ZEFERINO GOMES DA SILVA NETO

CURVA ROC PARA COMPARAÇÃO DE MODELOS DE PREDIÇÃO PARA VARIÁVEIS DICOTÔMICAS

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
S586c
2020

Silva Neto, Zeferino Gomes da, 1996-
Curva ROC para comparação de modelos de predição para variáveis dicotômicas / Zeferino Gomes da Silva Neto. – Viçosa, MG, 2020.
69f. : il. (algumas color.) ; 29 cm.

Inclui apêndices.
Orientador: Sebastião Martins Filho.
Dissertação (mestrado) - Universidade Federal de Viçosa.
Inclui bibliografia.

1. Análise de regressão logística. 2. Melhoramento genético. 3. Germinação. 4. Brusone. 5. *Capsicum chinense*. I. Universidade Federal de Viçosa. Centro de Ciências Exatas e Tecnológicas. Programa de Pós-Graduação em Estatística Aplicada e Biometria. II. Título.

CDD 22 ed. 519.36

ZEFERINO GOMES DA SILVA NETO

CURVA ROC PARA COMPARAÇÃO DE MODELOS DE PREDIÇÃO PARA VARIÁVEIS DICOTÔMICAS

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: 28 de fevereiro de 2020.

Assentimento:

Zeferino Gomes da Silva Neto
Zeferino Gomes da Silva Neto
Autor

Sebastião Martins Filho
Sebastião Martins Filho
Orientador

À minha família.

AGRADECIMENTOS

A Deus que pela sua infinita bondade me deu o dom da vida, e forças para continuar até aqui, e por cuidar sempre de cada simples detalhe em minha vida.

Aos meus pais pela educação, carinho, ensinamentos e pela motivação em todos os momentos da minha vida.

Às minhas irmãs Mara Luana e Ranuze Maria pelas palavras de conforto e apoio, pela torcida e por partilhar momentos importantes durante toda a minha trajetória.

A toda minha família, incluindo tios (as) e primos (as), que eu sei que estão sempre na "torcida" e que fazem muita falta no meu cotidiano.

À Natália Moreira Nunes, por compartilhar comigo momentos difíceis e momentos de alegria, por ser companheira, amiga e por tornar meus dias ainda melhores.

Ao meu orientador, prof. Dr Sebastião Martins Filho pela compreensão diante de minhas falhas e pela disposição em ajudar em todos os momentos. Sou grato pelo acolhimento, pela orientação durante o desenvolvimento deste trabalho e, principalmente, os ensinamentos durante esse pouco tempo de convivência.

Aos membros da banca, prof. Dr Antônio Policarpo Sousa Carneiro e prof. Dr Vinícius Silva dos Santos por estarem dispostos a dar suas contribuições para este trabalho.

Aos amigos Carlos Ayallas, Ithalo Coelho e Kaleo Pereira pelos momentos de descontração e alegria, durante a convivência nesse tempo que morei em Viçosa.

Ao meu grande amigo Antônio Alberto Ibiapina Filho pela amizade e parceria.

Aos colegas de Mestrado em Estatística, pelo companheirismo e pela ajuda prestada ao longo de dois anos de percurso.

Ao Manuel Zavala (Instituto Nacional de Pesquisa Florestal, Agrícola e Pecuária – INIFAP – MÉX, Pesquisador na área de hortaliças) por ter concedido o banco de dados para realização de parte das análises.

Ao Ministério Universidades Renovadas por me permitir fazer amizades em Deus que me ajudaram superar todas as dificuldades que apareceram.

À Universidade Federal de Viçosa, que pelo Departamento de Estatística me deu a oportunidade de realizar o mestrado.

À Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (CAPES) pela concessão da bolsa de estudo. O presente trabalho foi realizado com apoio da CAPES – Brasil – Código de Financiamento 001.

BIOGRAFIA

ZEFERINO GOMES DA SILVA NETO, filho de Antônia Rodrigues da Silva Gomes e de Antônio Gomes da Silva, nasceu em Teresina, no estado do Piauí, em 09 de abril de 1996.

Em abril de 2014, ingressou no curso de Bacharelado em Estatística na Universidade Federal do Piauí, Teresina – PI, graduando-se em julho de 2017.

Em março de 2018, iniciou o curso de mestrado do Programa de Pós-Graduação em Estatística Aplicada e Biometria na Universidade Federal de Viçosa, submetendo-se à defesa da dissertação em 28 de fevereiro de 2020.

"Nada te perturbes, nada te espante,
pois tudo passa. Só Deus não muda.
Tudo a paciência por fim alcança.
Quem a Deus tem, nada lhe falta.
pois só Deus basta"

Santa Teresa de Jesus

RESUMO

DA SILVA NETO, Zeferino Gomes, M.Sc., Universidade Federal de Viçosa, fevereiro de 2020. C**urva ROC para comparação de modelos de predição para variáveis dicotômicas**. Orientador: Sebastião Martins Filho.

A utilização de modelos de regressão logística e de seleção genômica ampla (GWS) tem elevada importância em ciências agrárias e, portanto, há necessidade de aplicação de metodologias eficientes para a avaliação do poder discriminatório destes modelos. Uma metodologia pouco difundida nesta área e em GWS é a *Receiver Operating Characteristic*, ou curva ROC. Neste trabalho, objetivou-se aplicar curva ROC para a seleção de modelos de regressão logística aplicados a dados de germinação de sementes de pimentas habanero e à GWS, considerando dados de resistência do arroz *Oriza sativa* à brusone. Os modelos testados para a predição da capacidade germinativa das sementes foram compostos dos fatores: variedade (laranja e vermelha), período de armazenamento do fruto (0, 7 e 14 dias), método de extração das sementes (manual e mecânico) e período de armazenamento das sementes (3, 6, 9 e 12 meses). O modelo que se mostrou mais adequado conteve a interação entre variedade, armazenamento do fruto e armazenamento da semente. Por sua vez, os métodos utilizados na GWS, para avaliação da resistência à brusone, foram o BRR (*Bayesian Ridge Regression*), Bayes Cπ e BLASSO. Esses métodos foram comparados pelos seguintes critérios: taxa de erro na validação, coeficiente de Spearman e viés com a área abaixo da curva ROC (AUC). Os valores de AUC para a seleção dos modelos mostraram-se equivalentes aos valores dos índices usuais, que indicaram os modelos BRR e Bayes Cπ como os melhores. Além disso, a representação gráfica das curvas ROC se mostrou ainda mais vantajosa por permitir a determinação da sensibilidade dos modelos em diferentes valores de especificidade. Abaixo de 0,25 de 1-especificidade os modelos BRR e Bayes Cπ foram mais sensíveis que o BLASSO, mas acima deste valor todos foram equivalentes. No entanto, o BRR apresentou menor tempo de execução (4h52min, 6h1min, 6h25min para o BRR, Bayes Cπ e BLASSO, respectivamente). Por fim, pode-se verificar que a análise ROC se mostrou eficiente para a avaliação de modelos de regressão logística e de GWS e, portanto, os resultados aqui apresentados indicam que a curva ROC pode ser utilizada como uma excelente ferramenta para seleção de modelos em ciências agrárias.

Palavras-chave: Regressão logística. Seleção Genômica Ampla. Germinação de sementes. Brusone do arroz. Pimenta habanero. Qualidade de ajuste de modelo.

ABSTRACT

DA SILVA NETO, Zeferino Gomes, M.Sc., Universidade Federal de Viçosa, February, 2020. **ROC curve for comparing prediction models for dichotomous variables**. Adviser: Sebastião Martins Filho.

The use of logistic regression models and genomics wide selection (GWS) is very important in agricultural sciences and, therefore, the application of efficient methodologies to assess the discriminatory power of these models is needed. A poorly-used methodology in this area and in GWS is the Receiver Operating Characteristic, or ROC curve. In this work, the aim was to apply a ROC curve for the selection of logistic regression models applied to germination data of habanero pepper seeds and to GWS, considering resistance data of rice Oriza sativa to blast. The tested models for the prediction of the germination capacity of the seeds were composed of the factors: variety (orange and red), period of storage of the fruit (0, 7 and 14 days), method of extraction of the seeds (manual and mechanical) and period of seed storage (3, 6, 9 and 12 months). The most suitable model contained the interaction between variety, fruit storage and seed storage. On the other hand, the methods used in GWS to assess blast resistance were BRR (Bayesian Ridge Regression), Bayes Cπ and BLASSO. These methods were compared using the following criteria: error rate in the validation, Spearman coefficient and bias with the area under the ROC curve (AUC). The AUC values for the selection of the models were equivalent to the values of the usual indices, which indicated the BRR and Bayes Cπ models as the best. In addition, the graphical representation of the ROC curves proved to be even more advantageous as it allows the determination of the sensitivity of the models in different specificity values. Below 0.25 of 1-specificity the BRR and Bayes Cπ models were more sensitive than the BLASSO, but above this value all the models were equivalent. However, the BRR had a faster execution (4h52min, 6h1min, and 6h25min for the BRR, Bayes Cπ and BLASSO, respectively). Finally, the ROC analysis proved to be efficient for the evaluation of logistic regression models and GWS and, therefore, the results presented here indicate that the ROC curve can be used as an excellent tool for selecting models in agrarian sciences.

Keywords: Logistic regression. Genomics Wide Selection. Seed germination. Rice blast. Habanero pepper. Model fitting quality.

SUMÁRIO

1. INTRODUÇÃO GERAL

Existem muitas situações em que alguns conjuntos de objetos, cenários ou ações podem ser classificados como pertencentes a uma de duas classes. Os métodos de classificação devem consistir em informações observadas sobre cada um (a) deles (as). No entanto, estes métodos não são perfeitos, na maioria das vezes acontecem erros, o que atrapalha uma correta classificação das classes. Por isso, deve-se avaliar a qualidade da realização dos métodos. Assim, é possível decidir se um modelo é bom o suficiente, tentar aperfeiçoá-lo ou, simplesmente, substituí-lo (KRZANOWSKI, HAND, 2009).

Para avaliar a qualidade do desempenho de modelos de classificação que visam responder a que classe pertence cada indivíduo em estudo existem várias metodologias que são amplamente utilizadas. Uma abordagem eficiente para a comparação de modelos de classificação é a curva *Receiver Operating Characteristic* (ROC). A curva ROC foi inicialmente desenvolvida na área da psicologia sensorial. O propósito original era mostrar a existência de uma associação empírica entre o corpo e a mente, conforme Gustav Theodor Fechner (1801-1887), filósofo alemão e médico de formação, apontado o precursor em psicometria. No decorrer da Segunda Guerra Mundial (1939-1945), a curva ROC foi usada para estimar a capacidade dos operadores de radares distinguirem um sinal de ruído (METZ, 2008). Desde então, a análise ROC tem sido utilizada em diversas áreas da ciência para a comparação de classificadores de dados.

Existem diversas técnicas que podem ser utilizadas para classificação de dados, advindas da estatística clássica ou da mineração de dados, tais como: classificação bayesiana, árvores de decisão, análise discriminante, máquinas de vetores de suporte, regressão logística, redes neurais artificiais, dentre outros. O livro de Hosmer et al. (2013) é tido como uma boa referência na área, pois foi produzido exclusivamente para abordar a aplicação adequada da regressão logística em diversos conjuntos de dados médicos. Atualmente a regressão logística tem sido usada nas áreas de biomedicina, ecologia, finanças, educação e meteorologia (OHYVER et al., 2017). No entanto, os procedimentos adotados ainda são pouco difundidos na área de ciências agrárias.

Por outro lado, uma metodologia muito utilizada nas ciências agrárias é a seleção genômica ampla (GWS – *Genome Wide Selection*), que visa a predição de valores genéticos genômicos (GEBV- *Genomic Estimated Breeding Value*) dos indivíduos por meio de marcadores moleculares (MEUWISSEN et al. 2001) e permite a seleção precoce de indivíduos geneticamente superiores. Isto faz da GWS uma ferramenta de relevância importância na área.

Desta forma, torna-se importante avaliar o desempenho tanto dos modelos de regressão logística na área de ciências agrárias (onde é mais empregada na área médica), bem como dos modelos de GWS. Neste sentido, a metodologia ROC pode consistir em uma ferramenta eficaz para a avaliação de tais modelos. Portanto, o presente trabalho teve como objetivo comparar modelos de regressão logística a partir das metodologias clássica e bayesiana para dados oriundos de um experimento de campo com variáveis categóricas (variedade dos frutos, período de armazenamento dos frutos, métodos de extração das sementes e período de armazenamento das sementes pós-extração) e para um conjunto de dados com informações genotípicas, respectivamente, utilizando curva ROC.

2. REVISÃO DE LITERATURA

2.1. Considerações Iniciais

As duas bases de dados utilizadas neste trabalho requerem técnicas de modelagem advindas de diferentes campos do conhecimento. A regressão logística normalmente é descrita como originária do campo de técnicas estatísticas clássicas e no ensino de técnicas de estatística multivariada (embora, possa ser utilizada para problemas univariados). Por sua vez, a seleção genômica ampla vem sendo cada vez mais utilizada devido aos avanços computacionais. Essa abordagem experimental possui inúmeras aplicações, e vem trazendo progressos para o melhoramento genético de animais domésticos e plantas. Gerando assim, mudanças positivas na capacidade de prever fenótipos e, com isso, aumentando a acurácia seletiva em idade precoce e maximizando o ganho genético por unidade de tempo (RESENDE et al., 2010).

2.2. Regressão Logística

A análise de regressão é uma das técnicas estatísticas mais utilizadas e consiste em um conjunto de ferramentas que permitem a modelagem e exame de associações entre variáveis relacionadas de maneira não determinística. A regressão linear tem bom desempenho quando a variável resposta é quantitativa contínua. Mas, na prática, existem muitas situações em que a variável de interesse é explicada por mais de uma variável explicativa, o que leva à regressão linear múltipla.

A análise de regressão linear é inadequada no caso de respostas dicotômicas, ou seja, quando admite duas respostas, pois quase nunca seriam satisfeitas as quatro suposições usuais da análise de regressão linear, as quais são: média zero, variância constante, normalidade e independência para os resíduos de tal modelo. Nestes casos, o modelo de regressão logística deve ser utilizado (MONTGOMERY, RUNGER, 2016; PARDOE 2006).

Quando se utiliza o método da regressão logística, o objetivo é o mesmo que de todas as técnicas de construção de modelos em estatística: obter o modelo que melhor se ajusta aos dados, o mais parcimonioso e que seja coerente para descrever a relação entre uma variável resposta e um conjunto de variáveis explicativas. A escolha da função logística se dá principalmente por ser extremamente flexível, facilmente utilizável e por permitir uma interpretação fundamentada (HOSMER; LEMESHOW; STURDIVANT, 2013).

Segundo Figueira (2006), o modelo de regressão logística binária é um caso particular dos modelos lineares generalizados (GLM). Estes, por sua vez, são especificados por três componentes. São elas:

- Uma componente aleatória, a qual identifica a distribuição de probabilidade da variável dependente;
- Uma componente sistemática que especifica uma função linear entre as variáveis independentes e
- Uma função de ligação que relaciona os valores esperados da componente aleatória com a componente sistemática.

A função de ligação na regressão logística é a função *logit*, que permite transformar as probabilidades dos fatores de uma variável com resposta categórica binária em uma escala contínua, que varia de $-\infty$ a $+\infty$. Após a transformação, a variável resposta pode ser modelada com regressão linear simples, pois irá satisfazer as propriedades desejáveis. O modelo de regressão logística pode ser utilizado da forma apresentada na Equação 1 ou pela sua forma equivalente (Equação 2).

$$\ln\left[\frac{\pi(x)}{1-\pi(x)}\right] = g(x) = \beta_0 + \beta_1 x \quad (1)$$

$$\pi(x) = \frac{e^{\beta_0+\beta_1 x}}{1+e^{\beta_0+\beta_1 x}} \quad (2)$$

onde $\pi(x)$ é o resultado mais importante (sucesso) que se pretende associar com as outras variáveis de interesse (x) e os parâmetros β_0 e β_1 tem significados similares aos seus análogos na regressão linear, onde representam o coeficiente de estimação das variáveis explicativas.

Em relação a estimação dos parâmetros β, na regressão linear, de acordo com Hosmer e Lemeshow (2000), o método mais utilizado é o Método dos Mínimos Quadrados (MMQ), onde são determinados valores para os parâmetros que minimizam a soma dos quadrados de desvios entre os valores observados e os valores preditos pelo modelo. Quando o MMQ é utilizado em modelos com resposta dicotômica, os estimadores ferem os pressupostos desejáveis.

Para resolver este problema, utiliza-se o Método da Máxima Verossimilhança (MMV), que gera valores para os parâmetros desconhecidos que maximizam a probabilidade de obtenção do conjunto de dados observado.

O completo desenvolvimento para estimar os parâmetros de um modelo de regressão logística, via o MMV pode ser encontrado em (HOSMER et al., 2013).

Uma vez obtidas as estimativas dos parâmetros de um modelo, as suas interpretações requerem que as inferências dos coeficientes estimados sejam feitas de maneira prática. Uma importante medida para a interpretação dos parâmetros encontrados no modelo de regressão logística é a *odds ratio* (OR) ou razão de chances. A OR quantifica a probabilidade de um acontecimento ser mais provável que outro, indicando o quão mais provável se torna a presença da variável resposta de acordo com o incremento da variável explicativa. Por exemplo, se a variável resposta for a presença ou ausência de uma determinada doença e a variável explicativa for se a pessoa é fumante ou não, assim uma razão de possibilidade = 2 estima que a doença é duas vais mais provável de acontecer entre fumantes em relação aos não fumantes. (SANTOS, 2016).

2.3. Curva ROC

A curva ROC (*Receiving Operator Characteristics*) foi inicialmente desenvolvida na área da psicologia sensorial. O propósito inicial era mostrar a existência de uma associação empírica entre o corpo e a mente. O filósofo alemão e médico de formação Gustav Theodor Fechner (1801-1887), apontado como precursor em psicometria, condicionava um determinado estímulo às pessoas, a fim de colher uma quantidade estável de respostas positivas. Ele desenvolveu graficamente a associação entre as respostas positivas e a medida física da intensidade do estímulo, obtendo assim uma função psicométrica (BRAGA, 2000).

Louis Leon Thurstone (1887-1955), precursor em psicometria e psicofísica, baseado nos resultados de Fechner, evoluiu novos métodos para avaliar as habilidades mentais. Ele é o criador dos conceitos de ruído, critério de decisão e ponto de corte (BRAGA, 2000).

No decorrer da Segunda Guerra Mundial (1939-1945), a curva ROC foi então utilizada para estimar a capacidade dos operadores de radares distinguirem um sinal de ruído (METZ, 2008). Isto é, cada vez que um radar identificava qualquer sinal competia ao operador definir a veracidade e relevância do que havia sido identificado, se um avião inimigo, um míssil ou, simplesmente, um bando de pássaros (MARTINEZ et al., 2003). Desde então, a análise ROC tem sido utilizada em diversas áreas.

A partir da década de 60, as curvas ROC foram utilizadas principalmente em psicologia experimental e na década de 70 estenderam-se pelos campos da medicina. Em medicina, inicialmente o objetivo foi auxiliar na classificação de indivíduos em doentes ou saudáveis (MARTINEZ et al., 2003). Posteriormente, a análise ROC foi introduzida por Spackman (1989) em Aprendizado de Máquina como uma ferramenta útil e poderosa para a avaliação de modelos de classificação.

Dado um classificador binário que classifica indivíduos em positivos e negativos e um conjunto de indivíduos avaliados, ao predizer a classe dos indivíduos desse conjunto e compara-los com a classe real, poderão ser apresentados quatro diferentes casos. Com esses casos é possível formar uma matriz de confusão, apresentada a seguir (Tabela 1):

Tabela 1 – Matriz de confusão correspondente as categorias resultantes da classificação de um conjunto de dados em duas classes distintas comparadas com a classe real observada.

Resultado	**Classe Real**	
Classe Predita	**Positivo**	**Negativo**
Positivo	Verdadeiro-positivo (VP)	Falso-positivo (FP)
Negativo	Falso-negativo (FN)	Verdadeiro-negativo (VN)

Fonte: Adaptado de Azevedo e Pereira, 2010.

Esta matriz de contingência apresenta quatro eventos possíveis: 1) Indivíduo positivo, corretamente predito, denominado verdadeiro positivo; 2) Indivíduo positivo avaliado incorretamente, definido como falso negativo; 3) Indivíduo negativo, mas avaliado como positivo ou falso positivo; e 4) Indivíduo negativo avaliado corretamente ou Verdadeiro Negativo. Pode-se então calcular, com base nas repostas do modelo, as várias probabilidades de interesse, do seguinte modo:

1) Taxa de verdadeiros positivos (VP) = $\frac{Número\ de\ verdadeiros\ positivos}{Número\ de\ casos\ positivos}$

2) Taxa de falsos negativos (FN) = $\frac{Número\ de\ falsos\ negativos}{Número\ de\ casos\ positivos}$

3) Taxa de falsos positivos (FP) = $\frac{Número\ de\ falsos\ positivos}{Número\ de\ casos\ negativos}$

4) Taxa de verdadeiros negativos (VN) = $\frac{Número\ de\ verdadeiros\ negativos}{Número\ de\ casos\ negativos}$

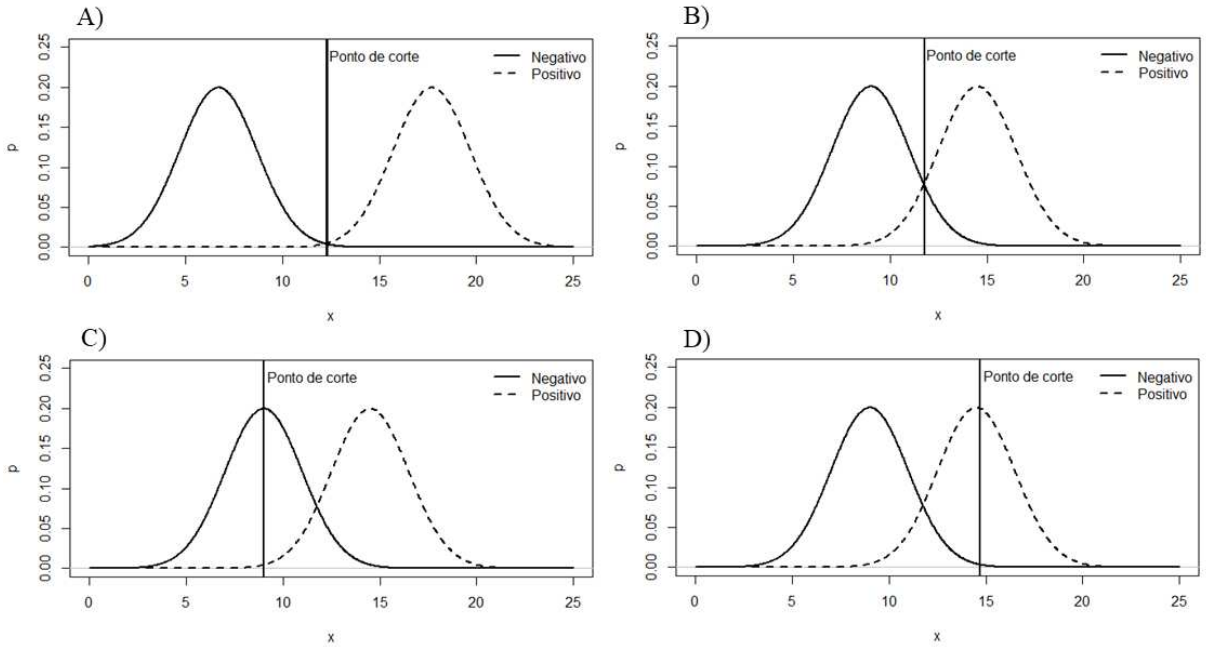
Com esses valores é possível definir medidas comuns e bastante utilizadas no contexto de curva ROC. São elas: a sensibilidade, especificidade e ponto de corte. As duas primeiras medidas são as principais técnicas de avaliação da capacidade de discriminação de um modelo de classificação. A sensibilidade (S_e) é a capacidade do modelo em identificar uma classificação positiva, dado que ela realmente é positiva. A especificidade (E_s) é definida como a capacidade do modelo em identificar um resultado negativo, dado que ele é realmente negativo (MARTINEZ et al., 2003). O cálculo da S_e e E_s é definido da seguinte forma:

$$Sensibilidade = S_e = \frac{VP}{VP+FN} \qquad Especificidade = E_s = \frac{VN}{FP+VN}$$

É possível perceber que a S_e e a E_s são medidas dependentes entre si, visto que são obtidas sob as mesmas investigações e em ambas se utiliza as classificações positivas e negativas. Assim, o acréscimo de uma das medidas implica geralmente o decréscimo da outra. Isso viabiliza a utilização de um ponto de corte que delimitará o nível de S_e e E_s esperado e que deverá ser escolhido de acordo com o desenvolvimento do estudo, podendo-se dar maior relevância a uma dessas medidas, ou à combinação das duas (PITANGA, 2004).

O ponto de corte equivale a um ponto de divisão que realiza a discriminação de uma variável, classificando a presença ou ausência de uma determinada característica investigada nos indivíduos por meio da observação de uma medida aplicada para realizar esta análise. O ponto de divisão demarca um valor para essa medida, determinando assim quais os indivíduos que estão acima ou abaixo desse ponto (MORANA, 2003). Para facilitar o entendimento, considera-se que um grupo de indivíduos é classificado como negativo se na sua análise resultar um valor excedente ao valor do ponto de corte e, caso contrário, será positivo. A Figura 1 explicita situações ilustrativas nas quais são apresentados variados pontos de corte, gerando assim diferentes delimitações de S_e e E_s.

Figura 1 – Representação de diferentes sobreposições de duas distribuições hipotéticas de indivíduos submetidos a mesma investigação. A distribuição com linha contínua representa os indivíduos positivos para determinada característica e a distribuição com linha tracejada representa os indivíduos negativos para esta característica. (A) apresenta o ponto de corte ideal que distingui perfeitamente os indivíduos e (B) apresenta o ponto de corte semelhante a realidade, onde pode-se verificar que a distinção não é perfeita. (C) e (D) apresentam valor maior e menor de ponto de corte, respectivamente.



Fonte: Adaptado de Cristiano, 2017.

Se for assumido que todos os indivíduos foram classificados corretamente, então a discriminação é dita perfeita (Figura 1A). No entanto, é bastante complicado encontrar um modelo para o qual não se consigam falsos resultados positivos ou negativos. Assim sendo, devemos analisar diferentes valores para o ponto de corte de modo a obter o valor que minimize a ocorrência de resultados falsos (FLUSS et al., 2005).

Tomando como referência o ponto de corte estabelecido no caso apresentado na Figura 1B, a classificação apresenta erros, uma vez que alguns diagnósticos negativos são classificados como positivos (os que estão à esquerda da reta do ponto de corte, abaixo da curva com linha tracejada), enquanto outros diagnósticos positivos são identificados como negativos (os que se encontram à direita da reta do ponto de corte, abaixo da curva com linha contínua).

Na maior parte dos casos, quanto menor o ponto de corte, maior será a S_e e menor será a E_s. Este entendimento é exemplificado por meio da Figura 1C. Assim, a área da região à esquerda da reta e abaixo da curva com linha tracejada (FN) diminui e a área da região à direita da reta e abaixo da curva com linha contínua (FP) aumenta. Em contrapartida, quanto maior for o ponto de corte, menor será a S_e, e maior será a E_s. Assim sendo, beneficiar uma medida prejudica a outra.

Na Figura 1D, em analogia às Figuras 1B e 1C, é constatado que de acordo com o aumento do valor do ponto de corte, a área da região à esquerda da reta e abaixo da curva com linha tracejada (FN) aumenta, logo a sensibilidade diminui e a área da região à direita da reta e baixo da curva com linha contínua (FP) também diminui, logo a especificidade aumenta.

Tanto a estatística como a informática visam, juntas, obter um modelo que tenha ao mesmo tempo alto poder de sensibilidade e alto poder de especificidade, pois o ponto de corte fixa um par S_e/E_s. Estes podem ser representados como coordenadas "x" e "y" dando origem à representação gráfica que compara os índices gerados a partir de diversos pontos de corte. Esta representação gráfica é denominada curva ROC. Ela possibilita fazer análises empíricas da eficácia e, portanto, da qualidade de um modelo diferenciando duas classes numa amostra (MARTINEZ et al., 2003).

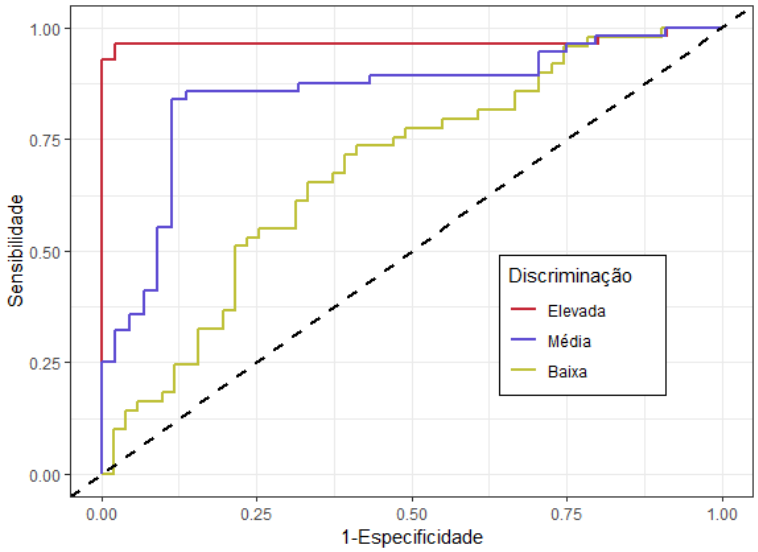
A curva ROC é feita por meio do cálculo da S_e e da E_s para cada ponto de corte e reproduzem-se graficamente os pontos de coordenadas (1-E_s, S_e). Por conveniência, 1-E_s é colocada no eixo das abcissas e a sensibilidade é indicada no eixo das ordenadas, e ambas variam de 0 a 1 (0-100%) (Figura 2).

De acordo com Olivera Ritta et al. (2015), quanto mais distante a curva ROC do modelo estiver em relação à diagonal principal ($x = y$), melhor será o poder classificatório das variáveis

no modelo. Quando a Área Abaixo a Curva (AUC) ROC fosse inferior a 0,7 o modelo seria completamente incapaz de classificar variáveis, seria aceitável com AUC entre 0,7 e 0,8, seria excelente com AUC entre 0,8 e 0,9; e fora de série (mas, extremamente rara) se AUC superior a 0,9 (HOSMER et al., 2013).

Assumindo que as curvas não se cruzam, podemos comparar curvas e avaliar o desempenho de modelos por meio de 3 tipos de discriminação: baixo, médio e elevado. Na Figura 2 estão apresentadas as curvas ROC dos diferentes tipos de discriminação.

Figura 2 – Curvas ROC dos diferentes tipos de discriminação (baixo, médio e elevado).



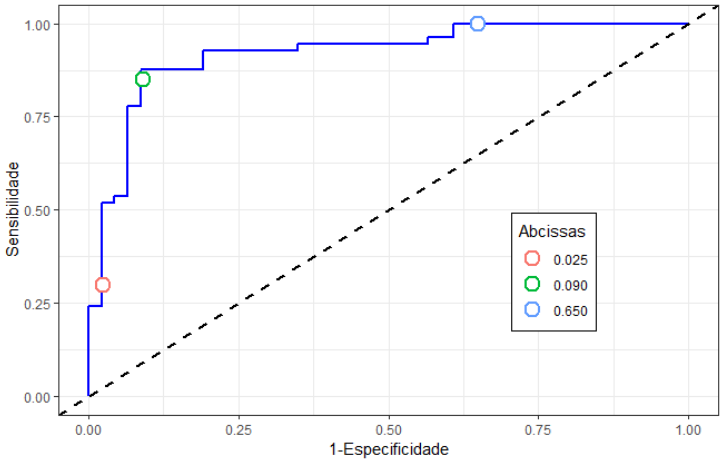
Fonte: Adaptado de Braga, 2000.

De acordo com o que se procura favorecer (sensibilidade, especificidade ou as duas) num modelo existem três critérios de decisão (BRAGA, 2000).

- Critério estrito: estabelece um ponto na curva ROC que se dispõe no canto inferior esquerdo do espaço ROC. Consiste em representar pequenas frações, tanto de verdadeiros positivos como de falsos positivos. Logo, corresponde a situações com baixa sensibilidade, mas elevada especificidade.
- Critério moderado: estabelece um ponto na curva ROC que se dispõe, aproximadamente, no meio do espaço ROC. Consiste em representar a fração de verdadeiros positivos relativamente superior a falsos positivos.
- Critério brando: estabelece um ponto na curva ROC que se dispõe no canto superior direito do espaço ROC. Consiste em representar uma grande fração de verdadeiros positivos e uma pequena fração de falsos positivos. Logo, corresponde a situações com elevada sensibilidade, mas pouca especificidade.

Estes critérios são retratados na Figura 3, em que os pontos com abcissas 0,025, 0,090, e 0,650 correspondem, respetivamente, aos critérios estrito, moderado e brando.

Figura 3 – Representação dos critérios de decisão em que os pontos com abcissas 0,025, 0,090, e 0,650 correspondem, respetivamente, ao critério estrito, moderado e brando.

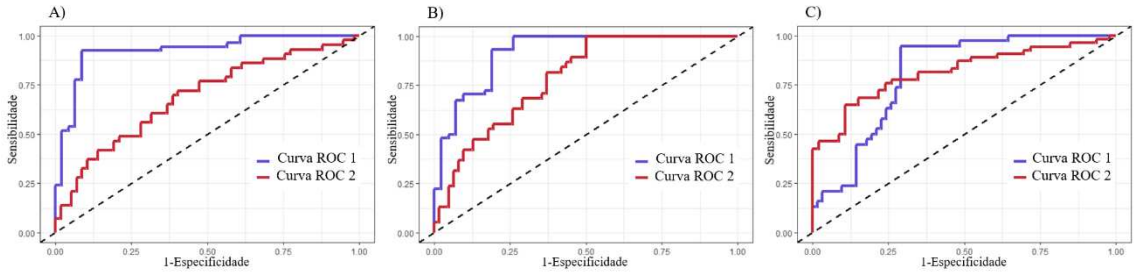


Fonte: Adaptado de Braga, 2000.

Os gráficos que reproduzem duas ou mais curvas ROC na mesma escala associadas a diferentes modelos possibilitam uma comparação empírica rápida e direta do desempenho teórico dos mesmos. A escolha do melhor modelo não resulta apenas do melhor desempenho, visto que tal desempenho pode ser mais dispendioso e levar mais tempo. Desta forma, não necessariamente o modelo matematicamente melhor será o escolhido.

Quando se comparam duas curvas de modelos diferentes podem acontecer os seguintes casos representados na Figura 4.

Figura 4 – Representação de três casos (A, B e C) de comparação de modelos por curva ROC.



Fonte: Adaptado de Cristiano, 2017.

Na Figura 4A, as curvas diferem totalmente e não se cruzam e a curva com maior área é a que avalia o modelo com melhor desempenho. Note-se que nestas situações, para qualquer valor de especificidade o modelo associado à curva ROC 1 obtém sempre melhor valor para a sensibilidade que o outro modelo que deu origem à curva ROC 2. Por sua vez, na Figura 4B as

curvas são diferentes num determinado intervalo, mas iguais em outro, sendo a curva com maior área a que traduz os modelos com melhor desempenho. Assim, nesta situação, pode verificar-se um melhor desempenho teórico da curva ROC 1. Note-se que nestes casos, para valores de especificidade entre 0 e 0,50 o modelo que gerou a curva ROC 1 obtém sempre melhor valor para a sensibilidade que o outro modelo que deu origem a curva ROC 2.

Já na Figura 4C, as curvas cruzam-se em determinado ponto e apesar das áreas diferirem, os modelos apresentam desempenhos melhores do que o outro dependendo do intervalo observado. O modelo correspondente à curva ROC 1 poderia ser utilizado primeiro para identificar a maioria dos verdadeiros positivos e, num segundo momento, utilizar o modelo correspondente à curva ROC 2 para os indivíduos que foram classificados como negativos inicialmente, uma vez que este segundo modelo é melhor para classificar os indivíduos que são de fato negativos. Esta mudança é feita próxima do ponto (0,30; 0,75).

Ao se comparar duas curvas ROC originadas de modelos diferentes que tenham combinações de sensibilidade e 1-especificidade significativos num intervalo de pontos de corte é, frequentemente, mais aproveitável utilizar métodos de comparação para esse intervalo representativo em vez de para toda a curva (PITANGA, 2004).

2.4. Seleção Genômica Ampla

A Seleção Genômica Ampla (Genome Wide Selection - GWS) baseia-se na predição conjunta dos efeitos genéticos de inúmeros marcadores genéticos (SNPs - polimorfismo de um único nucleotídeo, ou *single nucleotide polymorphisms*) espalhados em todo o genoma de um organismo, de modo a detectar os efeitos de todos os locs, de pequenos e grandes efeitos, explicar grande parte da variação genética de um caráter quantitativo e distinguir indivíduos geneticamente superiores de uma população (MEUWISSEN et al. 2001).

Por meio desse método, a predição e a seleção podem ser feitas em fases primárias das plantas, agilizando assim o processo de melhoramento genético. Portanto, a predição visa ser mais acurada, por examinar o real parentesco genético dos indivíduos avaliados, em desvantagem do parentesco médio esperado matematicamente (RESENDE et al., 2010).

Técnicas de estimação bayesiana foram incorporadas na seleção genômica ampla por Meuwissen et al. (2001). Esses procedimentos foram comparados com os usuais BLUP (Best Linear Unbiased Predictor) (HENDERSON, 1974) e LS (Least Squares) (LANDE e THOMPSON, 1990) e os resultados apresentaram maiores acurácias.

O modelo geral para seleção genômica, apresentado por Meuwissen et al. (2001) é dado por:

$$\boldsymbol{y} = \mathbf{1}\mu + \boldsymbol{X\beta} + \boldsymbol{e}$$

em que $\boldsymbol{y}$ é o vetor coluna que contém os valores da variável Y para cada indivíduo; $\mathbf{1}$ é o vetor coluna formado por $1s$ de dimensão $nx1$; μ é a média da variável y; $\boldsymbol{\beta}$ é o vetor de efeito aditivo dos marcadores com matriz de incidência $\boldsymbol{X}$ que relaciona efeitos de SNPs aos valores da característica y e $\boldsymbol{e}$ é o vetor de erro associado ao modelo.

Com efeito, os métodos bayesianos na GWS diferem de acordo com as suposições sobre as distribuições *a priori* dos efeitos de marcadores (i) e das suas respectivas variâncias (σ^2). Em seguida veremos, de maneira sucinta, algumas metodologias e suas ocorrências.

No método BRR (*Bayesian Ridge Regression*) é considerada pressuposição de homogeneidade das variâncias dos SNPs. Assim, tem-se apenas um valor assumido para σ^2. Os parâmetros de efeito de SNPs (β_i), variância dos marcadores (σ^2) e variância aditiva (σ_u^2) seguem respectivamente: $\beta_i|\sigma^2 \sim N(0, \sigma^2), \sigma^2 \sim \chi^{-2}(\mathrm{v}, S^2)$ e $\sigma_u^2 = 2\sigma^2 \sum_{i=1}^{m} p_i\,(1 - p_i)$, em que v representa os graus de liberdade, S^2 é o parâmetro de escala da distribuição e p_i denota as frequências alélicas. Meuwissen et al. (2001) consideram os valores 4,012 ou 4,2 para v e 0,002 e 0,0429 para S^2.

O BLASSO (*Bayesian Least Absolute Shrinkage and Selection Operator*), foi proposto por De Los Campos et al. (2009) a partir de uma interpretação bayesiana baseada no LASSO (TIBSHIRANI, 1996). Neste método, o modelo geral para seleção genômica é descrito como $y = 1\mu + Zm + e$, em que m é o vetor de efeitos aditivos dos marcadores com matriz de incidência Z que foi reparametrizada conforme Vitezica et al. (2013), a fim de se enquadrar na teoria de genética quantitativa. Assim:

$$Z = \begin{cases} se\ AA, então\ 2 - 2p_j, \\ se\ Aa, então\ 1 - 2p_j, \\ se\ aa, então\ 0 - 2p_j. \end{cases}$$

em que p_j é a frequência alélica do marcador j de AA, Aa e aa, que correspondem ao genótipo da planta i no marcador j, que pode ser homozigoto dominante, heterozigoto ou homozigoto recessivo, respectivamente.

As distribuições *a priori* dadas aos parâmetros dos efeitos de marcadores e componentes de variância deste modelo é dada por Pérez e De Los Campos (2014):

$$m_j|\sigma_\epsilon^2, \tau_j, \lambda^2 \sim N\left(0, \tau_j^2 \cdot \sigma_\epsilon^2\right); \tau_j \sim Exp\left(\frac{\lambda^2}{2}\right); \lambda^2 \sim gama(r, s)$$

Por sua vez, o método Bayes Cπ é semelhante ao método BLASSO, mas se difere na distribuição *a priori* dada ao método, que segundo Pérez e De Los Campos (2014) foi implementada como:

$$m_j|\sigma_m^2, \pi \sim \left[\pi \cdot N(0, \sigma_m^2) + (1 - \pi) \cdot \left(m_j = 0\right)\right]; \sigma_m^2 \sim \chi^{-2}(df_m, S_m); \pi \sim beta(p_0, \pi_0)$$

Essa diferença de distribuições *a priori* propicia a seleção de marcadores, uma vez que conduz uma quantidade (1-π) de marcadores a zero, onde π é tratado como uma incógnita com distribuição a priori uniforme.

Visando avaliar a qualidade do ajuste, e para que os efeitos dos marcadores não sejam superestimados devido à estimação e validação na mesma amostra (CRUZ et al., 2013), o procedimento de validação cruzada é constantemente utilizado em seleção genômica. O método consiste em dividir a população em k grupos. E em seguida, um grupo é utilizado como população de validação e k-1 grupos são utilizados como população de estimação. Na população de estimação, os efeitos dos marcadores são estimados e utilizados na população de validação a fim de obter as estimativas dos valores genéticos genômicos (GEBVs). Esse procedimento é executado até que cada um dos k grupos seja utilizado uma vez como população de validação.

2.5. Referências

AZEVEDO, L.; PEREIRA, A. da C. **Avaliação Crítica e Implementação Prática de Estudos Sobre e Validade de Testes Diagnósticos**: Parte II. Nascer e Crescer, v. 19, n. 4, p. 265–277, 2010.

BRAGA, A. C. da S. **Curvas ROC:** aspectos funcionais e aplicações. 2000. Tese (Doutorado em Engenharia de Produção e Sistemas) - Departamento de Produção e Sistemas,Universidade do Minho, Braga, 2000.

CRISTIANO, M. V. M. B. **Sensibilidade e Especificidade na Curva ROC:** Um Caso de Estudo. 2017. Dissertação (Mestrado em Gestão de Sistemas de Informação Médica Sensibilidade) - Faculdade de Medicina, Instituto Politécnico de Leiria, Leiria, 2017.

CRUZ, C. D.; SALGADO, C. S.; BHERING, L. L. **Genômica aplicada**. 1. ed. Visconde de Rio Branco, MG: Suprema Gráfica Editora, v. 1. 424p. 2013.

DE LOS CAMPOS, G.; NAYA, h.; GIANOLA, D. et al. **Predicting quantitative traits with regression models for dense molecular markers**. Genetics, Austin, v. 182, p. 375-385, 2009.

FIGUEIRA, C. V. **Modelos de regressão logística**. 2006. Dissertação (Mestrado em Matemática) - Departamento de Pós Graduação em Matemática, Universidade Federal do Rio Grande do Sul, Porto Alegre, RS 2006.

FLUSS, R.; FARAGGI, D.; REISER, B. **Estimation of the Youden Index and its associated cutoff point**. Biometrical journal. Biometrische Zeitschrift, v. 47, n. 4, p. 458–472, ago. 2005. DOI: 10.1002/bimj.200410135.

HENDERSON, C. R. **Aplications of linear models in animal breeding**. University of Guelph, Guelph, 462p., 1984.

HOSMER, D. W.; LEMESHOW, S. **Applied Logistic Regression**. 2. ed., John Wiley & Sons, Inc., 2000.

HOSMER J. R. D. W.; LEMESHOW, S.; STURDIVANT, R. X. **Applied Logistic Regression**. 3. ed. Hoboken, John Wiley & Sons, Inc., 2013.

KRZANOWSKI, W. J.; HAND, D. J. **ROC Curves for Continuous Data**. 1. ed. New York: Chapman & Hall/CRC, 2009.

LANDE, R.; THOMPSON, R. **Efficiency of marker-assisted selection in the improvement of quantitative traits**. Genetics, v. 124, p. 743-756, 1990.

MARTINEZ, E. Z. F.; LOUZADO-NETO, F.; PEREIRA, B. B. **A Curva ROC para testes diagnósticos**. Cadernos Saúde Coletiva, v. 11, n. 3, p. 7–31, 2003.

METZ, C. E. **ROC analysis in medical imaging**: a tutorial review of the literature. Radiological physics and technology, v. 1, n. 1, p. 2–12, jan. 2008. DOI: 10.1007/s12194-007-0002-1.

MEUWISSEN, T. H. E.; HAYES, B. J.; GODDARD, M. E. **Prediction of Total Genetic Value Using Genome-Wide Dense Marker Maps**. Genetics, v. 157, n. 4, p. 1819 LP – 1829, 2001. Disponível em: http://www.genetics.org/content/157/4/1819.abstract. Acesso em: 18 dez. 2019.

MONTGOMERY, D. C., RUNGER, G. C. **Estatística aplicada e probabilidade para engenheiros**. 6. ed., LTC, 2016.

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. 2003. Tese (Doutorado em Psiquiatria) - Faculdade de Medicina, Universidade de São Paulo, São Paulo, SP, 2003.

OHYVER, M.; MONIAGA, J. V.; YUNIDWI, K. R. et al. **Logistic Regression and Growth Charts to Determine Children Nutritional and Stunting Status:** A Review. Procedia

Computer Science, v. 116, p. 232–241, 1 jan. 2017. DOI: 10.1016/J.PROCS.2017.10.045. Disponível em: https://www.sciencedirect.com/science/article/pii/S1877050917320938. Acesso em: 14 dezembro 2019.

OLIVERA RITTA, C.; GORLA, M. C.; HEIN, N. **Modelo de regressão logística para análise de risco de crédito em uma instituição de microcrédito produtivo orientado**. Iberoamerican Journal of Industrial Engineering, v. 7, n. 13, p. 103–122, 2015.

PARDOE, I. **Applied Regression Modeling:** A Business Approach. John Wiley & Sons, 2006. DOI: 10.1002/9781118274415.

PÉREZ, P.; DE LOS CAMPOS, G. **Genome-wide regression and prediction with the BGLR statistical package**. Genetics, v. 198, n. 2, p. 483–495, out. 2014. DOI: 10.1534/genetics.114.164442. Disponível em: https://www.ncbi.nlm.nih.gov/pubmed/25009151. Acesso em: 12 dez. 2019.

PITANGA, F. **Sensibilidade e especificidade do índice de conicidade como discriminador do risco coronariano de adultos em Salvador, Brasil**. Revista Brasileira de Epidemiologia, p. 259–269, 2004.

PROVOST, F.; FAWCETT, T. **Robust classifier for imprecise environments**. Machine Learning, v. 42, n. 3, p. 203–231, 2001.

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.

SANTOS, R. **Probabilidades e conceitos associados**. p. 34-40, 2016.

SPACKMAN, K. A. **Signal detection theory**: Valuable tools for evaluating inductive learning". In. Proceedings of the 6th Int Workshop on Machine Learning (ICML'1989). Morgan Kaufmann, p. 160-163, 1989.

TIBSHIRANI, R. **Regression Shrinkage and Selection via the Lasso**. Journal of the Royal Statistical Society. Series B (Methodological), v. 58, n. 1, p. 267–288, 28 jan. 1996. Disponível em: http://www.jstor.org/stable/2346178. Acesso em 3 dez. 2019.

VITEZICA, Z. G.; VARONA, L.; LEGARRA, A. **On the additive and dominant variance and covariance of individuals within the genomic selection scope**. Genetics, v. 195, n. 4, p. 1223–1230, dez. 2013. DOI: 10.1534/genetics.113.155176.

CAPÍTULO 1

Curva ROC como ferramenta de comparação de modelos de regressão logística aplicados a dados de germinação de pimentas habanero

Resumo: Recentemente, as pimentas habanero tem ganhado destaque na produção de molhos, pastas e outros produtos alimentícios. Devido ao aumento na demanda dessas pimentas, torna-se ainda mais necessária a obtenção de sementes de qualidade a fim de se obter maior produtividade. Vários fatores podem influenciar na qualidade das sementes, de modo que, a utilização de modelos de regressão logística para predizer a percentagem de germinação das sementes pode trazer resultados importantes acerca das melhores condições de produção. O presente trabalho objetivou comparar modelos de regressão logística utilizando a metodologia de curva ROC para avaliar a capacidade germinativa de sementes de pimenta habanero com os fatores: variedade (laranja e vermelha), período de armazenamento do fruto (0, 7 e 14 dias), método de extração (manual e mecânico) e período de armazenamento das sementes (3, 6, 9 e 12 meses). A metodologia de curva ROC, que visa minimizar a ocorrência de falsos positivos e falsos negativos, foi empregada para avaliar a capacidade discriminatória dos modelos. Os valores do critério de AIC e os valores da área abaixo da curva ROC indicaram que o melhor modelo testado foi aquele com a interação entre "variedade", "armazenamento da semente" e "armazenamento do fruto", e sem o fator "método de extração". Além disso, a representação gráfica da curva ROC mostrou que o modelo escolhido tem melhor desempenho em relação aos demais em todos os valores de sensibilidade associados à especificidade. Assim, os resultados apresentados no presente trabalho demonstraram que a análise ROC é eficiente para a comparação de modelos de regressão logística.

Palavras-Chave: Modelagem Estatística. Germinação de sementes. Análise ROC. Qualidade de Ajuste.

1. Introdução

As espécies de pimentas do gênero *Capsicum* são oriundas das Américas e já eram consumidas há mais de 7.000 anos no México e atualmente são encontradas por todo o mundo. Elas são consumidas por um quarto da população mundial, principalmente como condimentos. Os frutos de pimenteiras são apreciados pela multiplicidade de formas, cores, tamanhos, sabores e pungência (ardência). Nas diversas regiões do Brasil, há uma ampla variedade de pimentas com características peculiares para a produção de molhos, conservas, pápricas, geleias, bombons, embutidos (salames, salsichas e linguiças), massas, patês, *ketchup* e maionese, como por exemplo as pimentas dedo-de-moça, bode, malagueta e habanero (RIBEIRO et al., 2008). Em consequência do aumento do mercado consumidor, tem sido observado crescimento na área plantada, o que consequentemente exige maior demanda por sementes de qualidade.

Em todas as espécies há uma limitação na disponibilidade de sementes de qualidade, principalmente pelo desconhecimento da melhor época de colheita, no qual a semente alcança a maturidade fisiológica, obtendo-se o máximo acúmulo de matéria seca. Entretanto, as sementes de pimenta apresentam dormência, uma característica intrínseca que impede a germinação após atingir seu maior vigor e qualidade.

Alguns estudos têm apontado que sementes armazenadas por determinado período no fruto após a colheita apresentam efeitos positivos no desempenho germinativo em relação a colheitas feitas previamente, dando continuidade ao processo de maturidade e atingindo níveis máximos de germinação e vigor (Barbedo et al., 1999, Castro et al., 2008, Queiroz et al., 2011, Sanchez et al., 1993, Vidigal et al., 2006).

Geralmente, sementes dispostas para germinar logo após a colheita do fruto, quando ainda não adquiriram maturidade fisiológica, apresentam menor porcentagem de germinação em relação àquales cujos testes de germinação são feitos após alguns dias de armazenamento. Isso também é percebido quando se faz o armazenamento dos frutos carnosos de algumas espécies como o pimentão, abóbora, melancia, mamão, berinjela, tomate e pepino (CASTRO et al., 2008).

Dessa forma, a modelagem estatística dos dados de germinação constitui uma técnica importante para alcançar a maior germinação possível das sementes, sendo a regressão logística um dos principais modelos utilizados quando se deseja calcular ou prever a probabilidade de um evento específico que admite duas ou mais categorias de classificação.

Em problemas envolvendo duas categorias, uma das maneiras mais completas de se analisar o desempenho de um modelo de classificação é por meio da análise ROC (PROVOST

e FAWCET, 1997). Apesar de ser uma forma de analisar o desempenho de classificação propagado desde a segunda guerra mundial (1939-1945), a análise ROC ainda tem seu uso pouco difundido na área das ciências agrárias.

Diante do exposto, o objetivo desse trabalho foi comparar modelos de regressão logística baseando-se na metodologia de curva ROC para avaliar a capacidade germinativa de sementes de pimenta habanero (*Capsicum chinense*) obtidas em frutos com diferentes colorações, submetidas ao repouso pós-colheita nos frutos e o posterior armazenamento depois da extração.

2. Material e métodos

Foram utilizadas sementes de pimenta habanero (*Capsicum chinense* Jacquin) colhidas em maio de 2016 em uma plantação localizada na Península de Yucatán (Motul), México, no sítio dos irmãos Hernández, localizado nas coordenadas 21°05'42N de latitude e 89°16'59W de longitude e 6 metros de altitude.

O delineamento experimental utilizado foi inteiramente casualizado (DIC), com quatro repetições, em esquema fatorial 2x3x2, correspondendo a duas variedades de pimenta Habanero, três tempos de armazenamento do fruto e dois métodos de extração da semente, tendo como variável resposta dicotômica a germinação das sementes. As análises estatísticas foram processadas com o auxílio dos pacotes tidyverse (WICKHAM, 2017), ROCR (SING et al., 2007) e ggplot2 (WICKHAM, 2009) do *software* R (R CORE TEAM, 2019) versão 3.6.1.

Na colheita, os frutos foram retirados das plantas com base na coloração apresentada e classificados como frutos laranja (N-MAYAPAN) e vermelho (R-CALAKMUL), totalizando aproximadamente 1200 frutos de cada coloração. Antes da extração das sementes, os frutos foram divididos em três grupos. No primeiro, as sementes foram extraídas um dia depois da colheita. No segundo, a extração foi realizada sete dias após a colheita e no terceiro grupo, foram extraídas após quatorze dias da colheita. A extração das sementes foi feita de forma manual (com faca) e mecânica (com liquidificador modificado). Em seguida, os frutos foram armazenados em geladeira a 10 ºC em embalagem plástica hermética, para evitar a alteração da umidade durante todo o armazenamento. Para separar as sementes de boa qualidade das chochas, foi utilizado o método de densidade em água.

Uma vez separadas as sementes, as de boa qualidade foram desinfestadas com cloro à 3% (v/v) por 3 minutos e em seguida foram lavadas com água para eliminar os resíduos de cloro. Por fim, as sementes foram colocadas em sacolas de tecido até alcançar o conteúdo de

água recomendado para a espécie 9 ± 3 (SAGARPA, 2014). Em laboratório, foi realizado teste de germinação em quatro repetições de 25 sementes, que foram dispostas em placa de Petri com duas folhas de papel germitest® umedecidas com água (em duas vezes o peso seco do papel). A germinação foi contabilizada todos os dias por até 7 dias. Foi estabelecido o critério de semente germinada o atingimento de 1 mm de comprimento da radícula (HERNÁNDEZ-VERDUGO et al., 2001).

A partir dos dados de germinação, foi estimada a porcentagem de sementes germinadas (PGF) utilizando-se a fórmula PGF = (n/N)×100, em que n é o número de sementes germinadas e N é o número total de sementes utilizadas em cada repetição (Silva et al., 2019).

Os dados foram modelados pela abordagem de regressão logística baseada nos passos propostos por Hosmer et al. (2013), conhecidos como *Purposeful Selection of Covariates*. Essa metodologia consiste em sete passos que abrangem a lógica que muitos pesquisadores utilizam quando examinam um conjunto de dados afim de construir um modelo de regressão múltiplo. Primeiramente, foi analisada cada variável explicativa separadamente aplicando-se análise univariada com nível de significância de 0,25. Em seguida, construiu-se o modelo múltiplo incluindo todas as variáveis significativas do primeiro passo, empregando nível de significância de 0,05. O modelo reduzido foi comparado com o modelo maior, até que foi obtido o modelo com as variáveis essenciais. Em seguida, foi realizada análise de variância para checar a presença de interações entre as variáveis do modelo.

A capacidade discriminatória dos diferentes modelos de regressão logística foi avaliada comparando a área sob as respectivas curvas ROC (*Receiving Operator Characteristics*). Considerou-se que a capacidade de discriminação de um modelo seria insuficiente se a Área Abaixo a Curva (AUC) ROC fosse inferior a 0,7, seria aceitável com AUC entre 0,7 e 0,8, seria excelente com AUC entre 0,8 e 0,9; e fora de série (mas, extremamente rara) se AUC superior a 0,9 (HOSMER et al. 2013).

3. Resultados e discussão

De acordo com os resultados obtidos no primeiro passo da elaboração do modelo de regressão logística de Hosmer et al. (2013), foi observado que os efeitos de todas as variáveis apresentaram associação significativa ($P<0,25$) pelo teste de Wald em relação à germinação das sementes de pimenta habanero (Tabela 1). Dessa forma, todas foram selecionadas para compor o modelo múltiplo, tendo em vista que o primeiro passo é uma análise univariada para cada variável explicativa.

Observando os dados apresentados na Tabela 1 podemos verificar que as porcentagens de germinação das sementes dos frutos com coloração vermelha (R-CALAKMUL) foram superiores em relação aos frutos com coloração laranja (N-MAYAPAN), e as sementes que ficaram em repouso nos frutos apresentaram um aumento na proporção de germinação conforme ficavam mais dias armazenadas. As sementes extraídas de maneira mecânica obtiveram maior percentual de germinação em relação às extraídas manualmente. Além disso, os períodos inicial (3 meses) e final (12 meses) do armazenamento das sementes após a extração apresentaram menores percentuais de germinação, sendo o período de 9 meses superior em relação aos demais tempos de armazenamento das sementes. Ressalta-se que embora tenha ocorrido menor taxa de germinação nas sementes que não foram armazenadas no fruto, e decréscimo na germinação aos 12 meses de armazenamento das sementes após a extração, as mesmas se encontram dentro do percentual mínimo recomendado para a comercialização de sementes básicas, que é de 70% (BRASIL, 2009).

Tabela 1 – Distribuição das 4.800 sementes de pimenta habanero avaliadas quanto a germinação em função da variedade dos frutos (laranja e vermelha), período de armazenamento dos frutos (0, 7 e 14 dias), método de extração das sementes (manual e mecânico) e tempo de armazenamento das sementes (3, 6, 9 e 12 meses); e aplicação do teste de Wald para análise univariada de cada variável independente.

	Sementes com germinação (%)	Sementes sem germinação (%)	Teste de Wald	Valor-P
Variedade				
Laranja	2054 (85,58)	346 (14,42)	-	-
Vermelha	2111 (87,96)	289 (12,04)	2,425	0,0153 [a]
Armazenamento do fruto				
0 dias	1241 (77,56)	359 (22,44)	-	-
7 dias	1402 (87,63)	198 (12,38)	7,413	<0,0001 [a]
14 dias	1522 (95,12)	78 (04,88)	13,248	<0,0001 [a]
Método de extração				
Manual	2065 (86,04)	335(13,96)	-	-
Mecânico	2100 (87,50)	300 (12,50)	1,49	0,1360 [a]
Armazenamento da semente				
3 meses	988 (82,33)	212 (17,67)	-	-
6 meses	1110 (92,50)	90 (7,50)	7,307	<0,0001 [a]
9 meses	1125 (93,75)	75 (6,25)	8,276	<0,0001 [a]
12 meses	942 (78,50)	258 (21,5)	-2,363	0,0181 [a]

[a] Teste de Wald significativo a 25% de significância.
Fonte: O autor.

As variáveis selecionadas para a construção do primeiro modelo múltiplo estão apresentadas na Tabela 2. Observa-se que a variável "Método de extração" não obteve significância estatística (P>0,05) e, portanto, será retirada do modelo nas análises posteriores. Deste modo, constatou-se que o método de extração, seja ele manual ou mecânico, não influencia na germinação das sementes. De acordo com Nascimento et al. (2006), uma das maiores dificuldades no cultivo e colheita de pimentas é devido ao ardume durante sua extração manual, o que dificulta a produtividade e o rendimento. Dessa forma, apesar de não ser significativo, o método mecânico pode ser utilizado sem perda da qualidade das sementes e com redução das dificuldades inerentes à extração manual. As demais variáveis foram significativas neste passo.

Tabela 2 – Análise de regressão logística múltipla para a germinação de sementes de pimenta habanero em função das variáveis: variedade dos frutos (laranja e vermelha), período de armazenamento dos frutos (0, 7 e 14 dias), método de extração das sementes (manual e mecânico) e tempo de armazenamento das sementes (3, 6, 9 e 12 meses), e seus respectivos valores de P obtidos pelo teste de Wald.

	Coef.	Erro padrão	Teste de Wald	Valor-P
Intercepto	0,6771	0,1085	6,239	<0,0001[a]
Variedade				
Laranja	-	-	-	-
Vermelha	0,2283	0,0897	2,545	0,0109[a]
Armazenamento do fruto				
0 dias	-	-	-	-
7 dias	0,7567	0,0994	7,614	<0,0001[a]
14 dias	1,7985	0,1328	13,545	<0,0001[a]
Método de extração				
Manual	-	-	-	-
Mecânico	0,1401	0,0896	1,565	0,1177
Armazenamento da semente				
3 meses	-	-	-	-
6 meses	1,0201	0,1362	7,491	<0,0001[a]
9 meses	1,2214	0,1441	8,478	<0,0001[a]
12 meses	-0,2625	0,1071	-2,451	0,0143[a]

[a] Teste Wald significativo a 5% de significância.
Fonte: O autor.

Após a retirada da variável método de extração notou-se que os coeficientes das demais variáveis continuaram significativos, como pode ser verificado na Tabela 3. Desta forma, o modelo múltiplo foi composto pelas seguintes variáveis: variedade dos frutos, período de armazenamento dos frutos e tempo de armazenamento das sementes, sem as interações entre elas. Este modelo foi denominado **Modelo 1** cujo AIC foi de 3342.

Tabela 3 – Análise de regressão logística múltipla entre a germinação de sementes de pimenta habanero e as variáveis: variedade dos frutos (laranja e vermelha), período de armazenamento dos frutos (0, 7 e 14 dias) e tempo de armazenamento das sementes (3, 6, 9 e 12 meses), e seus respectivos valores P obtidos no teste Z.

	Coef.	Erro padrão	Z	P
Intercepto	0,7462	0,0994	7,510	<0,0001[a]
Variedade (V)				
Laranja	-	-	-	-
Vermelha	0,2281	0,0897	2,544	0,0110[a]
Armazenamento do fruto (F)				
0 dias	-	-	-	-
7 dias	0,7561	0,0993	7,611	<0,0001[a]
14 dias	1,7975	0,1327	13,541	<0,0001[a]
Armazenamento da semente (S)				
3 meses	-	-	-	-
6 meses	1,0195	0,1361	7,489	<0,0001[a]
9 meses	1,2207	0,1440	8,475	<0,0001[a]
12 meses	-0,2623	0,1071	-2,450	0,0143[a]

[a] Teste Z significativo a 5% de significância.
Fonte: O autor.

O próximo passo do procedimento se deu pela investigação de possíveis interações duplas e triplas entre os efeitos principais. **Modelo 2**: modelo de regressão logística múltipla com a interação V*F; **Modelo 3**: modelo de regressão logística múltipla com a interação V*S; **Modelo 4**: modelo de regressão logística múltipla com interação F*S; **Modelo 5**: modelo de regressão logística múltipla com a interação tripla V*F*S.

Na Tabela 4 estão apresentados os resultados do **Modelo 2** com uma interação dupla entre a variedade e o armazenamento do fruto (AIC = 3332). Houve significância para a interação e também foi possível perceber que todas as variáveis foram significativas (P<0,05) com a inclusão desta interação no modelo.

Tabela 4 – Resumo da análise de variância do **Modelo 2** com uma interação dupla entre variedade e armazenamento do fruto dos dados referentes à germinação de sementes de pimenta habanero.

	GL	Deviance	Resid. Deviance	F	P
Variedade (V)	1	6,49	3327,60	6,49	0,0108 [a]
Arm. do Fruto (F)	2	234,67	3334,10	117,34	< 0,0001 [a]
Arm. da Semente (S)	3	182,15	3568,80	60,72	< 0,0001 [a]
V * F	2	13,42	3314,20	6,71	0,0012 [a]

[a] Teste F significativo a 5% de significância.
Fonte: Elaborada pelo autor.

Para o **Modelo 3** com interação entre variedade e armazenamento da semente (Tabela 5) houve significância (P <0,05) para a interação, bem como para as demais variáveis. O AIC deste modelo foi de 3219.

Tabela 5 – Resumo da análise de variância do **Modelo 3** com uma interação dupla entre variedade e armazenamento da semente dos dados referentes à germinação de sementes de pimenta habanero.

	GL	Deviance	Resid. Deviance	F	P
Variedade (V)	1	6,49	3327,60	6,49	0,0108 [a]
Arm. do Fruto (F)	2	225,99	3524,90	112,99	< 0,0001 [a]
Arm. da Semente (S)	3	190,84	3334,10	63,61	< 0,0001 [a]
V * S	3	128.73	3198,90	42,91	< 0,0001 [a]

[a] Teste F significativo a 5% de significância.
Fonte: O autor.

Na Tabela 6 estão apresentados os resultados do **Modelo 4** com interação dupla entre armazenamento do fruto e armazenamento da semente. Houve significância para esta interação. Todas as variáveis foram significativas com a inclusão desta interação. Este modelo obteve AIC = 3307.

Tabela 6 – Resumo da análise de variância do **modelo 4** com uma interação dupla entre armazenamento do fruto e armazenamento da semente referentes à germinação de sementes de pimenta habanero.

	GL	Deviance	Resid. Deviance	F	P
Variedade (V)	1	5,90	3745,00	5,90	0,0151 [a]
Arm. do Fruto (F)	2	235,04	3327,60	60,79	< 0,0001 [a]
Arm. da Semente (S)	3	182,38	3562,60	117,52	< 0,0001 [a]
F * S	6	46,95	3280,60	7,83	< 0,0001 [a]

[a] Teste F significativo a 5% de significância.
Fonte: O autor.

Na Tabela 7 está apresentado um resumo da ANOVA do **Modelo 5** com interação tripla (variedade do fruto * armazenamento do fruto * armazenamento da semente) e que pode ser verificado a significância estatística (P<0,05) para esta interação. Este modelo obteve AIC = 3140.

Tabela 7 – Resumo da análise de variância do **Modelo 5** com interação tripla entre variedade do fruto, armazenamento do fruto e armazenamento da semente referentes à germinação de sementes de pimenta habanero.

	GL	Deviance	Resid. Deviance	F	P
Variedade (V)	1	6,49	3327,60	6,49	0,0108 [a]
Arm. do Fruto (F)	2	225,99	3524,90	112,99	< 0,0001 [a]
Arm. da Semente (S)	3	190,84	3334,10	63,61	< 0,0001 [a]
V * F	2	31,75	3114,30	15,87	< 0,0001 [a]
V * S	3	134,63	3146,00	44,88	< 0,0001 [a]
F * S	6	46,95	3280,60	7,83	< 0,0001 [a]
V * F * S	6	22,35	3091,90	3,73	0,0010 [a]

[a] Teste F significativo a 5% de significância.
Fonte: O autor.

Após serem verificadas todas as interações foi estimada também a capacidade de discriminação por meio da área sob a curva ROC com seus respectivos intervalos de confiança (Tabela 8). Sabe-se que são desejáveis valores menores de AIC, enquanto deseja-se valores maiores para a área abaixo da curva ROC (AUC). Foi observado que os modelos com menores valores de AIC também são os melhores segundo a AUC. Hosmer et al. (2013) afirmam que é

considerada uma discriminação aceitável um AUC superior a 0,70. Portanto, todos os modelos apresentaram poder de discriminação aceitável.

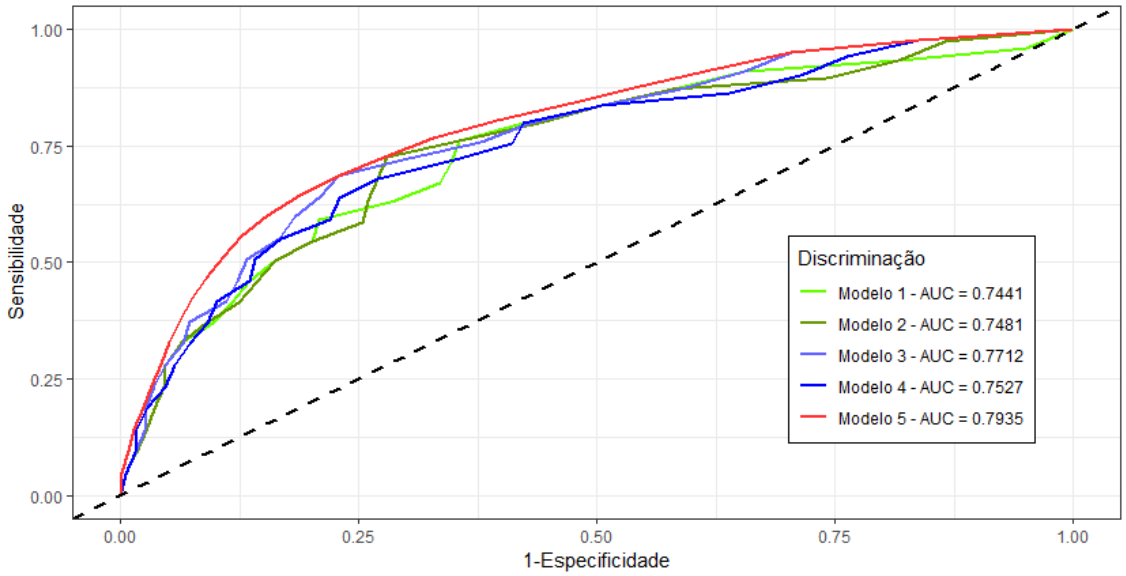
Tabela 8 – Critério de Akaike (AIC) e área sob a curva ROC dos diferentes modelos de regressão logística avaliados para os dados referentes à germinação de sementes de pimenta habanero armazenadas nos frutos em diferentes tempos e submetidas a diferentes períodos de armazenamento pós-extração.

	AIC	Área sob a Curva ROC (IC 95%)
Modelo 1	3342	0,7441 (0,7245-0,7637)
Modelo 2	3332	0,7481 (0,7285-0,7678)
Modelo 3	3219	0,7712 (0,7522-0,7902)
Modelo 4	3307	0,7527 (0,7364-0,7720)
Modelo 5	3140	0,7935 (0,7757-0,8113)

Fonte: O autor.

Levando em consideração o menor valor de AIC (3140) e maior valor de AUC (0,7935), optou-se por selecionar o modelo 5, considerando-o mais adequado para explicar a ocorrência de germinação, ou seja, maior capacidade de classificar corretamente sementes que germinam e as que não germinam. Para verificar o superior desempenho da interação contida no modelo 5, foi realizada sua comparação com os demais modelos. A Figura 1 mostra a curva ROC dos modelos avaliados. Percebeu-se que o modelo 5 apresenta o melhor poder de discriminação em todos os pontos de corte estabelecidos. É possível perceber ainda que os demais modelos se cruzam ou se sobrepõem entre si e em determinados intervalos de 1-especificidade alguns tem maiores valores de sensibilidade, ou vice versa. Enquanto o modelo 5 se mantém superior em todo o espaço ROC, evidenciando a importância da interação variedade * armazenamento do fruto * armazenamento da semente para explicar a germinação.

Figura 1 – Curvas ROC comparando os modelos de regressão logística múltipla sem interação (modelo 1), com interação V * F (modelo 2), com interação V * S (modelo3), com interação F * S (modelo 4) e com interação V * F * S (modelo 5).



Fonte: O Autor.

Na Tabela 9 estão apresentados os coeficientes estimados do modelo selecionado, bem como o erro padrão dos coeficientes, a estatística de teste, a significância das variáveis e a razão de chance com seus respectivos intervalos de confianças de 95%.

No ajuste do modelo de regressão logística utilizou-se como função de ligação a função *logit*. Assim, para que os parâmetros ajustados pudessem ser interpretados de uma forma mais direta foi necessário que eles fossem reescritos em sua escala natural, por meio da função inversa, neste caso a função *exp*, conforme resultados apresentados na Tabela 9. A função inversa *exp* coincidi com as razões de chances estimadas (OR).

Tabela 9 – Coeficientes, valores de erro padrão, valor de P do teste Z, *Odds Ratio* e intervalo de confiança de 95% para do modelo de regressão logística obtido a partir da retirada da variável "método de extração" e inclusão da interação entre variedade, armazenamento do fruto e armazenamento da semente.

Interação V*F*S		Coef.	Erro padrão	Teste de Wald	*OR* (IC 95%)
Laranja					
0 dias	3 meses	-	-	-	-
	6 meses	3,3770	0,4099	< 0,0001	29,28 (13,11-65,38)
	9 meses	1,6820	0,2373	< 0,0001	5,37 (3,375-8,56)
	12 meses	1,7560	0,2412	< 0,0001	5,79 (3,61-9,29)
7 dias	3 meses	-	-	-	-
	6 meses	1,4270	0,3447	< 0,0001	4,17 (2,12-8,18)
	9 meses	0,7658	0,2850	0,0072	2,15 (1,23-3,76)
	12 meses	-0,1446	0,2407	0,5480	0,87 (0,54-1,39)
14 dias	3 meses	-	-	-	-
	6 meses	0,7246	0,5104	0,1557	2,06 (0,76-5,61)
	9 meses	0,7246	0,5104	0,1557	2,06 (0,76-5,61)
	12 meses	-0,8056	0,3666	0,0280	0,45 (0,22-0,92)
Vermelha					
0 dias	3 meses	-	-	-	-
	6 meses	1,4185	0,7866	0,0713	4,13 (0,88-19,30)
	9 meses	2,7390	0,5315	< 0,0001	15,47 (5,46-43,81)
	12 meses	-0,8504	0,2188	0,0001	0,43 (0,28-0,66)
7 dias	3 meses	-	-	-	-
	6 meses	-0,5317	0,5264	0,3125	0,59 (0,21-1,65)
	9 meses	-1,8235	0,8739	0,0369	0,16 (0,03-0,90)
	12 meses	-2,4560	0,4444	< 0,0001	0,09 (0,04-0,21)
14 dias	3 meses	-	-	-	-
	6 meses	-1,9770	1,0730	0,0655	0,14 (0,02-1,14)
	9 meses	-1,4010	1,1220	0,2116	0,25 (0,03-2,22)
	12 meses	-2,9170	1,0330	0,0048	0,05 (0,01-0,41)

Fonte: O autor.

Analisando a interação observou-se que as sementes da variedade laranja cuja extração foi realizada logo após a colheita dos frutos foram avaliadas em diferentes meses de armazenamento tendo como referência o período de 3 meses. Aos 6 meses foi verificado a maior chance de germinação, sendo ela 29,28 vezes maior que aos 3 meses de armazenamento das sementes. Já as sementes desta mesma variedade que foram extraídas 7 dias após a colheita dos frutos tiveram comportamento decrescente na chance de germinação conforme o aumento do tempo de armazenamento. Para estas sementes o melhor tempo de armazenamento pós-colheita foi de 6 meses, período em que a chance de germinação foi 4,17 vezes maior em relação ao período de armazenamento durante 3 meses. Por sua vez, as sementes da variedade laranja armazenadas durante 14 dias nos frutos obtiveram o maior valor de *OR* (2,06) nos 6 e 9 meses

após a extração, de modo que o armazenamento ao longo desses meses em relação ao armazenamento pelo período de 3 meses aponta 2,06 mais chances de germinação das sementes. Tudo isso evidencia que com o aumento do tempo de armazenamento das sementes pode haver mais chance de perda na qualidade de germinação. Freitas et al. (2008) recomendam um período de repouso pós-colheita dos frutos de pimenta habanero entre 7 e 20 dias antes da extração, para que as sementes completem sua maturação ainda dentro dos frutos.

No que diz respeito à variedade vermelha, o comportamento se deu de maneira diferente com aumento do tempo de armazenamento após a extração nos diferentes períodos em que as sementes ficaram armazenadas nos frutos. Observou-se que quando as sementes foram extraídas logo após a colheita a maior chance (15,47 a mais) de ocorrência de germinação em relação as sementes que ficaram armazenadas durante 3 meses deu-se aos 9 meses de armazenamento. Além disso, foi verificado que as sementes desta variedade quando armazenadas no período de 7 dias dentro do fruto apresentaram uma diminuição da chance de germinação no decorrer dos meses em que foram submetidas ao armazenamento pós-colheita. Verificou-se que em todos os períodos a chance de germinação foi menor em relação aos 3 meses que foram tomados como referência, evidenciando que o tempo ideal de armazenamento após a extração seria de 3 meses. Este desempenho foi semelhante ao das sementes que foram armazenadas no fruto durante 14 dias após a colheita, no entanto a chance de germinação foi bem menor em relação aos 3 meses de armazenamento para todos os demais períodos.

4. Conclusões

A metodologia de curva ROC foi eficiente para avaliar a capacidade preditiva de modelos de germinação de sementes de pimenta habanero. O método de extração das sementes não interfere na capacidade germinativa, de modo que a extração mecânica pode ser utilizada sem perda na qualidade germinativa. As variáveis explicativas que afetam a germinação, foram, as diferentes colorações, a permanência em repouso nos frutos durante diferentes estádios de tempo e o armazenamento após a extração em determinados períodos, estes foram selecionadas para compor o modelo com interação tripla, o qual, obteve curva ROC com maior poder discriminativo.

5. Referências

BARBEDO, C. J.; BARBEDO, A. S. C.; NAKAGAWA, J. et al. **Efeito da idade e do repouso pós-colheita de frutos de pepino na semente armazenada**. Pesquisa Agropecuária Brasileira, v. 34, n. 5, p. 839–847, maio 1999. DOI: 10.1590/S0100-204X1999000500015. Disponível em:

http://www.scielo.br/scielo.php?script=sci_arttext
&pid=S0100-204X1999000500015&lng=pt&tlng=pt. Acesso em: 4 jan. 2020.

Brasil. Ministério da Agricultura e Reforma Agrária. Secretaria Nacional de defesa Agropecuária. Regras para análise de sementes. Brasília, 2009. 395p.

CASTRO, M. M.; GODOY, A. R.; CARDOSO, A. I. I. **Qualidade de sementes de quiabeiro em função da idade e do repouso pós-colheita dos frutos**. Ciência e Agrotecnologia, v. 32, n. 5, p. 1491–1495, out. 2008. DOI: 10.1590/S1413-70542008000500020. Disponível em: http://www.scielo.br/scielo.php?script=sci_arttext
&pid=S1413-70542008000500020&lng=pt&tlng=pt. Acesso em: 4 jan. 2020.

FREITAS R. A.; NASCIMENTO W.M.; CARVALHO S. I. C. **Produção de sementes**. In: RIBEIRO C. S. C.; LOPES C. A.; CARVALHO S. I. C.; HENZ G. P.; REIFSCHNEIDER (Org.). Pimentas Capsicum. 1 ed. Brasília: Embrapa Hortaliças, v. 1, p. 173-187, 2008.

HERNÁNDEZ-VERDUGO, S.; OYAMA, K.; VÁZQUEZ-YANES, C. **Differentiation in seed germination among populations of Capsicum annuum along a latitudinal gradient in Mexico**. Plant Ecology, v. 155, n. 2, p. 245–257, 2001. DOI: 10.1023/A:1013234100003. Disponível em: https://doi.org/10.1023/A:1013234100003. Acesso em 6 jan. 2020.

HOSMER J. R., D. W.; LEMESHOW, S.; STURDIVANT, R. X. **Applied Logistic Regression**. 3. ed. Hoboken, John Wiley & Sons, Inc., 2013.

NASCIMENTO, W. M.; DIAS, D. C. F. S.; FREITAS, R. A. **Produção de sementes de pimentas**. Informe Agropecuário: cultivo da pimenta, v. 27, n. 235, p. 30–39, 2006.

PROVOST, F.; FAWCET, T. **Analysis and Visualization of Classifier Performance: Comparison under Imprecise Class and Cost Distributions**. Proceedings of the Third International Conference on Knowledge Discovery and Data Mining, AAAI Press, p. 43–48, 1997.

QUEIROZ, L. A. F.; PINHO, É. V. de R. Von; OLIVEIRA, J. A. et al. **Época de colheita e secagem na qualidade de sementes de pimenta Habanero Yellow**. Revista Brasileira de Sementes, v. 33, n. 3, p. 472–481, 2011. DOI: 10.1590/S0101-31222011000300010. Disponível em: http://www.scielo.br/scielo.php?script=sci_arttext
&pid=S0101-31222011000300010&lng=pt&tlng=pt. Acesso em: 4 jan. 2020.

RIBEIRO, C. S. C.; LOPES, C. A.; CARVALHO, S. I. C. et al. **Pimentas Capsicum**. Embrapa Hortaliças, v. 1, p. 157–171, 2008.

SAGARPA, D. F. **Regla para la calificación de semilla de chile (Capsicum spp.)**. México, Servicio Nacional de Inspección y Certificación de Semillas, 2014.

SANCHEZ, V. M.; SUNDSTROM, F. J.; MCCLURE, G. N. et al. **Fruit maturity, storage and postharvest maturation treatments affect bell pepper (Capsicum annuum L.) seed quality**. Scientia Horticulturae, v. 54, n. 3, p. 191–201, 1 jun. 1993. DOI: 10.1016/0304-4238(93)90087-7. Disponível em: https://www.sciencedirect.com/
science/article/abs/pii/0304423893900877. Acesso em: 4 jan. 2020.

SILVA, L. J.; MEDEIROS, A. D.; OLIVEIRA, A. M. S. **SeedCalc, a new automated R software tool for germination and seedling length data processing**. Journal of Seed Science, v. 41, n. 2, 2019 - ISSN 2317-1537 versão online. Disponível em: https://doi.org/10.1590/2317-1545v42n2217267. Acesso em: 9 dez. 2019.

SING, T., SANDER, O., BEERENWINKEL, N. and LENGAUER, T. (2007). **ROCR**: Visualizing the performance of scoring classifiers. R package version 1.0-2

TEAM, R C. **R: A language and environment for statistical computing**. Vienna, R Foundation for Statistical Computing. , 2019.

VIDIGAL, D. de S.; DIAS, D. C. F. dos S.; NAVEIRA, D. dos S. P. C. et al. **Qualidade fisiológica de sementes de tomate em função da idade e do armazenamento pós-colheita dos frutos**. Revista Brasileira de Sementes, v. 28, n. 3, p. 87–93, dez. 2006. DOI: 10.1590/S0101-31222006000300013. Disponível em: http://www.scielo.br/scielo.php?script=sci_arttext&pid=S0101-31222006000300013&lng=pt&tlng=pt. Acesso em: 4 jan. 2020.

WICKHAM, H. **ggplot2**: elegant graphics for data analysis. Springer, New York, 2009.

WICKHAM, H. **tidyverse**: Easily Install and Load the 'Tidyverse'. R package version 1.2.1., 2017.

CAPÍTULO 2

Avaliação de modelos de predição genômica baseada em curva ROC

Resumo: A metodologia de curva ROC tem sido uma ferramenta importante para avaliação de modelos em diversas áreas, como na medicina, psicologia e economia. No entanto, ainda não foi empregada em seleção genômica. Neste trabalho, objetivou-se utilizar curvas ROC para a comparação dos modelos RR-BLUP, Bayes Cπ e BLASSO ajustados a dados de resistência de arroz *Oryza sativa* à doença brusone. A área abaixo da curva ROC (AUC) se mostrou equivalente aos índices usuais (taxa de erro na validação, correlação de Spearman e viés) para a avaliação de todos os modelos, dos quais o RR-BLUP e Bayes Cπ se mostraram mais acurados (AUC de 0,823 e 0,822 para RR-BLUP e Bayes Cπ, respectivamente e de 0,808 para o modelo BLASSO). O modelo RR-BLUP se mostrou melhor em relação ao tempo de execução (4h52min, 6h1min, 6h25min para o RR-BLUP, Bayes Cπ e BLASSO, respectivamente). Além disso, a representação gráfica das curvas ROC apresentou vantagens em relação aos valores pontuais dos índices, uma vez que permitiu a avaliação dos modelos em diferentes valores de especificidade. Abaixo de 0,25 de 1-especificidade os modelos RR-BLUP e Bayes Cπ foram mais sensíveis que o BLASSO, ao passo que acima deste valor todos os modelos tiveram igual desempenho (curvas ROC sobrepostas). A análise ROC somada ao tempo de execução indica que o melhor modelo para a seleção genômica para resistência de arroz à brusone foi o RR-BLUP. A metodologia de curva ROC pode ser utilizada como boa alternativa para a seleção de modelos de predição genômica, apresentando vantagens em relação aos métodos tradicionais.

Palavras-chave: Modelagem Estatística. Seleção Genômica Ampla. Análise ROC. Acurácia. Brusone do arroz.

1. Introdução

A Seleção Genômica Ampla (Genome Wide Selection - GWS) baseia-se na predição conjunta dos efeitos genéticos de inúmeros marcadores genéticos espalhados em todo o genoma de um organismo, de modo a detectar os efeitos de todos os *locus*, de pequenos e grandes efeitos, explicar grande parte da variação genética de um caráter quantitativo e distinguir indivíduos geneticamente superiores de uma população (MEUWISSEN et al., 2001).

O conhecimento da base genética das variações fisiológicas, de desenvolvimento e morfológicas do arroz (*Oryza sativa*) é essencial para melhorar a qualidade, o valor nutricional, a confiabilidade e a sustentabilidade deste suprimento alimentar mundial. O melhoramento genético deste cereal auxilia na redução da fome em diversas partes do mundo por ser a principal fonte de alimentação de pessoas de vários países.

A identificação da resistência desta espécie às doenças por meio de GWS, implica na estimação da probabilidade desta planta ser resistente ou suscetível. Essa característica qualitativa binária permite avaliar vários métodos de estimação, no que diz respeito à acurácia, além de avaliar a sensibilidade e especificidade na predição genômica. Uma boa ferramenta que pode ser utilizada para avaliar o desempenho de modelos de predição para duas classes (uma suscetível e outra resistente), classificando-as o mais corretamente possível, é a análise da curva ROC (*Receiver operator characteristic*).

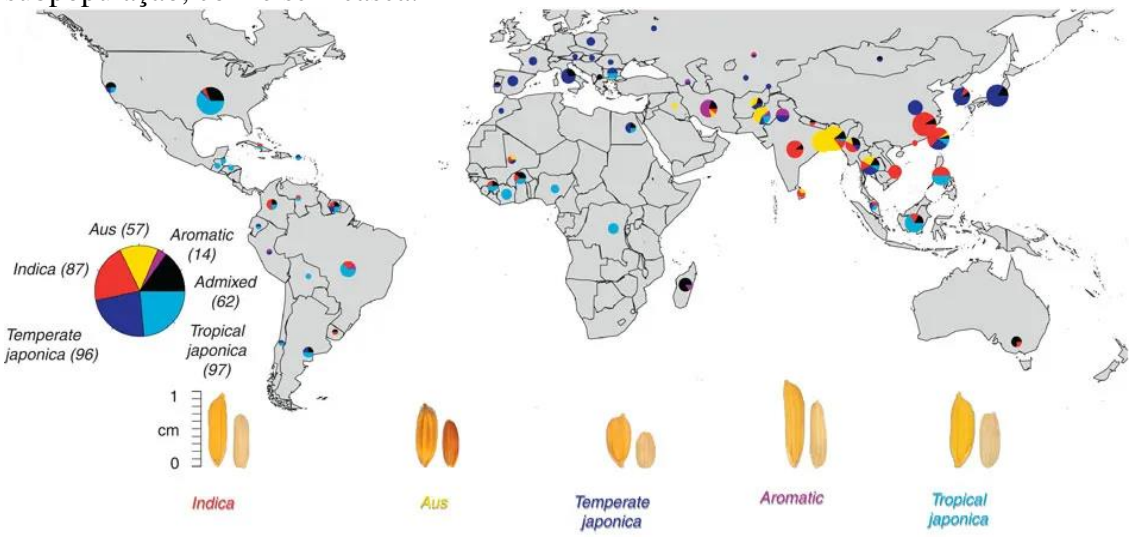
Neste contexto, objetivou-se estudar medidas para avaliação de acurácia na predição genômica, bem como propor a metodologia de curva ROC para a avaliação do desempenho dos métodos de estimação bayesianos RR-BLUP, Bayes Cπ e BLASSO na identificação da resistência do arroz à brusone.

2. Material e métodos

Para a condução do estudo foi utilizada uma população com 413 plantas de arroz (*Oryza sativa*), incluindo variedades locais, coletadas em 82 países (Figura 1), caracterizando todas as principais regiões produtoras de arroz no mundo (RESENDE et al., 2010). O banco de dados acerca das 413 plantas possuía informações referentes à genotipagem e fenotipagem, permitindo o estudo desses indivíduos. 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). Na Figura 1 pode-se observar a

distribuição das subpopulações amostradas em cada país e as cores em cada gráfico de pizza se referem à porcentagem de amostras em cada subpopulação.

Figura 1 – Distribuição de subpopulações nas 413 amostras de *Oryza sativa* (gráfico pizza maior) e distribuição das subpopulações em cada país em que foi realizada a amostragem (gráficos pizza menores. Países grandes foram divididos em regiões importantes de cultivo de arroz). Os gráficos são de diferentes tamanhos, que são proporcionais à porcentagem de amostragem em cada local. No inferior da figura estão representadas as sementes de cada subpopulação, com e sem casca.



Fonte: Zhao et al. (2011).

O banco de dados genotípicos possui um painel com 44.100 marcadores SNPs, que, após o controle de qualidade com *call rate* > 70% e MAF > 1% (MAF- *Minor Allele Frequency*) (RESENDE et al., 2014) levou a utilização de 36.901 SNPs. Após o *call rate*, os marcadores que tinham dados faltantes (4,33% do total de genótipos) foram imputados de acordo com a frequência alélica de cada marcador (WENG et al. 2012).

A avaliação fenotípica do arroz foi feita em Stuttgart (Arkansas, EUA) durante a os meses de maio a 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. No total, o conjunto de dados fenotípicos dispõe de 34 características relacionadas à morfologia das plantas, qualidade dos grãos, desenvolvimento das plantas, qualidade nutricional e ao grau de suscetibilidade do genótipo à doença.

Para este trabalho foi utilizada a parte do banco de dados referente à doença na folha do arroz que é causada pelo o fungo *Pyricularia oryzae*. A gravidade da doença foi inicialmente pontuada 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 (classe 1) e plantas pertencentes às classes 3, 4, 5, 6, 7, 8 e 9 foram classificadas como suscetíveis (classe 0). Além do mais, foram excluídas 28 plantas que não foram avaliadas quanto a resistência, resultando assim, em um cenário com 385 plantas, das quais 282 eram resistentes e 103 eram suscetíveis. Mais informações sobre a fenotipagem, genotipagem, controle de qualidade dos dados podem ser consultadas em Zhao et al. (2011).

A modelagem da resistência à brusone ajustou 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 é 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.

Essa função de ligação *probit* estima o valor de probabilidade de uma planta de arroz pertencer a cada uma das categorias da resistência à brusone. Dessa forma, realizou-se a reclassificação de cada planta considerando a classe de maior probabilidade. Ou seja, a planta foi classificada como resistente se $P[Y = 1|X] = P[\ell \leq \tau] \geq 0{,}5$, caso contrário ela foi suscetível.

Devido à necessidade da escolha de efeitos que melhor expliquem a variação genética da população, foram aplicadas três metodologias bayesianas (RR-BLUP Bayes, Bayes Cμ e BLASSO) a serem utilizadas na predição e seleção de indivíduos, utilizando os 36.901 marcadores SNPs.

No método RR-BLUP Bayes (*Bayesian Ridge Regression*) é considerada pressuposição de homogeneidade das variâncias dos SNPs. Assim, tem-se apenas um valor assumido para σ^2. Os parâmetros de efeito de SNPs (β_i), variância dos marcadores (σ^2) e variância aditiva (σ_u^2) seguem respectivamente: $\beta_i|\sigma^2 \sim N(0, \sigma^2), \sigma^2 \sim \chi^{-2}(\mathrm{v}, S^2)$ e $\sigma_u^2 = 2\sigma^2 \sum_{i=1}^{m} p_i\,(1 - p_i)$, em

que v representa os graus de liberdade, S^2 é o parâmetro da escala de distribuição e p_i denota as frequências alélicas. Meuwissen et al. (2001) consideram os valores 4,012 ou 4,2 para v e 0,002 e 0,0429 para S^2.

O BLASSO (*Bayesian Least Absolute Shrinkage and Selection Operator*), foi proposto por De Los Campos et al. (2009) a partir de uma interpretação bayesiana baseada no LASSO (TIBSHIRANI, 1996). Neste método, o modelo *Threshold* é descrito como $\ell = 1\mu + Zm + e$, em que m é o vetor de efeitos aditivos dos marcadores com matriz de incidência Z que foi reparametrizada conforme Vitezica et al. (2013), a fim de se enquadrar na teoria de genética quantitativa. Assim:

$$Z = \begin{cases} se\ AA, então\ 2 - 2p_j, \\ se\ Aa, então\ 1 - 2p_j, \\ se\ aa, então\ 0 - 2p_j. \end{cases}$$

em que p_j é a frequência alélica do marcador j de AA, Aa e aa, que correspondem ao genótipo da planta i no marcador j, que pode ser homozigoto dominante, heterozigoto ou homozigoto recessivo, respectivamente.

As distribuições *a priori* dada aos parâmetros dos efeitos de marcadores e componentes de variância deste modelo é dada por Pérez e De Los Campos (2014):

$$m_j|\sigma_\epsilon^2, \tau_j, \lambda^2 \sim N\left(0, \tau_j^2 \cdot \sigma_\epsilon^2\right); \tau_j \sim Exp\left(\frac{\lambda^2}{2}\right); \lambda^2 \sim gama(r, s)$$

Por sua vez, o método Bayes Cπ é semelhante ao método BLASSO, mas se difere na distribuição *a priori* dada ao método, que segundo Pérez e De Los Campos (2014) foi implementada como:

$$m_j|\sigma_m^2, \pi \sim \left[\pi \cdot N(0, \sigma_m^2) + (1 - \pi) \cdot \left(m_j = 0\right)\right]; \sigma_m^2 \sim \chi^{-2}(df_m, S_m); \pi \sim beta(p_0, \pi_0)$$

Essa diferença de distribuições *a priori* propicia a seleção de marcadores, uma vez que conduz uma quantidade (1-π) de marcadores a zero.

Visando avaliar a qualidade do ajuste, e para que os efeitos dos marcadores não fossem superestimados devido à estimação e validação na mesma amostra (CRUZ et al., 2013), uma técnica de validação cruzada foi adotada. O método de validação cruzada consistiu em dividir

a população em k grupos (k = 5). Em seguida, um grupo foi utilizado como população de validação e k-1 grupos foram utilizados como população de estimação. Na população de estimação, os efeitos dos marcadores foram estimados e utilizados na população de validação a fim de obter as estimativas dos valores genéticos genômicos (GEBVs). Esse procedimento foi executado até que cada um dos k grupos fosse utilizado uma vez como população de validação.

Dessa forma, os modelos foram validados pelo procedimento de validação cruzada via *Jacknife* (5-fold). Como haviam 385 plantas, a melhor maneira de fazer essa validação de forma a obter grupos com o mesmo número de indivíduos foi dividindo-os em 5 grupos de 77 plantas.

Ao longo de toda a fase de experimentação, que envolveu os passos descritos acima, a avaliação dos classificadores ocorreu a partir de medidas como: taxa de erro na validação, correlação de Spearman, viés, tempo de execução computacional e área abaixo da curva ROC (*Area Under the Curve* – AUC).

A taxa de erro na validação cruzada consistiu em somar os indivíduos que não foram reclassificados corretamente em cada população de validação do procedimento de validação cruzada e dividir pela quantidade de populações. A classificação incorreta dos indivíduos se deu a partir da observação dos valores preditos ($\hat{y}_i$) e observados (y_i), onde a predição foi errônea se $y_i \neq \hat{y}_i$.

Após a validação cruzada os GEBVs também foram correlacionados com os valores observados utilizando o coeficiente de correlação de Spearman. Este coeficiente é uma derivação do coeficiente de correlação de Pearson (medida importante na seleção genômica devido ao fato de informar a capacidade preditiva de um modelo), no entanto, não necessita da pressuposição de normalidade das variáveis correlacionadas, pois, seu cálculo é baseado em postos e é calculado da seguinte forma:

$$\rho = 1 - \frac{6\sum_{i=1}^{n} d_i^2}{n^3 - n}$$

onde d_i é a diferença de postos entre $\hat{y}_i$ e y_i e n é o número de pares $(\hat{y}_i, y_i)$. Se o posto de $\hat{y}_i$ for igual ao posto de y_i, então o numerador da equação acima será zero e, consequentemente, ρ será igual a 1 indicando uma máxima correlação entre as duas classificações.

As estimativas de viés foram obtidas a partir do coeficiente de regressão das variáveis respostas observadas em função dos GEBVs. O viés é uma medida de tendência que indica se as estimativas dos parâmetros estão sendo superestimadas, subestimadas ou não possui viés.

O tempo computacional foi obtido com o intuito de saber qual dos métodos possuem maior eficiência computacional combinado ao melhor valor de acurácia. O computador utilizado possui um processador Intel ® Core™ i7-4510U 2.60 GHz com 8,00 GB de memória RAM.

Por fim, os métodos também foram comparados por meio de análise ROC. A análise ROC foi realizada observando qual curva aparenta alcançar um melhor resultado, ou seja, a mais afastada da linha $x = y$.

As estimativas dos parâmetros foram obtidas utilizando o pacote BGLR (PÉREZ, DE LOS CAMPOS, 2014), e a análise ROC foi realizada pelos pacotes GGplot2 e pROC do software R (R CORE TEAM, 2019). Nesta análise, foram realizadas 100.000 iterações com *burn-in* de 20.000 para eliminar o período de aquecimento da cadeia e o *thin* de 10 para remover o efeito da autocorrelação. Os parâmetros foram preditos por meio de processo iterativo MCMC e a convergência das cadeias de Markov foram avaliadas por meio do critério de Geweke (GEWEKE, 1991). Este critério indica se a diferença entre a média das nA (10% das n iteração) primeiras iterações com a média das nB (50% das n iterações) últimas iterações seguem distribuição normal com média zero. Mais informações acerca dos métodos utilizados podem ser consultadas em Pérez, De Los Campos (2014).

3. Resultados e discussão

Na Tabela 1 estão apresentados os resultados das medidas de ajuste dos modelos cujos parâmetros foram estimados pelos métodos BRR, Bayes Cπ e BLASSO. A escolha do modelo cujos efeitos estão associados a resistência à doença brusone de maneira mais eficiente foi obtida de acordo com os valores mais apropriados para cada medida de ajuste calculada.

Tabela 1 – Valores de taxa de erro na validação, correlação de Spearman, viés, tempo de execução da análise e área abaixo da curva ROC (AUC) para os diferentes modelos testados.

Medida de ajuste	RR-BLUP	Bayes Cπ	BLASSO
Taxa de erro na validação	0,218	0,218	0,221
Correlação de Spearman	0,476	0,478	0,432
Viés	0,495	0,490	0,471
Tempo de Execução	4h52min	6h1min	6h25min
AUC	0,823	0,822	0,808

Fonte: O autor.

O modelo BLASSO obteve o maior valor de taxa de erro na validação cruzada, que foi igual a 0,221 (22,1%). Enquanto os modelos RR-BLUP e Bayes Cπ obtiveram valores de taxa de erro na validação cruzada iguais (21,8%) e muito próximos ao obtido pelo BLASSO. Quanto mais próxima de zero for a taxa de erro de validação, melhor será a acurácia do modelo. Mas, para que essa medida tenha uma interpretação igual à acurácia na seleção genômica, pode ser calculada a taxa de acerto, sendo dada por: Taxa de acerto = 1 - taxa de erro. Desta forma, os valores da taxa de acerto no RR-BLUP e Bayes Cπ foram 0,782 (78,2%) e no método BLASSO de 0,779 (ou 77,9%). Assim, por meio da análise desta medida o método BLASSO obteve menor acurácia em relação aos demais modelos. Biscarini et al. (2014) também utilizaram a taxa de erro na validação cruzada para avaliar o vigor da raiz de beterraba açucareira com dois níveis (alto ou baixo) utilizando o modelo *threshold* sob o método G-BLUP. A taxa de erro na validação cruzada encontrada por estes autores foi de 0,073%.

A estimativa da correlação de Spearman entre os valores genéticos reais e preditos pelo modelo RR-Blup foi semelhante à obtida pelo modelo Bayes Cπ. Para estes modelos foram encontrados valores de correlação de Spearman de aproximadamente 0,48. Entretanto, em comparação com a estimativa do modelo BLASSO, houve uma diminuição na correlação de Spearman. Para este último modelo foi encontrado valor de correlação de 0,432, indicando que há adequabilidade (em termos de comparação de métodos) com a taxa de erro na validação, tendo em vista que quanto mais próximo de 1, melhor será a correlação dos valores preditos pelos modelos com os valores genéticos reais.

Os valores do viés das estimativas em cada modelo foram obtidos em cada grupo da população de validação no procedimento de validação cruzada e posteriormente, foi obtida a média das populações. O viés de um modelo indica sua tendenciosidade e para que um modelo não seja tendencioso esse valor deve ser igual a 1. Quando o valor de viés é maior que 1, há indício de que o modelo apresenta subestimação à variável resposta e, quando for menor que 1, tem-se evidencia que o modelo está superestimando a variável resposta. Foi possível observar que todas os modelos superestimam a resistência a doença brusone. Entretanto, nos modelos RR-BLUP e Bayes Cπ, o viés foi mais próximo de 1, indicando que, quando a resistência à brusone foi avaliada pelo modelo BLASSO, as estimativas produzidas foram mais tendenciosas.

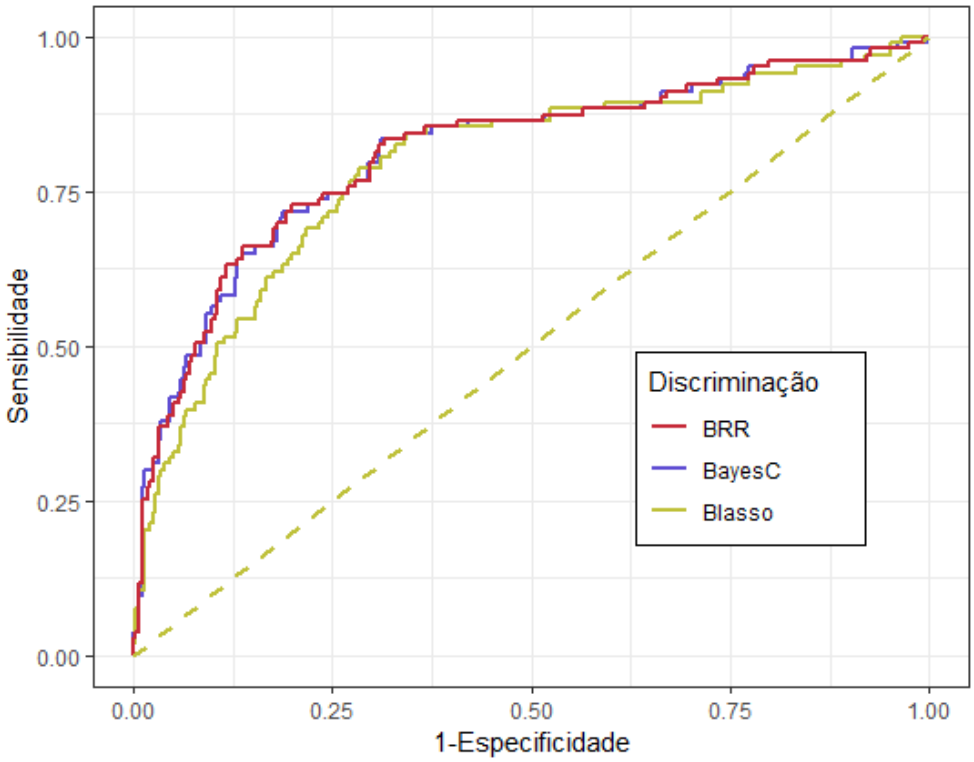
Por sua vez, o tempo de execução refere-se ao tempo gasto para os métodos estimarem os parâmetros do modelo. Em relação ao tempo de execução, foi obtido que RR-BLUP (4h52min) < Bayes Cπ (6h1min) < BLASSO (6h25min). Essa diferença no tempo de execução

é explicada pelo número de parâmetros que foram estimados em cada modelo. Como os métodos Bayes Cπ e BLASSO são a nível de marcador, o vetor de parâmetros a ser estimado foi $\boldsymbol{m}$ (36.901×1) com 36.901 parâmetros de efeitos de marcadores mais os parâmetros de variância. No entanto, o RR-BLUP é um método que pressupõe homogeneidade das variâncias, ao contrário dos outros dois métodos que supõem uma variância para cada marcador, assim, somente os parâmetros dos efeitos dos marcadores são estimados.

Os valores de AUC, assim como a taxa de erro, são variáveis aleatórias estimadas por validação cruzada. Foi possível observar por meio da AUC que todos os modelos têm elevado poder de discriminação (>0,80) (Hosmer et al. 2013). Os modelos RR-BLUP e Bayes Cπ apresentam melhor performance (AUC 0,823 e 0,822, respectivamente) quando comparado ao BLASSO (AUC = 0,808).

De acordo com Ling et al. (2003), a AUC tem menos deficiências do que a taxa de erro. Mas sempre que possível, é recomendável plotar e analisar a curva dos modelos (Figura 2).

Figura 2 – Curvas ROC comparando o desempenho dos modelos BRR – RR-BLUP, Bayes Cπ e BLASSO para predizer a resistência à doença brusone do arroz (*Aryza sativa*).



Fonte: O Autor.

Por meio das curvas ROC apresentadas na Figura 2, os modelos podem ser visualizados quanto à sua performance e podem ser avaliados de maneira mais segura, ao invés de serem

simplesmente comparados a partir de uma única medida. É possível perceber visualmente com o nível de 1-especificidade abaixo de 0,25 o modelo derivado do método BLASSO tem pior desempenho em relação aos modelos derivados dos métodos RR-BLUP e Bayes Cπ, pois apresenta a curva mais próxima da diagonal principal. Isto mostra que em altos valores de especificidade o modelo BLASSO apresenta valores mais baixos de sensibilidade.

Em relação aos outros dois métodos, percebe-se que com o nível de 1-especificidade acima de 0,25 os níveis de sensibilidade se igualam, visto que há sobreposição das curvas. Este comportamento mostra de forma mais explícita quanto aos níveis de sensibilidade e 1-especificidade dos três modelos e condiz com as demais medidas avaliadas. Trazendo benefícios quanto ao ajuste dos modelos.

4. Conclusões

A área abaixo da curva ROC se mostrou equivalente às medidas de ajuste usuais (taxa de erro na validação, correlação de Spearman e viés) para avaliar a acurácia dos modelos preditos. Os modelos RR-BLUP e Bayes Cπ foram mais acurados para a predição de resistência de arroz à brusone. Além disso, o modelo RR-BLUP se mostrou melhor em relação ao tempo de execução.

A representação gráfica das curvas ROC de cada modelo se mostrou uma boa ferramenta para a visualização do desempenho dos modelos. Pela análise gráfica foi possível perceber que o modelo BLASSO obteve menor desempenho que os demais modelos em altos níveis de especificidade (<0,75). Em contrapartida, para valores de especificidade abaixo de 0,75 os modelos apresentaram valores de sensibilidade similares.

5. Referências

BISCARINI, F.; STEVANATO, P.; BROCCANELLO, C. et al. **Genome-enabled predictions for binomial traits in sugar beet populations**. BMC Genetics, v. 15, n. 1, p. 87, 2014. DOI: 10.1186/1471-2156-15-87. Disponível em: https://doi.org/10.1186/1471-2156-15-87. Acesso em 28 nov. 2019.

CRUZ, C. D.; SALGADO, C. S.; BHERING, L. L. **Genômica aplicada**. 1. ed. Visconde de Rio Branco, MG: Suprema Gráfica Editora, v. 1. 424p. 2013.

DE LOS CAMPOS, G.; NAYA, H.; GIANOLA, D. et al. **Predicting Quantitative Traits With Regression Models for Dense Molecular Markers and Pedigree**. Genetics, v. 182, n. 1, p. 375 LP – 385, 2009. DOI: 10.1534/genetics.109.101501. Disponível em: http://www.genetics.org/content/182/1/375.abstract. Acesso em 28 nov. 2019.

GEWEKE, J. F. **Evaluating the accuracy of sampling-based approaches to the calculation of posterior moments**. **Bayesian Statistics**. Minneapolis, MN, USA: Federal Reserve Bank of Minneapolis, Research Department, 1991.

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.

LING, C. X.; HUANG, J.; ZHANG, H. **AUC**: a statistically consistent and more discriminating measure than accuracy. International Joint Conferences on Artificial Intelligence, p. 519–526, 2003.

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

MARCHETTI, M. A.; LAI, X. H.; 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. DOI: 10.1094/phyto-77-799. Disponível em: https://doi.org/10.1094/Phyto-77-799. Acesso em 27 nov. 2019.

MEUWISSEN, T. H. E.; HAYES, B. J.; GODDARD, M. E. **Prediction of Total Genetic Value Using Genome-Wide Dense Marker Maps**. Genetics, v. 157, n. 4, p. 1819-1829, 1 abr. 2001. Disponível em: http://www.genetics.org/content/157/4/1819.abstract. Acesso em: 7 janeiro 2020.

PÉREZ, P.; DE LOS CAMPOS, G. **Genome-wide regression and prediction with the BGLR statistical package**. Genetics, v. 198, n. 2, p. 483-495, out. 2014. DOI: 10.1534/genetics.114.164442. Disponível em: https://www.ncbi.nlm.nih.gov/pubmed/25 009151. Acesso em 28 nov. 2019.

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.

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.

RESENDE, M. D. V.; SILVA, F. F.; AZEVEDO, C. F. **Estatística matemática, biométrica e computacional: modelos mistos, multivariados, categóricos e generalizados (REML/BLUP), inferência bayesiana, regressão, aleatória, seleção genômica, QTL, GWAS, estatística espacial e temporal, competição, sobrevivência**. 1. ed., Viçosa-MG: Suprema Gráfica Editora, 882 p., 2014.

TEAM, R. C. **R: A language and environment for statistical computing**. Vienna, R Foundation for Statistical Computing, 2019.

TIBSHIRANI, R. **Regression Shrinkage and Selection via the Lasso**. Journal of the Royal

Statistical Society. Series B (Methodological), v. 58, n. 1, p. 267–288, 28 jan. 1996. Disponível em: http://www.jstor.org/stable/2346178. Acesso em 28 nov. 2019.

VITEZICA, Z. G.; VARONA, L.; LEGARRA, A. **On the additive and dominant variance and covariance of individuals within the genomic selection scope**. Genetics, v. 195, n. 4, p. 1223–1230, dez. 2013. DOI: 10.1534/genetics.113.155176.

WENG , Z.; ZHANG, Z.; XIANGDONG, D.; WEIXUAN, F.; MA, P.; WANG, C.; ZHANG, Q. **Application of imputation methods to genomic selection in Chinese Holstein cattle**. Journal of Animal Science and Biotechnology, v.3, p.6, 2012.

ZHAO, K.; TUNG, C.-W.; EIZENGA, G. C. et al. **Genome-wide association mapping reveals a rich genetic architecture of complex traits in Oryza sativa**. Nature Communications, v. 2, n. 1, p. 467, 2011. DOI: 10.1038/ncomms1467. Disponível em: https://doi.org/10.1038/ncomms1467. Acesso em 28 nov. 2019.

3. CONCLUSÕES GERAIS

Neste trabalho, a metodologia de curva ROC foi utilizada para a avaliação do desempenho de modelos de regressão logística aplicados a dados de germinação de pimentas habanero e de modelos de seleção genômica ampla (GWS) aplicados a resistência do arroz *Oriza Sativa* à doença brusone. Os resultados das análises de germinação de pimentas habanero indicaram que o método de extração das sementes (manual ou mecânico) não interfere na germinação e o modelo escolhido foi aquele com a interação entre variedade, armazenamento do fruto e armazenamento da semente. Por sua vez, a utilização do índice de área abaixo da curva ROC (AUC) na seleção de modelos de GWS (RR-BLUP, Bayes Cπ e BLASSO) se mostrou equivalente aos índices usuais (taxa de erro na validação, coeficiente de Spearman e viés). Considerando todos os índices os melhores modelos foram o RR-BLUP e Bayes Cπ. Além disso, a análise gráfica apresentou vantagens adicionais, uma vez que permitiu avaliar a sensibilidade dos modelos em diferentes valores de especificidade. Considerando o tempo de execução, o RR-BLUP foi notavelmente melhor que os demais.

Desta forma, os resultados apresentados neste trabalho evidenciam que a metodologia ROC pode ser empregada para a avaliação do poder discriminativo de modelos de regressão logística em ciências agrárias e para a avaliação de modelos de seleção genômica. Assim, o emprego desta metodologia em ciências agrárias pode contribuir para a seleção de modelos mais eficazes, minimizando a ocorrência de falsos positivos e falsos negativos.

APÊNDICES

Apêndice A – Algoritmos utilizados para elaboração das figuras

```
### Figura 1 – Pág. 16
library (RcmdrMisc)
par (mfrow=c(2,2))
x <- seq(0, 25, length.out=1000)
# A)
plotDistr (x, dnorm (x, mean=7, sd=2), lwd=2, ylim=c(0,0.5))
lines (x, dnorm (x, mean=17, sd=2), lty=2)
legend ('topright', c("Negativo","Positivo"),
       col=c("black","black"), lty=1:2, bty="n")
abline (v=12.5, lwd=2)
text (12.75, 0.48, "Ponto de corte", cex=1, col="black")
# B)
plotDistr (x, dnorm (x, mean=9, sd=2), ylim=c(0,0.5), lwd=2)
lines (x, dnorm (x, mean=14.5, sd=2), lty=2)
legend ('topright', c("Negativo","Positivo"),
       col=c("black","black"), lty=1:2, bty="n")
abline (v=12, lwd=2)
text (12.5, 0.48, "Ponto de corte", cex = 1, col="black")
# C)
plotDistr (x, dnorm (x, mean=9, sd=2), ylim=c(0,0.5), lwd=2)
lines (x, dnorm (x, mean=14.5, sd=2), lty=2)
legend ('topright', c("Negativo","Positivo"),
       col=c("black","black"), lty=1:2, bty="n")
abline(v=9.3, lwd=2)
text (9.5, 0.48, "Ponto de corte", cex=1, col="black")
# D)
plotDistr(x, dnorm (x, mean=9, sd=2), ylim=c(0,0.5), lwd=2)
lines (x, dnorm (x, mean=14.5, sd=2), lty=2)
legend ('topright', c("Negativo","Positivo"),
       col=c("black","black"), lty=1:2, bty="n")
```

```
abline (v=14.8, lwd=2)
text (15, 0.48, "Ponto de corte", cex = 1, col="black")
### Figura 2 – Pág. 18
# Pacotes
library(ggplot2)
library(pROC)
library(ROCR)
library(tidyverse)
# Leitura do conjunto de dados
dados <- read_csv2("curvas.csv")
curva1 <- roc(dados$labels1, dados$prob)
curva2 <- roc(dados$labels2, dados$prob)
curva3 <- roc(dados$labels3, dados$prob)
# Curva ROC
ggroc(list(Elevada = curva1, Média = curva2, Baixa = curva3),
cex = 1, legacy.axes=TRUE) +
  scale_color_manual(labels = c("Elevada", "Média", "Baixa"),
                     breaks = c("Elevada", "Média", "Baixa"),
                     values=c(Elevada = "#c62b3a", Média =
"#604ed3", Baixa = "#bfc13a")) +
  theme_bw() + theme(legend.position = c(0.75, 0.35),
legend.background = element_rect(colour = "black")) +
  labs(x= "1-Especificidade", y= "Sensibilidade", color=
"Discriminação") +
  geom_abline(intercept=0, slope=1, linetype="dashed", cex=1)

### Figura 3 – Pág. 19
# Pares de Sensibilidade e Especificidade
abcissas <- tibble(xe = c(0.025, 0.090,0.650), ye = c(0.3,
0.83,0.99), cor=c("0.025","0.090","0.650"))
# Curva ROC
ggroc(list(Média=curva2), cex=1, color="blue",legacy.axes=TRUE)
+ theme_bw() + labs(x = "1-Especificidade", y = "Sensibilidade",
color = "Abcissas") +
```

```
geom_abline(intercept = 0, slope = 1, linetype = "dashed", cex
= 1)  + geom_point(aes(x = xe, y = ye, color = cor), shape = 21,
stroke = 1.5, size = 4, fill = "white", data = abcissas) +
theme(legend.position  =  c(0.75,  0.35),  legend.background  =
element_rect(colour = "black"))

### Figura 4 – Pág. 19
# Leitura do conjunto de dados
dados <- read_csv2("curvas.csv")
curva1 <- roc(dados$labels1, dados$prob)
curva2 <- roc(dados$labels2, dados$prob)
curva3 <- roc(dados$labels3, dados$prob)
curva4 <- roc(dados$labels4, dados$prob)
curva5 <- roc(dados$labels5, dados$prob)
curva6 <- roc(dados$labels6, dados$prob)
par(mfrow=c(1,3))
# A)
ggroc(list("Curva ROC 1" = curva1, "Curva ROC 2" = curva2),
cex = 1.5, legacy.axes=TRUE) +
  scale_color_manual(labels = c("Curva ROC 1", "Curva ROC 2"),
                     breaks = c("Curva ROC 1", "Curva ROC 2"),
                     values=c("Curva ROC 1" = "#604ed3",
                              "Curva ROC 2" = "#c62b3a")) +
  theme_bw() + theme(legend.position = c(0.75, 0.35)) +
  labs(x="1-Especificidade", y="Sensibilidade", color=NULL) +
  geom_abline(intercept=0, slope=1, linetype="dashed", cex=1)
# B)
ggroc(list("Curva ROC 1" = curva3, "Curva ROC 2" = curva4),
cex = 1.5, legacy.axes=TRUE) +
  scale_color_manual(labels = c("Curva ROC 1", "Curva ROC 2"),
                     breaks = c("Curva ROC 1", "Curva ROC 2"),
                     values=c("Curva ROC 1" = "#604ed3",
                              "Curva ROC 2" = "#c62b3a")) +
  theme_bw() + theme(legend.position = c(0.75, 0.35)) +
```

```
  labs(x="1-Especificidade", y="Sensibilidade",color=NULL) +
  geom_abline(intercept=0, slope=1, linetype="dashed", cex=1)
# C)
ggroc(list("Curva ROC 1" = curva5, "Curva ROC 2" = curva6),
cex = 1.5, legacy.axes=TRUE) +
  scale_color_manual(labels = c("Curva ROC 1", "Curva ROC 2"),
                     breaks = c("Curva ROC 1", "Curva ROC 2"),
                     values=c("Curva ROC 1" = "#604ed3",
                              "Curva ROC 2" = "#c62b3a")) +
  theme_bw() + theme(legend.position = c(0.75, 0.35)) +
  labs(x="1-Especificidade", y="Sensibilidade",color=NULL) +
  geom_abline(intercept=0, slope=1, linetype="dashed", cex=1)
```

Apêndice B – Algoritmos utilizados para análise dos dados de pimenta

```
###Leitura do banco de dados e correção das variáveis
library(readxl)
pim <- read_xlsx("sementes_pimenta.xlsx")
attach(pim)

#Transformar variável binomial em bernoulli
pim <- rbind(cbind(case=1, pim[rep(1:192,RGER7),1:10]),
             cbind(case=0, pim[rep(1:192,N-RGER7),1:10]))
attach(pim)
#Modificar variável numérica para factor
pim$case<- as.factor(case)
pim$VARIEDAD<- as.factor(VARIEDAD)
pim$MADPOST<- as.factor(MADPOST)
pim$MEDEXT<- as.factor(MEDEXT)
pim$MESESEVAL<- as.factor(MESESEVAL)

### Análise Descritiva
table(VARIEDAD,case)
table(MADPOST,case)
table(MEDEXT,case)
table(MESESEVAL,case)
prop.table(table(VARIEDAD,case),margin = 1)
prop.table(table(MADPOST,case),margin = 1)
prop.table(table(MEDEXT,case),margin = 1)
prop.table(table(MESESEVAL,case),margin = 1)

### Construção dos modelos logísticos
#Hosmer, Lemeshow e Sturdivant (2013)
###PASSO 1
m_variedad<- glm(case~VARIEDAD, family=binomial(link ="logit"))
m_madpost<- glm(case~MADPOST, family=binomial(link ="logit"))
m_medext<- glm(case~MEDEXT, family=binomial(link ="logit"))
m_meseval<- glm(case~MESESEVAL, family=binomial(link ="logit"))
```

```
summary(model_variedad)
summary(model_madpost)
summary(model_medext)
summary(model_meseval)

###PASSO 2
#Modelo múltiplo com todas as variáveis
model_uni<- glm(case~VARIEDAD + MADPOST + MEDEXT + MESESEVAL,
family = binomial(link = "logit"))
summary(model_uni)

###Passo 3
#Retirada da variável MEDEXT - Metodo de extração - não
significativa
model_uni2<- glm(case~VARIEDAD + MADPOST + MESESEVAL, family =
binomial(link = "logit"))
summary(model_uni2)

###Passo 4
#Verificar multicolinariedade
install.packages("car")
library(car)
vif(model_uni2)

###PASSO 5
#Não há necessidade

###PASSO 6
#Verificar Interações
# Modelo com Interação entre Var. e arm. fruto
model_int2<- glm(case~MESESEVAL + MADPOST + VARIEDAD/MADPOST,
family = binomial(link = "logit"))
anova(model_int2,test = "F")
summary(model_int2)
```

```
# Modelo com Interação entre Var. e arm. semente
model_int3<- glm(case~VARIEDAD + MADPOST + VARIEDAD/MESESEVAL,
family = binomial(link = "logit"))
anova(model_int3,test = "F")
summary(model_int3)

# Modelo com Interação entre arm. fruto e arm. semente
model_int4<- glm(case~VARIEDAD + MESESEVAL + MESESEVAL/MADPOST,
family = binomial(link = "logit"))
anova(model_int4,test = "F")
summary(model_int4)

#Modelo com interação tripla
model_int5<-  glm(case~VARIEDAD/MADPOST/MESESEVAL,  family  =
binomial(link="logit"))
anova(model_int5,test = "F")
summary(model_int5)

### Curva ROC
library(ggplot2)
library(ROCR)
library(tidyverse)

curva1 <- roc(pim2$case, model_uni2$fitted.values)
curva2 <- roc(pim2$case, model_int2$fitted.values)
curva3 <- roc(pim2$case, model_int3$fitted.values)
curva4 <- roc(pim2$case, model_int4$fitted.values)
curva5 <- roc(pim2$case, model_int5$fitted.values)

ggroc(list(Modelo1 = curva1,
           Modelo2 = curva2,
           Modelo3 = curva3,
           Modelo4 = curva4,
           Modelo5 = curva5),
```

```
      cex = 1,legacy.axes=TRUE) +
  scale_color_manual(labels = c("Modelo 1 - AUC = 0.7441",
                                "Modelo 2 - AUC = 0.7481",
                                 "Modelo 3 - AUC = 0.7712",
                                 "Modelo 4 - AUC = 0.7527",
                                 "Modelo 5 - AUC = 0.7935"),
                     breaks = c("Modelo1",
                                 "Modelo2",
                                 "Modelo3",
                                 "Modelo4",
                                 "Modelo5"),
                     values=c(Modelo1 = '#66FF00',
                              Modelo2 = '#669900',
                              Modelo3 = '#6666FF',
                              Modelo4 = '#0000FF',
                              Modelo5 = '#FF3333')) +
  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 = 0,
              slope = 1,
              linetype = 2,
              cex = 1)

### Odds ratio - OR
OR=exp(model_int5$coefficients)

###Intervalos de Confiança 95% para Odds ratio - OR
ICbeta=exp(confint.default(model_int5,level = 0.95))
```

Apêndice C – Algoritmos utilizados para análise dos dados de arroz

```
###Leitura dos vetores de genótipos
gen <- read.table("plink.raw", h=T)
gen1 <- gen[,7:36907]
M1=gen1 # matriz de marcadores
dim(M1)

# frequências alélicas
p1=matrix(0,ncol(M1),1)
q1=matrix(0,ncol(M1),1)
for(i in 1:ncol(M1)) {
 p1[i,]=(2*length(which(M1[,i]==2))+length(which(M1[,i]==1)))/
(2* length(na.omit(M1[,i])))
 q1[i,]=(2*length(which(M1[,i]==0))+length(which(M1[,i]==1)))/
(2* length(na.omit(M1[,i])))
}
p1[1:5,1]
q1[1:5,1]

# 3- Imputação de faltantes
#============================
M=as.matrix(M1)
for(j in 1:ncol(M)){
  lo=is.na(M[,j])
  if(2*p1[j]<=0.5){M[lo,j]=0}
  if(2*p1[j]>0.5 & 2*p1[j]<=1.5){M[lo,j]=1}
  if(2*p1[j]>1.5 & 2*p1[j]<=2){M[lo,j]=2}
}
M[1:10,1:10]
M3=cbind(as.matrix(gen[,2]),M)
colnames(M3)=c("ID",colnames(M))
M3[1:5,1:10]
###Leitura dos vetores de fenótipos
fen <- read_xlsx("fenocomp.xlsx", na="NA")
```

```
fen[1:5,1:10]
fen1 <- fen[,c(2,33)]
head(fen1)
colnames(fen1) <- c("ID","BlastResistance")
dados <- merge(fen1, M3, by=intersect("ID","ID"))
dados <- dados[!is.na(dados$BlastResistance), ]
table(dados$BlastResistance)
dim(dados)

# Matriz de incidência aditiva
gen2 <- dados[,-c(1:2)]
M4 <- matrix(0, nrow(gen2), ncol(gen2))
for(i in 1:ncol(gen2)){
  M4[gen2[,i]==2,i] <- 2-2*p1[i]
  M4[gen2[,i]==1,i] <- 1-2*p1[i]
  M4[gen2[,i]==0,i] <- 0-2*p1[i]
}

# Transformando a resistência a brusone em dados dicotômicos
dados$BlastResistance[dados$BlastResistance==0] <- 1
dados$BlastResistance[dados$BlastResistance==1] <- 1
dados$BlastResistance[dados$BlastResistance==2] <- 1
dados$BlastResistance[dados$BlastResistance==3] <- 0
dados$BlastResistance[dados$BlastResistance==4] <- 0
dados$BlastResistance[dados$BlastResistance==5] <- 0
dados$BlastResistance[dados$BlastResistance==6] <- 0
dados$BlastResistance[dados$BlastResistance==7] <- 0
dados$BlastResistance[dados$BlastResistance==8] <- 0
dados$BlastResistance[dados$BlastResistance==9] <- 0

# Ajustando os modelos
# Pacotes
library(BGLR) #Bayesian Generalized Linear Regression
library(pROC) #Display and Analyze ROC Curves
```

```
library(boa) #Bayesian Output Analysis Program (BOA) for MCMC
library(glmnet) #Lasso and Elastic-Net Regularized Generalized
Lienar Models

# fenótipos
yBin<- dados$BlastResistance
yn<-as.matrix(yBin)

# SNPs
gen<- M4 # Genótipo
M1<-as.matrix(gen)

#==========================================
# Model 1 - BRR (Bayesian Ridge Regression)
#==========================================
ETABRR<- list(SNP=list(X=M1,model="BRR"))
fm.brr<-BGLR(y=dados$BlastResistance,response_type = 'ordinal',
ETA=ETABRR, saveAt='M3_', nIter=70000, burnIn=20000, thin = 10)
save(fm.brr,file = "Modelo_BRR.RData")
ROCBRR <- roc(dados$BlastResistance, fm.brr$probs[,2])

# Cross Validation
#=========================
Folds <- 5
set.seed(123)      #Set seed for the random number generator
sets <- rep(1:5,77)  # population 385 indiv divided for five
sets <- sets[order(runif(nrow(M1)))]
COR.CV <- rep(NA,times=(folds+1))
names(COR.CV) <- c(paste('fold=',1:folds,sep=''),'Pooled')
yHatCV <- numeric()
cor_vc = NULL
byg_vc = NULL
AUC = NULL
m_vc = matrix(0,ncol(M1),5)
```

```
gbv_vc = NULL
ROC = NULL

system.time(
  for(fold in 1:folds){
    yNa <- yn
    whichNa <- which(sets==fold)
    yNa[whichNa] <- NA
    ETA1 <- list(list(X=M1,model="BRR"))
    Fm <- BGLR(y=yNa,response_type="ordinal",saveAt = "BRR_B_",
ETA=ETA1, nIter=100000, burnIn=20000, thin=10)
valores1  <-  cbind(dados$ID[whichNa],  dados$BlastResistance
[whichNa], fm$yHat[whichNa], fm$probs[whichNa,])
  colnames(valores1)=c("ID","yObs","yHat","prob(0)","prob(1)")
  write.table(valores1,paste("valores1_",fold,".txt",sep=""),
sep = " ",quote=FALSE,row.names=F)

# SNPs Effects
    m_vc[,fold] = as.matrix(fm$ETA[[1]]$b)
    pred <- fm$probs[whichNa,2]

# AUC - area under curve
    ROC <- roc(as.factor(yn[whichNa]), as.numeric(pred))
    AUC[fold] <- as.matrix(ROC$auc)

arquivo1     =     c("valores1_1.txt",     "valores1_2.txt",
"valores1_3.txt", "valores1_4.txt", "valores1_5.txt")
for(k in arquivo1){
  resultados1 <- read.table(paste(k),h=T)
resultados1$pred1<- ifelse(resultados1[,4]>resultados1[,5],0,1)
  RightRate1 <- ifelse(resultados1$yObs==resultados1$pred1,1,0)
  print(table(RightRate1))
  print(mean(RightRate1))
```

```
  print(cor(resultados1$yObs,resultados1$pred1,
method="spearman"))
  print(lm(resultados1$yObs~resultados1$pred1))
  print(lm(resultados1$yObs~resultados1$yHat))
  write.table(resultados1,paste(k),row.names=FALSE)
}
AUC[6] <- mean(as.numeric(AUC[1:5]))
write.table(AUC,  paste("AUC1.txt",  sep=""),  sep  =  "    ",
quote=FALSE, row.names=F)

#=========================
# Model 2 - BayesCpi
#=========================
ETABC <- list(SNP=list(X=M1,model='BayesC'))
fm.bc<-BGLR(y=dados$BlastResistance,response_type  =  'ordinal',
ETA=ETABC, saveAt='M2_', nIter = 70000,burnIn = 20000,thin = 10)
save(fm.bc,file = "Modelo_Bayes_C.RData")
ROCBC <- roc(dados$BlastResistance, fm.bc$probs[,2])

# Cross Validation
#=========================
folds <- 5
set.seed(123)
sets <- rep(1:5,77)
sets <- sets[order(runif(nrow(M1)))]
COR.CV2 <- rep(NA,times=(folds+1))
names(COR.CV2) <- c(paste('fold=',1:folds,sep=''),'Pooled')
yHatCV2 <- numeric()
cor_vc2 = NULL
byg_vc2 = NULL
AUC2 = NULL
m_vc2 = matrix(0,ncol(M1),5)
gbv_vc2 = NULL
ROC2 = NULL
```

```
system.time(
  for(fold in 1:folds){
    yNa2 <- yn
    whichNa2 <- which(sets==fold)
    yNa2[whichNa2] <- NA
    ETA2 <- list(list(X=M1, model="BayesC"))
    fm2 <-BGLR(y=yNa2, response_type="ordinal", saveAt="BC_B_",
ETA=ETA2, nIter=100000, burnIn=20000, thin=10)
valores2  <-  cbind(dados$ID[whichNa2]  ,dados$BlastResistance
[whichNa2], fm2$yHat[whichNa2], fm2$probs[whichNa2,])
  colnames(valores2)=c("ID","yObs","yHat","prob(0)","prob(1)")
  write.table(valores2,paste("valores2_",fold,".txt",sep=""),
sep = "  ",quote=FALSE,row.names=F)

# SNPs Effects
    m_vc2[,fold] = as.matrix(fm2$ETA[[1]]$b)
    pred2 <- fm2$probs[whichNa2,2]

# AUC - area under curve
    ROC2 <- roc(as.factor(yn[whichNa2]), as.numeric(pred2))
    AUC2[fold] <- as.matrix(ROC2$auc)

arquivo2      =      c("valores2_1.txt",      "valores2_2.txt",
"valores2_3.txt", "valores2_4.txt", "valores2_5.txt")
for(k in arquivo2){
  resultados2 <- read.table(paste(k),h=T)
resultados2$pred2<- ifelse(resultados2[,4]>resultados2[,5],0,1)
  RightRate2 <- ifelse(resultados2$yObs==resultados2$pred2,1,0)
  print(table(RightRate2))
  print(mean(RightRate2))
  print(cor(resultados2$yObs,resultados2$pred2,
method="spearman"))
  print(lm(resultados2$yObs~resultados2$pred2))
  print(lm(resultados2$yObs~resultados2$yHat))
```

```
  write.table(resultados2,paste(k),row.names=FALSE)
}
AUC2[6] <- mean(as.numeric(AUC2[1:5]))
write.table(AUC2,  paste("AUC2.txt",  sep=""),  sep  =  "    ",
quote=FALSE, row.names=F)

#=========================
# Model 3 - Blasso
#=========================
ETABL <- list(SNP=list(X=M1,model='BL'))
fm.bl<-BGLR(y=dados$BlastResistance,response_type  =  'ordinal',
ETA=ETABL, saveAt='M1_', nIter = 70000,burnIn = 20000,thin = 10)
save(fm.bl,file = "Modelo_Blasso.RData")
ROCBL<-roc(dados$BlastResistance, fm.bl$probs[,2])

# Cross Validation
#=========================
folds <- 5
set.seed(123)
sets <- rep(1:5,77)
sets <- sets[order(runif(nrow(M1)))]
COR.CV3 <- rep(NA,times=(folds+1))
names(COR.CV3) <- c(paste('fold=',1:folds,sep=''),'Pooled')
yHatCV3 <- numeric()
cor_vc3 = NULL
byg_vc3 = NULL
AUC3 = NULL
m_vc3 = matrix(0,ncol(M1),5)
gbv_vc3 = NULL
ROC3 = NULL

system.time(
  for(fold in 1:folds){
    yNa3 <- yn
```

```
    whichNa3 <- which(sets==fold)
    yNa3[whichNa3] <- NA
    ETA3 <- list(list(X=M1, model="BL"))
    fm3 <-BGLR(y=yNa3, response_type="ordinal", saveAt="BL_B_",
ETA=ETA3, nIter=100000, burnIn=20000, thin=10)
valores3   <-   cbind(dados$ID[whichNa3],dados$BlastResistance
[whichNa3], fm3$yHat[whichNa3], fm3$probs[whichNa3,])
  colnames(valores3)=c("ID","yObs","yHat","prob(0)","prob(1)")
  write.table(valores3,paste("valores3_",fold,".txt",sep=""),
sep = "  ",quote=FALSE,row.names=F)

# SNPs Effects
    m_vc3[,fold] = as.matrix(fm3$ETA[[1]]$b)
    pred3 <- fm3$probs[whichNa3,2]

# AUC - area under curve
    ROC3 <- roc(as.factor(yn[whichNa3]), as.numeric(pred3))
    AUC3[fold] <- as.matrix(ROC3$auc)

arquivo3     =     c("valores3_1.txt",     "valores3_2.txt",
"valores3_3.txt", "valores3_4.txt", "valores3_5.txt")
for(k in arquivo3){
  resultados3 <- read.table(paste(k),h=T)
resultados3$pred3<- ifelse(resultados3[,4]>resultados3[,5],0,1)
  RightRate3 <- ifelse(resultados3$yObs==resultados3$pred3,1,0)
  print(table(RightRate3))
  print(mean(RightRate3))
  print(cor(resultados3$yObs,resultados3$pred3,
method="spearman"))
  print(lm(resultados3$yObs~resultados3$pred3))
  print(lm(resultados3$yObs~resultados3$yHat))
  write.table(resultados3,paste(k),row.names=FALSE)
}
```

```
AUC3[6] <- mean(as.numeric(AUC3[1:5]))
write.table(AUC3, paste("AUC3.txt", sep=""), sep = "    ",
quote=FALSE, row.names=F)

#### CURVA ROC
#============================================================
predicao<- read.table("predicao.txt",h=T)
names(predicao)

curvabrr <- roc(predicao$yObs,predicao[,3])
curvabc <- roc(predicao$yObs,predicao[,4])
curvabl <- roc(predicao$yObs,predicao[,5])

ggroc(list(BRR = curvabrr, BayesC = curvabc, Blasso = curvabl),
      cex = 1, legacy.axes=TRUE) +
  scale_color_manual(labels = c("BRR", "BayesC", "Blasso"),
                     breaks = c("BRR", "BayesC", "Blasso"),
                     values = c(BRR = "#c62b3a",
                                BayesC = "#604ed3",
                                Blasso = "#bfc13a")) +
  theme_bw()   +   theme(legend.position   =   c(0.75,   0.35),
legend.background = element_rect(colour = "black")) +
  labs(x = "1-Especificidade",
       y = "Sensibilidade",
       color = "Discriminação") +
  geom_segment(aes(x = 0, xend = 1, y = 0, yend = 1),
               linetype = "dashed", size = 1)

#============================================================
```