UNIVERSIDADE FEDERAL DE MINAS GERAIS
Instituto de Ciências Exatas
Programa de Pós-graduação em Estatística

Rafael Romero Nicolino

Análise do desempenho de diferentes modelos de predição de eventos binários em epidemiologia – Estudo com dados de retenção de placenta.

Belo Horizonte
2022

Rafael Romero Nicolino

Análise do desempenho de diferentes modelos de predição de eventos binários em epidemiologia – Estudo com dados de retenção de placenta.

Versão final

Monografia de especialização apresentada ao Instituto de Ciências Exatas da Universidade Federal de Minas Gerais, como requisito parcial à obtenção do título de Especialista em Estatística.

Área de Ênfase: Estatística

Orientador: Marcelo Azevedo Costa

Belo Horizonte

2022

Nicolino, Rafael Romero

N644a Análise do desempenho de diferentes modelos de predição de eventos binários em epidemiologia [recurso eletrônico]: estudo com dados de retenção de placenta. / Rafael Romero Nicolino — 2022.

1 recurso online (39 f. il, color.): pdf.

Orientador: Marcelo Azevedo Costa.

Monografia (especialização) - Universidade Federal de Minas Gerais, Instituto de Ciências Exatas, Departamento de Estatística.

Referências: 38-39

1. Estatística. 2. Epidemiologia 3. Análise de regressão .4. Aprendizado do computador. I. Costa, Marcelo Azevedo. II. Universidade Federal de Minas Gerais I. Instituto de Ciências Exatas, Departamento de Estatística. III. Título.

CDU 519.2 (043)

Ficha catalográfica elaborada pela bibliotecária Belkiz Inez Rezende Costa CRB 6/1510
Universidade Federal de Minas Gerais – ICEx

Universidade Federal de Minas Gerais **E-mail:**
r
Instituto de Ciências Exatas Tel: 3409-
9-5924
Departamento de Estatística
P Programa de Pós-Graduação / Especialização
Av. Pres. Antônio Carlos, 6627 - Pampulha
31270-901 – Belo Horizonte – MG

ATA DO 265º. TRABALHO DE FIM DE CURSO DE ESPECIALIZAÇÃO EM ESTATÍSTICA DE RAFAEL ROMERO NICOLINO.

Aos dezenove dias do mês de dezembro de 2022, às 14:00 horas, com utilização de recursos de videoconferência a distância, reuniram-se os professores abaixo relacionados, formando a Comissão Examinadora homologada pela Comissão do Curso de Especialização em Estatística, para julgar a apresentação do trabalho de fim de curso do aluno **Rafael Romero Nicolino**, intitulado: "**Análise do desempenho de diferentes modelos de predição de eventos binários em epidemiologia – Estudo com dados de retenção de placenta**.", como requisito para obtenção do Grau de Especialista em Estatística. Abrindo a sessão, o Presidente da Comissão, Professor Marcelo Azevedo Costa – Orientador, após dar conhecimento aos presentes do teor das normas regulamentares, passou a palavra ao candidato para apresentação de seu trabalho. Seguiu-se a arguição pelos examinadores com a respectiva defesa do candidato. Após a defesa, os membros da banca examinadora reuniram-se sem a presença do candidato e do público, para julgamento e expedição do resultado final. Foi atribuída a seguinte indicação: o candidato foi considerado **Aprovado** condicional às modificações sugeridas pela banca examinadora no prazo de 30 dias a partir da data de hoje por unanimidade. O resultado final foi comunicado publicamente ao candidato pelo Presidente da Comissão. Nada mais havendo a tratar, o Presidente encerrou a reunião e lavrou a presente Ata, que será assinada por todos os membros participantes da banca examinadora. Belo Horizonte, 19 de dezembro de 2022.

Prof. Marcelo Azevedo Costa (Orientador)
Departamento de Engenharia da Computação / Escola de Engenharia / UFMG

Prof. Guilherme Lopes de Oliveira
DECOM / CEFET-MG

Prof. João Paulo Amaral Haddad
Laboratório de Epidemiologia e Bioestatística / UFMG

Universidade Federal de Minas Gerais
Instituto de Ciências Exatas
Departamento de Estatística
Programa de Pós-Graduação / Especialização
Av. Pres. Antônio Carlos, 6627 - Pampulha
31270-901 – Belo Horizonte – MG

E-mail: pgest@ufmg.br
Tel: 3409-5923 – FAX: 3409-5924

DECLARAÇÃO DE CUMPRIMENTO DE REQUISITOS PARA CONCLUSÃO DO CURSO DE ESPECIALIZAÇÃO EM ESTATÍSTICA.

Declaro para os devidos fins que o **Rafael Romero Nicolino,** número de registro 2021679181, cumpriu todos os requisitos necessários para conclusão do curso de Especialização em Estatística, entregando a versão final do trabalho de conclusão de curso para seu orientador, o professor Marcelo Azevedo Costa, que aprovou a versão final. O trabalho foi apresentado no dia 19 de dezembro de 2022 com o título "*Análise do desempenho de diferentes modelos de predição de eventos binários em epidemiologia – Estudo com dados de retenção de placenta*".

Belo Horizonte, 15 de fevereiro de 2023.

Roberto da Costa Quinino:80871291720

Assinado de forma digital por Roberto da Costa Quinino:80871291720
Dados: 2023.02.15 19:00:58 -03'00'

Prof. Roberto da Costa Quinino
Coordenador do curso de Especialização em Estatística
Departamento de Estatística / UFMG

AGRADECIMENTOS

Para a minha família, minha esposa Natali, meus filhos Francisco e Rafaella por todas as ausências durante a realização do curso de especialização.

Ao Professor Marcelo Costa, por aceitar me orientar no desenvolvimento do trabalho.

A todos que me deram suporte nestes 2 anos de curso de especialização.

RESUMO

Técnicas estatísticas de classificação têm por objetivo, basicamente, predizer determinado comportamento de um evento com base em suas características que, de alguma forma, estão correlacionadas com a variável resposta de interesse. Modelos preditivos podem ser desenvolvidos usando métodos tradicionais de análise de regressão, por exemplo regressão logística ou até mesmo com métodos mais sofisticados como modelos de árvore de classificação (CART) e Floresta Aleatória. Este estudo teve como objetivo analisar o desempenho de três diferentes métodos de modelagem de dados: um método clássico de GLM, a Regressão Logística e dois métodos baseados em técnicas de aprendizado de máquina (Machine Learning), a Árvore de Classificação (CART) e o Modelo de Floresta Aleatória. A aplicabilidade dos modelos desenvolvidos foi avaliada por meio de índices de desempenho de classificação como a estatística AUC (Area Under the Curve). O banco de dados é relacionado ao evento de retenção de placenta em 3 propriedades leiteiras de Unaí, Minas Gerais. A retenção de placenta foi analisada através do preenchimento de fichas epidemiológicas pelos proprietários, gerando um total de 699 observações. Devido a um perfil muito específico do banco de dados, a análise buscou comparar os resultados de dois bancos de dados, denominados dt01 e dt02. Buscou-se verificar como esse desbalanceamento de dados, e um problema de dados faltantes relacionado a ordem de lactação e número de partos, específicos de uma propriedade, poderia atuar no desempenho dos modelos. Assim, avaliando individualmente cada modelo, o desempenho da regressão logística e do modelo CART, modelos menos complexos e de entendimento mais direto, obtiveram AUC maiores que o de Floresta Aleatória. Baseados na discussão anterior, da complexidade do evento estudado e variáveis levantadas, podemos considerar que o desempenho dos modelos mais simples foi minimamente satisfatório e superiores aos modelos mais complexos.

Palavras-chave: Epidemiologia, Regressão, Classificação, Machine Learning.

ABSTRACT

Statistical modeling for classification is basically aimed at predicting certain behavior of an event based on its characteristics that, in some way, are correlated with the response variable of interest. Predictive models can be developed using traditional methods of regression analysis, for example logistic regression or even with more sophisticated methods like classification tree models (CART) and Random Forest. This study aimed to analyze the performance of three different methods of data modeling: a classic GLM method, Logistic Regression and two methods based on machine learning techniques (Machine Learning), the Classification Tree (CART) and the Random Forest Model. The applicability of the models was evaluated using classification performance indices such as the AUC statistic (Area Under the Curve). The database is related to the retained placenta in 3 dairy farms in Unaí, Minas Gerais. Retained placenta was analyzed by self-completed epidemiological forms by the owners, generating a total of 699 observations. Due to a very specific profile of the database, the analysis was conducted to compare the results of two databases, called as dt01 and dt02. We aimed to verify how an imbalance data, and a problem of missing data related to the lactation order and number of lactations, specific to a property, could affect the performance of the models. Thus, evaluating each model individually, the performance of the logistic regression and the CART model, less complex models with a more direct understanding, obtained higher AUCs than the Random Forest. Based on the previous discussion, the complexity of the event studied, and the studie variables, we can consider that the performance of the simpler models was minimally satisfactory and superior to the more complex models.

Keywords: Epidemiology, Regression, Classification, Machine Learning.

LISTA DE ILUSTRAÇÕES

LISTA DE TABELAS

SUMÁRIO

1 INTRODUÇÃO

No campo da saúde e da medicina veterinária, a epidemiologia apresenta uma série de situações em que a resposta de interesse ao pesquisador é uma variável binária, (Y = 0 ou Y=1). Mais comumente este desfecho representa a ausência ou presença de uma doença ou o agravo em saúde (Dohoo et al., 2009). A modelagem estatística e a previsão em epidemiologia fornecem um método para entender por que e como as doenças e agravos se espalham e como elas podem ser prevenidas (Yadav e Akhter, 2021).

A modelagem de dados em epidemiologia também é feita para conhecer os riscos concorrentes das mortes por doenças infecciosas. A modelagem estatística ajuda a identificar possíveis fatores que de alguma forma visa limitar a extensão da infecção empregando algumas estratégias supressivas como quarentena, distanciamento social, abate de animais, rastreamento de contatos e vacinação quando disponível. Um dos pontos fracos em dados epidemiológicos que utilizam técnicas de modelagem é que os dados são, na grande maioria das vezes, escassos para as doenças infecciosas. A modelagem epidemiológica é crucial para conhecer as características salientes da dinâmica de infecção da doença.

Técnicas estatísticas de modelagens estatísticas de classificação têm por objetivo, basicamente, predizer determinado comportamento de um evento com base em suas características que, de alguma forma, estão correlacionadas com a variável resposta de interesse. Modelos preditivos podem ser desenvolvidos usando métodos tradicionais de análise de regressão, por exemplo regressão logística ou até mesmo com métodos mais sofisticados como modelos de árvore de classificação (CART) e Florestas Aleatórias (Segal, 2004).

A Retenção de Placenta (RP) causa um impacto econômico considerável nos sistemas de produção de leite do Brasil, sendo um agravo que se caracteriza pela expulsão da placenta após 12 horas do parto. A ocorrência de RP é epidemiologicamente complexa, sendo considerada um processo multifatorial. É possível destacar fatores de risco como: a distocia, parto gêmeo, parto induzido, deficiências nutricionais e processos infecciosos que reduzem a resposta imune, bem como problemas de manejo associados às condições ambientais, incluindo as altas temperaturas e volumes de chuva (Buso et al., 2018). A incidência de RP está relacionada com a estação do ano, mostrando que no verão essa incidência pode ser

até duas vezes maior do que no inverno (Dahl et al., 2020). A complexidade na definição das causas combinadas com a divergência na eficácia dos tratamentos enfatiza a importância da prevenção de RP (Beagley et al., 2010; Buso et al., 2018; Hernández-Castellano et al., 2019).

Este estudo teve como objetivo analisar o desempenho de três diferentes métodos de modelagem de dados: um método clássico de GLM, a Regressão Logística e dois métodos baseados em técnicas de aprendizado de máquina (Machine Learning), a Árvore de Classificação (CART) e o Modelo de Florestas Aleatórias. A aplicabilidade dos modelos desenvolvidos foi avaliada por meio de índices de desempenho de classificação como a estatística AUC (*Area Under the Curve*).

2 METODOLOGIA

2.1 A MODELAGEM ESTATÍSTICA

Segundo Breiman, (2001), toda a análise estatística tem início com o levantamento de dados, e esses dados sendo gerados a partir de um conjunto de variáveis de entrada, denominadas X, levam a uma resposta Y, a partir de diferentes funções de associação entre os preditores X e o desfecho Y. Assim são definidas duas abordagens nas análises desses dados e associações entre X e Y, denominados em seu artigo como a cultura de modelagem de dados e outra modelagem de algoritmos.

A cultura de modelagem de dados é chamada de caixa branca, usa os dados disponíveis para estimar os parâmetros do modelo e fornecer uma inferência estatística. Os parâmetros dos modelos de caixa branca podem ter significados físicos para o pesquisador. No entanto, os modelos de caixa branca (Figura 1) aplicam, em geral, estruturas matemáticas e estatísticas simples, como a equação de regressão linear e a logística. Consequentemente, os parâmetros dos modelos de caixa branca, também conhecidos como coeficientes, têm interpretações significativas e palpáveis ao pesquisador, que conseguem estimar associações entre X e Y, e como uma mudança no nível da variável X impacta na ocorrência do Y.

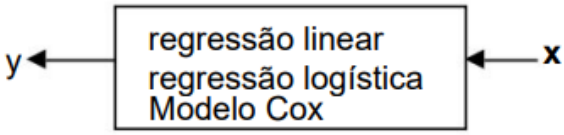


Figura 1. Exemplificação da Modelagem de dados. (Adaptado de Breiman, 2001)

A cultura de modelagem de algoritmos, denominado modelo caixa preta (Figura 2), está interessada em definir um modelo mais preditivo, que pode ter uma estrutura altamente complexa e utilizar um número de variáveis X para predizer Y. Em geral, os modelos de caixa preta compreendem aproximadores universais, que se baseiam em funções e/ou transformações não lineares. Em geral, os parâmetros dos modelos de caixa preta não têm uma interpretação direta ao pesquisador, quando são comparadas aos modelos de caixa branca e seus coeficientes. Exemplos de modelos de caixa preta são as Florestas Aleatórias e as Árvores de Classificação.

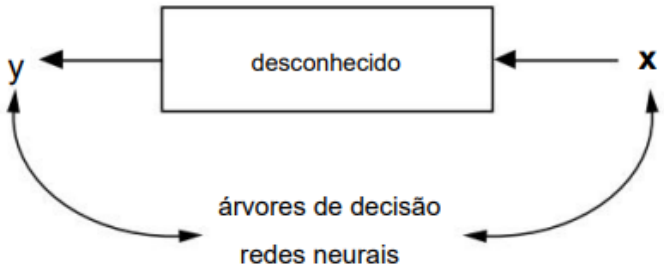


Figura 2. Exemplificação da Modelagem de algoritmos. (Adaptado de Breiman, 2001)

2.2 GESTÃO DA BASE DE DADOS E MÉTODOS COMPARATIVOS

Para a comparação do desempenho, foram criados três diferentes modelos, um modelo clássico de regressão logística, modelo CART (árvore de classificação) e Modelo de Florestas Aleatórias.

O banco de dados é relacionado ao evento de retenção de placenta em 3 propriedades leiteiras de Unaí, Minas Gerais. A retenção de placenta foi analisada através do preenchimento de fichas epidemiológicas pelos proprietários, gerando um total de 699 observações. A ficha define como caso de retenção (Y =1) a vaca que apresentava retenção de placenta 12 horas ou mais após o parto. Junto desta informação, foram levantados dados sobre a data que ocorreu o parto, o sexo do bezerro, se o parto era simples ou gemelar, o tipo de parto (Normal, Distócico, Aborto) e o número de partos da vaca.

A partir da data de parto, podemos categorizar o período como Chuvas (outubro a março) e Seca (abril a setembro). A partir do número de partos da vaca, categorizamos a vaca como: primeira, segunda, terceira e quarta ou mais lactações, baseados na literatura que demonstra que vacas com 4 ou mais lactações têm um aumento considerável no risco de reter a placenta (Van Werven et al., 1992; Nobre et al., 2012; Martins et al., 2013). As variáveis de estudo estão demonstradas na Tabela 1. Todas as análises foram realizadas no software estatístico R Core Team (2022).

Tabela 1. Resumo das variáveis com suas abreviaturas e os níveis de potenciais fatores de risco para placenta retida nas vacas leiteiras.

Variável	Abreviação	Níveis
Variável dependente		
Ocorrência de Retenção de Placenta	Ret_plac	0 = Não; 1 = Sim
Variáveis Independentes		
Mês do ano	Mes	1 até 12 (Janeiro até Dezembro)
Ano	Ano	2017 e 2018
Estação do ano	Estação	Chuvas_outubro_março Seca_abril_setembro
Número de lactações	N_partos	Dado discrete – 1 a 9 partos.
Ordem de Lactação (categorização da variável número de lactações)	Ordem_lact	1 = Primeira lactação 2 = Segunda Lactação 3 = Terceira Lactação 4 = Quarta ou mais Lactações
Tipo de parto Eutócico (sem ajuda externa) Distócico (com ajuda externa - manipulação)	tipoparto	Normal Distócico Aborto
Sexo	fêmea	1=fêmea; 0=macho
Simples ou gemelar	simples	Simples Gemelar

2.2.1 Limpeza e Codificação da base de dados

A base de dados de retenção contém 699 observações com 8 variáveis preditoras, conforme Tabela 1. A primeira análise a ser realizada foi a verificação do total de valores faltantes na base. chegando a um total de 22 dados faltantes para a variável parto gemelar (sim ou não) e 196 dados faltantes para o número de partos e ordem de lactação.

Realizando uma análise exploratória dos dados faltantes, verificamos que em especial os dados de número de partos é um problema relacionado especificamente a uma propriedade, a Fazenda C, conforme Tabela 2.

Tabela 2. Dados faltantes por Fazenda.

Fazenda	Total de observações	Parto foi Gemelar (% de dados faltantes)	Número de Partos (% de dados faltantes)
A	205	202 (1,5%)	205 (0%)
C	358	339 (5,3%)	165 (53,9%)
L	136	136 (0%)	133 (2,2%)
Total Geral	**699**	**677 (3,1%)**	**503 (28,4%)**

Desta maneira, devido a um perfil muito específico do banco de dados, a análise buscou comparar os resultados de dois bancos de dados, denominados **dt01** e **dt02**. Buscou-se verificar como esse desbalanceamento de dados, e um problema de dados faltantes relacionado a ordem de lactação e número de partos, específicos de uma propriedade, poderia atuar no desempenho dos modelos. Desta forma foram utilizadas as bases denominadas:

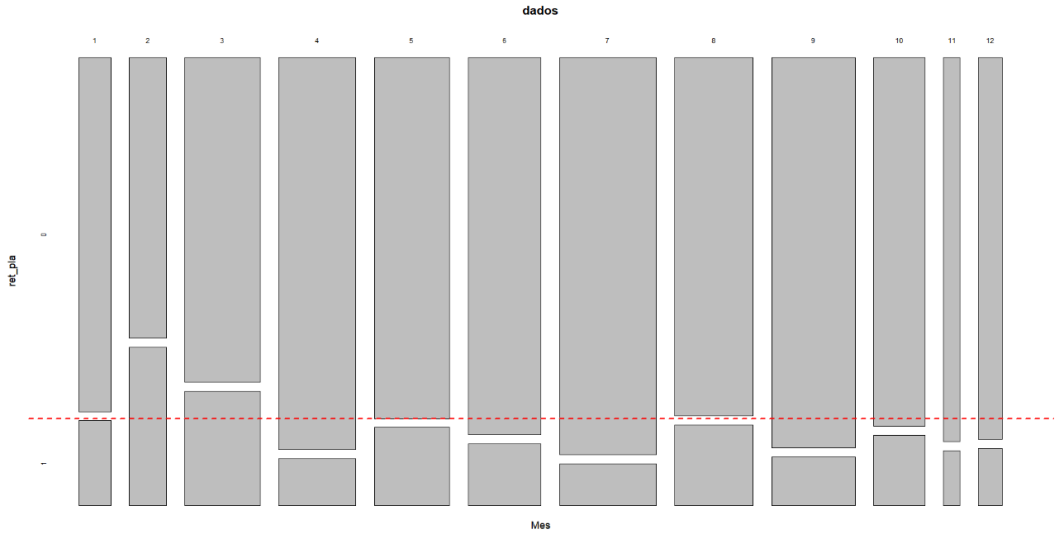
- Dt01 – Base com 677 observações (base sem dados de número de partos e ordem de lactação).
- Dt02 – Base com 493 observações (base com dados de número de partos e ordem de lactação).

As variáveis categóricas requerem atenção especial na análise de regressão porque, ao contrário das variáveis contínuas, elas não podem ser inseridas na equação de regressão exatamente como estão. Em vez disso, eles precisam ser recodificados em uma série de variáveis que podem ser inseridas no modelo de regressão. Há uma variedade de sistemas de codificação que podem ser usados ao recodificar variáveis categóricas.

A variável mês foi recodificada de maneira que se tenta maximizar seu poder discriminante. Ao inserir a variável mês (12 meses) como fator, se aumentou o número de variáveis, não sendo ideal usar todos os meses. Caso fossem codificados todos os meses, seriam 11 variáveis dummy e impactaria o desempenho do modelo.

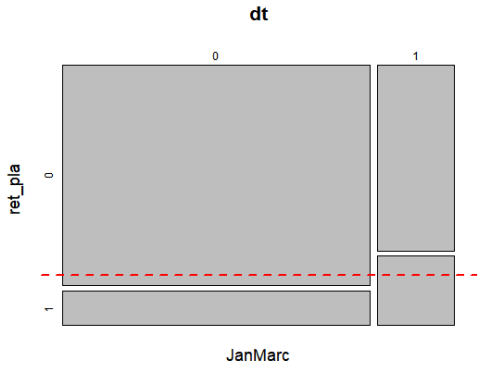
Inicialmente realizamos uma análise gráfica, através do gráfico de mosaico, conforme Figura 3, para tentar buscar algum padrão dos dados, que faça sentido para a epidemiologia do evento.

Figura 3. Gráfico de mosaico, variável mês por frequência de retenção de placenta. Em vermelho destaca-se a frequência média de retenção de placenta no banco de dados.



Podemos verificar que o aumento da frequência da retenção de placenta ocorre em especial nos meses de janeiro, fevereiro e março, sendo mais marcante nos últimos dois meses. Como existe uma explicação lógica, relacionada a picos de calor, estresse térmico dos animais e período de chuva com contaminação ambiental, foi criada uma variável dummy – JanMarc (Sim 1- ou Não- 0), conforme Figura 4.

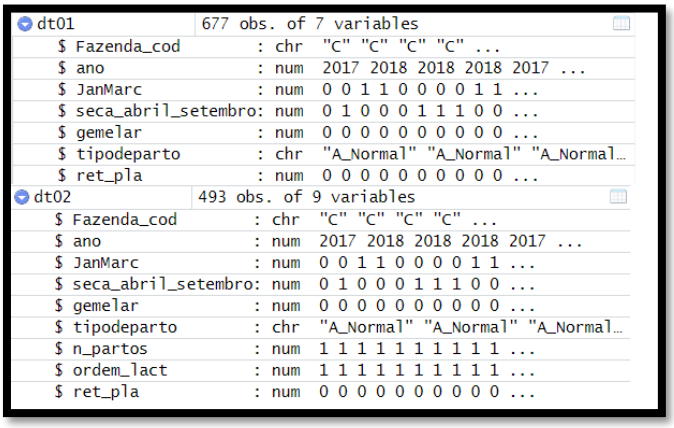
Figura 4. Gráfico de mosaico, variável JanMarc por frequência de retenção de placenta. Em vermelho destaca-se a frequência média de retenção de placenta no banco de dados.



Além da classificação de janeiro a março, buscou-se uma classificação mais generalista, como classificar em período de chuvas e secas, conforme demonstrado na Tabela 1.

A síntese dos dois bancos de dados utilizados nas análises é demonstrada na Figura 5.

Figura 5. Base de dados utilizadas no estudo, imagem do programa R.



2.2.2 Modelo de regressão logística

O modelo de regressão logística constitui um método de classificação supervisionado e trata-se de um caso particular dos modelos lineares generalizados, denominados modelos GLM (Mccullagh & Nelder, 2019), definido pela distribuição binomial com função de ligação canônica (logit), apropriado para a modelagem de resposta binária. A regressão logística tem como objetivo principal estudar a probabilidade de ocorrência de um evento, denominado como Y (no exemplo deste presente estudo ocorrência ou não de retenção de placenta), com base no comportamento de variáveis chamadas explicativas ou preditoras X. Assim, é definido uma expressão que relaciona de forma linear a probabilidade de ocorrência do evento e as covariáveis, da seguinte forma:

$$Z_i = \alpha + \beta_1.X_{1i} + \beta_2.X_{2i} + \cdots + \beta_k.X_{ki} ,$$

em que Z é conhecido como o logito, α representa a constante, β_j (j= 1, 2, 3, ..., K) são os parâmetros estimados para cada variável explicativa Xj. Importante frisar que Z não representa a variável Y. A regressão logística define o logito Z como o logaritmo natural da chance de ocorrência do evento, do modo que:

$$\ln\left(chance_{(Y=1)}\right) = Zi.$$

A chance é definida como:

$$Chance\ (odds)_{Y=1} = \frac{p_i}{1-p_i},$$

onde i (i = 1 ,2 ,3 ..., n) representa o índice da amostra.

Como o intuito é o de definir uma expressão para a probabilidade de ocorrência de um evento do estudo em função do logito Z, podemos matematicamente isolar p, chegando na fórmula que define a probabilidade de ocorrência de um evento como:

$$p_i = \frac{e^{Zi}}{1+ e^{Zi}} .$$

E a fórmula geral de probabilidade da ocorrência de um evento por uma observação i como:

$$p_i = \frac{1}{1+ e^{-(\alpha+ \beta_1+ \beta_2+\cdots+ \beta_K)}} .$$

Portanto podemos interpretar p_i como a probabilidade de um animal reter placenta, condicionada às características atribuídas, como época do ano, fazenda, se é parto gemelar ou não, e número de lactações do animal.

Segundo Hosmer e Lemeshow (1989), para um modelo ser considerado adequado, além de apresentar um bom ajuste, deve apresentar uma compreensão prática e ser parcimonioso. Para isso, existem métodos de seleção de variáveis que reduzem o número de variáveis explicativas, selecionando apenas as que mais contribuem para a explicação da variável resposta. Um algoritmo usado para seleção de variáveis a serem incluídas no modelo é o *stepwise*, que tem por objetivo selecionar variáveis que maximizam o ajuste com o menor número de variáveis empregadas. A medida de ajuste que foi utilizada no trabalho foi o critério de informação de Akaike (AIC), que pondera a qualidade de ajuste do modelo com a quantidade de parâmetros estimados no modelo (Aho et al., 2016). A função utilizada foi `stepAIC()` do pacote MASS (Venable & Ripley, 2002).

2.2.3 Modelo de Árvore de Classificação

Uma Árvore de Classificação é um modelo não-paramétrico que modela relações complexas entre as entradas e saídas de um problema de classificação ou regressão, sem a necessidade de assumir hipóteses *a priori*. Árvores de classificação são modelos de aprendizado que possuem a capacidade de tratar de atributos do tipo numérico, categórico ou ambos. Uma árvore de classificação implementa, intrinsecamente, a seleção de características, o que proporciona a este algoritmo uma

robustez na tratativa de casos em que haja variáveis irrelevantes ou que apresentem ruído. Além disto, árvores de classificação possuem fácil interpretabilidade quanto à suas regras de predição (Louppe, 2014). Estas particularidades fazem deste modelo, um algoritmo de aprendizado popular e muito difundido (Wu et al., 2008).

Árvores de classificação configuram métodos que utilizam uma representação gráfica baseada em árvores, cujo objetivo é identificar grupos de indivíduos com características de interesse em comum. Para tal, é utilizado um método recursivo que divide a amostra inicial em subamostras, baseando-se em resultados observados das variáveis explicativas e em suas interações. Formam-se, assim, grupos para os quais a variável resposta apresenta comportamento homogêneo dentro dos grupos e heterogêneo entre eles (Loh, 2011).

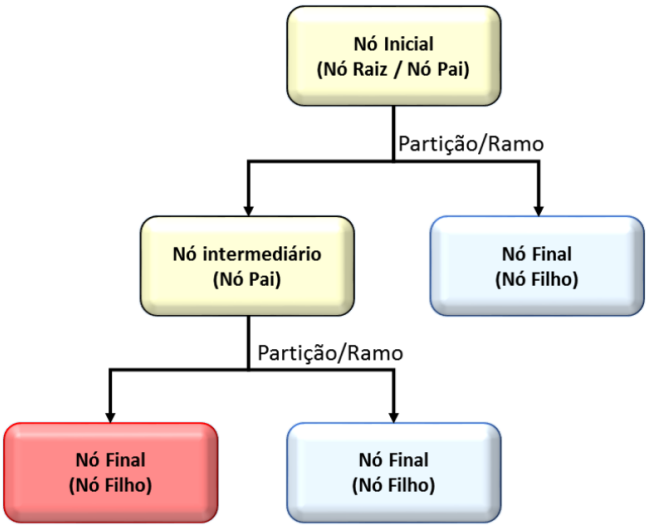
Uma Árvore de Decisão é chamada de Árvore de Classificação se a variável resposta for categórica, ou Árvore de Regressão, se numérica (Taconeli, 2008). O processo de indução de árvores é iniciado por meio de uma amostra, denominada nó raiz, que é dividida em subamostras, denominadas nós filhos ou nós intermediários. Essas subamostras quando subdivididas são chamadas de nós pais, pois geram nós filhos. Quando uma subamostra não puder mais ser subdivida segundo algum critério de parada, é então denominada de nó final ou nó folha. Esse processo é dito recursivo devido a cada subamostra gerar novas subamostras.

O CART (*Classification and Regression Trees*) é um algoritmo para indução de árvores de classificação (Breiman et al., 2017). O método de partição divide as variáveis de forma binária, ou seja, sempre em dois nós, com base em alguma medida de impureza, como por exemplo o índice de Gini ou a Entropia, para tornar os subconjuntos de dados cada vez mais homogêneos, em relação a variável resposta. Uma das suas vantagens é que não precisa realizar qualquer tipo de categorização, pois o algoritmo não faz restrições quanto às escalas das variáveis explicativas, podendo estas serem numéricas (discretas ou contínuas), ordinais e nominais. Por utilizar partições binárias, as variáveis podem aparecer em diferentes níveis do modelo, permitindo reconhecer diversas interações com outras variáveis.

O modelo CART constrói um processo conhecido como Poda de Custo-Complexidade, que ocorre no sentido reverso do crescimento de suas ramificações. Este mecanismo resulta em uma sequência de sub-árvores candidatas à modelo ótimo. Cada uma destas é obtida através da simples remoção de um nó interno, e

respectivas folhas, de uma sub-árvore anterior. Deste modo, a última árvore da série apresenta apenas uma única folha (Hastie et al., 2009). Este procedimento visa selecionar o melhor modelo entre as sub-árvores. A estrutura de uma árvore de classificação está exemplificada na Figura 6.

Figura 6. Estrutura de uma árvore de classificação. Fonte: Dantas, (2003).



2.2.4 Modelo de Florestas Aleatórias

O modelo de Floresta Aleatória é baseado em árvores de classificação, que lida bem com conjunto de dados de alta dimensão, e com presença de multicolinearidade (Hastie et al., 2009; Belgiu e Dragut, 2016). Este tipo de modelo é usualmente utilizado não apenas para classificação, mas também para regressão, estudo de importância e seleção de variáveis, e detecção de outliers (Verikas et al., 2011).

O modelo de Floresta Aleatória é usado em diversas aplicações que requerem o aprendizado a partir de dados, como as áreas de diagnóstico médico por imagem (Criminisi e Shotton, 2013), predição rápida de movimentos (Shotton et al., 2013), sensoriamento remoto (Belgiu e Dragut, 2016), análise de big data e detecção de falhas (Costa et al., 2019).

Essa seleção trata-se de um sorteio feito a cada nó da árvore, selecionando aleatoriamente algumas variáveis candidatas para dividir este nó. Com a utilização dessa técnica, diferentes conjuntos de variáveis poderão aparecer em níveis distintos

em cada uma das árvores. Com isso, a técnica se torna mais sensível a interações entre as variáveis, além de resultar em árvores de classificação correlacionadas, devido ao sorteio aleatório das variáveis candidatas a dividir o nó feito a cada partição (Breiman, 2001).

O algoritmo Random Forest gera B amostras bootstrap a partir dos dados originais. Para cada uma das amostras bootstrap, induzir uma árvore sem poda (tamanho máximo) com a seguinte modificação: em cada nó, selecionar aleatoriamente um número M das variáveis explicativas a serem candidatas para dividir o nó e, dentre estas, escolher a que melhor particiona. Predizer novos dados agregando as predições de todos os B modelos, o que pode ser feito com base na média das probabilidades estimadas fornecidas por cada modelo. Em caso de respostas qualitativas, também é usual a utilização da classificação por proporção de votos. Neste caso, conta-se quantas vezes o indivíduo foi classificado em cada classe. Se a proporção de vezes que o indivíduo for classificado em uma classe for maior que um valor preestabelecido (tal como 50%), então será classificado nela (Verikas et al., 2011).

2.3 AVALIAÇÃO DESEMPENHO DOS MODELOS

Para medir o desempenho dos modelos discriminantes no presente estudo, utilizaremos a métrica – Área sob a curva (AUC) ROC (Receiver Operating Characteristics). A curva ROC é uma medida de desempenho para os problemas de classificação em várias configurações de limite. ROC é uma curva de probabilidade e AUC representa o grau ou medida de separabilidade. Ele diz o quanto o modelo é capaz de distinguir entre as classes. Esta é uma métrica de performance dos modelos discriminantes, definindo como o modelo consegue discriminar com sucesso as observações positivas (1) e negativas (0). O uso da AUC, como uma medida de performance tem surgido com especial interesse na área de mineração de dados e aprendizado de máquinas (Rosset 2004).

A AUC demonstra particular robustez e qualidade em analisar dados desbalanceados e menos propenso em deixar de diferenciar modelos não equivalentes (Ling et al., 2003).

Os modelos discriminantes rotulam um evento como presente ou ausente (ou seja, classificação binária). Existem apenas quatro resultados possíveis para cada

ponto de dados que é previsto, ou seja, verdadeiro positivo, verdadeiro negativo, falso positivo e falso negativo.

I. Verdadeiro positivo - (VP) é quando a instância originalmente rotulada como presente é corretamente prevista como presente.
II. Verdadeiro Negativo - (VN) é quando a instância originalmente rotulada como ausente é corretamente prevista como ausente.
III. Falso positivo - (FP) é quando a instância originalmente rotulada como ausente é incorretamente prevista como presente.
IV. Falso Negativo - (FN) é quando a instância originalmente rotulada como presente é incorretamente prevista como ausente.

Uma matriz de confusão é desenhada com base nos resultados acima, é um resumo de todos os resultados previstos após a classificação. As previsões corretas e incorretas são distribuídas com valores de contagem e divididas para cada classe.

A AUC oferece uma medida agregada de desempenho em todos os limites de classificação possíveis. Uma maneira de interpretar a AUC é a probabilidade de o modelo classificar um exemplo positivo aleatório mais alto do que um exemplo negativo aleatório. O AUC varia no valor de 0 a 1. Um modelo com previsões 100% incorretas tem uma AUC de 0,0, e uma com previsões 100% corretas tem uma AUC de 1,0. E quando AUC é 0,5, significa que o modelo não tem capacidade de separação de classes. Por exemplo, quando a AUC é 0,7, significa que há 70% de chance de que o modelo seja capaz de distinguir entre classe positiva e classe negativa.

3 RESULTADOS

3.1 ANÁLISE DESCRITIVA E EXPLORATÓRIA

A base de dados original, com 699 observações, apresentou 114 casos de retenção, um percentual de 16,30%. A propriedade C apresentou a menor frequência do evento, 10,61%, seguidos da propriedade L (18,38%) e propriedade A (24,87%).

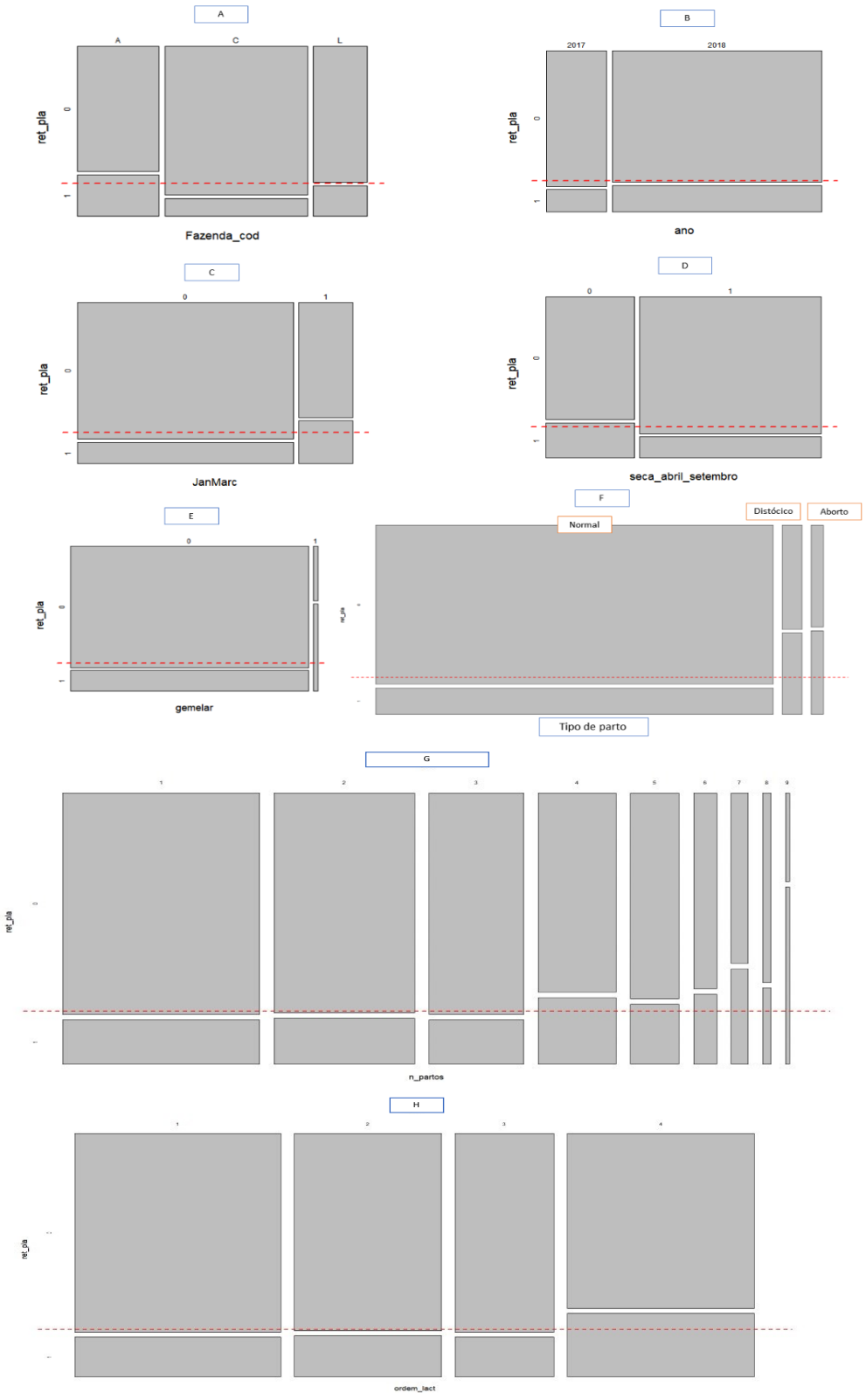
No banco de dados Dt01, com 677 observações, a frequência de retenção de placenta foi de 15,36%, sem grandes alterações com o dado original. Este cenário muda no Dt02, com 493 observações, tendo em vista que o grande problema de registro ocorreu especificamente na propriedade C, que possui a menor frequência do

evento. O banco Dt02 apresentou uma frequência de 19,06%. Esse problema de registro pode ser uma relação de uma menor preocupação do proprietário com o evento, já que ele apresentava a menor taxa de retenção.

A análise exploratória dos dados é demonstrada na Figura 7.

Os dados exploratórios demonstram, como discutido anteriormente, que existe uma relação de maior frequência do evento na propriedade A, (Fig. 7.A). Em relação aos meses do ano (Fig. 7.C e D), a classificação estação do ano, demonstra que no período chuvoso, existe uma maior frequência do evento, e que esse valor é maximizado quando é classificado de janeiro e março. O clima é epidemiologicamente um fator importante para o evento, uma justificativa para o aumento de retenção de placenta é o desafio ambiental e o estresse térmico impostos aos animais provocado pelas elevadas temperaturas, de 35 e 40ºC, e umidade relativa que varia de 60 a 100%, durante o período de chuvas na região de Unaí-MG (INMET,2018).

Figura 7. Análise Exploratória dos dados. Em vermelho destaca-se a frequência média de retenção de placenta no banco de dados.



Além disso, altas temperaturas e chuvas intensificam a contaminação do meio ambiente e o parto distócico (Fig. 7.F) expõe o ambiente uterino por um tempo excepcionalmente longo, facilitando a contaminação bacteriana (Hernández-Castellano et al., 2019).

Existe uma alta ocorrência de parto prematuro quando o parto é gemelar (Fig. 7. E), além da causar uma grande distensão uterina e alto estresse calórico dos animais, incorrendo em uma alta frequência retenção de placenta (Mur-Novales et al., 2018).

Somado às questões ambientais, o excesso de manuseio no parto distócico (Fig. 7.F) devido ao manuseio inadequada de profissionais não técnicos em países tropicais subdesenvolvidos (Hernández-Castellano et al., 2019) leva a um edema local excessivo, resultando em forte adesão à membrana fetal (Cruz et al., 2021).

Geralmente, as vacas com menos partos (Fig. 7. G e H) têm melhores condições endometriais e sistemas imunológicos mais eficientes, o que favorece a expulsão da placenta mais rapidamente em relação às vacas que tiveram mais de quatro parturições (Martins et al., 2013; Van Werven et al., 1992).

3.2 Análise dos Modelos de regressão Logística

Em um primeiro momento avaliou-se a AUC por variável (Tabela 3), buscando verificar, de forma individual, a contribuição individual de cada preditor, tendo como base os dados completos (699 observações), sem retirar dados faltantes, já que para a análise de forma individual, os dados faltantes em outras categorias não são levadas em consideração.

Tabela 3. AUC individual por variável – base de dados completo – 699 observações.

Variável	AUC
Fazenda_cod	**62,04%**
janeiro a março	57,94%
Tipo de parto	57,61%
Período de Seca	57,15%
Número de Partos	56,68%
Ordem de Lactação	56,40%
Parto Gemelar	53,34%
Ano	51,71%

A variável que discrimina melhor a retenção de placenta, é a Fazenda, demonstrando que existe uma importante relação, por exemplo, de manejo e da alimentação fornecida aos animais pelas propriedades que poderia estar explicando a ocorrência do evento, e muito provavelmente, todos esses fatores não foram devidamente levantados na ficha epidemiológica.

A segunda variável que melhor discrimina o evento é classificar a ocorrência da retenção de placenta nos meses de janeiro a março (Sim e Não), demonstrando a importância do estresse térmico e chuvas. A classificação de janeiro e março foi ligeiramente superior a classificação por período de seca e chuva. A classificação de período de seca e chuvas é uma forma mais ampla e generalista de classificar o evento. A classificação de janeiro a março teve como objetivo maximizar a discriminação do evento, mantendo uma relação lógica com o evento.

O tipo de parto foi a terceira variável que melhor discrimina a retenção, demonstrando a importância do parto distócico na ocorrência do evento. O número de partos (valor discreto) e a ordem de lactação (categorização do número de partos) apresentaram desempenho muito similar. A variável com menor poder de discriminar o evento foi o Ano, o que é esperado, por não ser um fator que por si só, levaria a modificação da frequência do evento, deveria ocorrer uma explicação como maior pluviosidade no ano, ou temperaturas, ou até mesmo mudanças de manejo ao decorrer do ano para que fosse uma variável importante.

Estes resultados demonstram que no geral a AUC individual foi baixa, e a variável que melhor poderia discriminar as diferentes características relacionadas ao manejo, clima, alimentação e a raça de animais, ao mesmo tempo, a Fazenda, foi a que apresentou maior AUC, demonstrando que existem características importantes ao desfecho que não estão contempladas na base de dados. E essa questão faz todo sentido, já que muitos estudos atualmente investigam, por exemplo, a questão relacionada a alimentação, como a suplementação de selênio e a retenção de placenta. Essas análises de minerais são bastante complexas e caras, o que não foi direcionado neste estudo (Ferreira, 2010; Rezende, 2013).

Em um segundo momento, foi ajustado um modelo de regressão logística com todas as covariáveis e, então, aplicou-se o método stepwise para seleção do modelo com menor AIC dentre todos os modelos possíveis para o conjunto de covariáveis. O AIC quantifica a quantidade de perda de informações devido a essa simplificação do modelo. O AIC é semelhante ao R-quadrado ajustado, pois também penaliza por adicionar mais variáveis ao modelo.

O modelo selecionado pelo método stepwise para o conjunto de dados Dt01 é apresentado na Figura 8.

Figura 8. Modelo retenção de placenta, banco Dt01.

```
Call:
glm(formula = ret_pla ~ Fazenda_cod + JanMarc + gemelar + tipodeparto,
    family = "binomial", data = dt01)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.3791  -0.6367  -0.3591  -0.3591   2.3553

Coefficients:
                    Estimate Std. Error z value Pr(>|z|)
(Intercept)          -1.4931     0.1890  -7.900 2.79e-15 ***
Fazenda_codC         -1.2161     0.2708  -4.491 7.07e-06 ***
Fazenda_codL         -0.3104     0.2876  -1.079 0.280462
JanMarc               0.9251     0.2602   3.555 0.000378 ***
gemelar               1.9558     0.6234   3.137 0.001706 **
tipodepartoDistócico  1.1557     0.4323   2.673 0.007511 **
tipodepartoF_Aborto   1.0013     1.1955   0.838 0.402286
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 580.78  on 676  degrees of freedom
Residual deviance: 523.29  on 670  degrees of freedom
AIC: 537.29

Number of Fisher Scoring iterations: 5
```

O modelo apresentou um problema de overfitting, com o teste da deviance sendo igual a 1. O overfitting de um modelo é uma condição em que um modelo estatístico começa a descrever o erro aleatório nos dados, em vez das relações entre as variáveis. Esse problema ocorre quando o modelo é muito complexo. Mesmo com vários testes no modelo, nenhum foi capaz de resolver o problema de overfitting.

Biologicamente, todos os coeficientes serem positivos, exceção da Fazenda, tem sentido com a epidemiologia da doença, apesar de não ser ponto central desta análise, não sendo então discutido com maior profundidade.

O teste Hosmer-Lemeshow (hltest) é um teste de qualidade de ajuste para modelos de classificação binária que informa o quão bem os dados se ajustam a um determinado modelo. Especificamente, o hltest calcula se as taxas de eventos observadas correspondem às taxas de eventos esperadas em subgrupos populacionais e podem ser usadas como um diagnóstico de suporte para aceitar ou rejeitar um determinado modelo. O teste apresentou p = 0,09, não demonstrando significância estatística, demonstrando que os dados se ajustam bem ao modelo. A hipótese nula é que as proporções observadas e esperadas são iguais, para cada subgrupo populacional. A hipótese alternativa é que as proporções observadas e esperadas não são as mesmas. Para cada subgrupo populacional, se observa a frequência observada e a esperada, que é gerada a partir do modelo.

O modelo apresentou uma AUC de 69,72%, valor considerado de baixa a moderada capacidade discriminante. Baseado na simplicidade dos dados e na natureza do evento, com uma complexa epidemiologia, o resultado encontrado é satisfatório. Porém em termos operacionais, como utilizar o modelo para prever o evento a campo, pode ser necessário outras variáveis.

O modelo do Dt02, realizou o mesmo procedimento de seleção, lembrando que o modelo Dt02 apresenta 493 observações e engloba os dados de ordem de lactação e número de partos, que por terem dados faltantes em uma das propriedades, foram retirados na primeira análise. A Figura 9 demonstra o modelo de regressão logística da base Dt02.

Figura 9. Modelo retenção de placenta, banco Dt02.

```
Call:
glm(formula = ret_pla ~ Fazenda_cod + JanMarc + gemelar + tipodeparto +
    n_partos, family = "binomial", data = dt02)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.4732  -0.6355  -0.5553  -0.4169   2.2298

Coefficients:
                    Estimate Std. Error z value Pr(>|z|)
(Intercept)          -1.88964    0.26523  -7.124 1.05e-12 ***
Fazenda_codC         -0.82383    0.32264  -2.553  0.01067 *
Fazenda_codL         -0.37346    0.29730  -1.256  0.20905
JanMarc               0.58768    0.29209   2.012  0.04422 *
gemelar               2.24817    0.71388   3.149  0.00164 **
tipodepartoDistócico  1.23164    0.46467   2.651  0.00804 **
tipodepartoF_Aborto   1.51558    1.52594   0.993  0.32061
n_partos              0.15720    0.06304   2.494  0.01265 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 480.37  on 492  degrees of freedom
Residual deviance: 440.96  on 485  degrees of freedom
AIC: 456.96

Number of Fisher Scoring iterations: 4
```

O modelo Dt02 apresentou uma AUC de 64,81%, ligeiramente inferior ao Dt01, lembrando que a diferença entre os bancos de dados ocorre com as variáveis número de partos e ordem de lactação. Podemos inferir que o Dt02 perde poder de predição, especialmente por ter menos observações e essa restrição do banco de dados estar direcionado a uma das fazendas. Assim como no modelo Dt01, o Dt02 apresentou overfitting. Biologicamente, todos os coeficientes serem positivos, exceção da Fazenda, tem sentido com a epidemiologia da doença, apesar de não ser ponto central desta análise, não sendo então discutido com maior profundidade. No hltest, o modelo de regressão logística Dt02, apresentou p = 0,1416, indicando um ajuste do modelo.

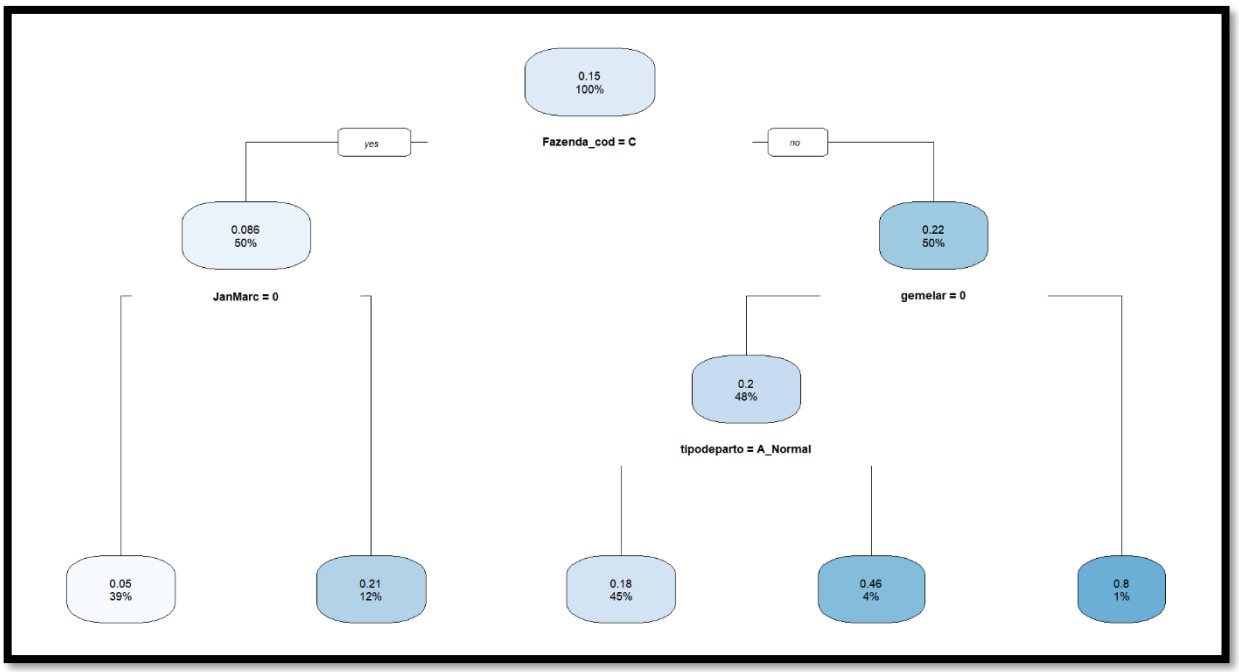
Houve uma mudança no desempenho dos modelos relacionado aos valores faltantes de uma propriedade. Usar o Dt02 foi uma decisão de como o modelo se comportaria sem usar duas variáveis que estavam faltantes em uma propriedade. Podemos verificar que a forma de categorizar as variáveis e os dados faltantes influenciaram o desempenho dos modelos.

3.3 MODELO DE ÁRVORES DE CLASSIFICAÇÃO

Neste trabalho, serão induzidas árvores de classificação, pois a variável resposta é dicotômica retenção de placenta SIM ou NÃO. O pacote utilizado para gerar as Árvores de Classificação foi o rpart, sem limitar o tamanho máximo que a árvore poderia ter.

O resultado gráfico do modelo CART relacionado ao Dt01 é demonstrado a seguir na Figura 10:

Figura 10. Modelo de Árvore de Classificação – Dt01.



O banco começa com 100% dos dados e uma frequência do evento de 15%. O primeiro nó seria a codificação das fazendas, sendo que Fazenda C – sim, que representa 50% dos dados e a frequência do evento é de 8,6%, as Fazendas L e A possuem outros 50% dos dados e 22% de frequência do evento. Assim como no dado individual da AUC, corroboramos com a tese que a fazenda, e todos suas questões não levantadas como manejo e alimentação, possuem influência no evento. Para o lado direito da árvore, relacionado as Fazendas L e A, o parto ser gemelar aumenta a frequência de retenção de placenta para 80%, apesar de apenas 1% destes serem gemelar, demonstrando que é uma condição importante no evento. Outro nó importante, ainda ao lado direito, é o parto ser normal, o parto distócico (normal - não) eleva a frequência para 46%. O parto ser normal, reduz de 22% a retenção de placenta nas fazendas L e A para 18%.

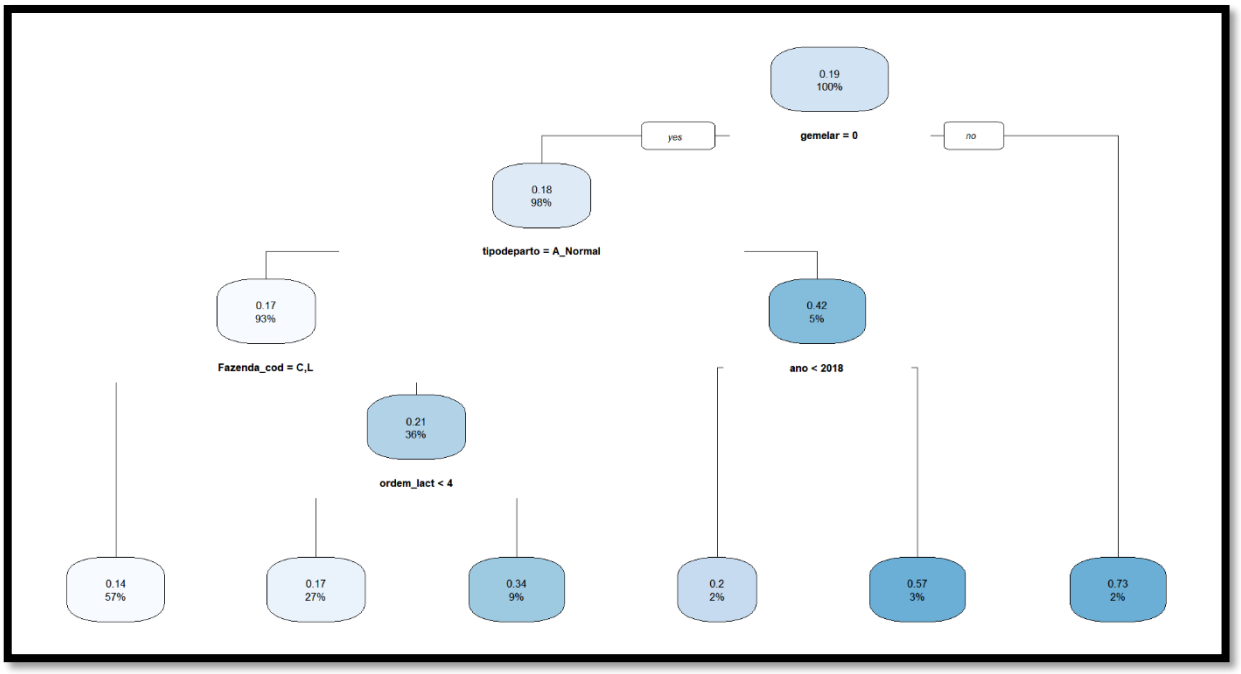
O lado esquerdo, relacionado a fazenda C, que é uma fazenda que apresenta menores frequências de retenção de placenta, o fato do parto ocorrer nos períodos de janeiro a março elevam para 21% a frequência do evento, enquanto nos demais

meses, fica em 5%. Corroborando assim o fator climático, em especial estresse térmico e chuvas.

A AUC do modelo de Árvore de classificação do banco Dt01 foi de 69,91%, similar ao modelo de regressão logística, que apresentou 69,72%.

O resultado gráfico do modelo CART relacionado ao banco Dt02 é demonstrado a seguir na Figura 11:

Figura 11. Modelo de Árvore de Classificação – Dt02.



Interessante que o banco Dt02, por ter os dados relacionados ao número de partos, leva a uma queda de dados da fazenda C, que agora tem 32% dos dados. Como o dado faltante de número de partos está relacionado à fazenda C, o banco Dt02 possui menor influência dela. Assim a fazenda deixa de ser o primeiro nó, que é o parto gemelar. Assim como no banco Dt01, o parto gemelar é uma importante variável, em que o evento sobre para 73% de retenção de placenta quando o parto é gemelar.

O tipo de parto continua sendo uma variável importante, em que, parto simples (primeiro nó), distócico, tem uma frequência de retenção de 42%. Interessante que o segundo nó é uma relação do ano ser 2018 (sim ou não) em que o não, logo ano de 2017, apresentou 57% de retenção de placenta, contra 20% em 2018. Essa

explicação poderia estar relacionado a mão de obra das propriedades no ano, e alguma modificação pode ter ocorrido, como uma mão de obra menos qualificada em 2017, que levava a maior manipulação dos animais e a retenção de placenta, ou até mesmo fatores climáticos mais desafiadores.

Ao lado esquerdo da árvore, ocorre a divisão das propriedades, pelas causas específicas já discutidas anteriormente, em que a Fazenda A (C e L não), vacas de 4 ou mais lactações tem maior frequência de retenção de placenta, 34%, contra 17% em vacas de 1, 2 e 3 lactação.

A AUC do modelo CART do Dt02 foi de 63,90%, ligeiramente inferior ao modelo de regressão logística, que apresentou 64,81%.

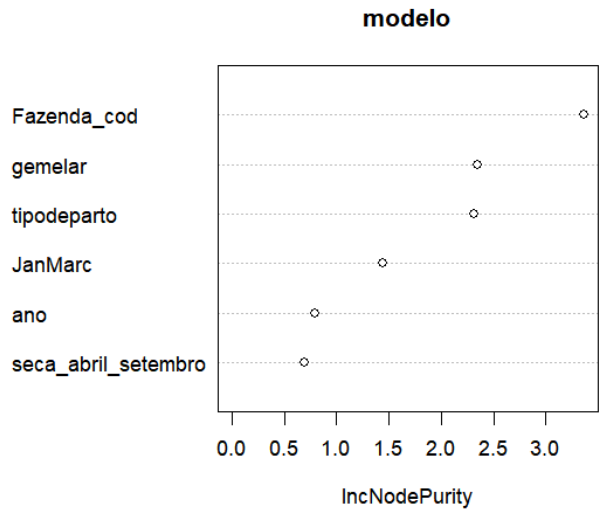
No geral, avaliando os modelos de regressão e CART, o desempenho foi muito similar, o que seria uma vantagem do modelo CART é conseguir demonstrar de uma forma gráfica e simples uma complexidade do modelo e do banco de dados, que o modelo de regressão não possibilita.

3.4 MODELO DE FLORESTA ALEATÓRIA

Essa seleção trata-se de um sorteio feito a cada nó da árvore, selecionando aleatoriamente algumas variáveis candidatas para dividir este nó. Com a utilização dessa técnica, diferentes conjuntos de variáveis poderão aparecer em níveis distintos em cada uma das árvores. Com isso, a técnica se torna mais sensível a interações entre as variáveis. O modelo de Floresta Aleatória foi gerado a partir do pacote *randomForest*, definindo um número máximo de florestas como 500.

O modelo de Florestas Aleatórias, por ser um modelo de caixa preta, não apresenta uma saída simples da relação entre as variáveis, uma forma de verificar a importância de cada variável no modelo é através do comando varImpPlot, que demonstra dentro das médias dos diversos modelos criados, como cada variável se comporta, conforme Figura 12.

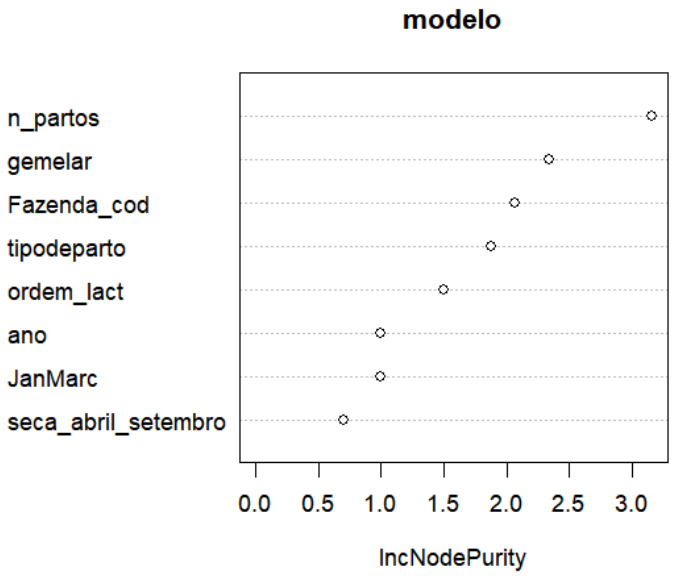
Figura 12. Importância das variáveis no modelo de Floresta Aleatória – Dt01.



Novamente é demonstrada a importância da Fazenda, partos gemelares e o tipo de parto. A AUC do modelo foi de 56,36%, tendo uma redução considerável dos modelos de regressão logística e CART.

O gráfico de importância das variáveis dentro do modelo de florestas aleatórias para o banco de dados Dt02, é demonstrado a seguir na Figura 13:

Figura 13. Importância das variáveis no modelo de Floresta Aleatória – Dt02.



A variável número de partos passa a ser a de maior importância no modelo, seguido da variável parto gemelar, código da fazenda e tipo de parto. Interessante, que para o banco de dados Dt02 a AUC do modelo de Florestas Aleatórias foi de 50,27%, um valor muito baixo,

demonstrando que o desempenho do modelo é praticamente nulo, tendo em vista que 50% seria um modelo sem poder de discriminar. Esses valores são consideravelmente menores que os 64,81% da regressão logística e 63,90% do CART.

3.5 ANÁLISE COMPARATIVA DOS 3 MODELOS

Na tabela 4, podemos observar a métrica de desempenho de cada modelo proposto.

Tabela 4. Valores de AUC para os modelos propostos nos dois bancos de dados.

Modelo	Banco	AUC
Regressão Logística	Dt01	**69,72%**
CART	Dt01	**69,91%**
Random Forest	Dt01	**53,36%**
Regressão Logística	Dt02	**64,81%**
CART	Dt02	**63,91%**
Random Forest	Dt02	**50,27%**

4 CONCLUSÃO

Este trabalho se propôs a investigar o desempenho de modelos de regressão logística, árvore de classificação e florestas aleatórias na discriminação do evento de retenção de placenta em um banco de dados de 3 propriedades leiteiras da região de Unaí, noroeste de Minas Gerais.

O banco de dados desbalanceado se demonstrou um grande obstáculo em gerar modelos com AUC satisfatórias. Uma explicação para o resultado de modelos com baixo poder de predição é que o número e o tipo de variáveis no estudo são incapazes de discriminar um evento tão complexo como a retenção de placenta. Um grande indicativo deste problema é que o código da fazenda sistematicamente foi uma das variáveis mais importantes em todos os modelos propostos. Estudos demonstram que fatores como alimentação e status sanitários dos animais são de grande importância ao prever a retenção de placenta, todavia, esses dados são complexos de serem levantados e com custos relativamente altos, o que não era o proposito inicial do banco de dados, que foi gerado a partir de uma ficha simples. A variável fazenda, como discutido anteriormente, dentro as variáveis levantadas, é aquela que conseguiria discriminar, ao mesmo tempo, um número alto de variáveis relacionadas a alimentação e manejo, que não foram levantadas.

Assim, avaliando individualmente cada modelo, o desempenho da regressão logística e CART, modelos menos complexos e de entendimento mais direto, obtiveram AUC maiores que o de Floresta Aleatória, com valores de 65 a 69%. Baseados na discussão anterior, da complexidade do evento estudado e variáveis levantadas, podemos considerar que o desempenho dos modelos mais simples foi minimamente satisfatório e superiores aos modelos mais complexos.

Importante discutir que o atual problema, se trata de uma base de dados relativamente pequena, com poucas variáveis preditoras. Em um universo com maior complexidade de dados e de grande quantidade de variáveis no modelo, as técnicas de Aprendizado de Máquina (Machine Learning), e modelos caixa preta, podem auxiliar a modelar tal complexidade, que seria extremamente complicado de ser realizado de forma manual, variável por variável.

5 REFERÊNCIAS

Beagley, J. C., et al. "Physiology and treatment of retained fetal membranes in cattle." Journal of veterinary internal medicine 24.2 (2010): 261-268.

Belgiu, Mariana, and Lucian Drăguţ. "Random forest in remote sensing: A review of applications and future directions." ISPRS journal of photogrammetry and remote sensing 114 (2016): 24-31.

Breiman, Leo, et al. Classification and regression trees. Routledge, 2017.

Breiman, Leo. "Statistical modeling: The two cultures (with comments and a rejoinder by the author)." Statistical science 16.3 (2001): 199-231.

Buso, Rodrigo R., et al. "Retained placenta and subclinical endometritis: prevalence and relation with reproductive performance of crossbred dairy cows." Pesquisa Veterinária Brasileira 38 (2018): 1-5.

Costa, Marcelo Azevedo, et al. "Failure detection in robotic arms using statistical modeling, machine learning and hybrid gradient boosting." Measurement 146 (2019): 425-436.

Criminisi, Antonio, and Jamie Shotton, eds. Decision forests for computer vision and medical image analysis. Springer Science & Business Media, 2013.

Cruz, Irene, et al. "Clinical disease incidence during early lactation, risk factors and association with fertility and culling in grazing dairy cows in Uruguay." Preventive veterinary medicine 191 (2021): 105359.

Dahl, Geoffrey E., Sha Tao, and Jimena Laporta. "Heat stress impacts immune status in cows across the life cycle." Frontiers in veterinary science 7 (2020): 116.

Dohoo, I. Martin, W. and Stryhn H. No Title. Veterinary Epidemiology. 1st ed., Atlantic Veterinary College, Charlottetown; 2003

Ferreira, A. de M. "Reprodução da fêmea bovina: fisiologia aplicada e problemas mais comuns (causas e tratamentos)." Juiz de Fora: Minas Gerais–Brasil (2010): 422.

Hastie, Trevor, et al. The elements of statistical learning: data mining, inference, and prediction. Vol. 2. New York: springer, 2009.

Hernández-Castellano, Lorenzo E., et al. "Dairy science and health in the tropics: challenges and opportunities for the next decades." Tropical Animal Health and Production 51.5 (2019): 1009-1017.

Hosmer, David W., Borko Jovanovic, and Stanley Lemeshow. "Best subsets logistic regression." Biometrics (1989): 1265-1270.

INMET. National Institute of Meteorology. 2021 n.d. https://portal.inmet.gov.br/ (accessed May 30, 2021).

Ling, Charles X., Jin Huang, and Harry Zhang. "AUC: a statistically consistent and more discriminating measure than accuracy." Ijcai. Vol. 3. 2003.

Loh, Wei-Yin. "Classification and regression trees." Wiley interdisciplinary reviews: data mining and knowledge discovery 1.1 (2011): 14-23.

Louppe, Gilles. "Understanding random forests: From theory to practice." arXiv preprint arXiv:1407.7502 (2014).

Martins, T. M., et al. "Reproductive and productive parameters of Holstein cows with normal or pathological puerperium." Arquivo Brasileiro de Medicina Veterinária e Zootecnia 65 (2013): 1348-1356.

McCullagh, Peter, and John A. Nelder. Generalized linear models. Routledge, 2019.

Mur-Novales, R., et al. "An economic evaluation of management strategies to mitigate the negative effect of twinning in dairy herds." Journal of dairy science 101.9 (2018): 8335-8349.

Nobre, M. M., et al. "Evaluation of incidence rate and risk factors of retained placenta of crossbred dairy cattle." Arquivo Brasileiro de Medicina Veterinária e Zootecnia 64 (2012): 101-107.

R Core Team (2022). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL https://www.R-project.org/.

Rezende, Estevão Vieira de. "Incidência da retenção de placenta e as consequências na produção de leite e na eficiência reprodutiva de vacas holandesas." (2013).

Rosset, Saharon. "Model selection via the AUC." Proceedings of the twenty-first international conference on Machine learning. 2004.

Segal, Mark R. "Machine learning benchmarks and random forest regression." (2004).

Taconeli, Cesar Augusto. "Árvores de classificação multivariadas fundamentadas em coeficientes de dissimilaridade e entropia." São Paulo: Universidade de São Paulo (2008).

Van Werven, T., et al. "The effects of duration of retained placenta on reproduction, milk production, postpartum disease and culling rate." Theriogenology 37.6 (1992): 1191-1203.

Venables, W. M., Ripley, B. D. Modern Applied Statistics with S, Fourth edition. Springer, New York. ISBN 0-387-95457-0, https://www.stats.ox.ac.uk/pub/MASS4/.

Verikas, Antanas, Adas Gelzinis, and Marija Bacauskiene. "Mining data with random forests: A survey and results of new tests." Pattern recognition 44.2 (2011): 330-349.

Wang, Hongwei, F. Richard Yu, and Hailin Jiang. "Modeling of radio channels with leaky coaxial cable for LTE-M based CBTC systems." IEEE Communications Letters 20.5 (2016): 1038-1041.

Wu, Xindong, et al. "Top 10 algorithms in data mining." Knowledge and information systems 14.1 (2008): 1-37.

Yadav, Subhash Kumar, and Yusuf Akhter. "Response: Commentary: Statistical Modeling for the Prediction of Infectious Disease Dissemination With Special Reference to COVID-19 Spread." Frontiers in Public Health 9 (2021).