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. A 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, que é relativamente curto, o que certamente preocupa produtores de leite (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 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, representadas pelos dias entre 30 dias antes do parto e 40 dias após este evento. Consideramos um efeito aleatório para acomodar a variação intra-animal, devido a repetição de medições das vacas em tempos diferentes. Suponha 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, $f_j(x)$ são funções suaves de $x_{ij(k-1,k-2)}$, $k-1$ e $k-2$ são referentes ao dias anteriores ao dia k e j é referente a variável explicativa. A função $f_j(x)$ será estimada através de

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

onde b_p são as funções bases e $\hat{B}_p$ seus respectivos coeficientes; e $u_i \sim N(0, \sigma^2)$ é referente aos efeitos aleatórios para os animais analisados.

Conceitualmente, a única diferença de um modelo com efeito aleatório para um modelo sem é o efeito da vaca, que em média não tem efeito algum, mas varia de animal para animal por uma quantidade. Dependendo do efeito estimado para uma vaca, ela pode começar acima ou abaixo do intercepto estimado para todos os animais. E com isso, pode-se obter uma melhora na predição, pois é levado em conta o histórico do animal. Além disso, quando especifica-se o uso de efeitos aleatórios a tendência é encontrar valores menos extremos para os efeitos do animal do que quando for especificado como um efeito fixo (Bell, 2019). Umas das suposições para o uso de efeitos aleatórios, especialmente nesse caso mais simples, é a de associação não-negativa entre as respostas intra-animal.

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 $f_j(x)$ foram estimadas através da expansão de bases de splines de regressão cúbica pela própria função *gamm*. Na função *gamm*, o modelo é ajustado através maximização da quasi-verossimilhança penalizada. Para mais informações sobre o ajuste com variáveis funcionais e efeitos aleatórios veja Wood (2017).

2 Validação cruzada

Para avaliação da capacidade preditiva dos modelos, foi realizada uma validação cruzada com 10 reparticionamentos e com 10 envelopes, cada um contendo 10% do dados, 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, em cada repartição. 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 curva ROC (Bradley, 1997) conjuntamente com a área sob a curva da mesma (Ballings and Van den Poel, 2013) e acurácia (Novaković et al. 2017). A área sob a curva da curva 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 quando na verdade foram observados como tal. Um modelo com capacidade de predição tão informativo quanto um classificador aleatório tem uma área igual a 0,5, já um modelo perfeito tem uma área igual a 1. Enquanto, a acurácia indica qual é a porcentagem de predições corretas que o modelo é capaz de realizar.

Como é provável que surjam novas vacas que irão produzir leite no local onde foi coletado os dados (bezerras nascidas no local que chegam a idade de reprodução, por exemplo), isso foi levado em conta no problema. Com isso, a separação dos envelopes da validação cruzada ocorreu de modo que no conjunto treino não podem haver vacas que estão no conjunto de teste, e vice-versa. Dessa forma, é possível simular um cenário semelhante ao do futuro. Além disso, para haver uma amostra mais representativa no conjunto de teste, a validação cruzada ocorreu de tal modo que as observações de um animal só poderiam aparecer uma vez nesse conjunto. Com isso, ocorreu uma única predição para cada uma das observações no conjunto de dados na parte de validação dos modelos.

As funções 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).

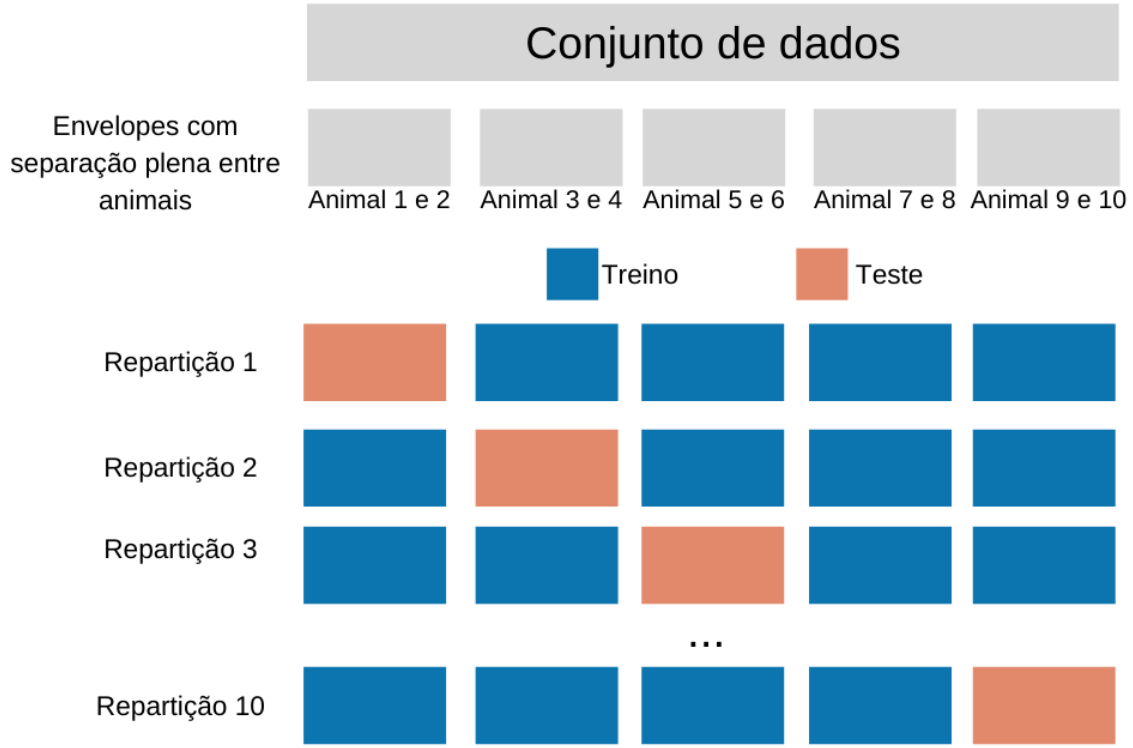

Figura 1: Diagrama simplificado da validação cruzada empregado neste trabalho. O conjunto de dados foi separado em 10 envelopes e não há observações de um mesmo animal em envelopes diferentes. Após isso ocorreu 10 repartições onde 9 envelopes eram utilizados para treino e 1 para o teste do modelo.

3 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. Para contornar esse problema tomou-se o seguinte procedimento: se o animal apresentou poucos dados faltantes para uma variável, esses seriam substituídos pela média daquela variável ao longo das medidas repetidas do animal. Para o animal sem nenhuma informação numa determinada variável, o valor imputado foi a mediana da variável para as vacas com informação

4 Resultados

Foram utilizadas duas abordagens para resolver o problema de classificação. A primeira faz o ajuste de um modelo a partir de uma única observação do estado de saúde do animal, como descrito na Seção 4.1. A segunda faz o ajuste de um modelo considerando medidas repetidas de cada animal, como descrito na Seção 4.2.

4.1 Resultados do ajuste do modelo sem medidas repetidas de saúde da vaca

Como mostra a Figura 2, as vacas ficaram doentes majoritariamente no período entre 5 até 20 dias após o parto.

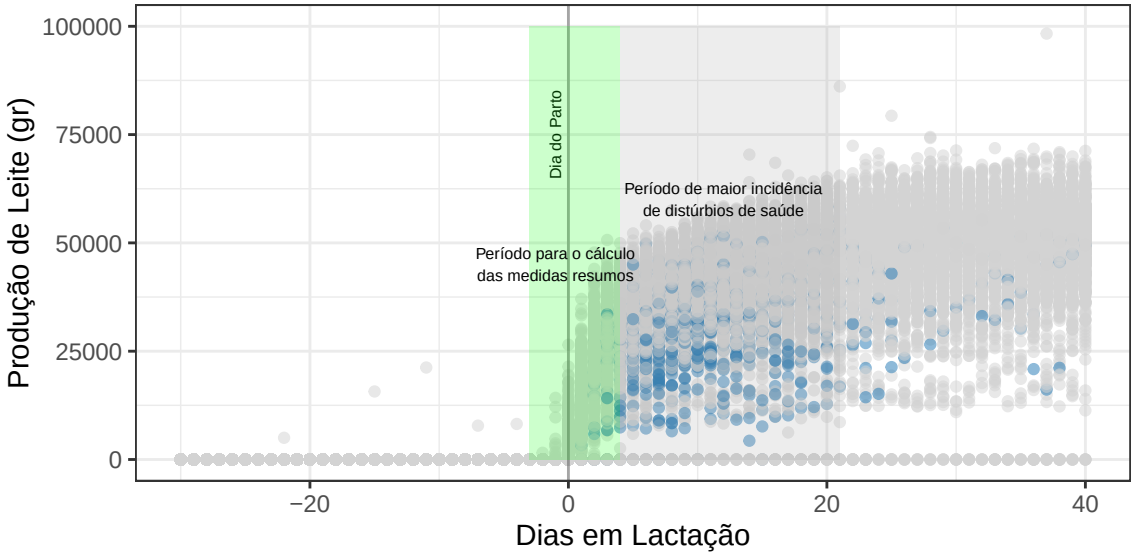

Figura 2: Produção de leite entre os dias em torno do parto. A cor azul indica distúrbio de saúde e o cinza, estado saudável.

Para o ajuste do modelo em (1) consideramos como variável resposta o estado de saúde da vaca no período de 5 a 20 dias após o parto. Caso a vaca tenha ficado doente pelo menos uma vez neste período, a variável resposta $Y_i = 1$ e caso contrário, $Y_i = 0$. Neste modelo, 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 variáveis na forma escalar e todas as observações. A partir de um mecanismo do tipo *stepwise* tendo o critério AIC (Bozdogan, 1987) para seleção de modelos, o modelo com menor AIC foi escolhido como melhor opção para essa Seção 3.1. O *stepwise* é um processo iterativo para seleção de variáveis explanatórias. Para mais informações sobre o método *stepwise*, veja Yamashita, Yamashita e Kamimura, 2007. O método *stepwise* foi empregado, 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.

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. Construiu-se a curva ROC e a acurácia para cada uma das iterações da validação cruzada, as quais encontram-se na Figura 3. O cálculo da acurácia depende do corte que está sendo utilizado. Como estamos usando o *logito*, temos uma probabilidade de estar doente associada a cada observação no banco de teste após a predição. Por exemplo, se fixarmos o corte em 0,3, todas as observações utilizadas para predição que obtiverem uma probabilidade de estar doente maior que 0,3 serão classificadas como doentes e caso contrário, como saudável.

Podemos ver que pela Figura 3 que o modelo tem uma acurácia em média de 62,5% a 75,0% em seu ápice, que é por volta de um corte de 0.4. Já as curvas ROC possuem uma variância considerável e uma área sob a curva de 0,735.

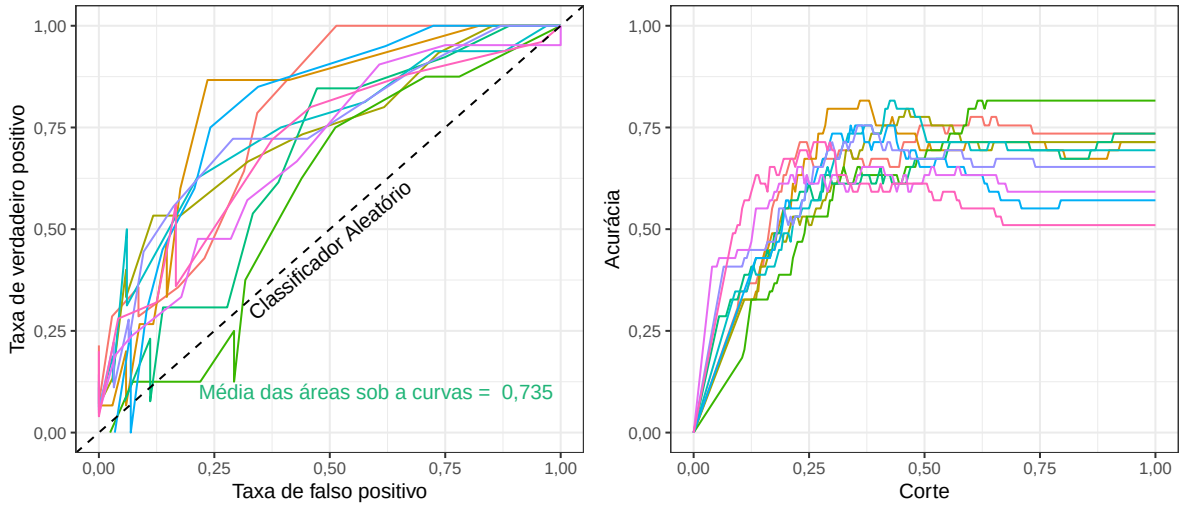

Figura 3: Curvas ROC e acurácias do modelo sem medidas repetidas alcançado por stepwise obtidas por meio de validação cruzada.

Os resultados da Tabela 3 mostram que, no modelo com menor AIC, ajustado agora com todas as observações do conjunto de dados, a média de porcentagem de gordura no leite e o número médio de dias que o estado da vaca é não grávida estão correlacionadas positivamente com a probabilidade do animal apresentar um distúrbio de saúde após o parto. Já a quantidade média de proteína no leite, o tempo médio de repouso, o tempo médio de ruminação e a média alimentação apresentam estar correlacionadas negativamente com a probabilidade do animal ter um problema de saúde.

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, deu-se preferência para as algumas variáveis que se mostraram significativas no modelo da Seção 4.1 e aquelas que apresentaram-se promissoras na análise exploratória, dessa forma, há uma quantidade menor de variáveis e, consequentemente, um menor processamento computacional é exigido.

Além disso, nas duas abordagens um efeito aleatório para cada vaca foi considerado para acomodar a correlação entre as medidas de cada animal. Chamaremos o modelo da primeira abordagem como modelo misto escalar e o da segunda de modelo misto escalar-funcional.

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 $k-1$. Porém, ainda pode-se considerar informações de outros dias passados (por exemplo, $k-2$ e $k-3$). Sendo assim, para o modelo em (2), consideramos o estado do animal i no dia k como função das covariáveis $x_{i,1,(k-1,k-2)}, \ldots x_{i,J,(k-1,k-2)}$, pois não houve uma melhora significativa na predição quando considerou se mais de dois dias para realizar a predição do estado de saúde. 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. Escolhe-se esse ponto de parada, pois os distúrbios de saúde ficam raros após isso, como podemos ver na Figura 2.

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 modelo misto escalar obtidas por meio de validações cruzadas indicam que o poder de classificação do modelo é satisfatório, pois a área sob a curva e a acurácia estão próximas de 0,9, como mostra a Figura 4. A variância das curvas é menor em comparação com o modelo da Seção 4.1, pois são utilizadas mais observações no teste nessa segunda abordagem, logo há uma estabilidade maior.

Para o modelo misto escalar-funcional, a variável referente aos dias de lactação foi utilizada como funcional devido a aparente não linearidade, como pode ser identificado na Figura 2. Há uma aparente

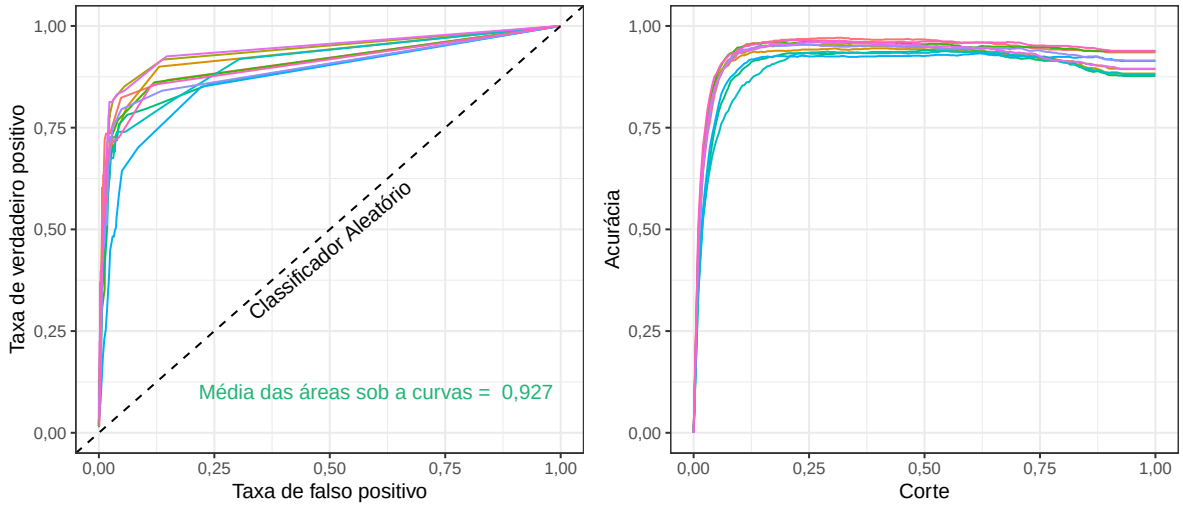

Figura 4: Curvas ROC e acurácias do modelo misto escalar obtidas por meio de validação cruzada.

menor quantidade de distúrbios de saúde entre os dias 0 a 5, uma maior quantidade entre os dias 5 a 15 e uma leve diminuição no número de ocorrências até o dia 20, como também mostra a Figura 2. 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 modelo misto escalar-funcional, na Figura 5, apontam que a capacidade de predição é muito boa, pois a área sob a curva e a acurácia são maiores que 0,9, superando o poder preditivo do modelo escalar. A acurácia atingiu seu ápice por volta do corte igual a 0,25 nos dois modelos.

As estimativas dos efeitos escalares do modelo misto escalar-funcional, agora com todos os dados do conjunto de dados, se encontram na Tabela 4. As variáveis referentes a ruminação durante as últimas 24 horas e se o animal apresentou estar doente nos dias anteriores tem efeito positivo no logito, indicando que a probabilidade da vaca ficar doente aumenta quando há um aumento na ruminação ou se a vaca ficou doente nos dois dias anteriores. Enquanto que as variáveis produção de leite e alimentação nas últimas 24 horas tem efeito negativo, indicando que a probabilidade de adoecer diminui quando há um acréscimo nessas variáveis. 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 com não estiver doente ou estiver doente em apenas um dia.

A Figura 7 exibe a curva estimada a variável funcional referente ao dia de lactação. A estimação dos coeficientes de suavização, $\hat{B}_p$, 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.

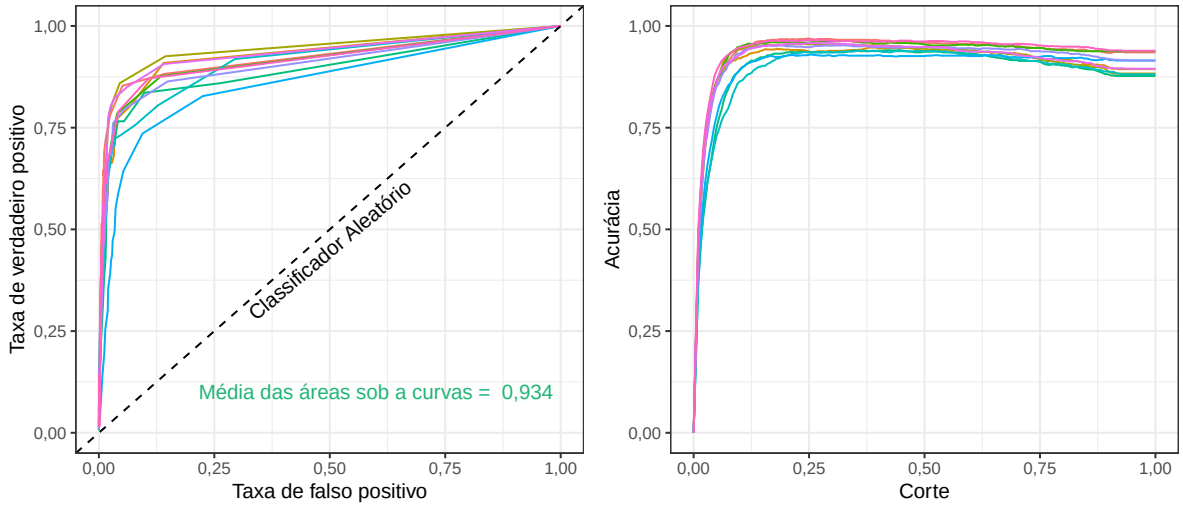

Figura 5: Curvas ROC e acurácias do modelo misto escalar-funcional obtidas por meio de validação cruzada.

Tabela 4: Estimativas dos parâmetros do ajuste do modelo misto escalar-funcional. <0,001 e >-0,001 são referentes à valores pequenos positivos e negativos, respectivamente. k1 e k2 são referentes à k-1 e k-2, respectivamente.

Termo	Estimativa	Erro padrão	Estatística do teste	P-valor
Intercepto	>-0,001	0,483	-6,219	<0,001
RuminacaoUltimas24h_k1	0,004	0,002	2,162	0,031
RuminacaoUltimas24h_k2	0,006	0,002	3,741	<0,001
AlimentacaoUltimas24h_k1	-0,001	0,002	-0,377	0,706
AlimentacaoUltimas24h_k2	0,004	0,002	1,734	0,083
doente_k1	2,697	0,163	16,591	<0,001
doente_k2	0,935	0,222	4,211	<0,001
ProdLeite_k1	>-0,001	<0,001	-0,408	0,683
ProdLeite_k2	>-0,001	<0,001	0,109	0,914
RuminacaoUltimas24h_k1:RuminacaoUltimas24h_k2	>-0,001	<0,001	-5,170	<0,001
AlimentacaoUltimas24h_k1:AlimentacaoUltimas24h_k2	>-0,001	<0,001	-2,786	0,005
doente_k1:doente_k2	0,478	0,289	1,653	0,098

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)	5,492	19,088	<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

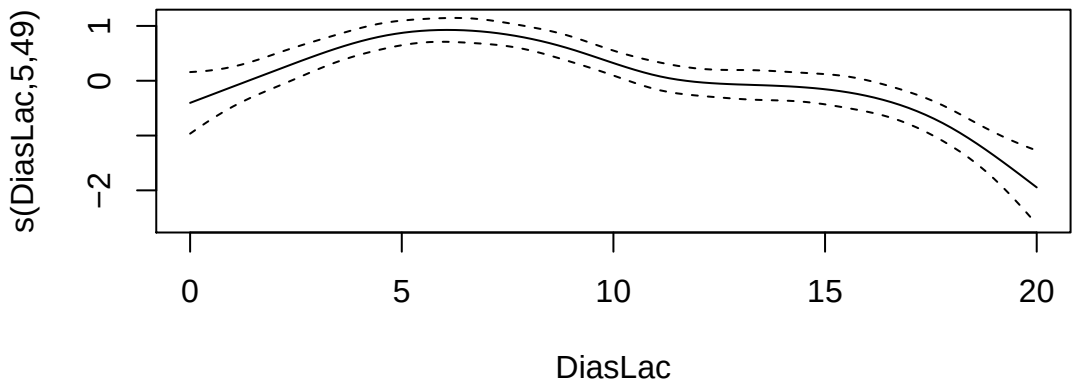

Figura 6: Curva estimada de dia de lactação (DiasLac)

também teria uma acurácia de 90%. Para investigar mais a fundo como ocorre a classificação no modelo misto escalar-funcional, construiu-se a Figura 8. Ao invés de analisar a acurácia geral do modelo, dividiu-se as observações do teste em intervalos referentes a quantos dias a vaca referente a essa observação apresentou ficar doente. 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. 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 modelo misto escalar-funcional 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

Os resultados indicam que a utilização de variáveis funcionais trouxe uma melhora para a capacidade de predição em comparação com o modelo misto escalar para essa abordagem, pois a área sob a curva e a acurácia apresentaram um aumento.

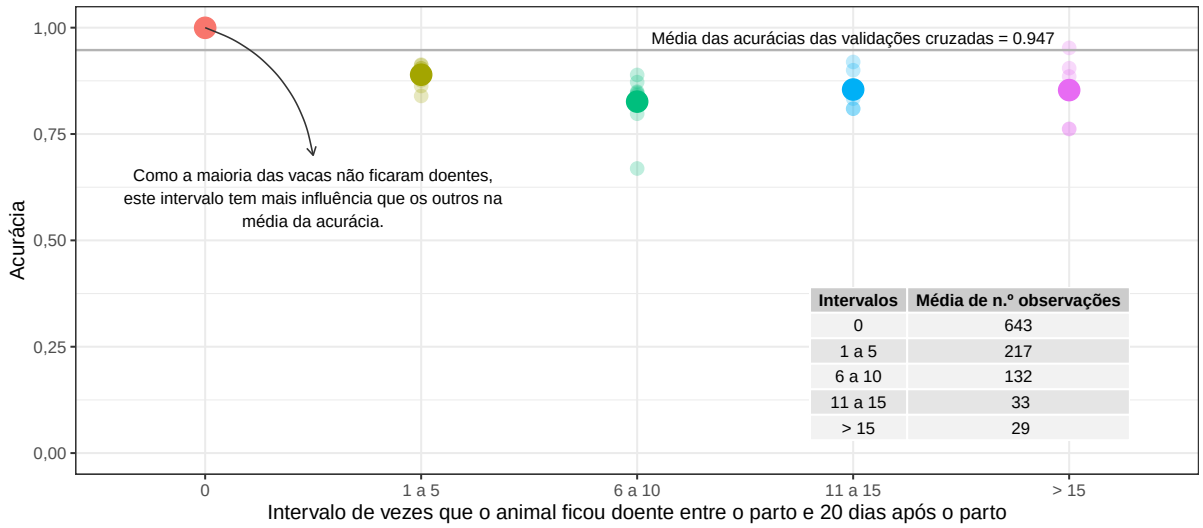

Figura 7: Acurácia em intervalos de quantas vezes as vacas do conjunto de teste ficaram doentes entre o dia do parto e os 20 dias posteriores, obtida pela validação cruzada.

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 uma 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 abril. 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.