ÉDIMO FERNANDO ALVES MOREIRA

TÉCNICAS DE APRENDIZADO ESTATÍSTICO APLICADAS À SELEÇÃO ENTRE FAMÍLIAS DE CANA-DE-AÇÚCAR

Tese 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 *Doctor Scientiae.*

VIÇOSA
MINAS GERAIS – BRASIL
2017

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

T
M838t
2017

Moreira, Édimo Fernando Alves, 1988-
Técnicas de aprendizado estatístico aplicadas à seleção entre famílias de cana-de-açúcar / Édimo Fernando Alves Moreira. – Viçosa, MG, 2017.
viii, 33f. : il. (algumas color.) ; 29 cm.

Orientador: Luiz Alexandre Peternelli.
Tese (doutorado) - Universidade Federal de Viçosa.
Referências bibliográficas: f.29-33.

1. Amostragem (Estatística). 2. Cana-de-açúcar - Melhoramento genético. 3. Cana-de-açúcar - Seleção. 4. Cana-de-açúcar - Classificação. I. Universidade Federal de Viçosa. Departamento de Estatística. Programa de Pós-Graduação em Estatística Aplicada e Biometria. II. Título.

CDD 22. ed. 519.5

ÉDIMO FERNANDO ALVES MOREIRA

TÉCNICAS DE APRENDIZADO ESTATÍSTICO APLICADAS À SELEÇÃO ENTRE FAMÍLIAS DE CANA-DE-AÇÚCAR

Tese 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 *Doctor Scientiae*.

APROVADA: 03 de outubro de 2017.

Moysés Nascimento

Carlos Henrique Osório Silva

Felipe Lopes da Silva

Clécio da Silva Ferreira

Luiz Alexandre Peternelli
Orientador

AGRADECIMENTOS

Aos meus Pais, José Alves Moreira e Maria Clélia Moreira, pelo amor, carinho, força e dedicação.

Á minha irmã e segunda mãe, Clenilda Alves Moreira, pelo esforço e carinho.

Aos professores Décio e Walderez, porta de entrada para a minha graduação.

Ás companheiras de estudo, Laís e Gabi, pela parceria.

Ao meu orientador, Luiz Alexandre Peternelli, pela orientação e confiança depositados desde a graduação, pelas oportunidades oferecidas, pela paciência e ensinamentos.

Aos professores do departamento de estatística, CHOS, Fabiano, Policarpo, José Ivo, Nerilson e Moysés, que tanto contribuíram para a minha formação.

Ao CNPq, FAPEMIG e CAPES pelo incentivo financeiro à pesquisa.

BIOGRAFIA

Édimo Fernando Alves Moreira, filho de Maria Clélia Moreira e José Alves Moreira, nasceu na Vila Santo Antônio, Minas Gerais, em 19 de fevereiro de 1988.

Em janeiro de 2012, graduou-se em Agronomia pela Universidade Federal de Viçosa, Viçosa-MG.

Em fevereiro de 2014 concluiu o curso de Mestrado em Estatística Aplicada e Biometria pela Universidade Federal de Viçosa, Viçosa-MG.

Em março de 2014 iniciou o curso de Doutorado em Estatística Aplicada e Biometria pela Universidade Federal de Viçosa, Viçosa-MG.

SUMÁRIO

RESUMO

MOREIRA, Édimo Fernando Alves, D.Sc., Universidade Federal de Viçosa, outubro de 2017. **Técnicas de aprendizado estatístico aplicadas à seleção entre famílias de cana-de-açúcar.** Orientador: Luiz Alexandre Peternelli.

Uma das grandes dificuldades dos programas de melhoramento de cana-de-açúcar é a seleção de genótipos nas fases iniciais. O uso de métodos estatísticos que visam a predição com base em informações tomadas a nível de campo pode contribuir para aumentar a probabilidade de identificação de genótipos potencialmente superiores. O objetivo deste trabalho é comparar as técnicas de classificação regressão logística (LR), análise discriminante linear (LDA), análise discriminante quadrática (QDA), K-*nearest neighboor* (KNN), rede neural artificial (ANN) de única camada intermediária, árvores de decisão com *random forests* (RF) e máquinas de vetor de suporte (SVM) como alternativas para seleção entre famílias de cana-de-açúcar. Os dados utilizados neste trabalho foram provenientes de 5 experimentos, com 22 famílias cada, no delineamento em blocos casualizados, com 5 repetições. Nestes experimentos foram coletados os caracteres de produção, número de colmos (NC), diâmetro de colmos (DC) e a altura de colmos (AC), bem como a produtividade real, expressa em tonelada de cana por hectare (TCHr). Para o treinamento dos métodos de classificação foram utilizados, como variáveis explicativas, os caracteres indiretos de produção NC, DC e AC. A variável resposta utilizada no treinamento foi a indicadora $Y = 0$, se a família não foi selecionada via TCHr, e $Y = 1$, caso contrário. Previamente à obtenção das regras de classificação, os valores de NC, DC e AC foram padronizados para média 0 e variância 1. Além disso, visando maior eficiência no treinamento dos modelos, foram produzidos dados sintéticos com base na simulação de valores de NC, DC, AC e TCHr para 1.000 famílias. A simulação foi feita utilizando a estrutura de médias e covariâncias fenotípicas de cada i-ésimo experimento. As análises foram processadas em 5 diferentes cenários de acordo com o experimento utilizado para simulação e treinamento dos dados. Foram ainda considerados dois modelos, um completo, com todos os preditores, NC, DC e AC, e um reduzido, onde foi excluída a variável AC. Para avaliação dos classificadores foram utilizadas a taxa de erro aparente (AER) e a taxa de verdadeiros positivos (TPR). Todas as técnicas apresentam alta concordância com a seleção via TCHr (AER média < 0,14), em ambos os modelos, completo e reduzido. No modelo completo, o melhor desempenho, menor AER média (AER=0,0886) e maior TPR média (TPR=0,9831), foi observado no classificador SVM. No modelo reduzido, os

classificadores ANN (AER média=0,0932; TPR média=0,9210), SVM (AER média=0,0977; TPR média=0,9417) e k-nearest neighboor (AER=0,1000, TPR=0,9167) apresentam os melhores resultados. O modelo reduzido pode ser preferido, pois apresenta resultados similares ao completo e tem a vantagem de ser operacionalmente mais simples.

ABSTRACT

MOREIRA, Édimo Fernando Alves, D.Sc., Universidade Federal de Viçosa, October, 2017. **Techniques of statistical learning applied the selecting among families of sugarcane.** Adviser: Luiz Alexandre Peternelli.

One of the great difficulties of breeding programs is the selection of genotypes in the early stages. The use of statistical methods for the prediction based on information taken at the field level can contribute to increase the probability of identifying potentially superior genotypes. The objective of this study is to compare the classification techniques, logistic regression (LR), linear discriminant analysis (LDA), quadratic discriminant analysis (QDA), K-nearest neighboor (KNN), single-layer neural network (ANN), decision trees (DT) with random forests and support vector machines (SVM) as alternatives for selection of sugarcane families. The data used in this study were from five experiments with 22 families each, in randomized block design with 5 repetitions. In these experiments were collected production traits, number of stalks (NS), stalk diameter (SD) and the stalk height (SH) and the real production, expressed in tons of cane per hectare (TCHr). For training of methods were used as explanatory variables the indirect production traits, NS, SD and SH. The output variable used in training was the indicator, $Y = 0$, if the family was not selected by real ton cane per hectare, and $Y = 1$, if the family has been selected. Prior to obtaining the classification rules, the values of NS, SD and SH were standardized for mean 0 e variance 1. Moreover, aiming at greater efficiency in training of models were produced synthetic data based on simulation values of NS, SD, SH and TCHr for 1,000 families. The simulation was done using the structure of phenotypic mean and covariance of each ith experiment. The analyzes were performed in five different scenarios according to the experiment used for simulation and training data. In addition to the different scenarios they were considered two models, full, with all the explanatory variables and reduced, which was excluded from the variable SH. All the techniques of statistical learning feature high agreement with the selection via TCHr (AER mean < 0.14), in both models, full and reduced. For the full model, the best performance, lower AER mean (AER=0.0886) and higher TPR mean (0.9831), was observed in the classifier SVM. In the reduced model, the ANN (AER mean=0.0932; TPR mean=0.9210), the SVM (AER mean=0.0977; TPR mean=0.9417) and the k-nearest neighboor (AER=0.1000, TPR=0.9167) how the best results. The reduced model may be preferred because it

presents similar results to the complete model and has the advantage of being operationally simpler.

1. Introdução

A cana-de-açúcar é uma das principais culturas agrícolas do Brasil. O país é o maior produtor de açúcar e etanol de cana do mundo, e conquista cada vez mais o mercado externo com o uso do biocombustível como alternativa energética (MAPA, 2016).

O melhoramento genético é fundamental no agronegócio de cana-de-açúcar. As variedades melhoradas permitem o aumento da produtividade em cana-de-açúcar e a melhoria da matéria prima para fabricação de açúcar, etanol e geração de energia (BARBOSA e SILVEIRA, 2015).

A principal dificuldade encontrada pelos melhoristas nos programas de melhoramento de cana-de-açúcar é a seleção de genótipos promissores nas fases iniciais (PETERNELLI et al., 2011). Esta dificuldade advém da enorme quantidade de genótipos que precisam ser avaliados.

Nestas fases não são usados delineamentos experimentais para indivíduos. Isto é, cada genótipo apresenta uma única repetição. Neste caso, as plantas são mantidas em um único ambiente e a seleção dos melhores genótipos geralmente é feita visualmente baseado no fenótipo do indivíduo (seleção massal) (OLIVEIRA et al., 2013).

Ainda que a seleção massal seja o método comumente empregado nas fases iniciais, esta estratégia tem sido questionada quanto a eficiência de seleção. Kimbeng e Cox (2003), Resende e Barbosa (2006), Stringer et al. (2011) e Oliveira et al. (2013) discutem que ao invés da seleção massal seja feita uma seleção das melhores famílias seguida de uma seleção individual dos melhores genótipos dentro destas. Esta estratégia apresentaria maior ganho genético que a simples seleção de clones pois a herdabilidade dos caracteres relacionados ao rendimento baseado em famílias é maior que em indivíduos.

A ideia básica desta estratégia é que dentro das melhores famílias existe uma maior probabilidade de se encontrar os melhores genótipos. Assim, ao indicar as melhores famílias, além do ganho genético, há um enorme ganho de tempo na seleção dos melhores genótipos.

De maneira geral, a seleção de famílias pode ser preferida quando a seleção é praticada com base em caracteres indiretos (FALCONER e MACKAY, 1996). Durante a seleção das melhores famílias, o diâmetro dos colmos (DC), o número de colmos (NC) e altura de colmos (AC) são os caracteres indiretos comumente usados para avaliar a produção de cana-de-açúcar (CHANG e MILLIGAN, 1992). Esses caracteres são comumente denominados componentes de produção de cana-de-açúcar (ESPÓSITO et al., 2012).

No que se refere aos métodos de estatística e genética utilizados para a seleção entre famílias destaca-se o BLUP (*Best Linear Unbiased Predictor*) individual (BLUPI) (RESENDE, 2002). Em especial para a cana-de-açúcar tem-se proposto o BLUP individual simulado (BLUPIS) (RESENDE e BARBOSA, 2006) ou o BLUP individual simulado modificado (CASTRO et al., 2016). Particularmente, para o caso de dados balanceados e sem o uso de pedigree, como neste trabalho, a seleção pode ser simplificada. Neste caso a seleção via BLUP coincide com a seleção das famílias que apresentam tonelada de cana por hectare real (TCHr) acima da média geral fenotípica do experimento.

É importante ressaltar que para a utilização das metodologias via BLUP é necessário pesar de forma integral todas as plantas contidas nas parcelas dos experimentos. Este fato faz com que, para a utilização destes métodos, seja necessário um esforço operacional muito grande.

Uma maneira de contornar o problema de pesagem no campo, e que será avaliado neste trabalho, é o uso de técnicas de aprendizado estatístico para seleção em cana-de-açúcar na fase de seleção de famílias.

No caso da utilização de uma técnica de aprendizado estatístico para seleção entre famílias haveria uma enorme redução no trabalho no campo. Isto porque seria necessário a pesagem de apenas uma pequena parte das famílias. Neste caso o processo de seleção seria otimizado, permitindo até mesmo a avaliação de uma maior quantidade de material.

O aprendizado estatístico pode ser visto com um conjunto de técnicas e procedimentos úteis para extrair informações de um conjunto de dados. A ideia básica é ajustar um modelo estatístico para predição ou estimação de uma variável de saída, ou variável resposta, em função de uma ou mais variáveis de entradas, também chamadas de variáveis explicativas (JAMES et al., 2013).

No contexto deste trabalho, as variáveis NC, DC e AC são as variáveis de entrada, enquanto que a variável indicadora *Y*, onde $Y = 0$, se a família não foi selecionada via TCHr, e $Y = 1$, caso contrário, corresponde à variável de saída. Vale ressaltar que, uma vez que a variável de saída está associada a duas classes, isto é, família selecionada ou não selecionada, temos um problema de classificação (HASTIE et al., 2009).

Há muitas técnicas de aprendizado estatístico possíveis para classificação (HASTIE et. al., 2009). Neste trabalho utilizaremos as seguintes técnicas: regressão logística, análise discriminante linear, análise discriminante quadrática, *k-nearest neighboor*, árvore de decisão com *random forests*, máquina de vetor de suporte e rede neural artificial de única camada intermediária.

Alguns exemplos de aplicações, na área de genética e/ou melhoramento genético, das técnicas de aprendizado estatístico citadas incluem, dentre outros: i) Uso de modelos de regressão logística para seleção em cana-de-açúcar (BRASILEIRO et al., 2015); ii) uso da análise discriminante linear para seleção em cana-de-açúcar (MOREIRA e PETERNELLI, 2015) iii) uso da análise discriminante quadrática para predição de *splice sites* (ZHANG e LUO, 2003); iv) uso do K-*nearest neighboor* em seleção genômica (GRINBERG et al., 2016); v) uso de árvores de decisão com *random forests* em seleção genômica (OGUTU et al., 2011; HESLOT et al., 2012; GRINBERG et al., 2016); vi) uso da máquina de vetor de suporte em seleção genômica (OGUTU et al., 2011; HESLOT et al., 2012); vii) uso de modelos de redes neurais para seleção em cana-de-açúcar (ZHOU et al., 2011.; BRASILEIRO et al., 2015).

O objetivo deste trabalho é comparar as técnicas de classificação, regressão logística, análise discriminante linear, análise discriminante quadrática, *k-nearest neighboor*, árvore de decisão com *random forests*, máquina de vetor de suporte e rede neural artificial de única camada intermediária, como alternativas ao procedimento baseado em pesagem da parcela e comumente utilizado para seleção entre famílias de cana-de-açúcar.

2. Material e Métodos

2.1. Conjunto de dados e avaliação fenotípica

Os dados utilizados são provenientes de cinco experimentos conduzidos no Centro de Pesquisa e Melhoramento de Cana-de-açúcar (CECA), da Universidade Federal de Viçosa, localizado no município de Oratórios, Minas Gerais, com latitude 20°25’S, longitude 42°48’W e altitude de 494 metros. Os experimentos foram instalados em blocos completos casualizados com 5 repetições e 22 famílias distintas em cada experimento. A unidade experimental foi constituída por 20 plantas, distribuídas em dois sulcos de 5 m de comprimento, espaçados em 1,40 m.

Os seguintes caracteres foram avaliados: altura média de colmos (AC) em metros, mensurando-se um colmo representativo de cada touceira, desde a base até o primeiro *dewlap* visível; diâmetro médio de colmos (DC) em centímetros, medido, naquele mesmo colmo, com paquímetro digital no terceiro internódio, contado da base do colmo para o ápice; número total de colmos por parcela (NC) e tonelada de colmos por hectare real (TCHr) medida pesando todas as plantas da parcela por meio de balança.

A seleção das melhores famílias foi realizada a partir do TCHr das famílias de cada um dos cinco experimentos. Aquelas famílias que apresentavam TCHr acima da média fenotípica geral de cada experimento foram classificadas como selecionadas.

Uma vez que 110 famílias poderia ser um número insuficiente para obtenção de regras de classificação com ampla capacidade de generalização, foram produzidos dados sintéticos a partir da simulação de valores de NC, AC, DC e TCHr para novas 1.000 famílias tomando-se como base a estrutura de médias e covariâncias de cada um dos cinco experimentos separadamente. As observações das famílias simuladas foram utilizadas para compor o conjunto de treinamento. Vale destacar que Moreira e Peternelli (2015) mostraram, para o mesmo conjunto de dados, que a utilização de dados sintéticos pode melhorar significativamente o desempenho de uma técnica de aprendizado. No trabalho citado foi utilizada a técnica de classificação análise discriminante linear e, para a obtenção dos melhores resultados, foram suficientes 1.000 famílias simuladas.

A simulação foi feita utilizando a decomposição de Cholesky na matriz de correlação entre as variáveis NC, DC, AC e TCHr. Este método é bastante utilizado para simulação com múltiplas variáveis correlacionadas (HAINING, 2005; CRESSIE, 1993; SANTOS e FERREIRA, 2003).

A decomposição de Cholesky de uma matriz simétrica positiva definida $\mathbf{A}$, caso de uma matriz de covariância, é uma decomposição na forma $\mathbf{A} = \mathbf{L}\mathbf{L}^T$, onde $\mathbf{L}$ é uma matriz triangular inferior e $\mathbf{L}^T$ é a matriz transposta conjugada de $\mathbf{L}$. Aplicando $\mathbf{L}$ a um vetor de amostras não correlacionadas produz-se um vetor com as propriedades de correlação e covariância do sistema que está sendo modelado.

Além da simulação, as variáveis de entrada foram padronizadas para a média zero e variância um. Este procedimento evita que a escala das variáveis cause interferências na construção das regras de classificação.

Finalmente, o conjunto de dados compreendeu valores padronizados de NC, DC, AC e TCHr para 1110 famílias em cinco diferentes cenários. Cada i-ésimo cenário foi composto das 110 famílias originais (22 famílias x 5 experimentos) acrescido de 1000 famílias simuladas como base no i-ésimo experimento, isto é, o cenário 1 foi composto das 110 famílias originais acrescido de 1000 famílias simuladas com base na estrutura de médias e covariâncias das 22 famílias do experimento 1 e assim por diante.

Em todos os cenários o conjunto de dados foi dividido em observações de treinamento e teste. As observações de treinamento são utilizadas para o ajuste das regras de classificação e as observações de teste são utilizadas para avaliação destas regras. Em cada i-ésimo cenário, as observações de treinamento foram dadas pelas informações das 22 famílias do i-ésimo experimento acrescido das informações das 1000 famílias simuladas com base neste experimento. As observações de teste foram compostas pelas 88 famílias correspondentes aos demais experimentos. Estes cenários podem ser melhor visualizados na Tabela 1.

Tabela 1. Observações de treinamento e teste em cinco diferentes cenários definidos em função do experimento utilizado para simulação e treinamento.

Cenários	Exp. de simulação	Simulação	Treinamento	Teste
1	Exp. 1	1000 famílias	Exp. 1 + Simulação	Exp. 2,3,4,5
2	Exp. 2	1000 famílias	Exp. 2 + Simulação	Exp. 1,3,4,5
3	Exp. 3	1000 famílias	Exp. 3 + Simulação	Exp. 1,2,4,5
4	Exp. 4	1000 famílias	Exp. 4 + Simulação	Exp. 1,2,3,5
5	Exp. 5	1000 famílias	Exp. 5 + Simulação	Exp. 1,2,3,4

*Exp. = Experimento

É importante ressaltar que, para a aplicação das técnicas de aprendizado, necessitaríamos da tomada de informações do NC, DC e AC em todos os experimentos. Os caracteres NC e DC são relativamente simples de serem obtidos em campo, mas a obtenção do AC tem-se mostrado um trabalho bastante moroso. Algumas razões que tornam a coleta do AC complexa são: grande diversidade na altura de colmos dentro da mesma família, tombamento de plantas e irregularidades de forma no colmo.

Assim, visando avaliar modelos de seleção que combinem eficiência de seleção e simplicidade operacional, as análises foram processadas em dois diferentes modelos. No modelo 1, denominado modelo completo, foram consideradas, para o ajuste, todas as variáveis explicativas, isto é, NC, DC e AC. No modelo 2, chamado de modelo reduzido, foram utilizadas as variáveis NC e DC, ou seja, excluiu-se a variável AC.

Importante ainda destacar que as variáveis NC, DC e AC são utilizadas como variáveis de entrada nas técnicas de classificação utilizadas. A variável TCHr, por sua vez, é utilizada para a obtenção da variável indicadora Y, onde $Y = 0,$ para famílias com TCHr abaixo da média geral do i-ésimo experimento ou família "não selecionada" e $Y = 1,$ para famílias com TCHr acima da média geral do i-ésimo experimento ou família "selecionada". A variável indicadora assim obtida é utilizada como variável de saída.

2.2. Modelagem via regressão logística múltipla

Na modelagem via regressão logística múltipla (LR) nosso objetivo é predizer a variável de saída Y ($Y = 0$, família não selecionada e $Y = 1$, família selecionada) usando múltiplas variáveis explicativas X_i (NC, DC e AC).

A ideia central da regressão logística múltipla é que iremos modelar a distribuição condicional de Y dado X_i usando a função logística (HASTIE et. al., 2019), isto é

$$p(Y \mid X_i) = \frac{e^{\beta_0+\beta_1 X_1+\beta_2 X_2+\beta_3 X_3}}{1+e^{\beta_0+\beta_1 X_1+\beta_2 X_2+\beta_3 X_3}} \ .$$

Note que o método primeiro prediz a probabilidade de uma observação pertencer a uma das categorias da variável qualitativa, para, depois fazer a classificação; neste sentido comporta-se como um método de regressão.

É conveniente lembrar que as estimativas dos parâmetros são feitas utilizando as observações de treinamento e as predições são feitas nas observações de teste. O método comumente utilizado na regressão logística para estimação dos parâmetros é o da máxima verossimilhança (BISHOP, 2006).

2.3. Modelagem via análise discriminante linear

Para a modelagem via análise discriminante linear (LDA) iremos pressupor que $\mathbf{x}$ seja a realização de uma variável aleatória $\mathbf{X}$ com p preditores (NC, DC, AC). Consideraremos ainda que $\mathbf{X}$ tem distribuição normal multivariada com vetor de médias $\boldsymbol{\mu}_k$ e matriz de covariância comum $\boldsymbol{\Sigma}$, isto é, $\mathbf{X} \sim N(\boldsymbol{\mu}_k, \boldsymbol{\Sigma})$ (BISHOP, 2006; FERREIRA, 2011). A função densidade de probabilidade normal multivariada para a k-ésima classe (k = 0, família não selecionada e k = 1, família selecionada) será escrita como

$$f_k(\mathbf{x}) = \frac{1}{(2\pi)^{\frac{p}{2}} |\boldsymbol{\Sigma}|^{\frac{1}{2}}} \exp\left(-\frac{1}{2} (\mathbf{x} - \boldsymbol{\mu}_k)^T \boldsymbol{\Sigma}^{-1} (\mathbf{x} - \boldsymbol{\mu}_k) \right).$$

Substituindo a função densidade normal multivariada $f_k(\mathbf{x})$ na expressão do teorema de Bayes teremos a função discriminante (JAMES et al., 2013), dada por:

$$\delta_k(\mathbf{x}) = \mathbf{x}^T \boldsymbol{\Sigma}^{-1} \boldsymbol{\mu}_k - \frac{1}{2} \boldsymbol{\mu}_k^T \boldsymbol{\Sigma}^{-1} \boldsymbol{\mu}_K + \log \pi_k .$$

Na prática não temos os parâmetros $\boldsymbol{\mu}_k, \boldsymbol{\Sigma}$ e π_k, e sim seus estimadores $\hat{\boldsymbol{\mu}}_k, \hat{\boldsymbol{\Sigma}}$ e $\hat{\pi}_k$ respectivamente. Substituindo os estimadores na função discriminante temos a função discriminante amostral (JAMES et al., 2013), dada por:

$$\hat{\delta}_k(\mathbf{x}) = \mathbf{x}^T \hat{\boldsymbol{\Sigma}}^{-1} \hat{\boldsymbol{\mu}}_k - \frac{1}{2} \hat{\boldsymbol{\mu}}_k^T \hat{\boldsymbol{\Sigma}}^{-1} \hat{\boldsymbol{\mu}}_k + \log \hat{\pi}_k .$$

Uma vez obtida a função discriminante amostral, uma nova observação $\mathbf{x}$, neste caso uma família de cana-de-açúcar, será alocada à classe k em que $\hat{\delta}_k(\mathbf{x})$ é um alto valor (JAMES et al., 2013).

Os métodos apresentados nos itens 2.2 (LR) e 2.3 (LDA), apresentam uma fronteira linear de separação entre as classes. A diferença entre as abordagens está nos procedimentos de ajuste. Na LR os estimadores são obtidos pelo método da máxima verossimilhança, enquanto na LDA os estimadores são obtidos usando o vetor de médias e a matriz de covariância de uma distribuição normal multivariada (BISHOP, 2006).

2.4. Modelagem via análise discriminante quadrática

Na modelagem via análise discriminante quadrática iremos assumir que uma observação $\mathbf{X} = \mathbf{x}$ (um vetor com 3 componentes, NC, DC e AC) de uma $k-ésima$ classe é da forma $\mathbf{X} \sim N(\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)$, onde $\boldsymbol{\mu}_k$ e $\boldsymbol{\Sigma}_k$ são, respectivamente, o vetor de médias e a matriz de covariância da $k-ésima$ classe (HASTIE et al., 2009). Note que neste caso, diferentemente do classificador LDA, cada classe tem uma matriz de covariância específica. Neste caso, a função densidade normal multivariada $f_k(\mathbf{x})$ pode ser escrita como

$$f_k(\mathbf{x}) = \frac{1}{(2\pi)^{\frac{p}{2}} |\boldsymbol{\Sigma}_k|^{\frac{1}{2}}} \exp\left(-\frac{1}{2} (\mathbf{x} - \boldsymbol{\mu}_k)^T \boldsymbol{\Sigma}_k^{-1} (\mathbf{x} - \boldsymbol{\mu}_k) \right).$$

Substituindo a função de densidade normal multivariada $f_k(x)$, na expressão do teorema de Bayes teremos a seguinte função discriminante

$$\delta_k(\mathbf{x}) = -\frac{1}{2} \mathbf{x}^T \boldsymbol{\Sigma}_k^{-1} \mathbf{x} + \mathbf{x}^T \boldsymbol{\Sigma}_k^{-1} \boldsymbol{\mu}_K - \frac{1}{2} \boldsymbol{\mu}_k^T \boldsymbol{\Sigma}_k^{-1} \boldsymbol{\mu}_k - \frac{1}{2} \log|\boldsymbol{\Sigma}_k| + \log(\pi_k)$$

(JAMES et. al., 2013). Note que, na expressão acima, $\mathbf{x}$ aparece como uma função quadrática e, por isso, o nome análise discriminante quadrática.

Substituindo na expressão anterior, os parâmetros pelos seus estimadores, podemos atribui uma observação de teste $\mathbf{x}$ em uma classe k para a qual $\hat{\delta}_k(\mathbf{x})$ é um alto valor (JAMES et. al., 2013).

É importante lembrar que, por acrescentar um termo quadrático na regra de classificação, o classificador QDA se adapta a uma amplitude maior de problemas quando comparado a LDA.

2.5. Modelagem via *K-nearest neighboors*

O classificador *K-nearest neighboor* (KNN), assim como vários outros métodos de classificação, é um método de aprendizado estatístico que objetiva estimar a distribuição condicional de Y dado X para classificar uma dada observação.

Para classificar uma observação de teste $\mathbf{x}$, um vetor tri-dimensional ($x_1 = NC, x_2 = DC, x_3 = AC$), o KNN primeiro identifica, no conjunto de treinamento, K pontos mais próximos de $\mathbf{x}$. Em seguida ele estima a probabilidade condicional da saída pertencer à classe j, dada a observação de teste, como uma fração de pontos em K cuja resposta é igual a j (JAMES et al., 2013)

$$P(Y = j \mid \mathbf{X} = \mathbf{x}) = \frac{1}{K}\left(\sum_{i \in K} I(y_i = j)\right) .$$

Feito isto, uma observação de teste $\mathbf{x}$ será alocada na classe com maior probabilidade (JAMES et. al., 2013).

Note que todos os classificadores vistos até o momento, LR, LDA, QDA e KNN tem como semelhança o fato de usarem a regra de Bayes para modelar a distribuição condicional de Y dado X, obtendo assim uma regra de classificação. O *K-nearest neighboor* difere-se dos demais por ser um procedimento não paramétrico. Isto é, nenhuma suposição é feita sobre a distribuição dos dados nem sobre a fronteira de separação entre as classes (DUDA et al., 2000)

Ainda que no classificador KNN nenhuma suposição seja feita sobre a fronteira de separação entre as classes, é esperado que este tenha desempenho superior às técnicas citadas em situações onde a fronteira de separação entre as classes for não-linear (DUDA et al., 2000). Vale ressaltar que LR e LDA são técnicas lineares de classificação e o classificador QDA pode ser visto como uma transição entre o linear e o não linear.

Na modelagem via K-*nearest neighboor* o valor de K é de fundamental importância. Como mencionado, este parâmetro define o número de observações de treinamento mais próximas a uma dada observação de teste $\mathbf{x}$. Quando da análise o K é escolhido verificando o valor que a partir deste não há um incremento na acurácia da classificação.

Vale ainda ressaltar que um valor baixo de K fará com que o classificador apresente baixo viés, mas alta variância na predição. Isto porque a mudança de uma simples observação no treinamento mudará toda a regra de classificação. Para maiores valores de K esperamos uma variância menor ao custo de um pouco de viés (DUDA et al., 2000).

2.6. Modelagem via árvores de decisão

A classificação via árvores de decisão passa pela estratificação do espaço de entrada dos preditores em um pequeno número de regiões na forma retangular. Em cada região é ajustado um modelo simples que é utilizado para predição de uma variável de saída qualitativa (BREIMAN et al., 1984).

A construção da regra de classificação, de maneira geral, pode ser dividida em dois passos (BREIMAN et al., 1984; JAMES et al., 2013):

i) O conjunto de possíveis valores das variáveis de entrada $X_1, X_2, ..., X_j$ é dividido em j regiões distintas e não sobrepostas, $R_1, R_2, ..., R_j$.

ii) Em cada região R_j haverá uma classe mais provável. Uma dada observação será sempre classificada de acordo com a classe mais provável da região em que ela está alocada.

Na construção da árvore, o nosso objetivo é encontrar as regiões $R_1, R_2, ..., R_j$ que minimizem uma "medida de impureza" $Q_m(T)$. Dentre as medidas $Q_m(T)$ utilizadas em problemas de classificação podemos citar o *gini index*, a *cross-entropy* e a *misclassification error*. Quando o objetivo é acurácia na predição, a mais recomendada é a *misclassification error*, que pode ser escrita como:

$$E = 1 - \max(\hat{p}_{mk})$$

onde $\hat{p}_{mk}$ é a proporção das observações de treinamento na m-ésima região que são da k-ésima classe (HASTIE et al., 2009).

Na obtenção da regra de classificação é preciso dividir o espaço de entrada dos preditores de modo a minimizar a *misclassification error*. Considerar todas as possíveis partições é computacionalmente inviável. Assim, é utilizada uma abordagem descendente conhecida como *recursive binary splitting* (RBS). Nesta, a cada passo da construção da árvore de decisão é feita a melhor divisão, de maneira a levar para a divisão seguinte a melhor árvore possível (BREIMAN et al., 1984).

As árvores obtidas via RBS apresentam excelente desempenho nas observações de treinamento, mas frequentemente pobre desempenho na classificação de novas observações. Isto ocorre porque a árvore resultante é frequentemente complexa, com várias partições e assim possui alta variância nas predições além de baixa interpretabilidade (JAMES et. al., 2013).

Uma estratégia para obter árvores com menos divisões (isto implica em menor variância e maior interpretabilidade) é deixar a árvore crescer até atingir um grande tamanho (T_0) e podá-la, a fim de obter um subárvore mais simples. Para escolher a melhor subárvore uma alternativa é estimar a taxa de erro de teste via *cross-validation*. Ainda assim, haveria um número muito grande de subárvores que tornaria computacionalmente inviável esta estratégia (BREIMAN et al., 1984).

O *cost complexity prunning* (CCP) é uma estratégia de poda que contorna este problema. Neste caso, será considerado um número menor de subárvores, todas indexadas por um parâmetro não negativo α. Para cada α haverá uma subárvore, que fará como que a expressão

$$(1-\max(\hat{p}_{mk}))+\alpha|T|$$

seja tão pequena quanto possível. Na expressão acima, $|T|$ indica o número de nós da subárvore T, e $(1-\max(\hat{p}_{mk}))$ é a *misclassification error*, medida de impureza utilizada (JAMES et al., 2013).

O parâmetro α tem a função de controlar o perde-ganha entre viés-variância. Note que quando α aumenta a expressão acima tende a ser minimizada por uma árvore menor, mais simples, com menor variância, ao custo de um pouco de viés. Quando $\alpha = 0$ a subárvore T é igual à árvore inicial T_O, isto é, uma árvore complexa, maior e com maior variância.

De acordo com James et al. (2013), um algoritmo para a construção de uma árvore de decisão pode ser descrito como a seguir:

i) aplicamos o RBS, nas observações de treinamento, para obter uma árvore com elevado número de nós. O processo para apenas quando cada nó atinge algum número mínimo de observações.

ii) aplicamos o CCP para obter uma sequência de subárvores em função do parâmetro α.

iii) utilizamos o *cross-validation* para escolher o valor de α que minimiza a *misclassification error*.

iv) em função do α obtido, voltamos ao passo 2 e selecionamos a subárvore correspondente.

A classificação via árvores de decisão é um método bastante simples, prático e de boa interpretabilidade. No entanto, a técnica sofre com a alta variância e baixa acurácia quando comparado a outras técnicas de aprendizado (BREIMAN et al., 1984).

Uma maneira de contornar tais problemas é utilizar estratégias como o *bagging* e o *random forests*. Nestas abordagens são produzidas múltiplas árvores que quando combinadas provocarão um incremento na acurácia às custas de perda na interpretabilidade (BREIMAN et al., 1984; BREIMAN, 1996; BREIMAN, 2001).

2.6.1 *Bagging* e *Random Forests*

O *Bagging* ou, *bootstrap agregation*, consiste em usar o *bootstrap* para tomar repetidas amostras do conjunto de treinamento até serem obtidos *n* conjuntos de treinamento que serão usados na construção de múltiplas árvores de decisão. Ao final, usaremos, para a predição, a média das múltiplas árvores geradas. Ao utilizar a média, essa técnica provoca uma grande redução na variância e com isso um aumento na acurácia de predição (BREIMAN, 1996).

Em cada passo da construção da árvore de decisão uma divisão na árvore é considerada. Na existência de uma variável de entrada que seja altamente correlacionada com a variável de saída, a maioria das árvores utilizará essa variável no topo da árvore, fazendo com que elas sejam similares entre si. Isto faz com que as predições usando estas árvores sejam altamente correlacionadas. Vale lembrar que quando tomamos a média de quantidades altamente relacionadas não há uma grande redução na variância (JAMES et al., 2013).

No *Random Forests*, a cada divisão na árvore será considerado somente um subconjunto de m preditores. Assim, uma parte das divisões não irá considerar a variável de entrada forte, e então outras variáveis de entrada terão a chance de aparecer no topo do árvore. Esse procedimento pode ser visto como uma maneira de quebrar a correlação existente entre as múltiplas árvores de decisão (BREIMAN, 2001).

Quando da utilização do *Random Forests* dois parâmetros são de extrema importância: o número de árvores (n) que serão combinadas, e o número de preditores (m) considerados em cada passo da construção da árvore. Neste trabalho, o n foi escolhido dentre os seguintes valores: 1, 250, 500, 1000 e 2000. Já o m foi obtido dentre as seguintes alternativas: $\sqrt{p}$; $\frac{p}{2}$ e p , onde p é o número de preditores. Vale destacar que quando $m = p$, o *random forests* corresponde ao *bagging*.

2.7. Modelagem via redes neurais artificiais

Redes neurais são modelos matemáticos inspirados na estrutura do cérebro humano e que adquirem conhecimento através de experiência, podendo ser utilizadas para resolver problemas de predição e classificação (HAYKIN, 2001).

O termo rede neural abrange uma enorme quantidade de modelos e de métodos de aprendizado. Neste trabalho utilizaremos a mais comum, chamada de *single hidden layer back-propagation network* (HASTIE et al., 2009). Esta é uma rede de múltiplas camadas (Multilayer Perceptron) com uma única camada intermediária, ou oculta, entre a camada de entrada e a camada de saída como mostra a Figura 1.

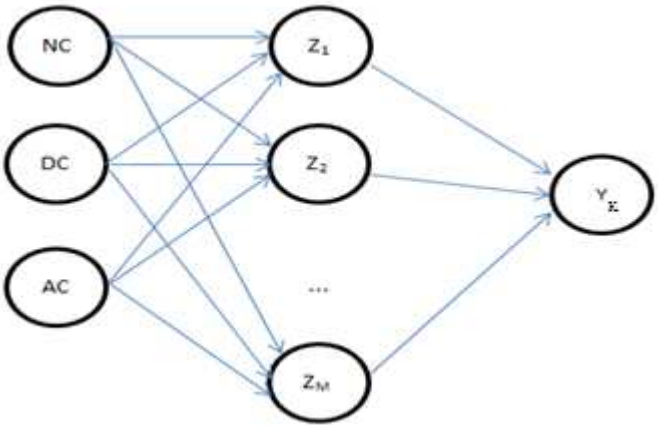

Figura 1. Esquema da *single hidden layer back-propagation network.* NC, DC e AC são as variáveis de entrada. Z_m, com $m = 1,2,...,M$, são funções de ativação. Y_k, com $k = 0,1$ é a saída da rede.

Na Figura 1, temos que NC, DC e AC são as variáveis de entrada. Z_m, com $m = 1,2,...,M$, são as funções de ativação, responsáveis pela soma ponderada das entradas. Esta ponderação é feita de acordo com os parâmetros da rede W_i, com $i = 1,2,...,I$. Y_k, com $k = 0,1$, é a saída da rede, ou seja, o resultado do processo de seleção via rede neural (HASTIE et al, 2009).

A função de ativação comumente utilizada é a função sigmoide logística, dada por:

$$f_k(x) = \frac{1}{1+e^{-wx}} \quad .$$

Esta tem sido a mais utilizada por ser uma função não-linear com comportamento levemente linear, conseguindo assim se adaptar a uma amplitude maior de problemas. Outra característica de interesse na função sigmoide logística é que ela é facilmente diferenciável, permitindo assim a estimação dos parâmetros da rede, ou pesos, de maneira mais simples (HAYKIN, 2001).

Os parâmetros da rede, ou pesos, foram estimados pela minimização da *cross-entropy*, escrita como

$$R(W) = -\sum_{i=1}^{N}\sum_{k=1}^{K} y_{ik} \log f_k(x_i) .$$

A minimização é feita pela aplicação de um algoritmo de gradiente decrescente conhecido como *back-propagation* (HASTIE et al., 2009). Este algoritmo pode ser descrito como a seguir:

i) atribuição de valores iniciais aos parâmetros da rede;

ii) apresentação à rede as observações de treinamento, isto é, os valores das variáveis de entrada e a correspondente variável de saída;

iii) verificação da saída nos neurônios da camada oculta;

iv) verificação da resposta real da rede, isto é, a resposta obtida na camada de saída;

v) obtenção da *cross-entropy*, obtida em função da resposta, ou saída, da rede ($f_k(x_i)$) e da saída desejada (y_{ik});

vi) retropropagação da *cross-entropy* para o ajuste dos pesos.

Vale destacar que, segundo Venables e Ripley (2002) os valores iniciais para os pesos podem ser escolhidos ao acaso em um intervalo que satisfaça a seguinte equação $LS * \max(|X|) \approx 1$, onde LS é o limite superior do intervalo, e $\max(|X|)$ é o maior valor em módulo do conjunto de treinamento.

Quando da modelagem via rede neural, dois são os parâmetros de interesse: o número de neurônios na camada oculta ou intermediária, que chamaremos de z, e o número de parâmetros da rede ou pesos (w). Estes serão obtidos em função de um erro máximo desejado que é pré-fixado.

2.8. Modelagem via máquinas de vetor de suporte

Em problemas linearmente separáveis o *support vector classifier* (SVC) é uma abordagem natural para classificação binária. A ideia central neste classificador é construir um hiperplano que maximize a margem de separação, menores distâncias entre as observações de treinamento e o hiperplano, podendo classificar incorretamente algumas observações (VAPNIK, 1998).

Estatisticamente, temos o seguinte problema de otimização

$$\begin{aligned} &\underset{\beta_0,\beta_1,...,\beta_p}{\text{maximizar}} \; M \\ &\text{sujeito a} \sum_{j=1}^{p} \beta_j^2 = 1 \\ &y_i\left(\beta_o + x_i^T \beta\right) \geq M(1-\varepsilon_i) \\ &\varepsilon_i > 0, \sum_{i=1}^{n} \varepsilon_i \leq C, \end{aligned}$$

onde M é a margem de separação, β são os parâmetros do hiperplano, C é um parâmetro de ajuste não-negativo, que pode ser obtido via *cross-validation* e $\varepsilon_1,...,\varepsilon_n$ são variáveis que permitem que observações individuais estejam do lado errado do hiperplano ou da margem de separação (JAMES et al., 2013).

O parâmetro C é que determina a violação à margem, e ao hiperplano, que será tolerada. Por violação ao hiperplano entenda-se classificar incorretamente uma observação. Se C = 0, nenhuma violação à margem é permitida, e então temos o chamado hiperplano de margem máxima. Obviamente este caso só ocorrerá se duas classes são perfeitamente separáveis. Se C aumenta, então um maior número de observações poderá estar do lado incorreto da margem de separação (VAPNIK, 1998).

Uma característica importante do problema de otimização apresentado é que somente as observações que estão do lado incorreto da margem de separação influenciam na obtenção do hiperplano. Estas observações são chamadas de *support vectors* (VAPNIK, 1998).

Uma vez que apenas os *support vectors* afetam o SVC, o parâmetro C desempenha um papel de controle no perde-ganha viés-variância do referido classificador. Se C é baixo, temos poucas observações influenciando na obtenção do classificador e então, este terá predições com baixo viés, mas alta variância. Se C é grande, um maior número de observações afeta a obtenção do hiperplano de separação e assim o classificador resultante terá predições com menor variância, ao custo de um pouco de viés (VAPNIK, 1998).

A solução desse problema de otimização passa diretamente pela obtenção do produto interno das observações de treinamento. De maneira geral, o produto interno entre duas observações $x_i, x_{i'}$ é dado por

$$\langle x_i, x_{i'} \rangle = \sum_{j=1}^{p} x_{ij} x_{i'j} .$$

Com base nisso, temos que o linear *support vector classifier* pode ser escrito como

$$f(x) = \beta_0 + \sum_{i=1}^{n} \alpha_i \langle x, x_i \rangle ,$$

onde cada observação de treinamento está associada a um parâmetro α. É importante ressaltar que no processo de estimação dos parâmetros são usados os produtos internos $\langle x_i, x_{i'} \rangle$ entre todos os pares de observações de treinamento e que a expressão $\langle x, x_i \rangle$ representa o produto interno entre uma nova observação de teste x e as observações de treinamento x_i (JAMES et al., 2013).

O *support vector classifier* descrito é adequado para casos onde a fronteira de separação entre as classes é linear no espaço de entrada. No entanto, na maior parte das vezes a fronteira de separação entre as classes é não linear e então toda técnica linear de classificação terá um fraco desempenho. Assim como acontece com outros métodos lineares, nós podemos tornar o classificador linear mais flexível expandindo os valores possíveis para as variáveis de entrada utilizando funções polinomiais e *splines*, entre outras (VAPNIK, 1998).

A máquina de vetor de suporte (SVM) é uma generalização do *support vector classifier* para resolver problemas onde a não linearidade está presente. A ideia central deste método é expandir o espaço de entrada das variáveis explicativas usando *kernels*. Através desta estratégia, o método consegue acomodar a não linearidade presente em problemas mais complexos (VAPNIK, 1998).

Combinando SVC com uma função kernel tem-se o classificador SVM. Este assume a forma $f(x) = \beta_0 + \sum_{i=1}^{n} \alpha_i K(x, x_i)$. Na expressão, $K(x, x_i)$ é uma função que é referida como *kernel*.

Uma *kernel*, de maneira bem simples, pode ser vista como uma função que quantifica a similaridade entre dois vetores de observações (JAMES et al., 2013). Três funções *Kernel* bastante utilizadas na literatura de SVM são (HASTIE et. al., 2009):

$$\text{Polinomial de grau d}: K(x_i, x_{i'}) = (1 + \sum_{j=1}^{p} x_{ij} x_{i'j})^d$$

$$\text{Radial basis}: K(x_i, x_{i'}) = \exp(-\gamma \sum_{j=1}^{p} (x_{ij} - x_{i'j})^2)$$

$$\text{Neural network}: K(x_i, x_{i'}) = \tanh(k_1 \langle x_i, x_{i'} \rangle + k_2).$$

Destas, a mais amplamente utilizada, e que também foi utilizada neste trabalho é a *kernel* radial basis. Neste caso, o desempenho do classificador SVM dependerá principalmente de dois parâmetros, o gamma (γ) e o parâmetro de ajuste C. Como já mencionado, C é um parâmetro de penalização que controla o perde-ganha viés-variância. Já o parâmetro γ representa a expansão no espaço de entrada dos preditores (LIU, 2013). O valor ótimo de γ foi escolhido dentre os seguintes valores, 0.01; 0.1; 0.5; 1; 2 e 3. Na otimização do parâmetro C a escolha foi entre os valores, 0.1; 1; 5, 10, 50 e 100.

Ao utilizar o classificador SVM, uma observação de teste qualquer x^*, ou seja, um vetor com valores de NC, DC e AC, é classificada em uma determinada classe de acordo o sinal de $f(x^*)$ (JAMES et al., 2013).

2.9. Avaliação e comparação das técnicas de classificação

Para o problema em questão, onde existem duas classes possíveis, selecionada ou não selecionada, qualquer regra de classificação está sujeita a dois possíveis erros:

i) selecionar uma família que não deveria ser selecionada; e

ii) não selecionar uma família que deveria ser selecionada.

Para comparação e avaliação dos classificadores é de nosso interesse obter estas medidas de erro. Um caminho natural para isso é utilizar a chamada matriz de confusão.

A matriz de confusão oferece uma medida efetiva do classificador utilizado ao mostrar o número de classificações corretas usando o método considerado ideal (seleção via TCHr) *versus* o número de classificações preditas pelo classificador em avaliação (LR, LDA, QDA, KNN, ANN, RF e SVM). Um esquema para a matriz de confusão é mostrado na Tabela 2.

Tabela 2: Esquema geral de uma matriz de confusão

	Predito		
Real	NS	S	Total
NS	VN	FP	N_2
S	FN	VP	P_2
Total	N_1	P_1	N

*NS= Não Selecionada; S=Selecionada; VN= Verdadeiro Negativo; FP= Falso Positivo; FN=Falso negativo; VP= Verdadeiro Positivo. O Real indica a seleção via o método tido como ideal (tonelada de cana por hectare real) e o predito mostra a seleção via um método de classificação alternativo (LR, LDA, QDA, KNN, ANN, RF e SVM).

Várias são as medidas que se originam da matriz de confusão para avaliação do classificador. A Tabela 3 mostra algumas destas medidas.

Tabela 3. Medidas de interesse para avaliação de classificadores derivadas das quantidades da Tabela 2.

Nome	Definição	Sinônimos
Taxa de falsos positivos	FP/N2	Erro tipo 1; Especificidade^C
Taxa de verdadeiros positivos	VP/P2	Erro tipo 2^C; Poder; Sensibilidade
Taxa de erro aparente	(FN+FP)/N	Acurácia^C

*O sobrescrito C indica o complemento.

Especificamente para o problema de seleção em cana-de-açúcar estamos interessados na taxa de erro aparente, que fornece o número de famílias que foram classificadas incorretamente pelo método de classificação considerado. Outra medida de interesse é a taxa de verdadeiros positivos, que nos informa a porcentagem de plantas que foram selecionadas corretamente pelo método de classificação estudado.

A taxa de falsos positivos, que representa as famílias que foram selecionadas incorretamente pela técnica de classificação, apesar de ser uma medida de erro, não representa um problema na prática. Isso porque famílias que foram selecionadas inadequadamente podem ser descartadas nas fases posteriores do programa.

Todas as análises foram feitas com o auxílio do software R (R Core Team, 2015). A Tabela 4 mostra as funções e pacotes utilizados na modelagem de cada um dos classificadores.

Tabela 4. Funções e pacotes utilizados para a modelagem dos classificadores regressão logística (LR), análise discriminante linear (LDA), análise discriminante quadrática (QDA), K-nearest neighboor (KNN), rede neural artificial de única camada intermediária (ANN), máquina de vetor de suporte (SVM) e árvores de decisão com *random forests* (RF).

Método	Função	Pacote
LR	`glm`	`Stats`
LDA	`lda`	`MASS`
QDA	`qda`	`MASS`
KNN	`knn`	`Class`
ANN	`nnet`	`Nnet`
SVM	`tune`	`e1071`
RF	`randomForest`	`randomForest`

3. Resultados e Discussão

Não há uma diferença marcante entre as variáveis NC, DC, AC e TCHr, nos cinco diferentes experimentos utilizados (Figura 2). De fato, o teste de Box (BOX, 1949) é não significativo ($\chi^2 = 48{,}992\,;\, \text{p} - \text{valor} = 0{,}1558$) para a igualdade das matrizes de covariâncias de NC, DC, AC, TCHr nos cinco experimentos. Isto justifica a utilização de qualquer um dos experimentos para simulação e ajuste das regras de classificação.

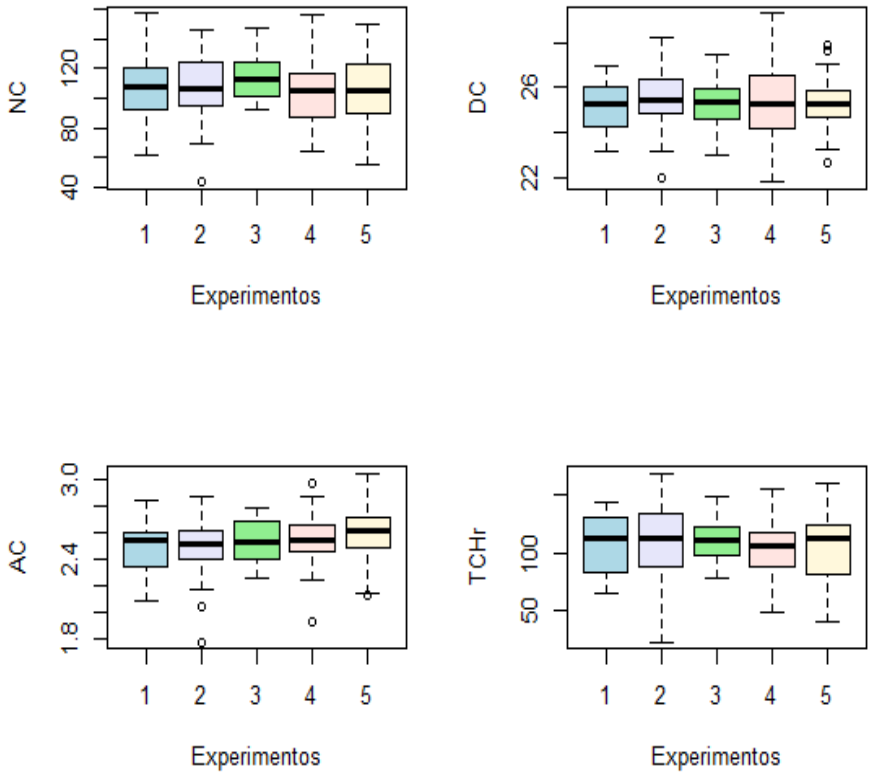

Figura 2. Boxplot para o número de colmos (NC), diâmetro médio de colmos (DC), em centímetros, altura média de colmos (AC), em metros, e tonelada de cana por hectare real (TCHr), nos cinco experimentos em estudo.

Todas as técnicas de aprendizado utilizadas, LR, LDA, QDA, KNN, ANN, SVM e RF, fornecem baixo valor para a taxa de erro aparente (AER < 0.14), quando o modelo completo (Modelo 1) é utilizado. Isto é, estes métodos apresentam alta concordância com o método de seleção tido como ideal, aplicado em função da variável TCHr (Tabela 5, Figura 3).

A grande vantagem de utilizar uma técnica de aprendizado para seleção entre famílias de cana-de-açúcar é que seria necessária a pesagem de apenas uma pequena parte do material. Neste estudo, com a pesagem de apenas um dos cinco experimentos (experimento de treinamento) houve uma excelente generalização para os quatro experimentos restantes (experimentos de teste). Assim, é evidente que tal estratégia poderia reduzir consideravelmente o trabalho de pesagem no campo, otimizando o processo de seleção.

Tabela 5. Taxa de erro aparente (AER) e taxa de verdadeiros positivos (TPR) para os classificadores, regressão logística (LR), análise discriminante linear (LDA), análise discriminante quadrática (QDA), K-nearest neighboor (KNN), rede neural artificial de única camada intermediária (ANN), máquina de vetor de suporte (SVM) e árvores de decisão com *random forests* (RF) em cinco diferentes cenários e em dois modelos (completo, ou modelo 1, com todos os preditores, número de colmos (NC), diâmetro de colmos (DC) e altura de colmos (AC), e reduzido, ou modelo 2, sem o preditor altura dos colmos (AC)).

			Cenários						
	Método	Modelo	1	2	3	4	5	Média	EP
AER	LR	1	0.125	0.125	0.1477	0.1477	0.1023	0.1295	0.0085
		2	0.125	0.0909	0.125	0.1364	0.1136	0.1182	0.0077
	LDA	1	0.125	0.125	0.1591	0.1477	0.1363	0.1386	0.0066
		2	0.1136	0.0909	0.125	0.125	0.1136	0.1136	0.0062
	QDA	1	0.125	0.125	0.1591	0.1477	0.0909	0.1295	0.0117
		2	0.1136	0.0909	0.125	0.125	0.1023	0.1114	0.0066
	KNN	1	0.1136	0.1136	0.1250	0.1364	0.0909	0.1159	0.0075
		2	0.1136	0.0795	0.1136	0.1023	0.0909	0.1000	0.0066
	ANN	1	0.0909	0.1023	0.125	0.1136	0.0909	0.1045	0.0066
		2	0.1023	0.0795	0.1136	0.1023	0.0682	0.0932	0.0084
	SVM	1	0.0909	0.0795	0.1023	0.0909	0.0795	0.0886	0.0043
		2	0.0909	0.1023	0.1023	0.1023	0.0909	0.0977	0.0028
	RF	1	0.1250	0.1023	0.1364	0.1364	0.0795	0.1159	0.0110
		2	0.1136	0.0795	0.1364	0.125	0.1136	0.1136	0.0095
TPR	LR	1	0.9149	0.898	0.8571	0.9167	0.9149	0.9003	0.0113
		2	0.8936	0.898	0.898	0.898	0.8723	0.892	0.0049
	LDA	1	0.9149	0.898	0.8367	0.9167	0.9149	0.8962	0.0152
		2	0.9149	0.8979	0.8979	0.9167	0.8723	0.8999	0.008
	QDA	1	0.9149	0.898	0.8367	0.9167	0.9574	0.9047	0.0196
		2	0.9149	0.898	0.898	0.898	0.8936	0.9005	0.0037
	KNN	1	0.9362	0.9184	0.898	0.9167	0.9574	0.9253	0.0101
		2	0.8936	0.9184	0.898	0.9375	0.9362	0.9167	0.0092
	ANN	1	0.9362	0.898	0.9184	0.9167	0.8936	0.9126	0.0077
		2	0.9167	0.9184	0.8979	0.9583	0.9362	0.921	0.0073
	SVM	1	0.9574	1	1	0.9782	0.9787	0.9831	0.0079
		2	0.9787	0.9592	0.9184	0.9375	0.9149	0.9417	0.0122
	RF	1	0.8936	0.898	0.8776	0.9167	0.9149	0.9001	0.0072
		2	0.8936	0.9388	0.8571	0.8958	0.8723	0.8915	0.0138

*EP=Erro Padrão

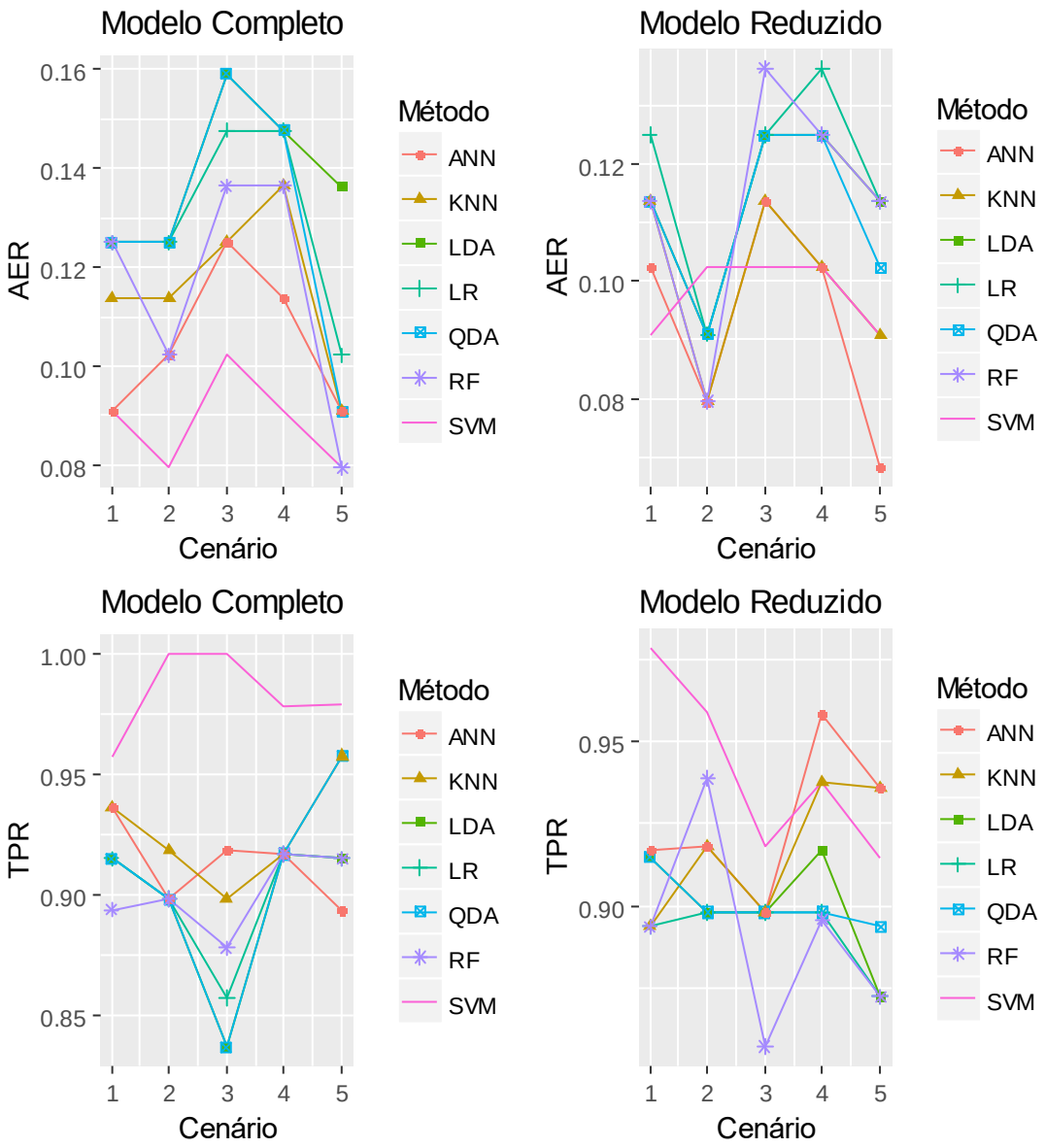

Figura 3. Taxa de erro aparente (AER) e taxa de verdadeiros positivos (TPR) para os classificadores, regressão logística (LR), análise discriminante linear (LDA), análise discriminante quadrática (QDA), K-nearest neighboor (KNN), rede neural artificial de única camada intermediária (ANN), máquina de vetor de suporte (SVM) e árvores de decisão com *random forests* (RF) em cinco diferentes cenários e em dois modelos (completo, com todos os preditores, número de colmos (NC), diâmetro de colmos (DC) e altura de colmos (AC) e reduzido, sem o preditor altura dos colmos (AC)).

Ainda que todas as técnicas de classificação apresentem bons resultados com relação à AER, vale destacar o melhor desempenho médio do classificador SVM. Para este classificador, a AER média foi 0.0886 no modelo completo, o que representa uma concordância, ou acurácia, de 91.14% com a seleção via TCHr (Tabela 5, Figura 4). Note também que o classificador SVM apresenta a menor AER em todos os cenários, quando o modelo completo é utilizado (Tabela 5, Figura 3). Para os classificadores LR, LDA, QDA, KNN, ANN e RF as AER médias foram de 0.1295, 0.1386, 0.1295, 0.1159, 0.1045 e 0.1159 respectivamente, quando da utilização do modelo completo (Tabela 5, Figura 4).

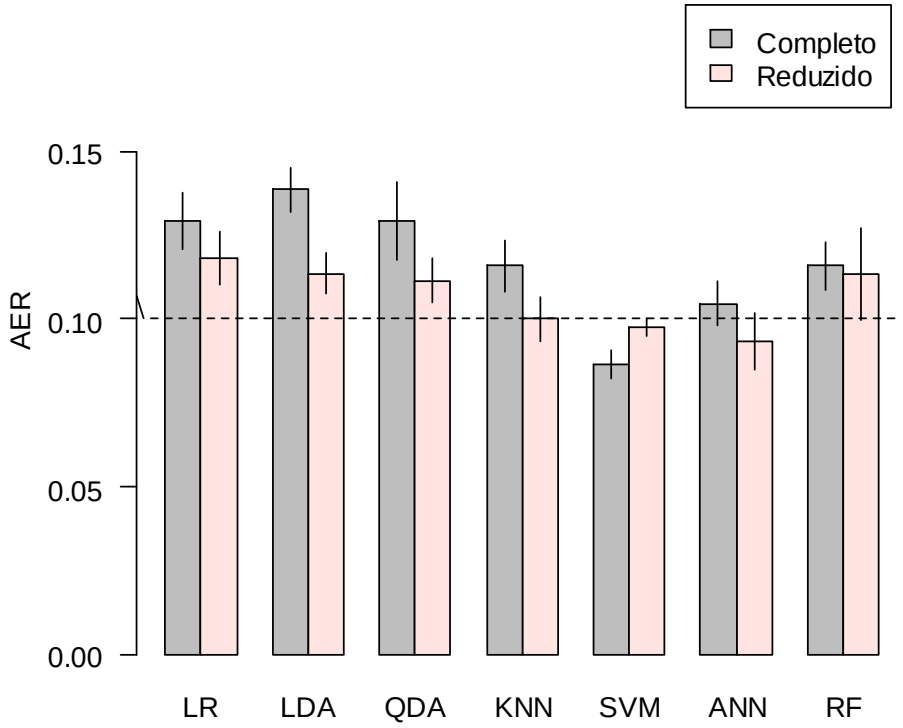

Figura 4. Taxa de erro aparente (AER) média e erro padrão para os classificadores regressão logística (LR), análise discriminante linear (LDA), análise discriminante quadrática (QDA), K-nearest neighboor (KNN), máquina de vetor de suporte (SVM), rede neural artificial de única camada intermediária (ANN) e árvores de decisão com *random forests* (RF) em dois modelos (completo, com todos os preditores e reduzido, sem o preditor altura de colmos). A linha pontilhada mostra o limite de acurácia igual a 90%.

Similarmente a estes resultados, diversos autores mostram a superioridade do classificador SVM nas mais diferentes áreas. Karimi et. al (2006) afirmam que o classificador SVM possui maior acurácia em problemas de agricultura de precisão quando comparado ao ANN. Ogutu et. al (2010) relatam a maior acurácia do classificador SVM quando comparado ao RF em seleção genômica. Maroco et. al (2011), em um estudo sobre predição de demência, mostram a superioridade, na acurácia, do SVM em relação aos classificadores LR, LDA, QDA, ANN e RF. Modaresi e Araghinejad (2014) mostram que o SVM, em um problema de classificação de qualidade de água, é superior, em termos de acurácia, aos classificadores ANN e KNN.

Com relação à TPR, todos os classificadores também apresentam bons resultados, quando da utilização do modelo completo. Novamente, vale destacar o desempenho do classificador SVM. Este apresenta maior valor para a TPR em todos os cenários (Tabela 5, Figura 3). A TPR média para este classificador foi de 0.9831. Ou seja, em média, 98.31% das famílias que foram selecionadas pelo método tido como ideal foram também selecionadas pelo classificador SVM (Tabela 5, Figura 5). Para os classificadores LR, LDA, QDA, KNN, ANN e RF os valores médios para a TPR foram respectivamente 0.9003, 0.8962, 0.9047, 0.9253, 0.9126 e 0.9001 (Tabela 5, Figura 5).

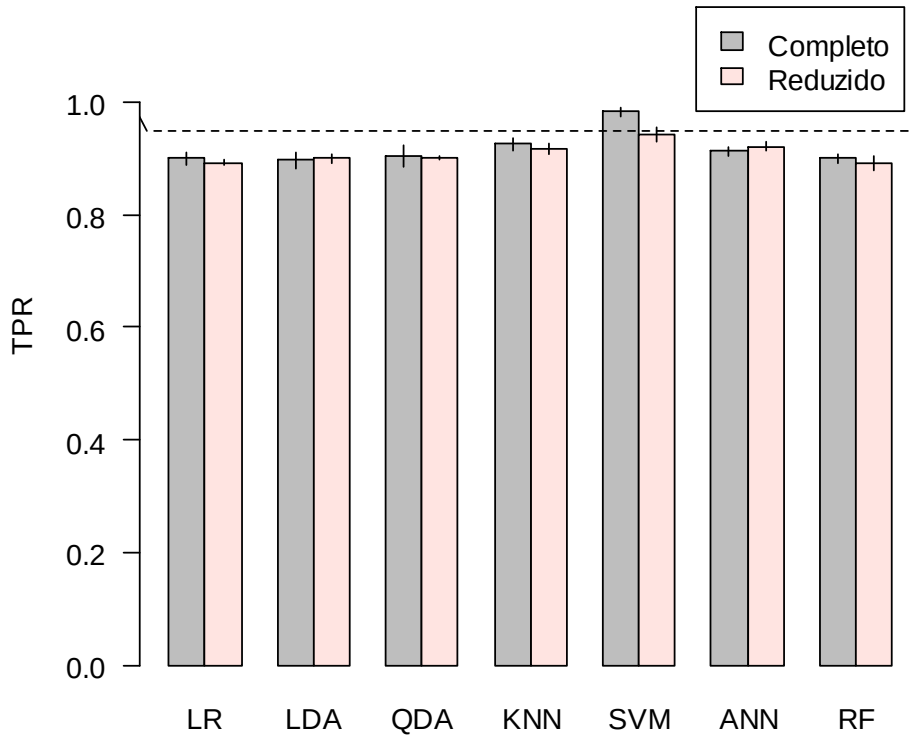

Figura 5. Taxa de verdadeiros positivos (TPR) média e erro padrão para os classificadores, regresão logística (LR), análise discriminante linear (LDA), análise discriminante quadrática (QDA), K-nearest neighboor (KNN), máquina de vetor de suporte (SVM), rede neural artificial de única camada intermediária (ANN) e árvores de decisão com *random forests* (RF) em dois modelos (completo, com todos os preditores e reduzido, sem o preditor altura de colmos). A linha pontilhada mostra o limite para a TPR igual a 95%.

Para o modelo reduzido, diferentemente do que se poderia pensar, os valores médios para a AER, em todos os classificadores, com exceção do SVM, foram ainda menores quando comparados aos obtidos no modelo completo (Tabela 5, Figura 4). Os melhores desempenhos, neste caso, foram observados para os classificadores ANN, SVM e KNN. Para estes classificadores as AER médias foram respectivamente 0.0932, 0.0977 e 0.1000, o que representa uma acurácia média igual ou superior aos 90% (Tabela 5, Figura 4). Os demais classificadores, LR, LDA, QDA e RF apresentaram respectivamente AER médias de 0.1182, 0.1136, 0.1114 e 0.1136 (Tabela 5, Figura 4).

Este resultado fornece indícios que a variável AC, ainda que seja considerada um componente de produção, pode ser desconsiderada quando da seleção das melhores famílias de cana-de-açúcar.

Com relação à TPR, quando o modelo reduzido é utilizado, os classificadores também apresentam bom desempenho (TPR > 0.89). Da mesma maneira que para a AER, o melhor desempenho é observado para o classificador SVM. A TPR média para este classificador foi de 0.9417, isto é, 94.17% das famílias que foram selecionadas via TCHr, foram também selecionadas pelo classificador SVM. Nos demais classificadores, LR, LDA, QDA, KNN, ANN e RF, as TPR médias foram, respectivamente, 0.8920, 0.899, 0.9005, 0.9167, 0.9210 e 0.8915 (Tabela 5, Figura 5).

Os bons resultados para a AER e a TPR, nas sete técnicas de classificação utilizadas, tanto no modelo completo quanto no modelo reduzido, nos mostram que tais técnicas poderiam ser utilizadas com sucesso na seleção entre famílias de cana-de-açúcar em ambos os modelos.

No entanto, devido a maior facilidade operacional, a utilização do modelo reduzido pode ser preferida. Neste caso, os classificadores com melhores desempenho são o SVM, o ANN e o KNN.

Em comparação aos classificadores LR, LDA e QDA acreditamos que o melhor desempenho do SVM, ANN e KNN se dê em razão da fronteira de separação entre as classes. Vale lembrar que um bom desempenho dos classificadores LR, LDA e QDA é esperado para problemas linearmente separáveis. Os classificadores SVM, ANN, KNN e RF são procedimentos não paramétricos e se adaptam a uma amplitude maior de problemas.

Quando comparada ao classificador RF o desempenho superior observado nos classificadores SVM, ANN e KNN pode ser atribuído, possivelmente, há uma maior habilidade na capacidade de extrair combinações lineares das variáveis de entrada, como sugere a Tabela 6.

Hastie et. al (2009) destaca que não há um método que é sempre melhor em todas as situações. O melhor método dependerá sempre de características particulares do conjunto de dados em estudo. Algumas características dos dados que nos permitem entender o desempenho dos classificadores SVM, ANN, KNN e RF são apresentadas na Tabela 6.

Tabela 6. Desempenho dos classificadores rede neural artificial (ANN), máquinas de vetor de suporte (SVM), k-nearest neighboors (KNN) e árvores de decisão com random forests (RF) de acordo com algumas características do conjunto de dados.

Características	ANN	SVM	KNN	RF
Desempenho na presença de diferentes tipos de variáveis no espaço de entrada	Ruim	Ruim	Ruim	Bom
Desempenho na presença de parcelas perdidas	Ruim	Ruim	Bom	Bom
Desempenho na presença de "outliers" no espaço de entrada	Ruim	Ruim	Bom	Bom
Habilidade para trabalhar com entradas irrelevantes	Ruim	Ruim	Ruim	Bom
Habilidade para extrair combinações lineares das variáveis de entrada	Bom	Bom	Regular	Ruim
Poder de predição	Bom	Bom	Bom	Bom

*Tabela adaptada de Hastie et. al (2009).

De fato, diferentemente do que sugere os resultados deste trabalho, alguns autores relatam a superioridade do RF quando comparado aos classificadores SVM e ANN. Alguns exemplos são Meher et. al (2016), em um estudo para predição de locais de *splicing* em arroz, e Liu et al. (2013), em um estudo com *eletronic tongue* (E-Tongue). Isto reforça a ideia de que, de maneira geral, o desempenho da técnica de classificação está atrelada a características intrínsecas do conjunto de casos.

É importante destacar ainda que o desempenho das técnicas de classificação utilizadas é dependente dos parâmetros utilizados nestas. A alteração destes parâmetros, ou mesmo da estrutura da técnica, levam a resultados distintos. Os parâmetros ótimos, obtidos neste trabalho, para os classificadores ANN, SVM, RF e KNN são apresentados na Tabela 7.

Tabela 7. Parâmetros ótimos para os classificadores, máquina de vetor de suporte (SVM), rede neural artificial de única camada intermediária (ANN), árvores de classificação com random forests (RF) e k-nearest neighboors (KNN) em cinco diferentes cenários e em dois modelos, modelo 1 ou completo (com todos os preditores, número de colmos (NC), diâmetro de colmos (DC) e altura de colmos (AC)) e modelo 2 ou reduzido (sem o preditor altura de colmos (AC)).

			Cenários				
Método	Parâmetros	Modelo	1	2	3	4	5
ANN	z	1	24	16	8	11	8
		2	7	19	24	14	8
	w	1	121	81	41	56	41
		2	29	77	97	57	33
SVM	C	1	1	5	5	1	5
		2	5	5	50	10	10
	γ	1	1	0.01	0.01	0.1	0.01
		2	0.01	0.01	0.01	0.01	0.01
RF	n	1	500	250	250	250	500
		2	250	500	1	2000	250
	m	1	p	$\sqrt{p}$	$p;\sqrt{p}$	$p;\sqrt{p};p/2$	$\sqrt{p}$
		2	$\sqrt{p}$	$p;\sqrt{p};p/2$	$\sqrt{p}$	$p/2$	$p/2$
KNN	K	1	12	4	8	3	12
		2	7	4	7	8	4

* z = número de neurônios na camada intermediária da rede ou oculta; w = número de parâmetros da rede ou pesos; C = parâmetro de ajuste do SVM; γ = parâmetro de penalização da kernel radial basis; n = número de árvores que foram combinadas no RF; m = número de preditores que foram considerados a cada passo da construção da árvore de decisão; p = número de preditores considerados, isto é, 3 no modelo completo e 2 no modelo reduzido e K = número de observações de treinamento mais próximas a uma dada observação (ou vetor de observações como neste caso) de teste.

Na prática, para utilização de uma técnica de aprendizado estatístico, como apresentamos neste trabalho, é necessário a pesagem de apenas parte das famílias do experimento. No trabalho utilizamos 20% das famílias para o ajuste das regras de classificação. Uma vez construído o método de aprendizado ele é aplicado às demais famílias. Nestas demais famílias, obviamente, necessitamos das informações do NC, DC e AC, se usado o modelo completo, e do NC e DC, caso do modelo reduzido. Assim, estamos assumindo que, em condições de difícil mecanização da colheita, tal coleta de dados é mais prática de ser realizada e que nesta situação o trabalho de campo é otimizado.

4. Conclusões

Todos os classificadores apresentam baixos valores para a taxa de erro aparente (AER) média (AER < 0.14) e altos valores para à taxa de verdadeiros positivos (TPR) média (TPR > 0.87) tanto no modelo completo, que contém os preditores, número de colmos, diâmetro de colmos e altura de colmos, quanto no modelo reduzido, que dispensa o preditor altura de colmos. Para o modelo completo, o melhor desempenho, menor taxa de erro aparente média (0.0886) e maior taxa média de verdadeiros positivos (0.9831), foi observado no classificador máquina de vetor de suporte. No modelo reduzido, os melhores resultados médios foram obtidos para os classificadores rede neural artificial de única camada intermediária (AER=0.0932; TPR=0.9210), máquina de vetor de suporte (AER=0.0977; TPR=0.9417) e k-nearest neighboor (AER=0.1000, TPR=0.9167). O modelo reduzido deve ser preferido, pois apresenta desempenho muito próximo do modelo completo e tem como vantagem o fato de ser operacionalmente mais simples. Na prática, qualquer técnica de classificação avaliada poderia ser usada de maneira satisfatória nas fases iniciais do programa de melhoramento de cana-de-açúcar.

5. Referências bibliográficas

BARBOSA, M.H.P.; SILVEIRA, L.C.I. (2015) Breeding Program and Cultivar Recomendations. In: Santos, F.; Borém, A. and Caldas, C. Editors. Sugarcane: Agricultural Production, Bioenergy, and Ethanol. Elsevier, 592 p.

BISHOP, C.M. (2006) Pattern Recognition and Machine Learning. Springer, New York, 729p.

BOX, G. E. P. (1949) A general distribution theory for a class of likelihood criteria. Biometrika, 36**:**317-346

BRASIL. Ministério da agricultura pecuária e abastecimento (2016). Disponível em:<<http://www.agricultura.gov.br/vegetal/culturas/cana-de-acucar>> Acesso em: 28/03/2016.

BRASILEIRO B.P.; MARINHO, C.D.; COSTA, P.M.A, CRUZ, C.D.; PETERNELLI, L.A.; BARBOSA, M.H.P. (2015) Selection in sugarcane families with artificial neural network. Crop Breeding and Applied Biotecnology, 15: 72-78.

BREIMAN, L.; FRIEDMAN, J. H.; OLSHEN, R.; STONE, C. J. (1984) Classification and Regression Trees. Wadsworth, California, 368p.

CASTRO, R.D. ; PETERNELLI, L.A. ; RESENDE, M.D.V. ; MARINHO, C.D. ; COSTA, P.M.A. ; BARBOSA, M.H.P. ; MOREIRA, E.F.A. (2016) Selection between and within full-sib sugarcane families using the modified BLUPIS method (BLUPISM). Genetics and Molecular Research, v. 15, p. gmr.15017334.

CHANG, Y.S.; MILLIGAN S.B. (1992) Estimating the potential of sugarcane families to procude elite genotypes using univariate cross prediction methods. Theoretical and Applied Genetics, 84: 662-671.

CONAB (2016). Companhia Nacional de Abastecimento. Acompanhamento da safra brasileira de cana-de-açúcar. Safra 2016/17.<<Disponível em: http://www.conab.gov.br/OlalaCMS/uploads/arquivos/16_12_27_16_30_01_boletim_cana_portugues_-3o_lev_-_16-17.pdf >> Acesso em: 14/02/2017.

CRESSIE, N. A. C. (1993) Statistics for Spatial data. Jonh Wiley & Sons. 900p.

DUDA, R.O.; HART, P.E.; STORK, D.G (2000) Pattern Classification. Jonh Wiley & Sons. 680p

ESPÓSITO D.P.; PETERNELLI L.A.; PAULA T.O.M.; BARBOSA M.H.P. (2012) Análise de trilha usando valores fenotípicos e genotípicos para componentes do rendimento na seleção de famílias de cana-de-açúcar. Ciência Rural, 42: 38-44

FALCONER D.S.; MACKAY T.F.C. (1996) Introduction to Quantitative Genetics. Malaysia: Longmans Green, 463p.

FERREIRA, D. F. (2011) Estatística Multivariada. 2.ed, Lavras: Ed. UFLA, 675p.

GRINBERG, N.F.; LOVATT, A.; HEGARTY, M.; LOVATT, A.; SKOT, K.P.; KELLY, R.; BLACKMORE, T.; THOROGOOD, D.; KING, R.D.; ARMSTEAD, I; POWELL, W.; SKOT, L. (2016) Implementation of Genomic Prediction in *Lolium perenne* (L.) Breeding Populations. Frontiers in Plant Sciense, 7: 133.

HAYKIN, S. (2001) Redes neurais princípios e prática. Porto Alegre: Bookman, 900p.

HAINING, R (2005) Spatial data analysis – theory and practice. Cambridge University Press. 432p.

HASTIE, T.; TIBSHIRANI, R.; FRIEDMAN J. (2009) The elements of statistical learning: data mining, inference, and prediction. Springer, New York, 745p.

HESLOT, N; YANG, H.P.; SORRELLS, M.E.; JANNINK, J.L. (2012) Genomic Selection in Plant Breeding: A Comparasion of Models, Crop Sciense, 52: 146-160.

HOGARTH, D. M.; COX, M. C.; BULL, J. K (1997) Sugarcane improvement: Past achievements and future prospects. In: Kang, M. S. Crop improvement for the 21st century. Baton Rouge: Louisiana State University, p. 29-56.

JAMES, G.; WITTEN, D.; HASTIE, T.; TIBSHIRANI, R. (2013) An introduction to statistical learning: with aplications in R. Springer, New York, 426p.

KARIMI, Y.; PRASHER, S.O.; PATEL, R.M; KIM, S.H. (2006) Application of support vector machine technology for weed and nitrogen stress detection in corn. Computers and Electronics in Agriculture 51: 99–109.

KIMBENG, C.A.; COX, M.C. (2003) Early generation selection of sugarcane families and clones in Australia: a review. Journal American Society of Sugarcane Technologists. 23:20-39.

LIU, M.; WANG, M.; WANG, J.; LI, D. (2013) Comparison of random forest, support vector machine and back propagation neural network for electronic tongue data classification: Application to the recognition of orange beverage and Chinese vinegar. Sensors and Actuators, 177: 970-980.

MATSUOKA, S.; GARCIA, A.A.F.; ARIZONO, H. (2005) Melhoramento da cana-de-açúcar. In: BORÉM, A. (Ed.) Melhoramento de espécies cultivadas. Viçosa: Ed. da UFV. 969p.

MAROCO, J.; SILVA, D.; RODRIGUES, A.; GUERREIRO, M.; SANTANA, I.; MENDONÇA, A. (2011) Data mining methods in the prediction of Dementia: A real-data comparison of the accuracy, sensitivity and specificity of linear discriminant analysis, logistic regression, neural networks, support vector machines, classification trees and random forest. BMC research notes, 4: 299.

MEHER, P. K.; SAHU, T. K.; RAO, A. R. (2016) Performance evaluation of neural network, support vector machine and random forest for prediction of donor splice sites in rice. The Indian Journal of Genetics. 76(2): 173-180

MOREIRA, E. F. A.; PETERNELLI, L. A. (2015) Sugarcane families selection in early stages based on classification by linear discriminant analysis. Revista Brasileira de Biometria, 33(4): 484-493.

MODARESI, F.; ARAGHINEJAD, S. (2014) Water Resour Manage, 28: 4095-4111.

OGUTU, J. O.; PIEPHO, H. P.; STREECK, T. S. (2011) A comparison of random forests, boosting and support vector machines for genomic selection. BMC Proceedings, 5(Suppl 3):S11.

OLIVEIRA, R. A.; DAROS, E.; RESENDE, M. D. V.; BESPALHOK-FILHO, J. C.; ZAMBON, J. L. C.; SOUZA, T. R.; LUCIUS, A. S. F. (2011) Procedimento BLUPIS e seleção massal em cana-de-açúcar. Bragantia, 70(4): 796-800.

PETERNELLI, L. A; RESENDE, M. D. V.; MENDES, T. O. P. (2011) Experimentação e análise estatística em cana-de-açúcar. In: Fernando Santos; Aluizio Borém; Celso Caldas. (Org.). Cana-de-açúcar. 2ed.Viçosa: Folha de Viçosa Ltda, v. 1, p. 333-353.

R Development Core Team (2015) R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria (http://www.r-project.org).

RESENDE, MDV (2002) Genética biométrica e estatística no melhoramento de plantas perenes. Brasília: Embrapa Informação Tecnológica, 975p.

RESENDE, M.D.V.; BARBOSA, M.H.P. (2006) Selection via simulated Blup base on family genotypic effects in sugarcane. Pesquisa Agropecuária Brasileira, 41(3): 421-429.

SANTOS, A. C.; FERREIRA, D. F. (2003) Definição do tamanho amostral usando simulação Monte Carlo para o teste de normalidade baseado em assimetria e curtose. II. Abordagem Multivariada. Ciência Agrotécnica, 27(1): 62-69.

STRINGER, J.K.; COX, M.C.; ATKIN, F.C.; WEI, X.; HOGARTH. (2011) Family Selection Improves the Efficiency and Effectiveness of Selecting Original Seedlings and Parents. Sugar Tech, 13(1):36-41.

VAPNIK, V. N. (1998) Statistical learning theory. Wiley, New York, 768p.

VENABLES W.N.; RIPLEY B.D. (2002) Modern applied statistics with S. Springer, New York, 493p.

WACLAWOVSKY, A. J.; SATO, P.M.; LEMBKE, C.G.; MOORE, P.H.; SOUZA, G.M. (2010). Sugarcane for bioenergy production: an assessment of yield and regulation of sucrose content. Plant Biotechnology Journal, 8(3): 263-276.

ZHANG, L; LUO, L. (2003) Splice site prediction with quadratic discriminant analysis using diversity measure, Nucleic Acids Research, 31 (21).

ZHOU M. M.; KINBENG C.A.; THEW T. L.; GRAVOIS K. A.; PONTIF M. J. (2011) Artificial Neural Network Models as a Decision Support Tool for Selection in Sugarcane: A Case Study Using Seedling Populations. Crop Science 51: 21:31.