MOMATE EMATE OSSIFO

SELEÇÃO DE VARIÁVEIS EM MODELO DE REGRESSÃO LOGÍSTICA PARA PREDIÇÃO DA RESISTÊNCIA À BRUSONE DO ARROZ

Dissertação apresentada à Universidade Federal de Viçosa, como parte das exigências do Programa de Pós-Graduação em Estatística Aplicada e Biometria, para obtenção do título de *Magister Scientiae*.

Orientador: Sebastião Martins Filho

VIÇOSA - MINAS GERAIS
2020

Ficha catalográfica preparada pela Biblioteca Central da Universidade Federal de Viçosa - Câmpus Viçosa

T
O84s
2020

Ossifo, Momate Emate, 1988-
Seleção de variáveis em modelo de regressão logística para predição da resistência à brusone do arroz / Momate Emate Ossifo. – Viçosa, MG, 2020.
55f. : il. (algumas color.) ; 29 cm.

Inclui apêndice.
Orientador: Sebastião Martins Filho.
Dissertação (mestrado) - Universidade Federal de Viçosa.
Referências bibliográficas: f.43-46.

1. Arroz - Resistência a doenças e pragas - Métodos estatísticos. 2. *Oryza sativa* L.. 3. Curva característica de operação do receptor. 4. Modelos matemáticos. 5. Probabilidades. I. Universidade Federal de Viçosa. Departamento de Estatística. Programa de Pós-Graduação em Estatística Aplicada e Biometria. II. Título.

CDD 22 ed. 633.18946

MOMATE EMATE OSSIFO

SELEÇÃO DE VARIÁVEIS EM MODELO DE REGRESSÃO LOGÍSTICA PARA PREDIÇÃO DA RESISTÊNCIA À BRUSONE DO ARROZ

Dissertação apresentada à Universidade Federal de Viçosa, como parte das exigências do Programa de Pós-Graduação em Estatística Aplicada e Biometria, para obtenção do título de *Magister Scientiae*.

APROVADA: 20 de fevereiro de 2020.

Assentimento:

Momate Emate Ossifo

Momate Emate Ossifo
(Autor)

Sebastião Martins Filho
(Orientador)

Aos meus pais, Emate Ossifo Salato (*in memoriam*) e Rosa Félix Luís.

Aos meus sogros, Jaime e Fatima Muchico

A minha esposa, Joaquina da Marcia Jaime Muchico.

Ao meu filho, Ramadane Jaime Emate Ossifo.

Aos meus irmãos, primos e sobrinhos.

Aos demais familiares.

Aos amigos.

Dedico.

AGRADECIMENTOS

A Deus, por me guiar em toda essa caminhada e me dar forças em todos os momentos.

Aos meus pais, pelos ensinamentos e por sempre incentivarem nos estudos e apoiarem nas minhas decisões.

Aos meus irmãos (Lagimo, Maria, Dado, Rabia, Shaquil e Yassin), pelo conforto e apoio na minha trajetória.

Aos meus primos, pela força e pelo carinho, em especial (Abdul Latibo Momade Mussa) por todo o apoio durante a minha graduação.

Aos meus sobrinhos, pelo apoio moral e pela força.

A minha esposa, pela paciência e carinho, e que mesmo distante não deixou de ser excelente companheira em todos os momentos.

Aos meus sogros, pelos conselhos, apoio e encorajamento.

À Rehana Amuji, amiga e colega de trabalho na Escola Superior de Desenvolvimento Rural (ESUDER), pelo apoio, pela forca, pelo carinho, e pela disponibilidade quando mais precisei.

Ao meu orientador, Professor Doutor Sebastião Martins Filho, pelos ensinamentos, pela paciência, simplicidade e generosidade.

Aos professores membros da banca, Professor Doutor Antônio Policarpo Sousa Carneiro e Professor Doutor Vinícius Silva dos Santos, pelas críticas e sugestões para melhoria do trabalho.

A todos os professores do programa de Pós-graduação em Estatística Aplicada e Biometria, pelos seus ensinamentos e apoio.

Ao secretário Júnior José Pires pela constante prontidão em auxiliar nas diversas atividades burocráticas do PPESTBIO/UFV.

Aos colegas de Mestrado em Estatística pela amizade e ajuda, sempre ajudando uns aos outros a superar cada degrau, e aos amigos pela confiança em mim depositada.

À Universidade Federal de Viçosa, Campus Viçosa, pelo acolhimento nesta caminhada.

À Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (CAPES), pela concessão da bolsa de estudos, durante o período de realização deste trabalho. O presente trabalho foi realizado com apoio da CAPES – Brasil – Código de Financiamento 001.

BIOGRAFIA

MOMATE EMATE OSSIFO, filho de Emate Ossifo Salato e de Rosa Felix Luís, nasceu em Moçambique, distrito de Mocuba – Província da Zambézia, no dia 11 de outubro de 1988.

Em fevereiro de 2009, ingressou no curso de Licenciatura em Ensino de Matemática da Universidade Pedagógica de Moçambique – Delegação de Quelimane – Moçambique, graduando-se em julho de 2013.

Em fevereiro de 2013, ingressou na Escola Secundaria Geral de Lioma – Gurué (Moçambique), como professor de Matemática para o ensino médio.

Em fevereiro de 2015, ingressou na Universidade Eduardo Mondlane (UEM) – Moçambique, como professor assistente estagiário nas disciplinas de analise matemática e estatística.

Em agosto de 2018 iniciou o curso de Mestrado no Programa de Pós-graduação em Estatística Aplicada e Biometria na Universidade Federal de Viçosa – Brasil, submetendo-se à defesa de dissertação em 20 de fevereiro de 2020.

RESUMO

OSSIFO, Momate Emate, M.Sc., Universidade Federal de Viçosa, fevereiro de 2020. **Seleção de variáveis em modelo de regressão logística para predição da resistência à brusone do arroz**. Orientador: Sebastião Martins Filho.

Ao longo da história, o arroz (*Oryza sativa* L.) tem sido um dos alimentos mais consumido no planeta, apresentando fundamental importância econômica e social. Dentre os fatores limitantes do potencial produtivo do arroz estão as doenças. Entre estas a brusone, causada pelo fungo *Pyricularia oryzae* é a maior causadora dos danos tanto na produtividade como na qualidade de grãos, podendo comprometer até 100% da produção. Neste trabalho, para a predição da resistência à doença foi utilizado o modelo de regressão logística, em que, a probabilidade da resposta predita pode classificar um indivíduo de acordo com um dos dois grupos suscetíveis ($\hat{y} = 0$) ou resistentes ($\hat{y} = 1$). Foi utilizada ainda a curva ROC, de forma a permitir a avaliação da capacidade preditiva do modelo usando o ponto de corte escolhido. Desse modo, teve-se como objetivo geral, selecionar um conjunto de variáveis que podem influenciar na resistência à brusone do arroz. A estratégia de seleção derivada da proposta de Collett foi utilizada para seleção das variáveis, e dentre as quinze variáveis iniciais utilizadas para avaliar a brusone do arroz, apenas cinco fizeram-se importantes, ao nível de significância de 0,10 pelo teste da razão de verossimilhança. Essas variáveis, fazem parte do modelo 2 (modelo de regressão logística múltipla com as variáveis $\mathrm{V4} + \mathrm{V8} + \mathrm{V11} + \mathrm{V15} + \mathrm{V4} \times \mathrm{V15}$) selecionado, uma vez que ele foi bem ajustado e não causou grandes alterações nas estimativas e ainda foi o modelo mais parcimonioso. Entretanto, apenas as variáveis largura da folha bandeira (V4), número da ramificação da panícula primária (V8) e quantidade de amilose presente nos grãos moídos (V15) foram as que mostraram significância estatística ao nível de 5% no modelo. Desta forma, foi então ajustado um novo modelo incluindo apenas essas três variáveis, tendo mostrado uma boa qualidade de ajuste pelo teste de Hosmer e Lemeshow ao nível de significância de 5%. Sendo que, para cada cm de V4, espera-se 279,3% de acréscimo na probabilidade da resistência à brusone do arroz, e para cada unidade de acréscimo em V8, espera-se em média o acréscimo de 31,9% na probabilidade da resistência à brusone do arroz, e ainda para cada unidade de acréscimo em V15, espera-se o acréscimo de 9,4% na probabilidade da resistência à brusone do arroz.

PALAVRAS-CHAVE: Arroz-Resistência a doenças e pragas-Métodos estatísticos. *Oryza sativa* L. Curva característica de operação do receptor. Modelos matemáticos. Probabilidades.

ABSTRACT

OSSIFO, Momate Emate, M.Sc. Universidade Federal de Viçosa, February 2020. **Selection of variables in a logistic regression model to predict rice blast resistance**. Adviser: Sebastião Martins Filho.

Throughout history, rice (*Oryza sativa* L.) has been one of the most consumed foods on the planet, presenting fundamental economic and social importance. Among the limiting factors of the productive potential of rice are diseases. Among these, blast, caused by the fungus *Pyricularia oryzae*, is the biggest cause of damage both in productivity and in grain quality, which can compromise up to 100% of production. In this work, the logistic regression model was used to predict disease resistance, in which the probability of the predicted response can classify an individual according to one of the two susceptible ($\hat{y} = 0$) or resistant ($\hat{y} = 1$). The ROC curve was also used, in order to allow the evaluation of the predictive capacity of the model using the chosen cutoff point. Thus, the general objective was to select a set of variables that can influence the resistance to rice blast. The selection strategy derived from Collett's proposal was used to select variables, and among the fifteen initial variables used to assess rice blast, only five became important, at the significance level of 0.10 by the ratio test. likelihood. These variables are part of model 2 (multiple logistic regression model with the variables $V4 + V8 + V11 + V15 + V4 \times V15$) selected, since it was well adjusted and did not cause major changes in the estimates and it was still the model more parsimonious. However, only the variables of the leaf width (V4), the number of the primary panicle branch (V8) and the amount of amylose present in the ground grains (V15) were those that showed statistical significance at the level of 5% in the model. In this way, a new model was then adjusted including only these three variables, having shown a good quality of adjustment by the Hosmer and Lemeshow test at the 5% significance level. Since, for each cm of V4, a 279.3% increase in the probability of resistance to rice blast is expected, and for each unit of increase in V8, an average increase of 31.9% in the probability is expected resistance to rice blast, and for each unit of increase in V15, a 9.4% increase in the probability of resistance to rice blast is expected.

Keywords: Rice-Disease and pest resistance-Statistical methods. *Oryza sativa* L. Receiver operating characteristic curve. Mathematical models. Probabilities.

.

SUMÁRIO

1. INTRODUÇÃO

Ao longo da história, o arroz (*Oryza sativa* L.) tem sido um dos alimentos mais consumido no planeta, apresentando fundamental importância econômica e social. O cereal supre, no mínimo, metade da caloria energética da população mundial, principalmente para as populações pobres dos países de regiões tropicais e subtropicais, e dos chamados países emergentes ou em fase de desenvolvimento. No continente asiático, onde 90% desse cereal é cultivado, o consumo médio é alto, de 78kg/pessoa/ano, enquanto que em países situados na América Latina o consumo médio fica em torno de 29 kg/pessoa/ano, com destaque para o Brasil, considerado um grande consumidor de arroz, com média de 32 kg/pessoa/ano (SOSBAI, 2018).

Em todas as fases de crescimento e desenvolvimento desta cultura, fatores bióticos e abióticos podem impactar diretamente sobre a disponibilidade deste alimento. Dentre os fatores limitantes da expressão do potencial produtivo do arroz estão as doenças, com destaque para a brusone, causada pelo fungo *Pyricularia oryzae*, maior causadora dos danos tanto na produtividade como na qualidade de grãos, comprometendo até 100% da produção Sosbai (2018). No Brasil, tanto a brusone nas panículas quanto nas folhas apresenta-se como um dos principais fatores que afetam a produtividade, tanto no sistema de cultivo de terras altas quanto no irrigado.

Na avaliação de doenças a regressão logística é o método estatístico mais empregado, quando se pretende verificar a relação entre uma variável resposta binária ou dicotômica e variáveis explicativas de interesse. Neste caso, a variável resposta (y), geralmente, apresenta duas possibilidades de resposta: resistência (atribui-se o valor $y = 1$), e ao resultado complementar suscetível (atribui-se o valor $y = 0$).

No modelo logístico a probabilidade da resposta predita pode consequentemente formar a base para se classificar um indivíduo de acordo com um dos dois grupos ($\hat{y} = 0$ ou $\hat{y} = 1$). Desta forma, é preciso estabelecer um ponto de corte em que valores acima dele o indivíduo é classificado como resistente e valores abaixo dele ou iguais o indivíduo é classificado como suscetível. Geralmente 0,5 é um valor razoável para este ponto de corte, entretanto, se os dois grupos não podem ser classificados como simétricos, um valor diferente de 0,5 deve ser considerado. Uma maneira de se determinar este ponto de corte é por meio da curva ROC (*Receiver Operating Characterisctic*), a qual permite avaliar a capacidade preditiva de um modelo usando o ponto de corte escolhido (ABREU, 2004).

A avaliação da capacidade preditiva de um modelo, após a classificação das observações em um dos dois grupos ($\hat{y} = 0$ ou $\hat{y} = 1$), é baseada nos conceitos de sensibilidade e

especificidade estatística, obtidas a partir da construção de matrizes de confusão, gerada pelo modelo. A sensibilidade é definida como a capacidade do modelo em identificar uma classificação positiva, dado que ela realmente é positiva. A especificidade é definida como a capacidade do modelo em identificar um resultado negativo, dado que ele é realmente negativo (MARTINEZ, et al., 2003).

Para a construção da Curva ROC, são calculadas a sensibilidade e a especificidade para todas as observações da amostra, considerando diferentes pontos de corte do modelo. A curva é obtida registrando em um gráfico "sensibilidade" *versus* "1 – especificidade" para os diversos pontos de corte. Em geral, o melhor ponto de corte produz valores para a sensibilidade e a especificidade que se localizam no ponto mais à esquerda e superior possível, isto é, no "ombro" da curva, ou próximo dele (OLIVEIRA, 2011).

2. OBJETIVOS

2.1. GERAL

Selecionar um conjunto de variáveis que podem influenciar na resistência à brusone do arroz, utilizando a estratégia de seleção derivada da proposta de Collett.

2.2. ESPECÍFICOS

Verificar os modelos preditivos que apoiem efetivamente na decisão sobre a característica resistência à brusone do arroz;

Avaliar a capacidade preditiva destes modelos por meio das medidas de sensibilidade e especificidade, utilizando para isto a área sob a Curva ROC.

3. REVISÃO DE LITERATURA

3.1. IMPORTÂNCIA DO ARROZ (*Oryza sativa* L.)

O arroz é um dos cereais mais cultivados e consumidos no mundo, sendo um dos alimentos mais importantes para a nutrição humana (SOSBAI, 2018). Aproximadamente 90% do arroz produzido no mundo são oriundos do continente asiático, sendo a China, Índia e Indonésia os maiores produtores mundiais, respectivamente. Além disso, o maior volume de exportação pertence à Tailândia e o de importação a Indonésia, ambos da Ásia (SOSBAI, 2018).

O Brasil é o nono país produtor em escala mundial, com a produção de 11,3 milhões de toneladas, sendo o maior produtor das Américas e apontado como um produtor primordial entre os países ocidentais (USDA, 2018). O estado do Rio Grande de Sul contribuiu com 8,4 milhões de toneladas na safra 2017/2018. É considerado o Estado mais produtivo dentre os Estados brasileiros, e o sistema de cultivo utilizado, o irrigado por inundação, apresenta uma elevada contribuição na resposta do rendimento (CONAB, 2018).

No continente asiático, o consumo médio é alto, de 78kg/pessoa/ano, enquanto que em países situados na América Latina o consumo médio fica em torno de 29 kg/pessoa/ano, ressaltando o Brasil que é considerado um grande consumidor de arroz, com a média de 32 kg/pessoa/ano. Em decorrência, desempenha papel estratégico na solução de questões de segurança alimentar. E, apesar do grande volume produzido, o arroz é um produto com pequeno comércio internacional (SOSBAI, 2018).

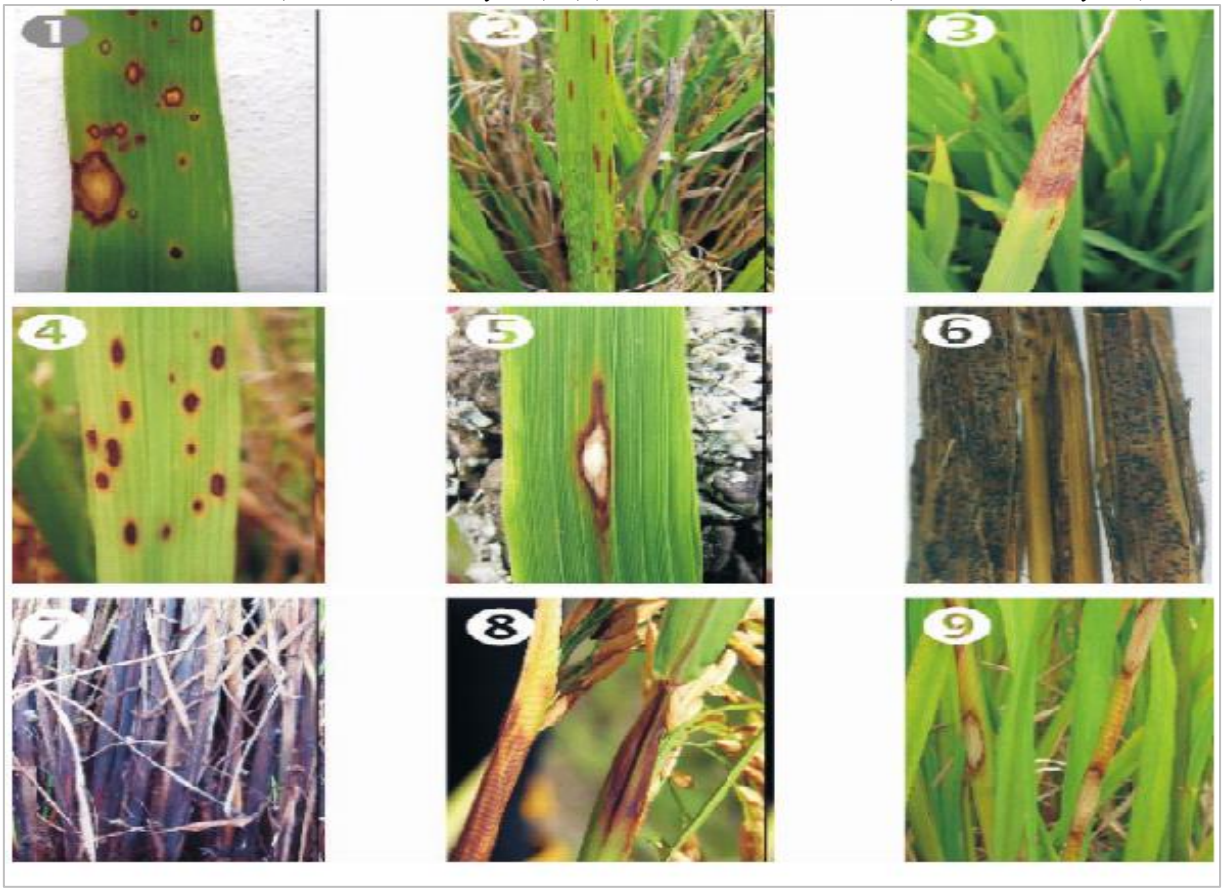
Entretanto, segundo Sosbai (2018), o Brasil com uma produção anual, base casca, entre 11 e 13 milhões de toneladas de arroz nas últimas safras, participa com 78% da produção do Mercosul (na média de 2009/10 até 2017/18), seguido pelo Uruguai, Argentina e, por último, o Paraguai, que na safra 2017/18 representou em torno de 6,00% do total produzido. Contudo, o volume exportado ainda é baixo, sendo que o Brasil possui condições para deixar de ser um ofertante residual de arroz no mercado internacional e se tornar um importante país exportador, tendo em vista a capacidade de expansão de áreas de cultivo, capacidade de inovar em aspectos ligados a tecnologia de produção, com várias instituições de pesquisas e universidades que disponibilizam tecnologias que aumentam potencial de produção da orizicultura.

3.2. PRINCIPAIS DOENÇAS DO ARROZ (*Oryza sativa* L.)

Dentre os fatores limitantes da expressão do potencial produtivo do arroz estão as doenças, como a mancha circular (*Alternaria padwickii*), mancha estreita (*Cercospora oryzae*),

escaldadura da folha (*Gerlachia oryzae*), mancha parda (*Bipolaris oryzae*), brusone (*Pyricularia oryzae*), podridão do colmo (*Sclerotium oryzae*), mal do pé (*Gaeumannomyces graminis* var. *graminis*), podridão da bainha (*Sarocladium oryzae*) e mancha das bainhas (*Rhizoctonia oryzae*), conforme pode ser visto na Figura 1.

Figura 1 – Doenças ocorrentes na cultura do arroz irrigado: (1) Mancha circular (*Alternaria padwickii*), (2) Mancha estreita (*Cercospora oryzae*), (3) Escaldadura da folha (*Gerlachia oryzae*), (4) Mancha parda (*Bipolaris oryzae*), (5) Brusone (*Pyricularia oryzae*), (6) Podridão do colmo (*Sclerotium oryzae*), (7) Mal do pé (*Gaeumannomyces graminis* var. *graminis*), (8) Podridão da bainha (*Sarocladium oryzae*), (9) Mancha das bainhas (*Rhizoctonia oryzae*)



Fonte: Sosbai, 2018, p. 153.

Das doenças acima descritas, a brusone, causada pelo fungo *Pyricularia oryzae*, é a maior causadora dos danos tanto na produtividade como na qualidade de grãos, podendo comprometer até 100% da produção (SOSBAI, 2018). Apesar de todos os esforços de pesquisa visando o desenvolvimento de cultivares resistentes à brusone, a doença ainda permanece como um dos principais fatores limitantes de produtividade do arroz no Mundo.

Os danos causados pela brusone são variáveis dependendo do grau de suscetibilidade da cultivar, das condições climáticas e das práticas culturais adotadas. Enquanto a brusone nas

folhas causa danos indiretos na produção de grãos, pela redução da taxa de fotossíntese e respiração, a brusone nas panículas afeta diretamente a formação e o peso dos grãos (SOSBAI, 2018).

3.3. MODELO DE REGRESSÃO LOGÍSTICA SIMPLES

O modelo de regressão logística é semelhante ao modelo de regressão linear. No entanto, no modelo logístico a variável resposta Y_i geralmente é binária, a qual assume dois valores, como por exemplo, $Y_i = 0$ e $Y_i = 1$ denominados "fracasso" e "sucesso", respectivamente. Assumindo $\pi_i = P(Y = 1)$ a probabilidade de "sucesso" e $1 - \pi_i = P(Y = 0)$ a probabilidade de "fracasso".

No modelo de regressão linear simples, tem-se que $Y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$. Assumindo $E(\varepsilon_i) = 0$, obtemos,

$$E(Y_i|X = x_i) = \beta_0 + \beta_1 x_i \,,$$

sendo as constantes β_0 e β_1 os parâmetros do modelo, onde β_0 é o parâmetro que representa o valor da variável resposta (Y_i) assumido $x = 0$, e β_1 representa o quanto aumenta ou diminui a variável resposta para a variação de cada unidade da variável independente x. Para qualquer valor x dentro do intervalo de $-\infty$ a $+\infty$, sempre poderá existir um valor esperado de Y, assim $-\infty < E(Y_i|X = x_i) < +\infty$.

Na regressão logística, devido à natureza da variável resposta (binaria ou dicotômica), sua média condicional deve ser maior ou igual a zero e menor ou igual a 1, ou seja, $0 \leq E(Y_i|X = x_i) \leq 1$. E, pela definição da variável aleatória discreta tem-se:

$$E(Y_i|X = x_i) = 1 \cdot P(Y_i = 1|X = x_i) + 0 \cdot P(Y_i = 0|X = x_i) = \pi_i$$

No entanto, assumindo $Y_i \sim Bern(\pi_i)$ a probabilidade de sucesso do modelo de regressão logística simples, também chamado de modelo $logit$, pode-se definir como:

$$\pi_i = \pi(x_i) = P(Y_i = 1|X = x_i) = \frac{\exp(\beta_0 + \beta_1 x_i)}{1 + \exp(\beta_0 + \beta_1 x_i)} \,, \qquad (1)$$

e a probabilidade de fracasso,

$$1 - \pi_i = 1 - \pi(x_i) = P(Y_i = 0|X = x_i) = \frac{1}{1 + \exp(\beta_0 + \beta_1 x_i)} \,, \qquad (2)$$

sendo, β_0 e β_1 dois parâmetros desconhecidos.

Para Hosmer e Lemeshow (1989), no modelo de regressão logística o valor da variável resposta (Y_i) pode ser expresso por $Y_i = \pi_i + \varepsilon_i$, sendo ε_i o erro do modelo, com uma

distribuição Binomial (1, π_i), com média zero e variância $\pi_i(1-\pi_i)$ e pode apenas assumir dois valores, isto é, $\varepsilon_i = 1-\pi_i$ para $Y_i = 1$ ou $\varepsilon_i = -\pi_i$ para $Y_i = 0$.

Sabendo que o modelo (1) é não linear, aplica-se uma transformação denominada $g(x)$ para tornar o modelo $logit$ linear em seus parâmetros contínuos e fazer com que assuma valores entre $-\infty$ a $+\infty$, dependendo do limite da variável x. Assim:

$$g(x_i) = \ln \frac{P(ocorrência\ do\ evento|x_i)}{P(não\ ocorrência\ do\ evento|x_i)} = \ln\left(\frac{\pi_i}{1-\pi_i}\right) = \beta_0 + \beta_1 x_i\,, \qquad (3)$$

sendo a razão $\frac{\pi_i}{1-\pi_i}$ chamada de Odds (chance) e $g(x) = \ln\left(\frac{\pi_i}{1-\pi_i}\right)$ é a função de ligação canônica para o modelo binomial.

3.3.1. ESTIMAÇÃO DOS PARÂMETROS DO MODELO

No modelo de regressão linear, o método mais usado para estimação dos parâmetros (β_0 e β_1) é dos mínimos quadrados. Sendo que a escolha desses parâmetros é dada pelos valores que maximizam a soma de quadrados dos desvios para os valores observados (y_i) em relação aos valores preditos ($\hat{y}_i$) baseado no modelo.

No entanto, para o modelo de regressão logística com os valores da variável independente x conhecidos, e apenas os parâmetros (β_0 e β_1) desconhecidos, a estimação desses parâmetros é realizada por meio do método da máxima verossimilhança, (HOSMER e LEMESHOW, 2000).

Segundo Mayer (1978), o método da máxima verossimilhança conduz as estimativas mais razoáveis para os dados dicotômicos. A função de probabilidade y_i para o modelo de regressão logístico simples com $y_i \sim Bern(\pi_i)$ é dada por:

$$f(y_i, \pi_i) = \pi_i^{y_i}(1-\pi_i)^{1-y_i}\,, \qquad (4)$$

como as observações são independentes, a Função de Verossimilhança é obtida pelo produto dos termos dados na equação (4), sendo:

$$L(\beta) = \prod_{i=1}^{n} \pi_i^{y_i}(1-\pi_i)^{1-y_i}\ , \beta \in R \qquad (5)$$

aplicando logaritmo natural (ln) em ambos os lados da equação, tem-se:

$$l(\beta) = \ln[L(\beta)] = ln\left[\prod_{i=1}^{n} \pi_i^{y_i}(1-\pi_i)^{1-y_i}\right] = \sum_{i=1}^{n}\left[y_i \ln\left(\frac{\pi_i}{1-\pi_i}\right) + \ln(1-\pi_i)\right]. \qquad (6)$$

Substituindo pelas equações (2) e (3), temos:

$$l(\beta) = \sum_{i=1}^{n} \left[y_i(\beta_0 + \beta_1 x_i) + \ln\left(\frac{1}{1 + \exp(\beta_0 + \beta_1 x_i)}\right)\right] \quad (7)$$

$$= \sum_{i=1}^{n} [y_i(\beta_0 + \beta_1 x_i) - \ln(1 + \exp(\beta_0 + \beta_1 x_i))]$$

Para determinar os valores de β que maximizam a função (7), bastará apenas aplicar a derivada parcial em função de cada parâmetro e os valores a serem encontrados para β são chamados Estimadores de Máxima Verossimilhança (EMV) e indicam a importância de cada variável independente para a ocorrência do evento de interesse (SICSÚ, 2010). Assim, aplicando as derivadas parciais temos:

$$\frac{\partial l(\beta)}{\partial \beta_0} = \sum_{i=1}^{n} \left[y_i - \frac{1}{1 + \exp(\beta_0 + \beta_1 x_i)} \exp(\beta_0 + \beta_1 x_i)\right] \quad e$$

$$\frac{\partial l(\beta)}{\partial \beta_1} = \sum_{i=1}^{n} \left[x_i y_i - \frac{1}{1 + \exp(\beta_0 + \beta_1 x_i)} \exp(\beta_0 + \beta_1 x_i) x_i\right],$$

igualando a zero, tem-se o seguinte sistema de equação:

$$\sum_{i=1}^{n} (y_i - \pi_i) = 0 \quad (8)$$

$$\sum_{i=1}^{n} x_i (y_i - \pi_i) = 0 \quad (9)$$

Uma vez que as equações (8) e (9) são não lineares nos parâmetros (β_0 e β_1), serão necessários métodos iterativos para sua resolução, pois, de acordo com Walker e Duncan (1967) citado por Martins (1998), a exata determinação dos parâmetros não é possível, em função da complexidade do problema resultante. Dentre os métodos iterativos, o de Newton-Raphson é o mais empregado, pois, apresenta a vantagem de convergir rapidamente para a solução.

Nesse método, o primeiro passo consiste no uso de uma solução inicial para os valores que maximizam a função de verossimilhança. A função é aproximada em uma vizinhança da solução inicial por um polinômio do segundo grau. A segunda solução, é o ponto de máximo do valor do polinômio, e assim por diante. Dessa forma, para Figueira (2006), o método gera uma sequência de soluções que convergem para o ponto de máximo da função de verossimilhança. Para maiores detalhes do método ver seção 3 do Capitulo XII de Casella e Berger (2010).

3.3.2. INTERPRETAÇÃO DOS COEFICIENTES

Uma vez ajustado o modelo de regressão logística, em seguida deve-se avaliar a significância dos coeficientes estimados, e interpretar os seus valores. Para que possamos interpretar os valores associados aos coeficientes do modelo de regressão logística, é conveniente proceder à análise de acordo com a natureza das variáveis independentes, as quais podem ser: dicotômicas, politômicas (nominais ou ordinais com mais de duas categorias) ou contínuas.

Considerando a variável independente (x) dicotômica, codificada em 0 e 1, a chance da resposta para $x = 1$ é definida como $\pi(1)/1 - \pi(1)$ e para $x = 0$ como $\pi(0)/1 - \pi(0)$. Sendo o algoritmo da razão de chance dado por:

$$g(1) = \ln\left[\frac{\pi(1)}{1-\pi(1)}\right] \qquad e \qquad g(0) = \ln\left[\frac{\pi(0)}{1-\pi(0)}\right]$$

Uma vez que $\pi(x)$ assume os valores $\pi(0)$ e $\pi(1)$, então a variável resposta (y_i) tem a distribuição de probabilidade apresentada na Tabela 1 abaixo.

Tabela 1 – Valores do modelo de regressão logística para variável independente binária ou dicotômica

	$x = 0$	$x = 1$
$y = 0$	$1 - \pi(0)$	$1 - \pi(1)$
$y = 1$	$\pi(0)$	$\pi(1)$

Fonte: Adaptado de Hosmer e Lemeshow, 2000.

Sendo:

$$\pi(0) = \frac{\exp(\beta_0)}{1+\exp(\beta_0)} \qquad e \qquad \pi(1) = \frac{\exp(\beta_0+\beta_1)}{1+\exp(\beta_0+\beta_1)}$$

$$1-\pi(0) = \frac{1}{1+\exp(\beta_0)} \qquad e \qquad 1-\pi(1) = \frac{1}{1+\exp(\beta_0+\beta_1)}$$

Uma medida muito utilizada em regressão logística é a Razão de chance (*Odds ratio*), que denotaremos por OR, e que é estimada da seguinte forma:

$$\text{OR} = \frac{\pi(1)/1-\pi(1)}{\pi(0)/1-\pi(0)}, \qquad (10)$$

aplicando logaritmo natural (ln) na equação (10) obtemos:

$$\ln \text{OR} = \ln\left[\frac{\pi(1)/1-\pi(1)}{\pi(0)/1-\pi(0)}\right] = g(1) - g(0)\,,$$

a esta diferença chamamos de **diferença logit**.

A partir da expressão (10), substituindo os valores de $\pi(0)$ e $\pi(1)$, de $1-\pi(0)$ e de $1-\pi(1)$ pelas expressões anteriormente apresentadas, temos:

$$\text{OR} = \frac{\left[\frac{\exp(\beta_0+\beta_1)}{1+\exp(\beta_0+\beta_1)}\right]\left[\frac{1}{1+\exp(\beta_0)}\right]}{\left[\frac{\exp(\beta_0)}{1+\exp(\beta_0)}\right]\left[\frac{1}{1+\exp(\beta_0+\beta_1)}\right]} = \frac{\exp(\beta_0+\beta_1)}{\exp(\beta_0)} = \exp(\beta_1),$$

assim, a **diferença logit** será:

$$\ln \text{OR} = \ln[\exp(\beta_1)] = \beta_1 \qquad (11)$$

De acordo com Ayres et al. (2005), a razão de chance é uma medida de fácil interpretação e com propriedades estatísticas fundamentais em diversos estudos.

3.4. MODELO DE REGRESSÃO LOGÍSTICA MÚLTIPLA

Assim como no modelo de regressão linear, podemos ajustar um modelo para a variável resposta levando em conta mais de uma variável explicativa, o que chamamos de Modelo de Regressão Logística Múltipla.

Considerando um conjunto de p variáveis independentes denotadas como um vetor $x_i' = (x_{i0},\ x_{i1},\ x_{i2}, \dots,\ x_{ip})$, da i-ésima linha da matriz (X) das variáveis explicativas, em que cada elemento da matriz corresponde ao ij-ésimo componente (x_{ij}), com $i = 1, 2, \dots, n$ e $j = 0, 1, 2, \dots, p$, sendo $x_{i0} = 1$. Denota-se por $\beta' = (\beta_0,\ \beta_1,\ \beta_2, \dots,\ \beta_p)$, o vetor de parâmetros desconhecidos e β_j é o j-ésimo parâmetro associado a variável explicativa x_j.

No modelo de regressão logística múltipla, a probabilidade de sucesso é dada por:

$$\pi_i = \pi(x_i) = P(Y=1|X=x_i) = \frac{\exp\left(\beta_0+\beta_1 x_{i1}+\cdots+\beta_p x_{ip}\right)}{1+\exp\left(\beta_0+\beta_1 x_{i1}+\cdots+\beta_p x_{ip}\right)}$$

$$\pi_i = \pi(x_i) = P(Y=1|X=x_i) = \frac{\exp(x_i'\beta)}{1+\exp(x_i'\beta)} \qquad (12)$$

e a probabilidade de fracasso dada por:

$$1-\pi_i = 1-\pi(x_i) = P(Y=0|X=x_i) = \frac{1}{1+\exp\left(\beta_0+\beta_1 x_{i1}+\cdots+\beta_p x_{ip}\right)}$$

$$1 - \pi_i = 1 - \pi(x_i) = P(Y = 0|X = x_i) = \frac{1}{1 + \exp(x_i'\beta)} \quad (13)$$

Assumindo que a variável resposta (Y_i) segue uma distribuição de Bernoulli com parâmetro π_i, a função *logit* para o modelo de regressão logística múltipla será dada pela equação:

$$g(x_i) = \ln\left[\frac{\pi_i}{1 - \pi_i}\right] = x_i'\beta = \beta_0 + \sum_{j=1}^{p} \beta_j x_{ij} \quad (14)$$

3.4.1. ESTIMAÇÃO DOS PARÂMETROS DO MODELO

No modelo de regressão logística múltipla, o método para estimação dos parâmetros é o de máxima verossimilhança, similar ao da regressão logística simples, com a função de verossimilhança idêntica à expressão (5).

Assim, sendo o vetor β' de parâmetros relacionados com a probabilidade condicional $P(Y_i = 1|x_i) = \pi(x_i)$ para $i \in \{1, 2, \dots, n\}$, considerando a expressão (12) e a partir da expressão (5), podemos obter o estimador de β, denotado por $\hat{\beta}$, tendo como soluções as seguintes equações:

$$\sum_{i=1}^{n} (y_i - \pi_i) = 0 \qquad e \qquad \sum_{i=1}^{n} x_i(y_i - \pi_i) = 0 \quad (15)$$

No entanto, tem-se $p + 1$ equações de verossimilhança que são obtidas determinando as derivadas parciais em função do parâmetro β do logaritmo da função de verossimilhança, dada por:

$$L(\beta) = \sum_{i=1}^{n} [y_i \ln \pi_i + (1 - y_i) \ln(1 - \pi_i)], \quad (16)$$

com respeito a cada um dos $p + 1$ coeficientes. A expressão (16), é obtida a partir do logaritmo natural (ln) na expressão (5), fazendo o uso das propriedades de somatório e de logaritmo. Aplicando derivadas em (15), temos:

$$\frac{\partial L}{\partial \beta_0} = \sum_{i=1}^{n} y_i - \sum_{i=1}^{n} \hat{\pi}_i = 0 \quad e$$

$$\frac{\partial L}{\partial \beta_j} = \sum_{i=1}^{n} x_{ij} y_i - \sum_{i=1}^{n} x_{ij} \hat{\pi}_i = 0 \, , para \, j \in \{1, 2, \dots, p\} \, , \quad (17)$$

sendo $\hat{\pi}_i$ o estimador de máxima verossimilhança de π_i. Assim, representando em notação matricial todas as $p + 1$ equações de verossimilhança temos:

$$\frac{\partial L(\beta)}{\partial \beta} X'(Y - \Pi) = 0\,, \tag{18}$$

em que,

$$Y = (y_1, \dots, y_n)'_{1\times n}$$

$$\Pi = (\pi_1, \dots, \pi_n)'_{1\times n}$$

$$\beta = \left(\beta_0, \dots, \beta_p\right)'_{1\times(p+1)}$$

$$X = \begin{bmatrix} 1 & x_{11} & \dots & x_{1p} \\ 1 & x_{21} & \dots & x_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ 1 & x_{n1} & \dots & x_{np} \end{bmatrix}_{n\times(p+1)}$$

$$\Sigma = \begin{bmatrix} \pi_1(1-\pi_1) & 0 & \dots & 0 \\ 0 & \pi_2(1-\pi_2) & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & \pi_n(1-\pi_n) \end{bmatrix}_{n\times n}$$

Sendo Y e Π matrizes $n \times 1$, β é um vetor $(p+1) \times 1$, X é uma matriz $n \times (p+1)$ e Σ é uma matriz de variâncias e covariâncias $n \times n$.

Desta forma, como no modelo de regressão logística simples, as equações obtidas com a derivação da função de verossimilhança não são lineares, fazendo-se necessário utilizar os métodos iterativos, como o de Newton - Raphson para a resolução do sistema de equações resultantes.

3.5. TESTE DE SIGNIFICÂNCIA DOS PARÂMETROS

Após o ajuste dos modelos (estimação dos parâmetros), deve-se testar a significância das variáveis no modelo. Nesse processo está envolvido o teste de hipóteses estatísticas, o qual determina se as variáveis independentes (x_i) no modelo estão "significativamente" relacionadas com a variável resposta (Y_i).

Os testes mais utilizados para testar a qualidade do modelo ajustado e a significância de cada parâmetro ou de um conjunto de parâmetros do modelo são: teste de escore, teste da razão de verossimilhança e teste de Wald. Sendo que esses dois últimos testes os que foram abordados neste trabalho.

3.5.1. TESTE DA RAZÃO DE VEROSSIMILHANÇAS (TRV)

Uma vez ajustado o modelo, foi necessário testar a significância do modelo estimado. Uma medida para testar essa significância é o teste da razão de verossimilhança (TRV). Assim, foi

possível testar simultaneamente se os coeficientes de regressão associados ao vetor β são todos nulos com exceção de β_0.

Para a comparação entre os valores preditos e observados usando a função de verossimilhança, usa-se "menos duas vezes o logaritmo do quociente desses máximos" sendo que a sua distribuição equivale ao *Qui-Quadrado* com $n - p$ graus de liberdade, e é baseada na seguinte expressão:

$$D = -2\ln\left(\frac{\text{verossimilhança do modelo corrente}}{\text{verossimilhança do modelo saturado}}\right),$$

assim, o modelo corrente corresponde ao modelo com apenas as variáveis desejadas para o estudo e o modelo saturado corresponde ao modelo com todas as variáveis.

Segundo Nelder e Wedderburn (1972), essa estatística D é chamada de desvio *(deviance*) e avalia o valor ajustado na regressão logística, desempenhando o mesmo papel que tem a soma de quadrados dos resíduos na regressão linear. Quanto menor a *deviance* melhor é o ajuste do modelo.

Considerando o modelo com as proporções estimadas $\hat{\pi}_i$, a *deviance* pode ser escrita como:

$$D = -2\sum_{i=1}^{n}\left[y_i \ln\left(\frac{\hat{\pi}_i}{y_i}\right) + (1 - y_i)\ln\left(\frac{1-\hat{\pi}_i}{1-y_i}\right)\right] \quad (19)$$

Para estimar a significância de uma variável independente (x_i), faz-se necessário comparar o valor de D com e sem variável independente na equação. Sendo que a alteração no valor de D esperada pela inclusão da variável independente no modelo é obtida através de:

$$G = \mathrm{D}(\text{para o modelo sem a variável}) - \mathrm{D}(\text{para o modelo com a variável}),$$

esta estatística, também é comumente expressa:

$$\mathrm{G} = -2\ln\left(\frac{\text{verossimilhança do modelo sem a variável}}{\text{verossimilhança do modelo com a variável}}\right)$$

Para o caso de uma variável independente, verifica-se facilmente que se esta variável não está no modelo, o estimador de máxima verossimilhança de β_0 será $\hat{\beta}_0 = \ln\left(\frac{n_1}{n_0}\right)$, para $n_1 = \sum_{i=1}^{n} y_i$ e $n_0 = \sum_{i=1}^{n}(1 - y_i)$ sendo o valor predito dada pela constante $\frac{n_1}{n}$.

Considerando a informação acima, o valor de G pode ser escrito como:

$$\mathrm{G} = -2\ln\left[\frac{\left(\frac{n_1}{n}\right)^{n_1}\left(\frac{n_0}{n}\right)^{n_0}}{\prod\hat{\pi}_i{}^{y_i}(1-\hat{\pi}_i)^{(1-y_i)}}\right] \tag{20}$$

As hipóteses a serem testadas para o caso em que pretendemos analisar se pelo menos uma das variáveis explicativas é significativa no modelo em estudo, são:

$$\mathrm{H_0}: \beta_1 = \beta_2 = \cdots = \beta_\mathrm{p} = 0 \qquad vs \qquad \mathrm{H_1}: \text{pelo menos um } \beta_\mathrm{j} \neq 0$$

Ao rejeitarmos a hipótese nula podemos concluir que, pelo menos um dos coeficientes é estatisticamente diferente de zero. Assim, a estatística G terá uma distribuição de qui-quadrado (χ^2) com grau de liberdade igual à diferença do número de parâmetros dos modelos que estão sendo comparados. Essa estatística G com o valor de χ^2, poderá concluir se as variáveis podem, ou não, ser retiradas do modelo, a um nível de significância pré-estabelecido.

3.5.2. TESTE DE WALD

O teste de Wald testa se cada coeficiente é significativamente diferente de zero. Deste modo, o teste de Wald averigua se uma determinada variável independente apresenta uma relação estatisticamente significativa com a variável dependente. Pois, de acordo com Wald (1943) citado por Colosimo e Giolo (2006), esta estatística é baseada na distribuição assintótica de $\hat{\beta}$ e é uma generalização do teste t de Student, sendo geralmente mais usada no caso de hipóteses relativas a um único parâmetro β_j.

As hipóteses do teste são:

$$H_0: \boldsymbol{\beta} = 0 \qquad vs \qquad H_1: \boldsymbol{\beta} \neq 0\,,$$

sendo a estatística do teste dada pela expressão:

$$W = \widehat{\boldsymbol{\beta}}' I\left(\widehat{\boldsymbol{\beta}}\right)\widehat{\boldsymbol{\beta}}\,, \tag{21}$$

em que $I\left(\widehat{\boldsymbol{\beta}}\right)$ é a matriz de Informação de Fisher avaliada em $\hat{\beta}$, em que sob H_0 a estatística W apresenta uma distribuição qui-quadrado com número de graus de liberdade igual ao número de parâmetros.

No caso em que β é um escalar, a expressão (21) se reduz a:

$$W = \frac{\hat{\beta}}{\widehat{SE}(\hat{\beta})}$$

O teste de Wald, em alguns casos, costuma não rejeitar a hipótese de nulidade quando esta deveria ser rejeitada. Hauck e Donner (1977) e Jennings (1986) examinando a eficiência do

teste de Wald, recomendam que o TRV seja usado quando há dúvidas de que o teste de Wald tenha falhado.

3.6. CURVA ROC (*Receiver Operating Characteristic*)

3.6.1. PERSPECTIVA HISTÓRICA

A curva ROC foi criada na área da psicologia sensorial, tendo como objetivo comprovar a existência de uma relação empírica entre o corpo e a mente, segundo Gustav Theodor Fechner (1801-1887), filósofo alemão e médico de formação, considerado o pioneiro da psicometria. Fechner, sujeitava as pessoas a determinado estímulo até obter um valor consideravelmente estável de respostas positivas. Reproduziu graficamente a relação entre as respostas positivas e a medida física da intensidade do estímulo adquirindo, assim, uma função psicométrica, (BRAGA, 2000).

Louis Leon Thurstone (1887-1955), pioneiro em psicometria e psicofísica, com base nos resultados de Fechner, desenvolveu novas técnicas para quantificar as qualidades mentais. Publicou diversas escalas de atitude onde procurou medir a influência de preconceitos de propaganda do homem. Tinha especial interesse pela medição da aprendizagem e tentou exprimir, através de unidades absolutas, o desenvolvimento da aprendizagem. Thurstone é o criador dos conceitos ruído, critério de decisão e ponto de corte (BRAGA, 2000).

Durante a Segunda Guerra Mundial (1939-1945), a curva ROC foi usada para quantificar a capacidade dos operadores de radares distinguirem um sinal de ruído, (METZ, 2008). Ou seja, quando um radar detectava algum movimento cabia ao operador determinar a veracidade e relevância do que havia sido detectado (se um míssil, avião inimigo ou, simplesmente, um bando de pássaros).

De acordo com Martinez et al. (2003), na década de 60, a curva ROC foi usada essencialmente em psicologia experimental e na década de 70, expandiu pelos campos da investigação biomédica, cujo objetivo se tornou, basicamente, auxiliar a classificação de indivíduos em doentes ou saudáveis.

Assim, o nome ROC surge, em teoria, com o objetivo de detectar a presença, ou não, de um sinal particular.

Hoje é um dos testes mais importantes na biotecnologia médica, e pouco utilizado em ciências agrárias.

3.6.2. MATRIZ DE CONFUSÃO E PONTO DE CORTE

A matriz de confusão é uma tabela 2×2, entre a classificação predita através de um único ponto de corte e a condição real e conhecida de cada indivíduo, onde os valores da diagonal principal compreendem as classificações corretas e os valores fora dessa diagonal correspondem aos erros de classificação, Brocco (2006). A Tabela 2 ilustra uma matriz de confusão para um exemplo de classificação.

Tabela 2 – Matriz de confusão para valores observados e preditos

		Valores Preditos		
		0	1	
Valores Observados	0	a	b	$a+b$
	1	c	d	$c+d$
Total		$a+c$	$b+d$	$a+b+c+d$

Fonte: Autor.

Em que:

- a é a quantidade de plantas classificadas como suscetíveis dado que elas são suscetíveis;
- b é a quantidade de plantas classificadas como resistentes dado que elas são suscetíveis;
- c é a quantidade de plantas classificadas como suscetíveis dado que elas são resistentes;
- d é a quantidade de plantas classificadas como resistentes dado que elas são resistentes;
- $a+b$ é o número de plantas suscetíveis na amostra;
- $c+d$ é o número de plantas resistentes na amostra;
- $a+c$ é o número de plantas classificadas como suscetíveis na amostra;
- $b+d$ é o número de plantas classificadas como resistentes na amostra;
- $a+b+c+d$ é o número total de observações na amostra.

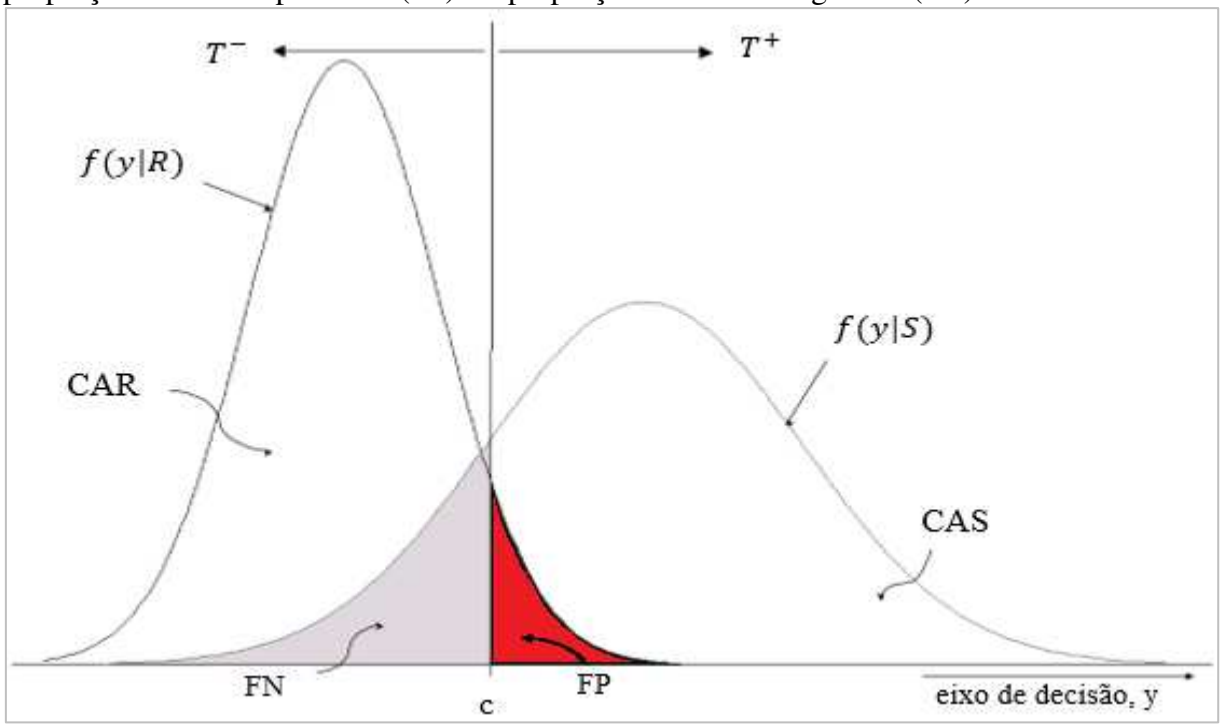
O ponto de corte corresponde a um ponto de separação na identificação dos indivíduos como zero (suscetíveis) ou um (resistentes). Esse ponto de separação representa um valor definido para essa medida, estabelecendo assim quais os indivíduos que estão acima ou abaixo desse ponto (MORANA, 2003).

Considerando a variável em estudo por y e que valores baixos de y favorecem a decisão "resistente" (T^-) e valores elevados de y favorecem a decisão "suscetível" (T^+). Assim, designando $f(y|S)$ por distribuição dos valores de y para os casos suscetíveis, y_S, e $f(y|R)$ por distribuição dos valores de y para os casos resistentes, y_R, ou seja, a distribuição de y_S está centrada à direita da de y_R.

Graficamente, a situação descrita é ilustrada pela Figura 2 abaixo. Assim, pode-se verificar que as distribuições de y_S e y_R sobrepõem-se, isto significa que, alguns dos casos inicialmente identificados como resistentes poderão ter leituras como suscetíveis, e por outro lado, alguns casos inicialmente identificados como suscetíveis poderão ter leituras como resistentes.

Para qualquer teste diagnostico é fixado um valor de corte para a variável em estudo, valor que determina a classificação dos indivíduos como suscetíveis ou resistentes. Esse teste é avaliado pela comparação relativa da capacidade preditiva de plantas suscetíveis (CAS) ou especificidade, capacidade preditiva de plantas resistentes (CAR) ou sensibilidade, proporção dos falsos positivos (FP) e dos falsos negativos (FN).

Figura 2 – Sobreposição de duas distribuições hipotéticas, para classificação dos indivíduos "resistente" (T^-) ou "suscetível" (T^+), a capacidade preditiva de plantas suscetíveis (CAS) ou especificidade, a capacidade preditiva de plantas resistentes (CAR) ou sensibilidade, a proporção dos falsos positivos (FP) e a proporção dos falsos negativos (FN)



Fonte: Adaptado de Braga, 2000.

Em geral, um teste de diagnóstico tende a ser avaliado pelas medidas da sensibilidade e da especificidade, assim, define-se sensibilidade como a capacidade do modelo em identificar uma classificação positiva ($y = 1$), dado que ela realmente é positiva. A especificidade é definida como a capacidade do modelo em identificar um resultado negativo ($y = 0$), dado que ele é realmente negativo Martinez, et al. (2003). Assim, de acordo com a Tabela 1, essas medidas são dadas por:

$$CAS = \frac{a}{a+b} \quad e \quad CAR = \frac{d}{c+d}$$

Essas duas definições (CAS e CAR) conduzem-nos a outras três diretamente relacionadas, a proporção de falsos positivos (FP) e falsos negativos (FN), e ainda a capacidade total de predições corretas ou acurácia (AC) que é também uma das medidas comumente utilizadas em problemas de classificação binaria ou dicotômica, pois esta permite determinar a proporção de predições corretas, sem levar em consideração o que é positivo ou o que é negativo. Essas três definições podem ser obtidas por:

$$FP = \frac{b}{a+b}\ ; \qquad FN = \frac{c}{c+d} \qquad e \qquad AC = \frac{a+d}{a+b+c+d}$$

Desse modo, valores de corte elevados, conduzem a um teste muito sensível e pouco especifico, por outro lado, valores de corte baixos, conduzem a um teste pouco sensível e muito especifico. E ainda, pode-se explicar a relação entre o valor de ponto de corte e as proporções dos FP e FN, sendo que diminuir a FP conduz a um aumento de FN (Figura 2).

Assumindo que todos os casos podem ser diagnosticados como resistentes ou suscetíveis (no que diz respeito a uma determinada doença), então, o número de decisões corretas mais o número de decisões incorretas deverá ser igual ao número de casos com esse estado atual. Assim, verifica-se que:

$$CAS + FN = 1 \qquad e \qquad CAR + FP = 1$$

Em termos práticos, é desejável ter um teste que seja ao mesmo tempo altamente sensível e especifico, pois um valor de corte fixa um par de sensibilidade/especificidade. De uma maneira geral, quando menor for o ponto de corte, maior será a habilidade do teste em classificar os doentes como positivos, ou seja, maior será a sensibilidade. Por outro lado, poderá ser inevitável que alguns indivíduos sadios sejam classificados como positivos, o que significa uma menor especificidade (MARTINEZ, at al. 2003).

Para Martinez, at al. (2003) a escolha do melhor ponto de corte muitas vezes é representada pelo ponto onde a sensibilidade e a especificidade são simultaneamente maiores, o que, nem sempre é adequado. Em algumas situações, quando não se pode correr o risco de não diagnosticar, é melhor privilegiar sensibilidade.

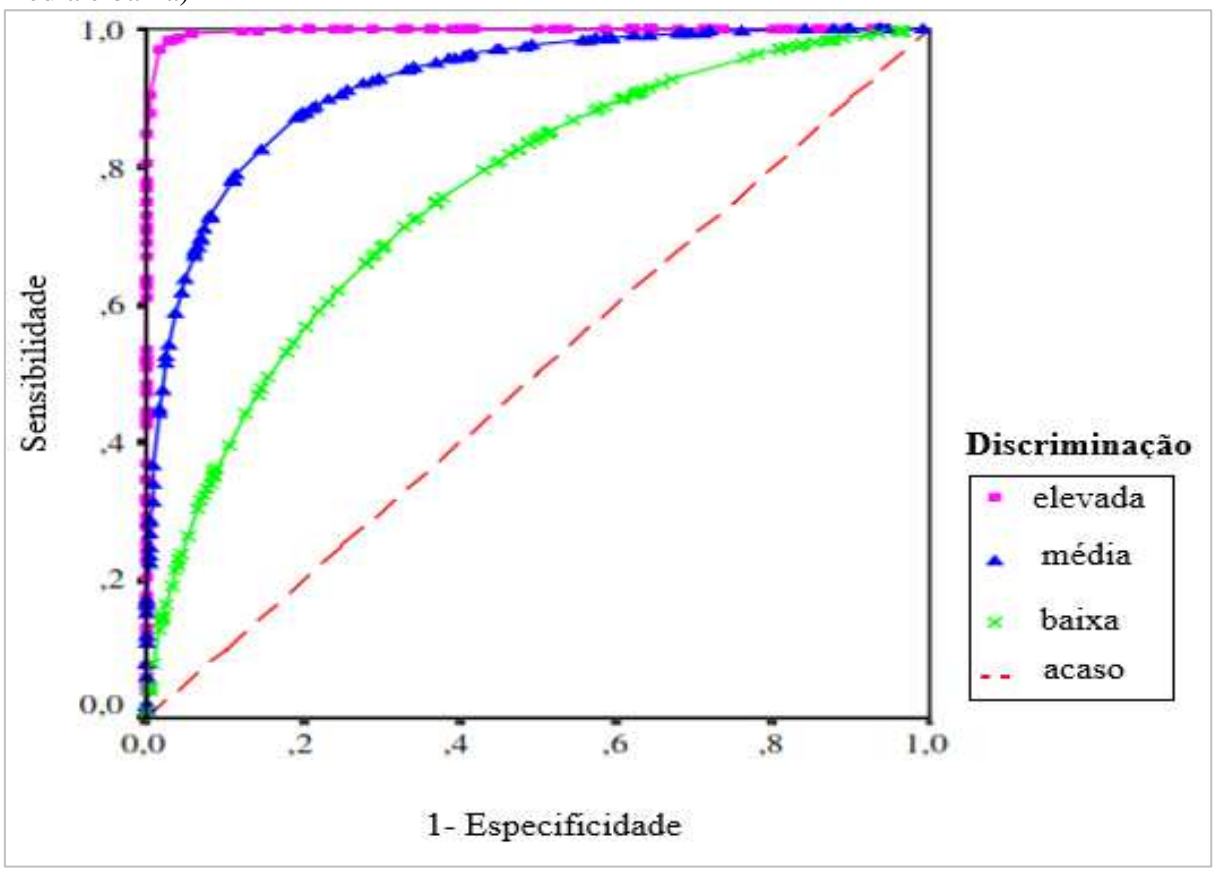
.

3.7. ÁREA SOB A CURVA ROC (AUC)

A Curva ROC constitui uma técnica útil para avaliar a capacidade preditiva dos modelos e está baseada nos conceitos da sensibilidade e da especificidade, obtidas a partir da construção da matriz de confusão 2×2, gerada pelo modelo. Dessa forma, de acordo com Martinez, at al. (2003), é necessario empregar uma regra de decisão baseada em buscar um ponto de corte que resume a quantidade de resposta dicôtomica, de forma que um individuo com mensurações menores ou iguais ao ponto de corte seja classificado como não doente, e, analogamente, um individuo com uma resposta ao teste maior que o ponto de corte seja classificado como doente (ou vice-versa). Assim, no contexto de classificação da resistência do arroz, se a probabilidade estimada para uma determinada planta for menor ou igual ao ponto de corte, esta é classificada como suscetível. Caso contrário, é classificada como resistente.

A área sob a Curva ROC mede a capacidade de discriminação do modelo, e quanto maior a área sob a curva, maior é a capacidade de acerto nas classificações dos indivíduos de acordo com um dos dois grupos ($\hat{y} = 0$ ou $\hat{y} = 1$). Na Figura 3, são apresentados três graus de discriminação possíveis fornecidos pelas curvas ROC.

Figura 3 – Curvas ROC representativas de três graus de capacidade de discriminação (elevada, média e baixa)



Fonte: Adaptado de Braga, 2000.

Observando a Figura 3, verifica-se uma linha diagonal indicando uma classificação aleatória, ou seja, um modelo que seleciona aleatoriamente saídas como zero ou um. Uma curva perfeita corresponderia a uma linha horizontal no topo do gráfico, com elevada capacidade de discriminação, porem dificilmente se verifica. As curvas que se apresentam entre a linha diagonal e a linha perfeita são consideradas de média e baixa capacidade de discriminação, assim quanto mais a curva estiver distante da diagonal principal, melhor o desempenho de modelo associado a ela. Deste modo, sugere-se que quanto maior for a área entre a curva ROC produzida e a diagonal principal, melhor será o desempenho global do modelo.

De acordo com Hosmer e Lemeshow (1989) a regra geral para avaliação do resultado da área sob a Curva ROC é:

- Se $\mathrm{AUC} = 0{,}5$: não há discriminação;
- Se $0{,}7 \leq \mathrm{AUC} \leq 0{,}8$: discriminação aceitável;
- Se $0{,}8 \leq \mathrm{AUC} \leq 0{,}9$: excelente discriminação;
- Se $\mathrm{AUC} \geq 0{,}9$: discriminação pendente.

Modelos de alto poder discriminatório concentram-se no canto superior esquerdo da Curva ROC, pois à medida que a sensibilidade aumenta há pouca perda de especificidade (Figura 3). Já os modelos com menor poder discriminatório aproximam-se da diagonal. Esta diagonal revela a relação entre as taxas de resultados da sensibilidade (verdadeiro-positivos) com o complementar da especificidade (falsos-positivos) OLIVEIRA (2011).

3.8.TESTE DE HOSMER E LEMESHOW

Após a seleção das variáveis que compõem o modelo, verificou-se a qualidade de ajuste do modelo, isto é, se o modelo foi eficiente para descrever a relação entre as variáveis preditoras e a variável resposta.

Segundo Hosmer e Lemeshow (1989), citado por Oliveira (2011), o teste de Hosmer e Lemeshow corresponde a um teste de Qui-quadrado com $g - 2$ graus de liberdade e consiste em dividir o número de observações em dez grupos ($g = 10$) e, em seguida comparar as frequências preditas com as observadas. A finalidade desse teste é verificar se existem diferenças significativas entre as classificações realizadas pelo modelo e a realidade observada.

As hipóteses a serem testadas foram:

$$\mathrm{H}_0: \mathrm{O}_\mathrm{j} = \mathrm{E}_\mathrm{j} \quad e \quad \mathrm{H}_1: \text{Pelo menos uma } \mathrm{O}_\mathrm{j} \text{ é diferente de } \mathrm{E}_\mathrm{j}\ ; \ \forall_\mathrm{j} = 1, \cdots, \mathrm{g}$$

A estatística do teste sob a hipótese de nulidade é dada por:

$$\chi^2_{gl} = \sum_{j=1}^{g} \frac{\left(O_j - E_j\right)^2}{E_j\left(1 - \frac{E_j}{n_j}\right)} = \sum_{j=1}^{g} \frac{\left(O_j - \bar{p}_j\right)^2}{n_j \bar{p}_j\left(1 - \bar{p}_j\right)} \sim \chi^2_{g-2}$$

sendo,

n_j o número de observações pertencentes ao grupo j, verificando-se $n = \sum_{j=1}^{g} n_j$;

O_j a frequência observada de sucesso no grupo j, onde $o_j = \sum_{j=1}^{n_j} y_{ij}$ e y_{ij} é a i-èsima observação do grupo j;

E_j a frequência esperada de sucesso no grupo j, onde $E_j = n_j \bar{p}_j$ e $\bar{p}_j = \frac{\sum_{j=1}^{n_j} \hat{p}_{ji}}{n_j}$;

$\hat{p}_j$ a probabilidade predita correspondente à i-ésima observação do grupo j.

A um nível de significância pré-estabelecido, busca-se não rejeitar a hipótese de nulidade. Entretanto, se houver diferenças significativas entre as classificações preditas pelo modelo e as observadas, então o modelo não representa a realidade de forma satisfatória. Nessa situação, o modelo não será capaz de reproduzir estimativas e classificações confiáveis Hosmer e Lemeshow (1989). Sendo que, para avaliar a qualidade de ajuste do modelo foi utilizada a função "logitgof" implementado por meio do pacote *generalhoslem* de (Jay, 2019) no software R. E, o código usado encontra-se no apêndice.

4. MATERIAL E MÉTODOS

4.1. DESCRIÇÃO DOS DADOS

Uma população de 413 plantas de arroz (*Oryza sativa* L.), e com informações referentes a genotipagem e fenotipagem, coletadas em 82 países, foram utilizadas, o que caracterizou todas as principais regiões produtoras de arroz no mundo (RESENDE, et al., 2010). Foi realizada a classificação de 5 subpopulações (Indica, Aus, Japonica temperado, Japonica tropical e Aromático) bem diferenciadas que resumem a variação genética global das plantas. Tal caracterização foi feita por meio de análise de componentes principais (PCA) por Price, et al., (2006).

A avaliação fenotípica do arroz foi feita em Stuttgart (Arkansas, EUA) durante os meses de maio à outubro nos anos de 2006 e 2007. Duas repetições por ano foram cultivadas em um delineamento de blocos inteiramente casualizados em parcelas de 5 m com espaçamento de 25 cm entre as plantas e 0,50 m entre as fileiras. O conjunto de dados fenotípicos dispõe de um total de 16 características relacionadas à morfologia das plantas, qualidade dos grãos, desenvolvimento das plantas, qualidade nutricional e grau de suscetibilidade do genótipo à doença, para avaliar a resistência da brusone do arroz. A gravidade da doença foi inicialmente classificada em uma escala de "0" (sem lesões da doença) a "9" (morte total da planta) quando as plantas tinham entre três e quatro semanas de idade, como descrito por Marchetti, et al., (1987). Esta escala foi convertida para tipos de reação (resistente e suscetível) de acordo com o tamanho e as características das lesões conforme apresentado por Mackill e Bonman (1992). Dessa forma, plantas pertencentes às classes 0, 1 e 2, foram classificadas como resistentes ($y = 1$) e plantas pertencentes às classes 3, 4, 5, 6, 7, 8 e 9 foram classificadas como suscetíveis ($y = 0$). Foram assim, excluídas 137 plantas que não foram avaliadas quanto a resistência, resultando em um cenário com 276 plantas, das quais 71 eram resistentes e 205 eram suscetíveis. Mais detalhes com relação ao banco de dados são encontrados em Zhao et al., (2011).

As variáveis explicativas utilizadas neste trabalho estão apresentadas na Tabela 3.

4.2. SELEÇÃO DE VARIÁVEIS

Para análise da regressão logística, foi necessário selecionar dentro do conjunto de variáveis candidatas aquelas que realmente podem explicar a resistência das plantas.

Hosmer e Lemeshow (1989), afirmam que para obter um modelo com o menor número de variáveis possível, mas que mantenha a eficiência nos resultados de previsão é necessário que

se tenha um plano de escolha das variáveis iniciais que serão testadas no modelo e um método que auxilie na seleção e adequação dessas variáveis.

Alguns métodos comumente utilizados para seleção de variáveis são *forward*, *backward* ou *stepwise*, cujos algoritmos estão implementados em programas computacionais. No entanto, esses métodos apresentam algumas desvantagens, pois tendem a identificar um particular conjunto de variáveis, em vez de possíveis conjuntos igualmente bons para explicar a variável resposta, impossibilitando o pesquisador a escolher o mais relevante em sua área de aplicação (COLOSIMO e GIOLO, 2006).

Nesse trabalho foi utilizada uma estratégia de seleção de variáveis derivada da proposta de Collett (2003), em que as informações do pesquisador podem ser incluídas no processo de decisão, o que envolve uma participação mais ativa do estatístico e pesquisador em cada passo do processo de seleção, podendo, por exemplo, incluir variáveis importantes no estudo, independente da significância estatística.

Tabela 3 – Código, descrição e tipos de variáveis utilizadas para a predição da resistência à brusone (*Blast*) do arroz (*Oryza sativa* L.)

Código	Descrição	Tipo
Y	Gravidade da doença na folha do arroz causada pelo fungo *Pyricularia oryzae* (0: suscetibilidade, 1: resistência)	Binária
V1	Hábito do colmo: ângulo médio do colmo das plantas na maturidade	Contínua
V2	Pubescência da folha (0: Ausência, 1: Presença)	Binária
V3	Comprimento da folha da bandeira (cm)	Contínua
V4	Largura da folha bandeira (cm)	Contínua
V5	Número médio de panículas (inflorescências) por planta	Contínua
V6	Altura da planta (cm)	Contínua
V7	Comprimento da panícula (cm)	Contínua
V8	Número médio de ramificação da panícula primária	Contínua
V9	Número médio de sementes por panícula	Contínua
V10	Comprimento da semente com casca (mm)	Contínua
V11	Largura da semente com casca (mm)	Contínua
V12	Volume da semente com casca	Contínua
V13	Área da superfície da semente com casca	Contínua
V14	Relação comprimento / largura da semente	Contínua
V15	Quantidade de amilose presente nos grãos moídos (%)	Contínua

Fonte: Autor.

Os passos utilizados no processo de seleção são apresentados a seguir, segundo Colosimo e Giolo (2006):

Passo 1: Ajustar todos os modelos contendo uma única variável. Incluir todas as variáveis que foram significativas ao nível de 0,10. Nesse passo, é necessário utilizar o teste de Razão de Verossimilhança.

Passo 2: As variáveis significativas no passo 1 são, então, ajustadas conjuntamente. Na presença de certas variáveis, outras podem deixar de ser significativas. Consequentemente, ajusta-se modelos reduzidos, excluindo uma única variável de cada vez. Verifica-se as variáveis que provocam um aumento estatisticamente significativo na estatística de Razão de Verossimilhança. E somente aquelas que atingirem significância permanecem no modelo.

Passo 3: Ajusta-se um novo modelo com as variáveis retidas no passo 2. Neste passo, as variáveis retidas no passo 2 retornam ao modelo para confirmar que elas não são estatisticamente significativas.

Passo 4: As eventuais variáveis significativas no passo 3 são incluídas ao modelo juntamente com aquelas do passo 2. Neste passo, retorna-se com as variáveis excluídas no passo 1 para confirmar que elas não são estatisticamente significativas.

Passo 5: Ajusta-se um modelo incluindo as variáveis significativas no passo 4. Neste passo é testado se alguma delas pode ser retirada do modelo.

Passo 6: Utilizando as variáveis que restaram no passo 5, ajusta-se o modelo final para os efeitos principais. E para completar a modelagem, deve-se verificar a possibilidade de inclusão de termos de interação dupla entre as variáveis incluídas no modelo. O modelo final fica determinado entre os efeitos principais identificados no passo 5 e os termos de interação significativos identificados neste passo.

Para Colosimo e Giolo (2006), ao ser utilizado este procedimento de seleção deve-se evitar ser muito rigoroso ao testar cada nível individual de significância. E para decidir se um termo deve ser incluído, o nível de significância não deve ser muito baixo, sendo recomendado um valor próximo de 0,10.

4.3. CONSTRUÇÃO DO MODELO

Com as variáveis selecionadas no passo 6 do item 4.2, foram construídos modelos para estimar e avaliar a resistência das plantas de arroz à brusone. Para a modelagem da resistência à brusone foram ajustados os tipos de reação (respostas dicotômicas) utilizando o modelo *threshold*, descrito por Gianola (1982). Esse modelo visa estimar a probabilidade da planta

pertencer a uma das duas reações. Essa probabilidade foi estabelecida de acordo com a estimativa de uma variável latente dada pela equação:

$$\ell = 1\mu + Xg + e$$

em que ℓ é o vetor de variáveis latentes (ou *liabilities*) em escala gaussiana cujo valor está ligado a uma variável categórica por meio da função de ligação *probit*, g é o vetor de efeito aditivo dos marcadores com matriz de incidência X que relaciona efeitos de SNP's aos valores da resistência à brusone e e é o vetor de erro associado ao modelo.

A função de ligação *probit* modela o valor de probabilidade de uma planta de arroz pertencer a cada uma das categorias da resistência à brusone. Dessa forma, realizou-se a categorização considerando a classe de maior probabilidade. Ou seja, as plantas foram classificadas como resistentes para $P[Y = 1|X] = P[\ell \leq \tau] > 0{,}5$. Caso contrário, foram classificadas como suscetíveis.

4.4. CONSTRUÇÃO DA CURVA ROC

Após a construção dos modelos para estimar e avaliar a resistência das plantas, foram selecionados os modelos que melhor descreveram a relação entre as variáveis preditoras e a variável resposta, baseando-se no teste de Hosmer e Lemeshow, que corresponde a teste de Qui-quadrado com $g - 2$ graus de liberdade. Foi adotado um nível de significância de 0,05 e buscou-se não rejeitar a hipótese de nulidade.

No entanto, dos modelos que melhor descreveram a relação entre as variáveis preditoras e a variável resposta, foi selecionado um modelo que melhor se adequou na predição da resistência à brusone do arroz (*Oryza sativa* L.), sendo este o modelo mais parcimonioso. Pois, de acordo com Bozdangan (1987), se dois ou mais modelos são bem ajustados e apresentam uma boa capacidade preditiva, deve-se preferir o modelo mais parcimonioso, isto é, o modelo que envolva o mínimo de parâmetros possíveis a serem estimados e que explique bem o comportamento da variável resposta.

Assim, de maneira a avaliar a capacidade preditiva dos modelos selecionados, foi construída a curva ROC. Para tal, plotou-se os valores de (1 – especificidade) no eixo das abscissas – eixo x e de sensibilidade no eixo das ordenadas – eixo y, que foram obtidos a partir da matriz de confusão 2×2, geradas pelos modelos (ver item 3.6.2, Tabela 2).

Foi utilizada como limiar de referência 0,5. Sendo que, para observações com valores de probabilidade estimados maiores que 0,5 foram classificadas como resistentes. Caso contrário, foram classificadas como suscetíveis.

Para construção da curva ROC, primeiramente foram calculados os valores do critério de informação de Akaike (AIC), baseado no conceito de entropia da informação, esse critério oferece uma medida relativa da informação perdida quando um modelo é usado para descrever a realidade (AKAIKE, 1974), calculada à partir da expressão:

$$AIC = -2\log\ L(\hat{\beta}) + 2k\ ,$$

sendo, $L(\hat{\beta})$ o valor máximo da função de verossimilhança e k o número de parâmetros do modelo. O funcionamento do AIC se baseia na inevitável perda de informação genérica, devido ao uso de um dos possíveis modelos para representar o "verdadeiro" modelo. Então devemos selecionar o modelo que minimize a quantidade esperada de informação perdida.

Na prática, dado um conjunto de candidatos para modelar um determinado evento, o melhor modelo é o que possuir menor AIC, ainda que, somente recompensa a precisão do ajuste, mas também inclui uma penalidade que aumenta a cada parâmetro incluído no modelo. Essa penalidade visa desencorajar o excesso de ajuste (*overfitting*). Portanto, o AIC proporciona um meio eficaz para comparação de modelos (BURNHAM e ANDERSON, 2004).

E, a partir dos valores do AIC, foram então encontrados os valores da área abaixo da curva ROC – AUC (do inglês *Area Under Curve*), para melhor comparação dos modelos. Para a construção da curva ROC foi utilizado o pacote *ROCR* de Sing at al. (2005) auxiliado pelos pacotes *ggplot2* de Wickham (2016), *tidyverse* de Wickham at al. (2019) e *pROC* de Robin at al. (2011), todos desenvolvidos para o software R (R DEVELOPMENT CORE TEAM, 2019). O código usado encontra-se no apêndice.

5. RESULTADOS E DISCUSSÃO

Para a construção do modelo de regressão logística, foi utilizada a estratégia de seleção de variáveis derivada da proposta de Collett (2003). Os resultados dessa seleção estão apresentados na Tabela 4.

Primeiramente, no passo 1, foram ajustados todos os modelos contendo uma única variável. Pelo teste da razão de verossimilhança verificou-se que as variáveis V1, V4, V7, V8, V9, V10, V11, V14 e V15 foram significativas ao nível de 0,10, isto é, mostraram ter alguma influência em relação a característica avaliada neste estudo (resistência à brusone do arroz).

No passo 2, as variáveis significativas anteriormente (passo 1) foram ajustadas conjuntamente. Segundo Colosimo e Giolo (2006), na presença de certas variáveis, outras podem deixar de ser significativas. Assim, foram ajustados modelos reduzidos, excluindo uma única variável de cada vez (do modelo de referência, em negrito). Verificou-se que apenas as variáveis V4, V8, V10, V11, V14 e V15 provocaram um aumento significativo na estatística da razão de verossimilhança. Assim, somente essas variáveis permanecerão no modelo.

No passo 3, com as variáveis que ficaram retidas no passo 2, ajustou-se um novo modelo (modelo de referência, em negrito) e as variáveis que foram excluídas no passo 2 retornaram ao modelo (variáveis em negrito) para confirmar que não foram estatisticamente significativas.

No passo 4, uma vez que as variáveis incluídas no modelo (V1, V7, V9) uma única de cada vez no passo 3 não mostraram significância, foi mantido o modelo de referência do passo 3 e retornou-se com as variáveis excluídas no passo 1 (variáveis em negrito), uma única de cada vez para confirmar se elas não eram estatisticamente significativas.

Foi ajustado então (passo 5), um modelo incluindo as variáveis (V3, V6, V12, V13) significativas no passo 4 (modelo de referência, em negrito) e foi testado se alguma delas poderia ser retirada do modelo. Observou-se que as variáveis (V10, V14, V3, V6, V12, V13) não apresentaram significância estatística e foram retirados do modelo.

Finalmente, no passo 6, com as variáveis selecionadas ajustou-se então o modelo para as variáveis que sobreviveram no passo 5. E para completar a modelagem foi verificada a possibilidade de inclusão dos termos de interação dupla entre as variáveis já incluídas no modelo. Observou-se que as interações $V4 * V15$ e $V11 * V15$ foram significativas ao nível de 0,10. Assim, na etapa final, chegou-se a três modelos.

Tabela 4 – Seleção de variáveis usando modelo de regressão logística para predição da resistência à brusone do arroz (*Oryza sativa* L.)

Passos	Modelo	-2log L(β)	Estatística de teste TRV	Valor p
Passo 1	**Nulo**	**314,7266**	**-**	**-**
	V1	308,9986	5,7280	0,0167*
	V2	313,0265	1,7001	0,1923
	V3	313,7101	1,0166	0,3133
	V4	300,4386	14,288	0,0002*
	V5	314,2728	0,4539	0,5005
	V6	314,6075	0,1191	0,7300
	V7	310,6362	4,0904	0,0431*
	V8	298,4124	16,3143	0,0001*
	V9	310,4151	4,3116	0,0379*
	V10	308,8089	5,9177	0,0150*
	V11	307,0889	7,6377	0,0057*
	V12	313,2573	1,4694	0,2254
	V13	314,7111	0,0155	0,9009
	V14	306,4341	8,2926	0,0040*
	V15	306,3386	8,3880	0,0038*
Passo 2	**V1+V4+V7+V8+V9+V10+V11+V14+V15**	**265,9581**	-	-
(Sem V1)	V4+V7+V8+V9+V10+V11+V14+V15	266,6430	0,6849	0,4079
(Sem V4)	V1+V7+V8+V9+V10+V11+V14+V15	268,6786	2,7205	0,0991*
(Sem V7)	V1+V4+V8+V9+V10+V11+V14+V15	267,4885	1,5304	0,2161
(Sem V8)	V1+V4+V7+V9+V10+V11+V14+V15	275,4432	9,4851	0,0021*
(Sem V9)	V1+V4+V7+V8+V10+V11+V14+V15	267,5873	1,6292	0,2018
(Sem V10)	V1+V4+V7+V8+V9+V11+V14+V15	272,9615	7,0033	0,0081*
(Sem V11)	V1+V4+V7+V8+V9+V10+V14+V15	273,3302	7,3721	0,0066*
(Sem V14)	V1+V4+V7+V8+V9+V10+V11+V15	271,8139	5,8558	0,0155*
(Sem V15)	V1+V4+V7+V8+V9+V10+V11+V14	270,6782	4,7201	0,0298*
Passo 3	**V4+V8+V10+V11+V14+V15**	**271,1824**	**-**	**-**
	V4+V8+V10+V11+V14+V15**+V1**	269,2947	1,8877	0,1695
	V4+V8+V10+V11+V14+V15**+V7**	268,7549	2,4276	0,1192
	V4+V8+V10+V11+V14+V15**+V9**	268,6713	2,5111	0,1131
Passo 4	**V4+V8+V10+V11+V14+V15**	**271,1824**	**-**	**-**
	V4+V8+V10+V11+V14+V15**+V2**	271,1223	0,0601	0,8063
	V4+V8+V10+V11+V14+V15**+V3**	265,1750	6,0074	0,0142*
	V4+V8+V10+V11+V14+V15**+V5**	269,4139	1,7685	0,1836
	V4+V8+V10+V11+V14+V15**+V6**	266,4102	4,7722	0,0289*
	V4+V8+V10+V11+V14+V15**+V12**	265,8614	5,3210	0,0211*
	V4+V8+V10+V11+V14+V15**+V13**	264,1504	7,0320	0,0080*
Passo 5	**V4+V8+V10+V11+V14+V15+V3+V6+V12+V13**	**259,1169**	**-**	**-**
(Sem V4)	V8+V10+V11+V14+V15+V3+V6+V12+V13	262,6172	3,5004	0,0614*
(Sem V8)	V4+V10+V11+V14+V15+V3+V6+V12+V13	267,7432	8,6265	0,0033*

...continuação

(Sem V10)	V4+V8+V11+V14+V15+V3+V6+V12+V13	259,3160	0,1992	0,6554
(Sem V11)	V4+V8+V10+V14+V15+V3+V6+V12+V13	266,7422	7,6255	0,0058*
(Sem V14)	V4+V8+V10+V11+V15+V3+V6+V12+V13	260,9080	1,7912	0,1808
(Sem V15)	V4+V8+V10+V11+V14+V3+V6+V12+V13	265,0047	5,8879	0,0153*
(Sem V3)	V4+V8+V10+V11+V14+V15+V6+V12+V13	261,1983	2,0815	0,1491
(Sem V6)	V4+V8+V10+V11+V14+V15+V3+V12+V13	260,1793	1,0626	0,3026
(Sem V12)	V4+V8+V10+V11+V14+V15+V3+V6+V13	259,2079	0,0911	0,7627
(Sem V13)	V4+V8+V10+V11+V14+V15+V3+V6+V12	260,3089	1,1922	0,2749
Passo 6	**V4+V8+V11+V15**	**281,2497**	**-**	**-**
	V4+V8+V11+V15+**V4*V8**	281,2217	0,0280	0,8671
	V4+V8+V11+V15+**V4*V11**	278,7223	2,5274	0,1119
	V4+V8+V11+V15+**V4*V15**	278,4753	2,7744	0,0958*
	V4+V8+V11+V15+**V8*V11**	281,0476	0,2021	0,6531
	V4+V8+V11+V15+**V8*V15**	280,3919	0,8578	0,3544
	V4+V8+V11+V15+**V11*V15**	274,2641	6,9856	0,0082*
Etapa	V4+V8+V11+V15**+V4*V15+V11*V15**	268,4995		
Final	V4+V8+V11+V15**+V4*V15**	278,4753		
	V4+V8+V11+V15**+V11*V15**	274,2641		

Fonte: Autor
*Valor - p < 0,10.

Considerando o conjunto de variáveis (etapa final), os possíveis modelos para a estimativa da probabilidade de ocorrência da característica resistência à brusone do arroz (*Oryza sativa* L.) foram:

Modelo 1: Modelo de regressão logística múltipla com as variáveis V4+V8+V11+V15+V4*V15+V11*V15.

Modelo 2: Modelo de regressão logística múltipla com as variáveis V4+V8+V11+V15+V4*V15.

Modelo 3: Modelo de regressão logística múltipla com as variáveis V4+V8+V11+V15+V11*V15.

De forma a avaliar o ajuste dos modelos e decidir qual deles deve ser usado, foi utilizado o teste de Hosmer e Lemeshow para testar a qualidade do ajuste e também a área sob a Curva ROC (AUC) para avaliar a capacidade preditiva dos modelos. De acordo com Hosmer e Lemeshow (1989), se as diferenças entre as classificações observadas e as preditas pelo modelo forem significativas, o grau de acurácia do modelo não é bom.

Além de analisar a qualidade de ajuste dos modelos pelo teste de Hosmer e Lemeshow, foram utilizados os valores de AIC e AUC para a comparação dos modelos. Uma vez que, segundo Akaike (1974), o AIC oferece uma medida relativa da informação perdida quando o

modelo é usado para descrever a realidade, sendo que AUC mede a capacidade de discriminação do modelo. E, o melhor modelo será aquele com menor AIC e maior AUC. A Tabela 5 abaixo mostra o resultado do teste para os três modelos ao nível de 5% e as suas respetivas medidas de desempenho.

Tabela 5 – Teste de Hosmer e Lemeshow e medidas de desempenho dos modelos: Modelo 1 (com duas interações entre as variáveis V4*V15 e V11*V15), Modelo 2 (com interação entre as variáveis V4*V15) e Modelo 3 (com interação entre as variáveis V11*V15)

	Qui-Quad.	GL	Valor-p	AIC	AUC
Modelo 1	14,58	8	0,068	282,4995	0,7553
Modelo 2	14,97	8	0,060	290,4753	0,7259
Modelo 3	16,27	8	0,039	286,2641	0,7391

AIC - Critério de informação de Akaike; AUC - Área abaixo da curva

Observando a Tabela 5, ao nível de 5% podemos verificar que não houve diferenças significativas (valor-p > 0,05) entre os valores preditos e observados para os modelos 1 e 2, o que indica que esses modelos foram capazes de produzir classificações confiáveis. Ainda, de acordo com Hosmer e Lemeshow (1989), esses modelos têm uma discriminação aceitável ($0{,}7 \leq \text{AUC} \leq 0{,}8$). E, o melhor modelo, de acorde com AIC foi o modelo 1, segundo Burnham e Anderson (2004). Em relação ao modelo 3, verificou-se diferença significativa (valor-p < 0,05) entre os valores preditos e observados, indicando uma classificação não confiável e, portanto, este foi retirado das análises posteriores.

Uma vez que a probabilidade da variável dependente estimada assumi valores entre zero e um, foi utilizado um ponto de corte de 0,5, de forma que as amostras com resultados inferiores ou iguais a esse valor foram classificadas como suscetíveis e superiores a esse valor classificadas como resistentes. Na Tabela 6, são apresentadas as medidas das capacidades preditivas dos modelos.

Tabela 6 – Medidas de capacidade preditiva dos modelos ajustados, Modelo 1 (com duas interações entre as variáveis V4*V15 e V11*V15) e Modelo 2 (com interação entre as variáveis V4*V15)

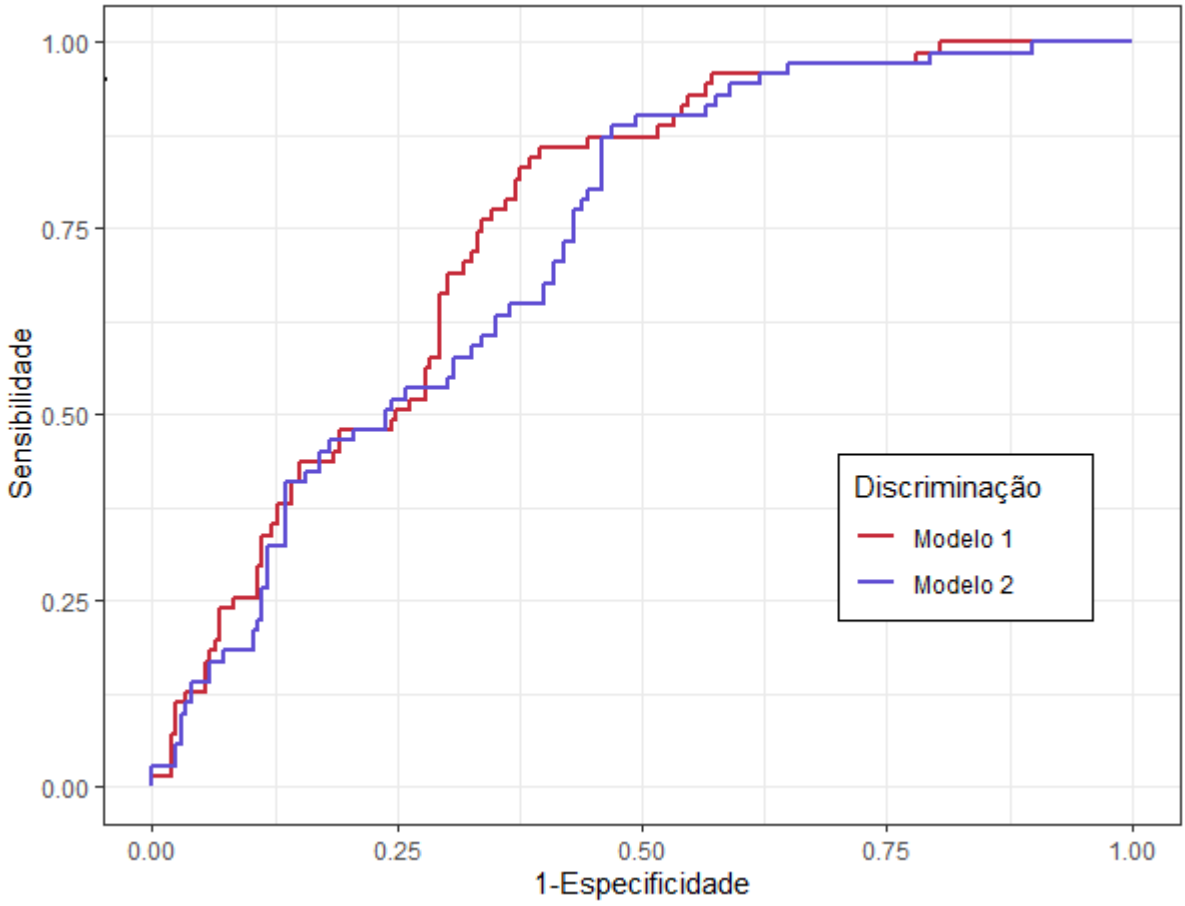
	% CAS	% CAR	% AC
Modelo 1	93,7	19,7	74,6
Modelo 2	96,1	14,1	75,0

%CAS: porcentual da capacidade preditiva de plantas suscetíveis; %CAR: porcentual da capacidade preditiva de plantas resistentes; %AC: porcentual da capacidade total de predições corretas ou acurácia.

De acordo com a análise das capacidades preditivas dos modelos, notou-se que o porcentual da sensibilidade (% CAS) do modelo 2 foi maior que do modelo 1 (96,1% e 93,7%, respetivamente). E o porcentual de especificidade (% CAR) do modelo 2 foi menor que a do modelo 1 (14,1% e 19,7%, respetivamente). E quanto ao porcentual da acurácia (% AC), esses modelos mostraram uma variação porcentual muito próximo (74,6% e 75,0%, respetivamente).

De forma a avaliar a capacidade preditiva dos modelos, foi construída a Curva ROC dos dois modelos de acordo com a Sensibilidade e 1-Especificidade. A Figura 4 abaixo mostra as Curvas ROC dos dois modelos ajustados.

Figura 4 – Comparação da área abaixo da curva dos modelos, Modelo 1 (com duas interações entre as variáveis V4*V15 e V11*V15) e Modelo 2 (com interação entre as variáveis V4*V15)



Observando a Figura 4, apesar das áreas serem próximas, verifica-se desempenhos diferentes dos modelos (1 e 2) em certos pontos de (1-Especificidade) e Sensibilidade, como pode-se ver próximo aos pontos (0,25; 0,50) e (0,50; 0,875) onde os modelos (1 e 2) mostram maior diferença. Entretanto, o modelo selecionado para a predição da resistência à brusone do arroz (*Oryza sativa* L.) foi o modelo 2, uma vez que este foi bem ajustado (Tabela 5) de acordo

com teste de Hosmer e Lemeshow e por ser o modelo mais parcimonioso, de acordo com Bozdangan, (1987).

O modelo selecionado (Modelo 2), foi composto pelas variáveis: largura da folha bandeira (V4), número médio da ramificação da panícula primária (V8), largura da semente com casca (V11), quantidade de amilose presente nos grãos moídos (V15) e a interação largura da folha da bandeira e quantidade de amilose presente nos grãos moídos (V4*V15). Assim, obtiveram-se estimativas dos parâmetros, com o intuito de identificar quais foram os efeitos significativos, pelo teste de Wald. Na Tabela 7 são apresentadas as estimativas dos parâmetros, seus erros-padrão, as estatísticas do teste de Wald e os valores das razões de chances.

Tabela 7 – Estimativas do modelo de regressão logística selecionado para a predição da resistência à brusone do arroz (*Oryza sativa* L.) pelo método de Collett (2003)

Variáveis	Coeficiente	Erro padrão	Z	Valor-P	OR	95%IC
Intercepto	-10,796	4,426	-2,439	0,015	-	-
V4	6,403	3,001	2,127	0,033	603,916	(4,274 - 85334,878)
V8	0,257	0,094	2,743	0,006	1,293	(1,108 - 1,509)
V11	-0,725	0,438	-1,658	0,097	0,484	(0,236 - 0,994)
V15	0,358	0,186	1,925	0,054	1,430	(1,053 - 1,941)
V4*V15	-0,233	0,140	-1,663	0,096	0,792	(0,629 - 0,997)

OR - Razão de chance; 95%IC - Intervalo de confiança a 95% para OR.

Analisando a Tabela 7, observa-se que ao nível de significância de 5%, as variáveis incluídas no modelo, V4 e V8 apresentaram significância estatística (valor-p $\leq$ 0,05), e a variável V15 por mostrar um nível de significância próximo de 5% (com um nível de confiança de 94,6%) foi também incluída no modelo, o que indica que estas variáveis têm influência sobre a doença. No entanto, a variável V11 e a interação V4*V15 não apresentaram significância estatística (valor-p > 0,05), indicando que estas variáveis não contribuem para o aumento ou redução à brusone do arroz (*Oryza sativa* L.*)*, podendo ser retiradas do modelo.

De acordo com os resultados obtidos na Tabela 7, foi então ajustado um novo modelo incluindo apenas as três variáveis (V4, V8 e V15) que mostraram significância ao nível de 5%. Uma vez ajustado o modelo com essas três variáveis, foi então utilizado o teste de Hosmer e Lemeshow para testar a qualidade do ajuste, e ao nível de significância de 5% verificou-se que não houve diferenças significativas (valor-p > 0,05) entre os valores preditos e observados para

o novo modelo, indicando que o modelo foi capaz de produzir classificações confiáveis. Foram então obtidas as estimativas dos parâmetros, seus erros-padrão, as estatísticas do teste de Wald e os valores das razões de chances do modelo, cujos valores estão apresentados na Tabela 8.

Tabela 8 – Estimativas do novo modelo de regressão logística selecionado para a predição da resistência à brusone do arroz (*Oryza sativa* L.) com as variáveis largura da folha bandeira (V4), número médio da ramificação da panícula primária (V8) e quantidade de amilose presente nos grãos moídos (V15)

Variáveis	Coeficiente	Erro padrão	Z	Valor-P	OR	95%IC
Intercepto	-7,370	1,289	-5,717	0,000	-	-
V4	1,333	0,653	2,042	0,041	3,793	(1,296 – 11,097)
V8	0,277	0,092	3,000	0,003	1,319	(1,133 - 1,535)
V15	0,090	0,033	2,760	0,006	1,094	(1,037 - 1,155)

OR - Razão de chance; 95%IC - Intervalo de confiança a 95% para OR.

Analisando os resultados da Tabela 8, observa-se que as variáveis largura da folha bandeira (V4), número médio da ramificação da panícula primária (V8) e quantidade de amilose presente nos grãos moídos (V15) incluídas no modelo, apresentaram significância estatística ao nível de 5%, indicando que estas variáveis influenciam a resistência a esta doença. Para cada unidade de acréscimo em cm de V4, espera-se 279,3% de acréscimo na probabilidade da resistência à brusone, e para cada unidade de acréscimo em V8, espera-se em média o acréscimo de 31,9% na probabilidade da resistência, e ainda para cada unidade de acréscimo em V15, espera-se o acréscimo de 9,4% na probabilidade da resistência à brusone do arroz. Assim, quanto maior for o valor dessas variáveis (V4, V8 e V15), maior será a resistência à doença do arroz (*Oryza sativa* L.*)*.

Vários autores (AGAHI et al., 2007; ADITYA e BHARTIYA, 2013) constataram a importância da largura da folha bandeira, pela correlação com o rendimento dos grãos de arroz. Dalchiavon et al., (2012); Fageria, (2000) e Guimarães et al., (2006) constataram importância do número de panículas por metro quadrado e espiguetas por panícula numa estreita correlação com a produtividade de grãos e os componentes da produção, como os que mais contribuíram para a produtividade do arroz e ainda, Ong e Blanshard, (1995) constataram a importância do teor de amilose no arroz, de forma que os grãos com maior teor de amilose apresentam textura mais firme após o cozimento, e por isso essa característica é também importante de ser avaliada durante o desenvolvimento de cultivares.

6. CONCLUSÕES

Neste trabalho, as quinze variáveis independentes iniciais, utilizadas para predição da resistência à brusone do arroz, apenas cinco (largura da folha bandeira (V4), número médio da ramificação da panícula primária (V8), largura da semente com casca (V11), quantidade de amilose presente nos grãos moídos (V15) e a interação $V4 \times V15$) mostraram importância, significativas a 0,10 pelo teste da razão de verossimilhança. Essas variáveis, fazem parte do modelo 2 (modelo de regressão logística múltipla com as variáveis $V4 + V8 + V11 + V15 + V4 \times V15$) selecionado, uma vez que foi bem ajustado e não causou grandes alterações nas estimativas e ainda foi o modelo mais parcimonioso.

Na análise das estimativas do modelo e razão de chance (OR), observou-se que ao nível de significância de 5%, as variáveis incluídas no modelo, V4, V8 e V15 apresentaram significância estatística (valor-p $\leq 0{,}05$), indicando que estas variáveis tem influência sobre a doença. Desta forma, foi então ajustado um novo modelo incluindo apenas essas variáveis, que mostrou uma boa qualidade de ajuste pelo teste de Hosmer e Lemeshow ao nível de significância de 5%. Observou-se que para cada unidade de acréscimo em cm de V4, espera-se 279,3% de acréscimo na probabilidade da resistência à brusone, e para cada unidade de acréscimo em V8, espera-se em média o acréscimo de 31,9% na probabilidade da resistência, e ainda para cada unidade de acréscimo em V15, espera-se o acréscimo de 9,4% na probabilidade da resistência à brusone do arroz. Assim, quanto maior for o valor dessas variáveis (V4, V8 e V15), maior será a resistência à doença do arroz (*Oryza sativa* L.*)*.

7. REFERÊNCIAS BIBLIOGRÁFICAS

ABREU, H. J., **Aplicação da Análise de Sobrevivência em um Problema de Credit Scoring e Comparação com Regressão Logística**. 2004. 118f. Dissertação (Mestrado em Estatística) – Universidade Federal de São Carlos. 2004.

ADITYA, J. P. E BHARTIYA, A., Genetic variability, correlation and path analysis for quantitative characters in rainfed upland rice of Uttarakhand Hills. **Journal of Rice Research**, 6, 24-34. 2013.

AGAHI, K., FOTOKIAN, M. E FARSHADFAR, E., Correlation and path coefficient analysis for some yield-related traits in rice genotypes (Oryza sativa L.). **Asian Journal of Plant Sciences**, 6, 513-517. http:// dx.doi.org/10.3923/ajps.2007.513.517. 2007.

AKAIKE, H. A new look at the statistical model identification. **IEEE Transactions on Automatic Control,** 1974. 19 (6): 716–723.

AYRES, M.; AYRES Jr, M.; AYRES, D. L.; dos SANTOS, A. S. ***BioEstat 4.0:*** **Aplicações estatísticas nas áreas das ciências biológicas e médicas.** Belém: Sociedade Civil Mamirauá; Brasília: CNPq, 2005.

BOZDONGAN. H. Model selection and Akaike's Information Criterion (AIC): The general theory and its analytical extensions. **Psychometrika**. v.52, n.3, p. 345-370, 1987.

BRAGA, A. C. S. **Curvas ROC: Aspectos Funcionais e Aplicações**. 2000. 267f. Tese (Doutorado em Engenharia de Produção e Sistemas) – Universidade do Minho, 2000.

BROCCO, J. B., **Ponderação de Modelos com Aplicação de Regressão Logística binária**. 2006. 78f. Dissertação (Mestrado em Estatística) – Universidade Federal de São Carlos, 2006.

BURNHAM, K. P.; ANDERSON, D. R. Multimodel inference: understanding AIC and BIC in Model Selection. **Sociological Methods and Research** (33), 2004. 261-304.

CASELA, G.; BERGER, R. L. **Statistical Inference**. 2nd edition., 2010.

COLLETT, D. **Modelling Survival Data In Medical Research**; Champman and Hall, London, 2003.

COLOSIMO, E. A., GIOLO, S. R. **Analise de Sobrevivência Aplicada**; ABE – Projeto Fisher, São Paulo: Edgar Blücher, 2006.

COMPANHIA NACIONAL DE ABASTECIMENTO (CONAB). **Produção de Arroz**. 2018. Disponível em: <http://www.conab.gov.br>. Acesso em 19 de set de 2019.

DALCHIAVON, F. C.; CARVALHO, M. P.; COLETTI, A. J.; CAIONE, G.; SILVA, A. F.; ANDREOTTI, M. Correlação linear entre componentes da produção e produtividade do arroz de terras altas em sistema de plantio direto. **Semina: Ciências Agrárias,** Londrina, v. 33, n. 5, p. 1629-1642, set./out. 2012.

FAGERIA, N. K. Eficiência do uso de potássio pelos genótipos de arroz de terras altas. **Pesquisa Agropecuária Brasileira**, Brasília, v. 35, n. 10, p. 2115-2120, 2000.

FIGUEIRA, C. V., **Modelos de regressão logística**. 2006. 138f. Dissertação (Mestrado em Matemática) – Universidade Federal do Rio Grande do Sul, 2006.

GUIMARÃES, C. M.; STONE, L. F.; CASTRO, E. M. Comportamento de cultivares de arrozde terras altas no sistema plantio direto em duas profundidades de adubação. **Bioscience Journal**, Uberlândia, v. 22, n. 1, p. 53-59, 2006.

HAUCK, W. W., DONNER, A. Wald's Tests Applied to Hypotheses in Logit Analysis, **Journal of the American Statistical Association**, v. 72, 1977.

JENNINGS, D. E. Judging Inference Adequacy in Logistic Regressin; **Journal of the American Statistical Association**; v.81; 1986.

GIANOLA, D. Theory and Analysis of Threshold Characters. **Journal of Animal Science**, v. 54, n. 5, p. 1079–1096, 1 maio 1982. DOI: 10.2527/jas1982.5451079x. Disponível em: https://doi.org/10.2527/jas1982.5451079x. Acesso em 27 nov. 2019.

MACKILL, A. O.; BONMAN, J. M. Inheritance of blast resistance in near-isogenic lines of rice**. Phytopathology**, St. Paul, v.82, p.746-749, 1992.

MARCHETTI, M. A., LAI, X.; BOLLICH, C. N. Inheritance of resistance *to Pyricularia oryzae* in rice cultivars grown in the United States; **Phytopathology**, v. 77, n. 6, p. 799-804, 1987.

MARTINEZ, E. Z., LOUZADO-NETO, F., PEREIRA, B. B., "A curva ROC para testes diagnósticos" in **Cadernos Saúde Coletiva**, Rio de Janeiro, p. 7-9, 2003.

MARTINS, C. A. C. **Analise de regressão logística**. 1998. 53 f. Dissertação (Mestrado em Biometria) – Universidade Federal de Pernambuco, 1998.

Matthew Jay, **Generalhoslem: Goodness of Fit Tests for Logistic Regression Models**, 2019.

MAYER, P. L. **Probabilidade: Aplicação à estatística**; Rio de Janeiro: Livros Técnicos e Científicos; 1978.

METZ C. E., "ROC analysis in medical imaging: a tutorial review of the literature" **Radiol Phys Technol**, vol. 1, p. 2-12, 2008.

MORANA H. C. P., **Identificação do ponto de corte para a escala PCL-R (Psychopathy Checklist Revised) em população forense brasileira: caracterização de dois subtipos de personalidade; transtorno global e parcial**, Faculdade de Medicina, 2003.

NELDER, J. A., WEDDERBURN, R. W. M. Generalized linear model**. Journal of the Royal Statistical Society**, London, v. 135, p. 370-384, 1972.

HOSMER, D. W., LEMESHOW, S. **Applied Logistic Regression**. New York: John Wiley & Sons, 1989.

HOSMER, D. W., LEMESHOW, S. **Applied Logistic Regression**. 2nd ed. New York: John Wiley & Sons, 2000.

OLIVEIRA, L. S. **Seleção de Covariáveis para Ajuste de Regressão Logística na Análise da Abundância de invertebrados Edáficos em Diferentes Agroecossistemas**. 2011. 63f. Dissertação (Mestrado em Estatística) – Universidade Federal de Viçosa, 2011.

ONG, M. H.; BLANSHARD, J. M. V. Texture determinants in cooked, parboiled rice I: rice starch amylose and the fine structure of amylopectin. **Journal of Cereal Science**, v.21, p.251-260, 1995.

PESKE, S.T. & BARROS, A.S.S.A. Produção de Sementes de Arroz. In: **Produção de arroz irrigado**. Pelotas. UFPel, 1998. 659 p.

PRICE, A. L.; PATTERSON, N. J.; PLENGE, R. M.; *et al.* Principal components analysis corrects for stratification in genome-wide association studies. **Nature Genetics**, v. 38, n. 8, p. 904-909, 2006. DOI: 10.1038/ng1847. Disponível em: https://doi.org/10.1038/ng1847. Acesso em 6 jan. 2020.

R DEVELOPMENT CORE TEAM. **R: A language and environment for statistical computing**. R Foundation for Statistical Computing, Vienna, Austria, 2019. Disponível em: <http://www.R-project.org>.

RESENDE, M. D. V. de.; RESENDE JÚNIOR, M. F. R.; AGUIAR, A. M.; *et al.* **Computação da Seleção Genômica Ampla (GWS)**. Série Documentos da EMBRAPA Florestas, n. 209, p. 78, 2010.

ROBIN X., TURCK N., HAINARD A., TIBERTI N., LISACEK F., SANCHEZ J., MÜLLER M., pROC: an open-source package for R and S+ to analyze and compare ROC curves. **BMC Bioinformatics**, v.12, p. 77. DOI: 10.1186/1471-2105-12-77, 2011.

SICSÚ, A. L. **Credit Scoring: desenvolvimento, implantação, acompanhamento**. São Paulo: Blucher, 2010.

SING T., SANDER O., BEERENWINKEL N., LENGAUER T., **ROCR: visualizing classifier performance in R**, 2005.

SOSBAI; **Arroz irrigado: Recomendações técnicas da pesquisa para o sul do Brasil.** Congresso Brasileiro de Arroz Irrigado. Reunião da Cultura do Arroz Irrigado, 28. Porto Alegre, 2018. 188 p.

UNITED STATES DEPARTAMENT OF AGRICULTURE (USDA). **Publications rice**. 2018. Disponível em: <http://www.ers.usda.gov>. Acesso em 21 de set. de 2018.

WALD, A., Tests of statistical hypotheses concerning several parameters when the number of observations is large, **Transactions of the American Mathematical Society,** v. 54, p. 426-482, 1943.

WALKER, S. H., DUNCAN D. B., "Estimation of the probability of an event as a function of several independent variables." **Biometrika**, v. 54, p. 167-179, 1967.

WICKHAM H., AVERICK M., BRYAN J., CHANG W., MCGOWAN L. D., FRANÇOIS R., GROLEMUND G., HAYES A., HENRY L., HESTER J., KUHN M., PEDERSEN T. L., MILLER E., BACHE S. M., MÜLLER K., OOMS J., ROBINSON D., SEIDEL D. P., SPINU V., TAKAHASHI K., VAUGHAN D., WILKE C., WOO K., YUTANI H., Welcome to the tidyverse. **Journal of Open Source Software**, v. 4, n. 43, 1686 p., DOI: 10.21105/joss.01686, 2019.

WICKHAM H., ggplot2: Elegant Graphics for Data Analysis. **Springer-Verlag New York**, 2016.

ZHAO, K., TUNG, C. W., EIZENGA, G. C., WRIGHT, M. H.; ALI, M. L., PRICE, A. H., NORTON, G. J., ISLAM, M. R., REYNOLDS, A., MEZEY, J., MCCLUNG, A. M.Ç, BUSTAMANTE, C. D., MCCOUCH, S.R. Genome-wide association mapping reveals a rich genetic architecture of complex traits in Oryza sativa. **Nature communications**, v. 2, 467 p., 2011.

APÊNDICE – CÓDIGO R

```
dados<-read.table("data.txt", h=T)
dados<-na.omit(dados) # Omite todas as linhas que tem NA
dados$v2<-factor(dados$v2) # Variavel categorica
str(dados)
attach(dados)
table(dados$blast)

############################ Seleção de variáveis ################################
###### Passo1
fit0=glm(blast~1, family=binomial(link="logit"))      # Modelo nulo
summary(fit0)
print((dv0=fit0$null.deviance),digits=7)

fit1=glm(blast~v1, family=binomial(link="logit"))     # Hábito colmo
summary(fit1)
print((dv1=fit1$deviance),digits=7)
print((X21<-dv0-dv1),digits=5)
print((pchisq(X21, df=1, lower.tail=FALSE)),digits=4)

fit2=glm(blast~v2, family=binomial(link="logit"))     # Pubescência da folha
summary(fit2)
dv2=fit2$deviance;dv2
print((X22<-dv0-dv2),digits=5)
print((pchisq(X22, df=1, lower.tail=FALSE)),digits=4)

fit3=glm(blast~v3, family=binomial(link="logit")) #Comprimento da folha bandeira
summary(fit3)
print((dv3=fit3$deviance),digits=7)
print((X23<-dv0-dv3),digits=5)
print((pchisq(X23, df=1, lower.tail=FALSE)),digits=4)

fit4=glm(blast~v4, family=binomial(link="logit"))     # Largura da Folha bandeira
summary(fit4)
print((dv4=fit4$deviance),digits=7)
print((X24<-dv0-dv4),digits=5)
print((pchisq(X24, df=1, lower.tail=FALSE)),digits=4)

fit5=glm(blast~v5, family=binomial(link="logit"))  # Número médio de panículas por planta
summary(fit5)
print((dv5=fit5$deviance),digits=7)
print((X25<-dv0-dv5),digits=4)
print((pchisq(X25, df=1, lower.tail=FALSE)),digits=4)

fit6=glm(blast~v6, family=binomial(link="logit"))     # Altura da planta
summary(fit6)
print((dv6=fit6$deviance),digits=7)
print((X26<-dv0-dv6),digits=4)
print((pchisq(X26, df=1, lower.tail=FALSE)),digits=4)

fit7=glm(blast~v7, family=binomial(link="logit"))     # Comprimento da panícula
summary(fit7)
print((dv7=fit7$deviance),digits=7)
print((X27<-dv0-dv7),digits=6)
```

```
print((pchisq(X27, df=1, lower.tail=FALSE)),digits=4)

fit8=glm(blast~v8, fit8=glm(blast~v8, family=binomial(link="logit")) #Número médio da ramificação
da panícula primária
summary(fit8)
print((dv8=fit8$deviance),digits=7)
print((X28<-dv0-dv8),digits=6)
print((pchisq(X28, df=1, lower.tail=FALSE)),digits=4)

fit9=glm(blast~v9, family=binomial(link="logit"))   #Número médio de sementes por panícula
summary(fit9)
print((dv9=fit9$deviance),digits=7)
print((X29<-dv0-dv9),digits=5)
print((pchisq(X29, df=1, lower.tail=FALSE)),digits=3)

fit10=glm(blast~v10, family=binomial(link="logit"))   # Comprimento da semente com casca
summary(fit10)
print((dv10=fit10$deviance),digits=7)
print((X210<-dv0-dv10),digits=5)
print((pchisq(X210, df=1, lower.tail=FALSE)),digits=3)

fit11=glm(blast~v11, family=binomial(link="logit"))   # Largura da semente com casca
summary(fit11)
print((dv11=fit11$deviance),digits=7)
print((X211<-dv0-dv11),digits=5)
print((pchisq(X211, df=1, lower.tail=FALSE)),digits=3)

fit12=glm(blast~v12, family=binomial(link="logit"))   #Volume da semente com casca
summary(fit12)
print((dv12=fit12$deviance),digits=7)
print((X212<-dv0-dv12),digits=5)
print((pchisq(X212, df=1, lower.tail=FALSE)),digits=4)

fit13=glm(blast~v13, family=binomial(link="logit")) #Área da superfície da semente com casca
summary(fit13)
print((dv13=fit13$deviance),digits=7)
print((X213<-dv0-dv13),digits=5)
print((pchisq(X213, df=1, lower.tail=FALSE)),digits=4)

fit14=glm(blast~v14, family=binomial(link="logit")) #Relação comprimento/largura da semente
summary(fit14)
print((dv14=fit14$deviance),digits=7)
print((X214<-dv0-dv14),digits=5)
print((pchisq(X214, df=1, lower.tail=FALSE)),digits=4)

fit15=glm(blast~v15, family=binomial(link="logit"))  #Quantidade de amilose presente nos grãos moíd
os
summary(fit15)
print((dv15=fit15$deviance),digits=7)
print((X215<-dv0-dv15),digits=6)
print((pchisq(X215, df=1, lower.tail=FALSE)),digits=4)

###### Passo 2
fit16=glm(blast~v1+v4+v7+v8+v9+v10+v11+v14+v15, family=binomial(link="logit"))
```

```
summary(fit16)
print((dv16=fit16$deviance),digits=7)

fit17=glm(blast~v4+v7+v8+v9+v10+v11+v14+v15, family=binomial(link="logit"))
summary(fit17)
print((dv17=fit17$deviance),digits=8)
print((X217<-dv17-dv16),digits=6)
print((pchisq(X217, df=1, lower.tail=FALSE)),digits=4)

fit18=glm(blast~v1+v7+v8+v9+v10+v11+v14+v15, family=binomial(link="logit"))
summary(fit18)
print((dv18=fit18$deviance),digits=8)
print((X218<-dv18-dv16),digits=6)
print((pchisq(X218, df=1, lower.tail=FALSE)),digits=3)

fit19=glm(blast~v1+v4+v8+v9+v10+v11+v14+v15, family=binomial(link="logit"))
summary(fit19)
print((dv19=fit19$deviance),digits=8)
print((X219<-dv19-dv16),digits=6)
print((pchisq(X219, df=1, lower.tail=FALSE)),digits=4)

fit20=glm(blast~v1+v4+v7+v9+v10+v11+v14+v15, family=binomial(link="logit"))
summary(fit20)
print((dv20=fit20$deviance),digits=8)
print((X220<-dv20-dv16),digits=6)
print((pchisq(X220, df=1, lower.tail=FALSE)),digits=4)

fit21=glm(blast~v1+v4+v7+v8+v10+v11+v14+v15, family=binomial(link="logit"))
summary(fit21)
print((dv21=fit21$deviance),digits=8)
print((X221<-dv21-dv16),digits=5)
print((pchisq(X221, df=1, lower.tail=FALSE)),digits=4)

fit22=glm(blast~v1+v4+v7+v8+v9+v11+v14+v15, family=binomial(link="logit"))
summary(fit22)
print((dv22=fit22$deviance),digits=8)
print((X222<-dv22-dv16),digits=5)
print((pchisq(X222, df=1, lower.tail=FALSE)),digits=4)

fit23=glm(blast~v1+v4+v7+v8+v9+v10+v14+v15, family=binomial(link="logit"))
summary(fit23)
print((dv23=fit23$deviance),digits=8)
print((X223<-dv23-dv16),digits=5)
print((pchisq(X223, df=1, lower.tail=FALSE)),digits=4)

fit24=glm(blast~v1+v4+v7+v8+v9+v10+v11+v15, family=binomial(link="logit"))
summary(fit24)
print((dv24=fit24$deviance),digits=8)
print((X224<-dv24-dv16),digits=5)
print((pchisq(X224, df=1, lower.tail=FALSE)),digits=4)

fit25=glm(blast~v1+v4+v7+v8+v9+v10+v11+v14, family=binomial(link="logit"))
summary(fit25)
print((dv25=fit25$deviance),digits=8)
```

```
print((X225<-dv25-dv16),digits=5)
print((pchisq(X225, df=1, lower.tail=FALSE)),digits=4)

############ Passo 3
fit26=glm(blast~v4+v8+v10+v11+v14+v15, family=binomial(link="logit"))
summary(fit26)
print((dv26=fit26$deviance),digits=7)

fit27=glm(blast~v4+v8+v10+v11+v14+v15+v1, family=binomial(link="logit"))
summary(fit27)
print((dv27=fit27$deviance),digits=8)
print((X227<-dv26-dv27),digits=5)
print((pchisq(X227, df=1, lower.tail=FALSE)),digits=4)

fit28=glm(blast~v4+v8+v10+v11+v14+v15+v7, family=binomial(link="logit"))
summary(fit28)
print((dv28=fit28$deviance),digits=7)
print((X228<-dv26-dv28),digits=5)
print((pchisq(X228, df=1, lower.tail=FALSE)),digits=4)

fit29=glm(blast~v4+v8+v10+v11+v14+v15+v9, family=binomial(link="logit"))
summary(fit29)
print((dv29=fit29$deviance),digits=7)
print((X229<-dv26-dv29),digits=5)
print((pchisq(X229, df=1, lower.tail=FALSE)),digits=5)

####### Passo 4
fit30=glm(blast~v4+v8+v10+v11+v14+v15, family=binomial(link="logit"))
summary(fit30)
print((dv30=fit30$deviance),digits=7)

fit31=glm(blast~v4+v8+v10+v11+v14+v15+v2, family=binomial(link="logit"))
summary(fit31)
print((dv31=fit31$deviance),digits=7)
print((X231<-dv30-dv31),digits=5)
print((pchisq(X231, df=1, lower.tail=FALSE)),digits=5)

fit32=glm(blast~v4+v8+v10+v11+v14+v15+v3, family=binomial(link="logit"))
summary(fit32)
print((dv32=fit32$deviance),digits=8)
print((X232<-dv30-dv32),digits=5)
print((pchisq(X232, df=1, lower.tail=FALSE)),digits=5)

fit33=glm(blast~v4+v8+v10+v11+v14+v15+v5, family=binomial(link="logit"))
summary(fit33)
print((dv33=fit33$deviance),digits=7)
print((X233<-dv30-dv33),digits=5)
print((pchisq(X233, df=1, lower.tail=FALSE)),digits=5)

fit34=glm(blast~v4+v8+v10+v11+v14+v15+v6, family=binomial(link="logit"))
summary(fit34)
print((dv34=fit34$deviance),digits=7)
print((X234<-dv30-dv34),digits=5)
print((pchisq(X234, df=1, lower.tail=FALSE)),digits=5)
```

```
fit35=glm(blast~v4+v8+v10+v11+v14+v15+v12, family=binomial(link="logit"))
summary(fit35)
print((dv35=fit35$deviance),digits=7)
print((X235<-dv30-dv35),digits=6)
print((pchisq(X235, df=1, lower.tail=FALSE)),digits=5)

fit36=glm(blast~v4+v8+v10+v11+v14+v15+v13, family=binomial(link="logit"))
summary(fit36)
print((dv36=fit36$deviance),digits=7)
print((X236<-dv30-dv36),digits=6)
print((pchisq(X236, df=1, lower.tail=FALSE)),digits=5)

######## Passo 5
fit37=glm(blast~v4+v8+v10+v11+v14+v15+v3+v6+v12+v13, family=binomial(link="logit"))
summary(fit37)
print((dv37=fit37$deviance),digits=7)

fit38=glm(blast~v8+v10+v11+v14+v15+v3+v6+v12+v13, family=binomial(link="logit"))
summary(fit38)
print((dv38=fit38$deviance),digits=7)
print((X238<-dv38-dv37),digits=6)
print((pchisq(X238, df=1, lower.tail=FALSE)),digits=3)

fit39=glm(blast~v4+v10+v11+v14+v15+v3+v6+v12+v13, family=binomial(link="logit"))
summary(fit39)
print((dv39=fit39$deviance),digits=7)
print((X239<-dv39-dv37),digits=5)
print((pchisq(X239, df=1, lower.tail=FALSE)),digits=3)

fit40=glm(blast~v4+v8+v11+v14+v15+v3+v6+v12+v13, family=binomial(link="logit"))
summary(fit40)
print((dv40=fit40$deviance),digits=8)
print((X240<-dv40-dv37),digits=5)
print((pchisq(X240, df=1, lower.tail=FALSE)),digits=4)

fit41=glm(blast~v4+v8+v10+v14+v15+v3+v6+v12+v13, family=binomial(link="logit"))
summary(fit41)
print((dv41=fit41$deviance),digits=7)
print((X241<-dv41-dv37),digits=5)
print((pchisq(X241, df=1, lower.tail=FALSE)),digits=3)

fit42=glm(blast~v4+v8+v10+v11+v15+v3+v6+v12+v13, family=binomial(link="logit"))
summary(fit42)
print((dv42=fit42$deviance),digits=8)
print((X242<-dv42-dv37),digits=5)
print((pchisq(X242, df=1, lower.tail=FALSE)),digits=4)

fit43=glm(blast~v4+v8+v10+v11+v14+v3+v6+v12+v13, family=binomial(link="logit"))
summary(fit43)
print((dv43=fit43$deviance),digits=7)
print((X243<-dv43-dv37),digits=5)
print((pchisq(X243, df=1, lower.tail=FALSE)),digits=4)
```

```
fit44=glm(blast~v4+v8+v10+v11+v14+v15+v6+v12+v13, family=binomial(link="logit"))
summary(fit44)
print((dv44=fit44$deviance),digits=7)
print((X244<-dv44-dv37),digits=5)
print((pchisq(X244, df=1, lower.tail=FALSE)),digits=4)

fit45=glm(blast~v4+v8+v10+v11+v14+v15+v3+v12+v13, family=binomial(link="logit"))
summary(fit45)
print((dv45=fit45$deviance),digits=7)
print((X245<-dv45-dv37),digits=5)
print((pchisq(X245, df=1, lower.tail=FALSE)),digits=4)

fit46=glm(blast~v4+v8+v10+v11+v14+v15+v3+v6+v13, family=binomial(link="logit"))
summary(fit46)
print((dv46=fit46$deviance),digits=7)
print((X246<-dv46-dv37),digits=4)
print((pchisq(X246, df=1, lower.tail=FALSE)),digits=4)

fit47=glm(blast~v4+v8+v10+v11+v14+v15+v3+v6+v12, family=binomial(link="logit"))
summary(fit47)
print((dv47=fit47$deviance),digits=7)
print((X247<-dv47-dv37),digits=5)
print((pchisq(X247, df=1, lower.tail=FALSE)),digits=4)

######## Passo 6
fit48=glm(blast~v4+v8+v11+v15, family=binomial(link="logit"))
summary(fit48)
print((dv48=fit48$deviance),digits=7)

fit49=glm(blast~v4+v8+v11+v15+v4*v8, family=binomial(link="logit"))
summary(fit49)
print((dv49=fit49$deviance),digits=7)
print((X249<-dv48-dv49),digits=4)
print((pchisq(X249, df=1, lower.tail=FALSE)),digits=4)

fit50=glm(blast~v4+v8+v11+v15+v4*v11, family=binomial(link="logit"))
summary(fit50)
print((dv50=fit50$deviance),digits=7)
print((X250<-dv48-dv50),digits=5)
print((pchisq(X250, df=1, lower.tail=FALSE)),digits=4)

fit51=glm(blast~v4+v8+v11+v15+v4*v15, family=binomial(link="logit"))
summary(fit51)
print((dv51=fit51$deviance),digits=7)
print((X251<-dv48-dv51),digits=5)
print((pchisq(X251, df=1, lower.tail=FALSE)),digits=3)

fit52=glm(blast~v4+v8+v11+v15+v8*v11, family=binomial(link="logit"))
summary(fit52)
print((dv52=fit52$deviance),digits=7)
print((X252<-dv48-dv52),digits=4)
print((pchisq(X252, df=1, lower.tail=FALSE)),digits=4)

fit53=glm(blast~v4+v8+v11+v15+v8*v15, family=binomial(link="logit"))
```

```
summary(fit53)
print((dv53=fit53$deviance),digits=7)
print((X253<-dv48-dv53),digits=4)
print((pchisq(X253, df=1, lower.tail=FALSE)),digits=4)

fit54=glm(blast~v4+v8+v11+v15+v11*v15, family=binomial(link="logit"))
summary(fit54)
print((dv54=fit54$deviance),digits=7)
print((X254<-dv48-dv54),digits=5)
print((pchisq(X254, df=1, lower.tail=FALSE)),digits=4)

################################## Etapa final ###################################
fit55=glm(blast~v4+v8+v11+v15+v4*v15+v11*v15, family=binomial(link="logit"))
summary(fit55)
print((dv55=fit55$deviance),digits=7)

fit56=glm(blast~v4+v8+v11+v15+v4*v15, family=binomial(link="logit"))
summary(fit56)
print((dv56=fit56$deviance),digits=7)

fit57=glm(blast~v4+v8+v11+v15+v11*v15, family=binomial(link="logit"))
summary(fit57)
print((dv57=fit57$deviance),digits=7)

###### Novo modelo ajustado com apenas variáveis que mostraram importância em fit56 ######
fit58=glm(blast~v4+v8+v15, family=binomial(link="logit"))  # Modelo sem interação
summary(fit58)
print((dv58=fit58$deviance),digits=7)

######################## Teste de Hosmer & Lemeshow ################################
Library(generalhoslem)

hlfit55<-logitgof(dados$blast, fitted(fit55))
cbind(hlfit55$expected, hlfit55$observed)
hlfit55

hlfit56<-logitgof(dados$blast, fitted(fit56))
cbind(hlfit56$expected, hlfit56$observed)
hlfit56

hlfit57<-logitgof(dados$blast, fitted(fit57))
cbind(hlfit57$expected, hlfit57$observed)
hlfit57

hlfit58<-logitgof(dados$blast, fitted(fit58))
cbind(hlfit58$expected, hlfit58$observed)
hlfit58

########AIC dos Modelos
aic1<- fit55$aic
aic2<-fit56$aic
aic3<-fit58$aic
aic<- rbind(aic1, aic2, aic3)
```

```
################### Medidas de associação (razões de chance) e ICde 95% ############
OR1=exp(fit55$coefficients)
OR2=exp(fit56$coefficients)
OR3=exp(fit58$coefficients)

ICbeta1=confint.default(fit55, level = 0.90)
ICbeta2=confint.default(fit56, level = 0.90)
ICbeta3=confint.default(fit58, level = 0.90)

########### Intervalos de Confiança 95% para Odds ratio - OR ######################
ICOR1= exp(ICbeta1)
ICOR2= exp(ICbeta2)
ICOR3= exp(ICbeta3)

round((cbind(OR1,ICOR1)),3)
round((cbind(OR2,ICOR2)),3)
round((cbind(OR3,ICOR3)),3)

########################### Construção da curva ROC ############################
library(tidyverse)
library(ggplot2)
library(pROC)
library(ROCR)

pred1 = predict(fit55,dados, type = "response")
pred2 = predict(fit56,dados, type = "response")
pred3 = predict(fit58,dados, type = "response")

pred_v1 = cbind(dados, pred1)
pred_v2 = cbind(dados, pred2)
pred_v3 = cbind(dados, pred3)

pred.val1 = prediction(pred1 ,pred_v1$blast)
pred.val2 = prediction(pred2 ,pred_v2$blast)
pred.val3 = prediction(pred3 ,pred_v3$blast)

########################### Obtendo valores de AIC e AUC ######################
auc1 = performance(pred.val1,"auc")
auc2 = performance(pred.val2,"auc")
auc3 = performance(pred.val3,"auc")

auc1<- unlist(slot(auc1, "y.values"))
auc2<- unlist(slot(auc2, "y.values"))
auc3<- unlist(slot(auc3, "y.values"))

auc1<- round(auc1,4)
auc2<- round(auc2,4)
auc3<- round(auc3,4)
auc<-rbind(auc1, auc2, auc3)
medidas<- cbind(aic,auc)
medidas
```

```
############################## Construção da curva ROC ########################

curva1 <- roc(dados$blast, fit55$fitted.values)
curva2 <- roc(dados$blast, fit56$fitted.values)
ggroc(list(Modelo1 = curva1,
        Modelo2 = curva2),
            cex = 1, legacy.axes = TRUE)  +
 scale_color_manual(labels = c("Modelo 1",
                               "Modelo 2"),
                    breaks = c("Modelo1",
                               "Modelo2"),
                    values=c(Modelo1 = "#c62b3a",
                             Modelo2 = "#604ed3")) +
 theme_bw() +
 theme(legend.position = c(0.8, 0.35),
      legend.background = element_rect(colour = "black")) +
 labs(x = "1-Especificidade",
     y = "Sensibilidade",
     color = "Discriminação") +
 geom_abline(intercept = 1,
           slope = 1,
           linetype = 2,
           cex = 1)
```