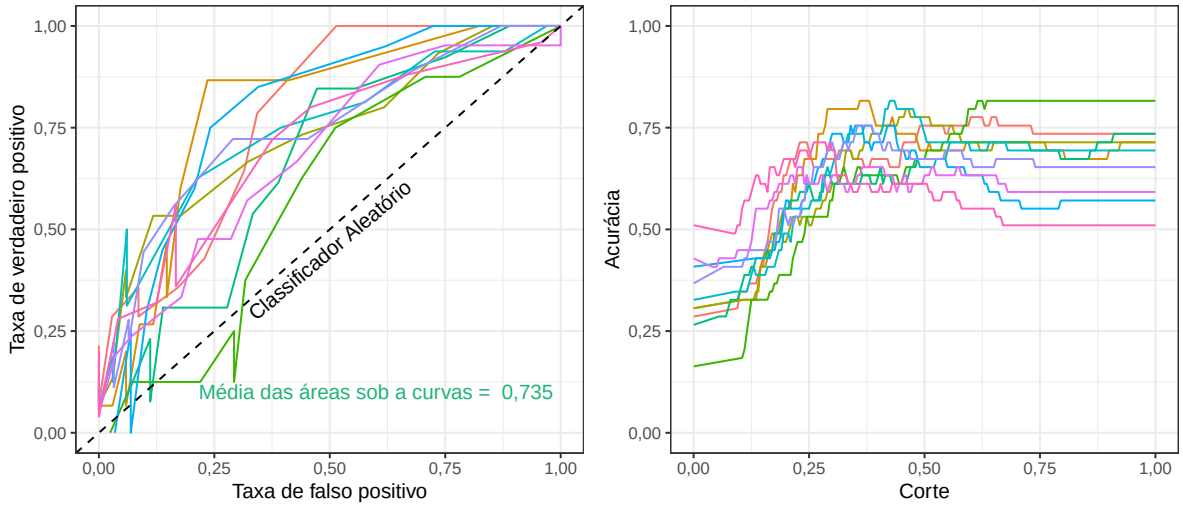


Figura 3: Curvas ROC e acurácias do modelo sem medidas repetidas alcançado por stepwise obtidas por meio de validação cruzada.

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 média de proteína no leite, o tempo médio de repouso, o tempo médio de ruminação e a média alimentação diminuem 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 do modelo com medidas repetidas de saúde da vaca

Ajustamos o modelo (2) considerando duas abordagens: uma delas considera as variáveis somente na forma escalar e a outra considera as variáveis na forma escalar ou funcional. Como o processamento computacional para o ajuste desses modelos é alto, principalmente com o uso de variáveis na forma funcional, consideramos como covariáveis a ruminação, alimentação e estados de saúde dos dois dias

antereiores, pois mostraram-se significativas no modelo da Seção 4.1 e apresentaram-se serem promissoras na análise exploratória. Além disso, nas duas abordagens um efeito aleatório para cada vaca foi considerado para acomodar a correlação 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 k, do ponto de vista biológico, fez sentido considerar o efeito das covariáveis sobre o estado de saúde do animal que foram medidas no dia $k-1$ e $k-2$. Não ocorreu nenhuma melhora significativa ao utilizar dias anteriores à $k-2$. Sendo assim, para o modelo em (2), consideramos o estado do animal i no dia k como função das covariáveis $x_{i,1,(k-1,k-2)}, \ldots x_{i,J,(k-1,k-2)}$. Desta forma, para o dia do parto utilizamos informações das variáveis medidas nos dois dias anteriores e assim sucessivamente até o vigésimo dia após o parto. Escolheu-se esse ponto de parada, pois, após o vigésimo dia, os distúrbios de saúde representam cerca de apenas 1% das observações, o que dificulta a modelagem e a predição.

Como utilizou-se informações de dias diferentes para realizar a predição do estado de saúde, logo foi possível adicionar ao modelo a interação entre variáveis de mesma natureza desses dias, por exemplo, interação entre estado de saúde no dia $k-1$ e $k-2$.

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 mostra a Figura 4.

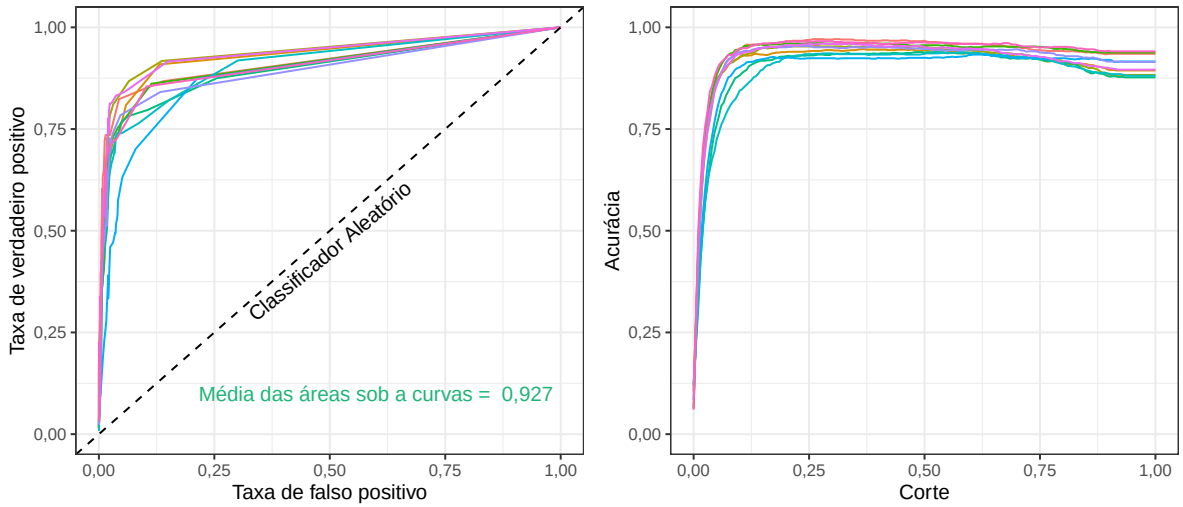


Figura 4: Curvas ROC e acurácias do MME obtidas por meio de validação cruzada.

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 são doentes. Entre o quinto e o décimo dia, 16,7% são doentes. Entre o décimo e o décimo quinto, 11,4% são doentes. E entre o décimo quinto e o vigésimo dia após o parto, apenas 4,8% são observações doentes, 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 encaixa melhor no modelo. As outras variáveis apresentaram comportamentos lineares quando analisadas separadamente em um modelo individual, ou seja, apenas com uma variável. Logo, elas 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 usadas apresentaram serem maiores que 0,9, indicando um bom poder preditivo que supera o do MME em 0,007 na média da área sob a curva e 0,0012 na média de acurácia.

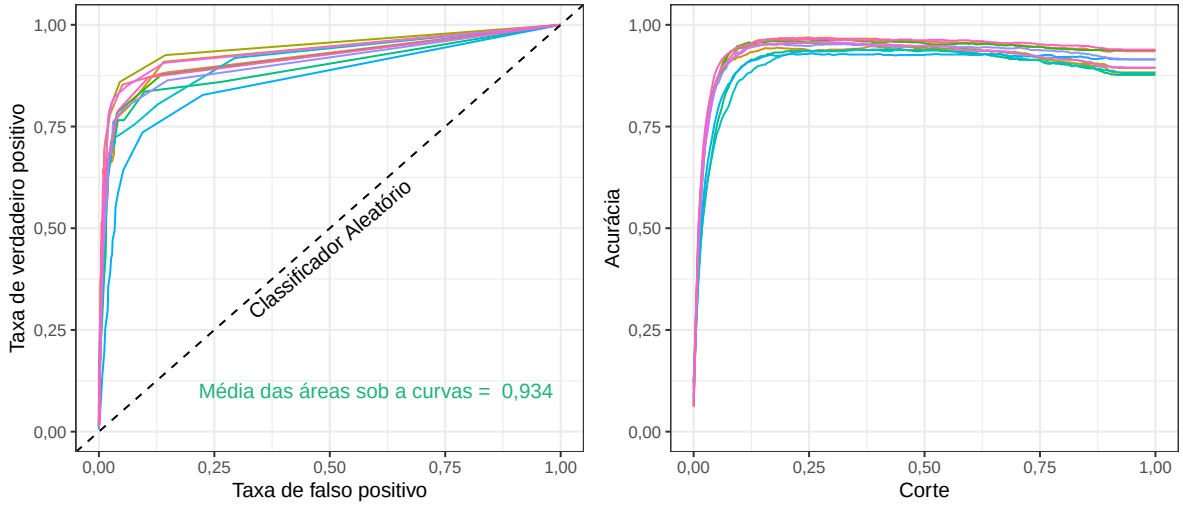


Figura 5: Curvas ROC e acurácias do MMEF obtidas por meio de validação cruzada.

As estimativas dos efeitos escalares do MMEF se encontram na Tabela 4. As variáveis referentes a ruminação durante as últimas 24 horas e estado de saúde nos dois dias anteriores tem efeito positivo no logito, 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, indicando que a probabilidade de adoecer 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 no logito, 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 dia.

A Figura 6 exibe a curva estimada a variável funcional referente ao dia de lactação. A estimativa dos coeficientes de suavização, $\hat{\theta}_p$, 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

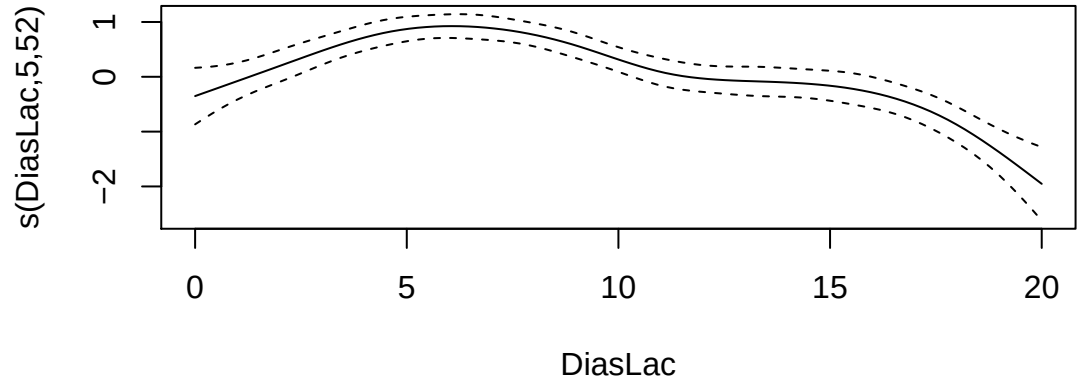


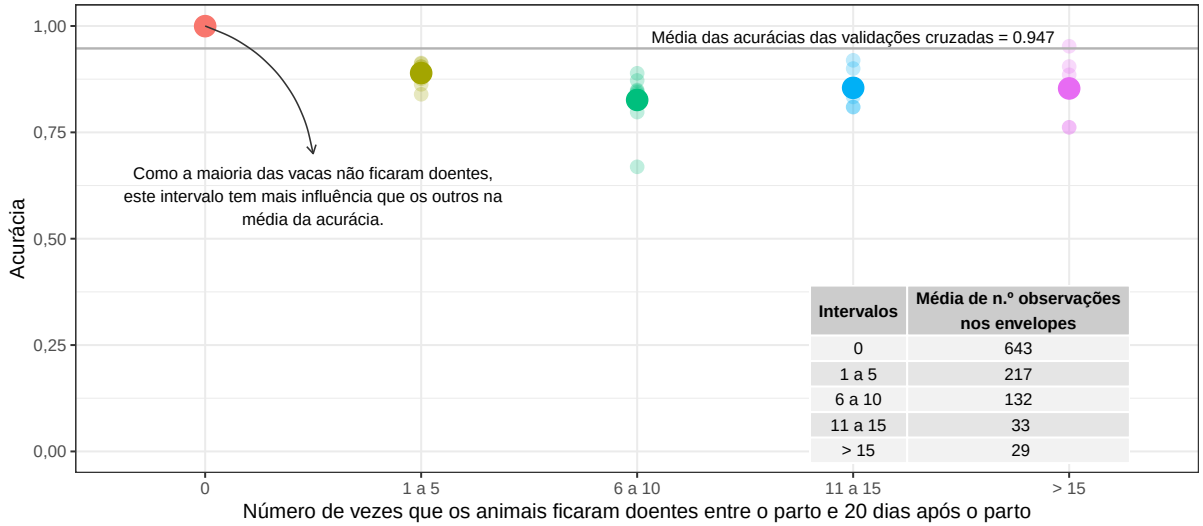
Figura 6: Curva estimada de dia de lactação (DiasLac)

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

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

teste em intervalos referentes a quantos dias a vaca apresentou ficar doente. Por exemplo, o intervalo 0 contém observações de animais que não ficaram doentes nos dias estudados, já o intervalo de 1 a 5 contém observações de animais que ficaram doentes entre 1 a 5 vezes no período analisado, e assim por diante. Cada ponto mais transparente é a acurácia no intervalo em uma repartição da validação cruzada, já os pontos opacos são a média de acurácia no intervalo no geral. Podemos ver que o modelo possui uma acurácia praticamente perfeita entre as vacas que não ficaram doentes em nenhum dos dias considerados, e cai quando se avança nos intervalos. Entretanto, a média acurácia se manteve maior que 75% nos demais intervalos. Com isso, o modelo ajustado apresenta ser superior a um classificador simplório, i.e. classifica tudo igual, pois ele consegue predizer corretamente, com uma boa acurácia, as observações marcadas com distúrbio de saúde presente. Na Figura 7, há ainda uma linha demarcando a acurácia média do MMEF no corte que otimiza essa medida. A tabela nessa figura 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, então, por exemplo, em um envelope há em média 643 observações de animais que não ficaram doentes durante esse período.



Intervalos	Média de n.º observações nos envelopes
0	643
1 a 5	217
6 a 10	132
11 a 15	33
> 15	29

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 resultados obtidos neste trabalho apresentam duas abordagens com uso de modelos para predição do estado de saúde animal no contexto do período periparturiente em vacas leiteiras. A escolha de qual abordagem dependerá das exigências do usuário. Se a pessoa interessada querer predizer com uma antecedência maior o estado do animal, é mais provável que ela tenha preferência pela abordagem sem medidas repetidas. Já se a preferência for por um modelo com uma capacidade preditiva maior e a antecedência do resultado curta não seja um empecilho, é mais provável que se opte pelo modelo com medidas repetidas.

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, Alan. 2003. *Categorical Data Analysis*. Vol. 482. John Wiley & Sons.

Ballings, Michel, and Dirk Van den Poel. 2013. *AUC: Threshold Independent Performance Measures for Probabilistic Classifiers.* https://CRAN.R-project.org/package=AUC.

Bell, Andrew, Malcolm Fairbrother, and Kelvyn Jones. 2019. "Fixed and Random Effects Models: Making an Informed Choice." *Quality & Quantity* 53 (2): 1051–74. 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: 345–70.

Bradley, Andrew P. 1997. "The Use of the Area Under the Roc Curve in the Evaluation of Machine Learning Algorithms." *Pattern Recognition* 30 (7): 1145–59.

Drackley, James K. 1999. "Biology of Dairy Cows During the Transition Period: The Final Frontier?" *Journal of Dairy Science* 82 (11): 2259–73. https://doi.org/https://doi.org/10.3168/jds.S0022-0302(99)75474-3.

Goff, JP, and RL Horst. 1997. "Physiological Changes at Parturition and Their Relationship to Metabolic Disorders1, 2." *Journal of Dairy Science* 80 (7): 1260–8.

Kaufman, EI, VH Asselstine, SJ LeBlanc, TF Duffield, and TJ DeVries. 2018. "Association of Rumination Time and Health Status with Milk Yield and Composition in Early-Lactation Dairy Cows." *Journal of Dairy Science* 101 (1): 462–71.

Kohavi, Ron. 1995. "A Study of Cross-Validation and Bootstrap for Accuracy Estimation and Model Selection," IJCAI'95,, 1137–43.

Lago, Alexandre Vaz AND Susin, 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 (October): 1544–9. http://www.scielo.br/scielo.php?script=sci_arttext&pid=S1516-35982001000600023&nrm=iso.

Novaković, Jasmina Dj., Alempije Veljović, Siniša S. Ilić, Željko Papić, and Tomović Milica. 2017. "Evaluation of Classification Models in Machine Learning." *Theory and Applications of Mathematics &Amp; Computer Science* 7 (1): Pages: 39. https://uav.ro/applications/se/journal/index.php/TAMCS/article/view/158.

R Core Team. 2020. *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing. https://www.R-project.org/.

Wickham, Hadley, Mara Averick, Jennifer Bryan, Winston Chang, Lucy D'Agostino McGowan, Romain François, Garrett Grolemund, et al. 2019. "Welcome to the tidyverse." *Journal of Open Source Software* 4 (43): 1686. https://doi.org/10.21105/joss.01686.

Wilcoxon, Frank. 1945. "Individual Comparisons by Ranking Methods." *Biometrics Bulletin* 1 (6): 80–83. http://www.jstor.org/stable/3001968.

Wood, Simon N. 2017. *Generalized Additive Models: An Introduction with R*. CRC press.

Yamashita, Toshie, Keizo Yamashita, and Ryotaro Kamimura. 2007. "A Stepwise Aic Method for Variable Selection in Linear Regression." *Communications in Statistics - Theory and Methods* 36 (13): 2395–2403. https://doi.org/10.1080/03610920701215639.