UNIVERSIDADE CATÓLICA DOM BOSCO
PROGRAMA DE PÓS-GRADUAÇÃO *STRICTO SENSU* EM
CIÊNCIAS AMBIENTAIS E SUSTENTABILIDADE AGROPECUÁRIA

# Aprendizagem de Máquina Aplicada ao Melhoramento de Plantas

Autor: Leandro Skowronski
Orientador: Dr. Reginaldo Brito da Costa

Campo Grande
Mato Grosso do Sul
Fevereiro - 2020

UNIVERSIDADE CATÓLICA DOM BOSCO
PROGRAMA DE PÓS-GRADUAÇÃO *STRICTO SENSU* EM
CIÊNCIAS AMBIENTAIS E SUSTENTABILIDADE AGROPECUÁRIA

# Aprendizagem de Máquina Aplicada ao Melhoramento de Plantas

Autor: Leandro Skowronski
Orientador: Dr. Reginaldo Brito da Costa

Tese apresentada, como parte das exigências para obtenção do título de DOUTOR EM CIÊNCIAS AMBIENTAIS E SUSTENTABILIDADE AGROPECUÁRIA, no Programa de Pós-Graduação *Stricto Sensu* em Ciências Ambientais e Sustentabilidade Agropecuária da Universidade Católica Dom Bosco - Área de concentração: Sustentabilidade Ambiental e Produtiva Aplicada ao Agronegócio e Produção Sustentável

Campo Grande
Mato Grosso do Sul
Fevereiro - 2020

Dados Internacionais de Catalogação na Publicação (CIP)
Universidade Católica Dom Bosco
Bibliotecária Mourâmise de Moura Viana - CRB-1 3360

S628c Skowronski, Leandro
Aprendizagem de máquina aplicada ao melhoramento de plantas/ Leandro Skowronski; Orientador Prof. Dr. Reginaldo Brito da Costa..-- Campo Grande, MS : 2020.
66 f.: il.; 30 cm

Tese (doutorado em ciências ambientais e sustentabilidade agropecuária) - Universidade Católica Dom Bosco, Campo Grande, 2020
Inclui bibliografia da p. 20 até a p. 24

1. Inteligência artificial - Sustentabilidade agropecuária. 2. Algoritmos - Melhoramento genético. 3. Plantas - Melhoramento genético. I.Costa, Reginaldo Brito da. II. Título.

CDD: Ed. 21 -- 581.3

UNIVERSIDADE CATÓLICA DOM BOSCO
*Inspira o futuro*

## Aprendizagem de Máquina Aplicada ao Melhoramento de Plantas

**Autor**: Leandro Skowronski
**Orientador:** Prof. Dr. Reginaldo Brito da Costa

**TITULAÇÃO:** Doutor em Ciências Ambientais e Sustentabilidade Agropecúaria
**Área de Concentração:** Sustentabilidade Ambiental e Produtiva.

APROVADO em 13 de fevereiro de 2020.

Prof. Dr. Reginaldo Brito da Costa - UCDB

Prof. Dr. Michel Ângelo Constantino de Oliveira - UCDB

Profa. Dra. Paula Martin de Moraes - UFGD

Prof. Dr. Mario Luiz Teixeira de Moraes - UNESP

Prof. Dr. Wesley Nunes Gonçalves - UFMS

# AGRADECIMENTOS

Agradeço primeiramente a Deus por sua direção, iluminando meus passos neste desafio, e por suas bênçãos.

À minha esposa e filhos pelo amor, carinho, incentivo, compreensão e orações.

Aos meus familiares pelo apoio e palavras de ânimo.

Aos amigos e companheiros de trabalho do NEPPI pelo apoio e orientações, especialmente ao Professor Antônio Brand (*in memorian*) pela convivência e incentivo à minha formação.

À Universidade Católica Dom Bosco pela oportunidade deste curso e pelo apoio mediante a bolsa de capacitação docente.

Aos professores e colegas do Programa de Pós-Graduação Stricto Sensu em Ciências Ambientais e Sustentabilidade Agropecuária pela contribuição na minha formação.

Ao Professor Reginaldo Brito da Costa, pela orientação, paciência e amizade, e a Dra. Paula Martins pela coorientação e companheirismo.

A todos que de alguma forma colaboraram nesta minha etapa de formação.

# SUMÁRIO

Página

# LISTA DE TABELAS

Página

# LISTA DE FIGURAS

Página

REVISÃO BIBLIOGRÁFICA



CAPÍTULO 1



CAPÍTULO 2

CAPÍTULO 3

# LISTA DE ABREVIATURAS

LDA – *Linear Discriminant Analysis*

AM - Aprendizagem de máquina

RNA - Redes neurais artificiais

DNA – Ácido desoxirribonucleico

BLUPIS - Melhor preditor imparcial linear individual

DArT - *Diversity Array Technology*

kNN - K-vizinhos mais próximos

RF - Floresta Aleatória

SVM - Máquina de Vetor de Suporte

RNA/MLP - Rede Neural Perceptron Multicamadas

MLP - Perceptron Multicamadas

DBSCAN - *Density-Based Spatial Clustering of Applications with Noise*

SOM - *Self-Organizing Maps*

APC - *Affinity Propagation Clustering*

SEQ - Soma dos erros quadráticos

WCSS – *Within Cluster Sum of Squares*

UPGMA - *Unweighted Pair Group Method using Arithmetic averages*

# RESUMO

A identificação da variabilidade genética em plantas tem grande importância para fins de melhoramento genético, nos estudos populacionais e na conservação da diversidade, sendo as abordagens multivariadas da estatística clássica utilizadas com frequência. Contudo, novas metodologias baseadas em aprendizagem de máquina (AM) têm-se mostrado promissoras. Assim, o objetivo deste trabalho foi avaliar o desempenho de técnicas de AM, tanto supervisionadas como não supervisionadas, frente a metodologias estatísticas convencionais, e desenvolver uma ferramenta computacional que facilite estes estudos. Para a comparação dos métodos de AM, foram utilizados dados fenotípicos simulados de populações com diferentes graus de similaridade genética. Os métodos de AM supervisionada testados para classificação foram: *Naive Bayes*, Árvore de decisão, K-vizinhos mais próximos (kNN), Floresta Aleatória (RF), Máquina de Vetor de Suporte (SVM) e Rede Neural Perceptron Multicamadas (RNA/MLP), sendo comparados com as técnicas de análise discriminante propostas por Anderson e por Fisher. Os métodos de AM não supervisionados testados para agrupamento foram: k-*Means*, *Density-Based Spatial Clustering of Applications with Noise* (DBSCAN), *Self-Organizing Maps* (SOM) e *Affinity Propagation Clustering* (APC), comparados com o método de agrupamento hierárquico de Ward, em dois cenários: com procedimento convencional de definição do número de grupos e com a indução da formação de 11 grupos. Os métodos de classificação baseados em algoritmos de AM se mostraram superiores às funções discriminantes de Fisher e de Anderson, permitindo obter alta acurácia em condições de maior similaridade entre as populações, sendo kNN, RF, SVM e Naive Bayes aqueles que apresentaram maior acurácia, superando os algoritmos de Árvore de Decisão e RNA/MLP. Para o teste de agrupamento, no primeiro cenário, os métodos de aprendizagem de máquina não supervisionados mostraram-se superiores ao método Ward. No segundo cenário os métodos k-Means e Ward mostraram-se melhores, e permitiu concluir que a utilização de outros métodos de estimação do número de grupos pode levar a resultados mais robustos. DBSCAN apresentou baixo índice de precisão nestas condições**,** com populações de alta similaridade, não conseguindo diferenciar os materiais mais próximos geneticamente. O método APC foi consistente nos dois cenários, apresentando bons índices, sendo considerado promissor por não necessitar de definição prévia do número de grupos, facilitando seu uso. O trabalho gerou também um produto, o software ICMGen que pode ser utilizado na avaliação de métodos de AM para classificação e agrupamento de genótipos de plantas, permitindo que pesquisadores da área do melhoramento, não familiarizados com a programação e com a linguagem R possam utilizar desta ferramenta para estes estudos. O ICMGen é gratuito e está disponível para download no site http://www.agroeco.com.br.

**Palavras-chave**: Aprendizagem de máquina, inteligência artificial, melhoramento de plantas, métodos supervisionados, métodos não supervisionados.

# ABSTRACT

The genetic variability identification in plants has great importance for genetic improvement, in population studies and in the diversity conservation, and multivariate approaches from classical statistics are frequently used. Nonetheless, new methodologies based on machine learning (ML) have shown to be promising to that purpose. Thus, the objective of the present study was to evaluate the performance of ML techniques, both supervised and unsupervised, against conventional statistical methodologies, and to develop a computational tool that facilitate these studies. To compare ML methods, simulated phenotypic data of populations with different degrees of genetic similarity were used. The supervised ML methods tested for classification were the following: Naive Bayes, Decision Tree, k-Nearest Neighbors (kNN), Random Forest (RF), Support Vector Machine (SVM) and Multi-layer Perceptron Neural Networks (MLP/ANN), which were compared with the discriminant analysis techniques proposed by Anderson and by Fisher. The unsupervised ML methods tested for grouping were the following: k-Means, Density-Based Spatial Clustering of Applications with Noise (DBSCAN), Self-Organizing Maps (SOM) and Affinity Propagation Clustering (APC). They were compared with the Ward grouping method in the following two scenarios: with a conventional procedure defining the number of groups and with the induction of the formation of 11 groups. Classification methods based on ML algorithms proved to be superior to Fisher's and Anderson's discriminant functions, allowing to obtain high accuracy in conditions of higher similarity between populations. The kNN, RF, SVM and Naive Bayes methods showed greater accuracy, surpassing the Decision Tree and MLP/ANN algorithms. For the grouping test, in the first scenario, unsupervised machine learning methods proved to be superior to the Ward method. In the second scenario the k-Means and Ward methods showed to be the best, and allowed concluding that the use of other methods of estimation for number of groups can lead to more accurate results. DBSCAN showed a low precision rate in these conditions, with high similarity populations not being able to differentiate genetically closer material. The APC method was consistent in the two scenarios, showing good rates. It was considered promising for not requiring previous definition of the number of groups, which facilitates its use. The work has generated a product, the ICMGen software, which can be used to evaluate ML methods for plant genotype classification and grouping, allowing that researchers of the improvement area, unacquainted with programming and with the R language, may use this tool for these studies. ICMGen is free of charge and is available for download in the website http://www.agroeco.com.br.



**Keywords**: Machine learning, artificial intelligence, plant breeding, supervised methods, unsupervised methods.

# INTRODUÇÃO

A atividade do melhoramento de plantas tem contribuído de forma decisiva para o desafio de produzir alimentos, com qualidade e quantidade, para suprir a demanda crescente da população mundial. Para isso, a atividade enfrenta desafios constantes em dispor de informações sobre a diversidade genética, com o objetivo de eleger os genótipos que potencialmente podem ser utilizados na obtenção de novas cultivares com as características almejadas. Estas informações são obtidas, na sua maioria, por meio de técnicas estatísticas tradicionais, que apresentam certas limitações, especialmente quanto a exigência de qualidade dos dados (Cruz *et al*., 2011).

Frente ao advento das novas metodologias computacionais, como a inteligência artificial, e em especial a área de aprendizagem de máquina, surge a possibilidade de mudança de paradigmas nas análises empregadas no melhoramento de planta, com a chance de integração destas novas técnicas com as convencionais, melhorando seus resultados, ou até mesmo substituindo estas. Contudo, os estudos da aplicação de métodos de aprendizagem de máquina no melhoramento vegetal são muito recentes, e ainda apresentam diversas lacunas que ainda inibem sua utilização (Cruz; Nascimento, 2018).

Assim, buscando contribuir para o melhor entendimento da aplicação da aprendizagem de máquina ao melhoramento, propõe-se no presente trabalho investigar o desempenho de diferentes métodos, alguns ainda não testados para esta finalidade, de modo a disponibilizar informações que possam justificar, ou não, a utilização destas técnicas.

A tese está estruturada da seguinte forma: o primeiro capítulo apresenta um trabalho de comparação de desempenho de algoritmos supervisionado de aprendizagem de máquina com as técnicas de análise discriminante tradicionais, na classificação de populações simuladas com diferentes graus de dificuldade de discriminação; no segundo capítulo é realizada a comparação do desempenho de algoritmos não supervisionado de aprendizagem de máquina com uma técnica de análise de agrupamento hierárquica clássica, no agrupamento de genótipos, utilizando

dados simulados; no terceiro capítulo apresenta-se uma ferramenta computacional que facilita o trabalho do melhorista em comparar e selecionar métodos de aprendizagem de máquina, tanto supervisionados como não supervisionados, para realizar as tarefas de classificação ou agrupamento de genótipos, de forma facilitada e sem a necessidade de aprendizado de linguagens de programação específicas, como as do Matlab e R.

# OBJETIVOS

## Objetivo Geral

O objetivo geral deste trabalho é avaliar o desempenho de algoritmos de aprendizagem de máquina, tanto supervisionadas como não supervisionadas, frente a metodologias estatísticas convencionais utilizadas no melhoramento de plantas, e desenvolver uma ferramenta computacional que facilite estes procedimentos.

## Objetivos Específicos

- Avaliar a eficácia dos algoritmos de aprendizagem de máquina:
  I. supervisionados: *Naive Bayes*, Árvore de decisão, K-vizinhos mais próximos, Floresta Aleatória, Máquina de Vetor de Suporte e Rede Neural Perceptron Multicamadas, na classificação e discriminação de populações com diferentes graus de similaridade, comparando-os com as técnicas de análise discriminante tradicionais, propostas por Anderson e por Fisher.
  II. não supervisionados: *k-Means, DBSCAN, Self-Organizing Maps e Affinity Propagation Clustering*, na formação de grupos de populações com diferentes graus de similaridade, comparando-os com técnica de análise de agrupamento hierárquico com método Ward.
- Desenvolver um aplicativo que facilite as análises, por meio das técnicas de aprendizagem de máquina, de dados do melhoramento de plantas.

# REVISÃO BIBLIOGRÁFICA

## Melhoramento de plantas

O melhoramento de plantas, impulsionado pelos desafios do crescimento populacional e a crescente preocupação da sociedade com o meio ambiente, tem contribuído de forma significativa na melhoria das atividades agrícolas. Dentre tantos objetivos, tomando apenas o aumento de produtividade, o melhoramento contribuiu para: aumento médio de 4% ao ano na produtividade de açúcar e 20,8% na produtividade de etanol em uma década (Ramalho *et al*., 2012); aumento de 10 $m^3$/ha/ano de produtividade na cultura do eucalipto (Squilassi, 2003), o que corresponde a 50% do ganho em produtividade total; para diversas outras culturas, como soja, milho, arroz, trigo, algodão, entre outras, têm-se conseguido aumentos de produtividade anuais por volta de 2% nos últimos anos (Borém, 2017).

Além dos ganhos financeiros provenientes do aumento de produtividade, esse aumento também tem o efeito positivo na redução do preço dos alimentos. Costa e Freitas (2006), avaliando a contribuição do melhoramento para a redução de preço dos alimentos, estimaram que o aumento da produtividade provocou um valor superior a um de proporcionalidade na redução dos preços de soja e trigo, e um pouco menor que um para o milho, portanto, em média cada unidade percentual de aumento de produtividade teve impacto na redução de uma unidade percentual no preço.

Dentre outros benefícios do melhoramento, pode-se ainda citar: aumento da qualidade ou a quantidade de proteínas, óleos, açúcares, vitaminas, minerais, conservação pós-colheita; obtenção de cultivares resistentes às doenças e às pragas; aumento da tolerância às condições adversas de clima e solo; e a introdução de caracteres exóticos, ou seja, características inexistentes nas espécies, como a produção de biofármacos, resistência a herbicida, e outros (Borém, 2017).

Contudo, a obtenção de cultivares que possam propiciar estes benefícios demandam recursos e muito trabalho pelos programas de melhoramento, já que é necessário neste processo ter área experimental que represente as condições de

adaptação da cultura, recurso financeiro, recurso humano e, especialmente, genótipos de ampla variabilidade ou diversidade.

A atividade que serve de elo entre os estudos da diversidade e a exploração desta nos programas de melhoramento é denominada de pré-melhoramento, e envolve a identificação de genes e características de interesse em germoplasma exótico ou em populações que não foram submetidas a qualquer processo de melhoramento, e sua posterior incorporação em materiais-elites (Nass; Paterniani, 2000).

## Diversidade genética

A variabilidade é o principal pré-requisito para se iniciar um programa de melhoramento genético de determinada espécie. Uma grande diversidade de germoplasma pode ser encontrada para a maioria das espécies. Estima-se que existem cerca de 6 milhões de acessos das mais variadas espécies vegetais, no mundo. Entretanto, pequena porcentagem desses materiais tem sido usada nos programas de melhoramento (Lopes; Carvalho, 2008).

Os programas de melhoramento, muitas vezes, procurando alcançar objetivos de maneira mais rápida e cômoda, utilizam de forma recorrente genótipos já bem conhecidos, e acabam por gerar cultivares com base genética estreita. Isso está levando vários programas, até mesmo aqueles de espécies com várias décadas de melhoramento, a retomar atividades de pré-melhoramento para reintrodução de variabilidade desejada nos materiais.

O fato de uma grande quantidade de genótipos com variabilidade genética estarem sendo conservados, não é um indicativo do seu potencial de utilização. Somente as atividades de caracterização e avaliação preliminar poderão disponibilizar esse acervo genético de modo a ser útil para a atividade dos melhoristas e permitir que se obtenha ganhos genéticos mais promissores no melhoramento (Carvalho, 2016).

O processo inicial de avaliação da diversidade genética visa à identificação de genitores adequados à obtenção de híbridos com maior efeito heterótico e que proporcione maior segregação em recombinações, possibilitando o aparecimento de genótipos que se destaquem.

Desse modo, a importância dos estudos sobre a diversidade genética para o melhoramento reside no fato de cruzamentos envolvendo genitores geneticamente diferentes são os mais convenientes para produzir alto efeito heterótico e, também, maior variabilidade genética em gerações segregantes, para isso busca-se população base para seleção que alie ampla variabilidade genética com alta média para o caráter a ser selecionado (Cruz *et al*., 2011).

## Biometria baseada em informações fenotípicas

A biometria é a aplicação da estatística ao campo biológico, sendo essencial ao planejamento, à avaliação e à interpretação de todos os dados obtidos em pesquisa na área biológica. No melhoramento de plantas a avaliação da diversidade por meio de análises biométricas pode se dar a partir de conjuntos de informações que podem ser fenotípicas, genotípicas ou geográficas (Cruz, 2006b).

Embora o volume de informações genéticas provenientes de marcadores moleculares tenha aumentado em grandes proporções para os estudos de diversidade genética, continua-se a dar ênfase aos estudos da diversidade por meio das características fenotípicas, principalmente de natureza quantitativa (Cruz *et al*., 2011), pois ainda são de fundamental importância para a seleção fenotípica direta.

Prova disso é o crescente movimento em torno de uma nova área do melhoramento denominada “fenômica”, que consiste em desenvolver métodos e tecnologias para fenotipar rapidamente grande quantidade de indivíduos para muitos caracteres no decorrer de todo o ciclo da cultura, de modo não destrutivo e com alta precisão e acurácia (Fritsche-Neto; Borém, 2015)

Técnicas de análise multivariada podem ser aplicadas na caracterização da diversidade genética vegetal, agrupando genótipos similares, de maneira que as maiores diferenças ocorram entre os grupos formados. Johnson e Wichern (2007) apontam como principais métodos de análise multivariada de distinção entre grupos a análise discriminante e análise de agrupamento.

## Análises discriminantes

Na análise discriminante, procura-se obter funções que permitam classificar um indivíduo, a partir das informações de um conjunto de características mensuradas, em uma entre várias populações conhecidas. Assim, devem-se obter funções que permitam alocar um indivíduo na população a qual ele pertence, e para tanto há a necessidade de se trabalhar com informações de populações previamente conhecidas, e após a constatação da eficácia da discriminação, as funções podem ser utilizadas para classificar novos indivíduos (Cruz *et al*., 2011).

Uma das mais conhecidas função discriminante é a análise discriminante linear de Fisher (Fisher, 1936), também conhecida como análise de discriminantes lineares (LDA), que busca realizar uma combinação linear das variáveis independentes com objetivo de maximizar a separação de grupos preditos em um espaço reduzido bidimensional e ainda permitir que novas observações sejam classificadas ou não dentro dos grupos conhecidos a priori.

Considerando que as matrizes de variância e covariâncias das variáveis que caracterizam as populações sejam homogêneas, a função discriminante linear de Fisher é dada por:

$$D_{ii,}(\tilde{x}) = \alpha'\tilde{x} = (\mu_i - \mu_{i'})' \sum^{-1} \tilde{x}$$

Em que:

$\alpha'$ : vetor discriminante;
$\tilde{x}$ : vetor aleatório das características das populações;
$\mu$ : vetor de médias;
$\sum$ : matriz comum de covariância das populações.

Assim, a função discriminante $D_{ii,}(\tilde{x})$ é uma combinação linear do conjunto de caracteres que possibilita alocar um determinado indivíduo, com vetor de observações $\tilde{x}$, em uma população i, ou i', com máxima probabilidade de acerto (Cruz *et al*., 2011).

Para tomada de decisão, aloca-se um indivíduo com vetor de observações $\tilde{x}$ na população i se $D_{ii,}(\tilde{x}) \geq m_{ii'}$ , e aloca-se na população i' se $D_{ii,}(\tilde{x}) < m_{ii'}$, onde $m_{ii'}$ é o ponto médio entre as populações i e i', expresso por:

$$m_{ii'} = \frac{1}{2}[D(\mu_1) + D(\mu_2)]$$

Na análise discriminante proposta por Anderson (1958) os objetivos e os pré-requisitos são os mesmos da análise discriminante de Fisher, ou seja, promover a melhor discriminação entre indivíduos, alocando-os em suas devidas populações, por meio de funções geradas a partir de informações de indivíduos sabidamente pertencentes a diferentes populações, e tendo como suposição a homogeneidade das matrizes de variância e covariâncias das variáveis que caracterizam as populações.

A função discriminante de Anderson é dada por:

$$D_j(\tilde{x}) = \ln(p_j) + \left(\tilde{x} - \frac{1}{2}\mu_j\right)\sum^{-1}\mu_j$$

Onde $\tilde{x}$ é o vetor de variáveis representativas dos caracteres envolvidos na análise; $p_j$ é a probabilidade *a priori*, de os indivíduos pertencerem a população *j*; e $\mu_j$ é o vetor de médias dos p caracteres avaliados na população *j.*

Assim, classifica-se o t-ésimo material genético, com vetor de média $\tilde{x}_i$, na população *j*; se e somente se $D_j(\tilde{x})$ for o maior entre os elementos do conjunto $\{D_1(\tilde{x}_i), D_2(\tilde{x}_i), D_3(\tilde{x}_i)\}$.

## Análises de agrupamento

A diversidade genética pode ser avaliada de forma simultânea em relação às várias características, e quando se dispõe de variáveis quantitativas contínuas ou discretas, recomenda-se a utilização de medidas de dissimilaridade. A análise de agrupamento permite a formação de grupos, não conhecidos previamente, por meio de técnicas de agrupamento aplicadas sobre as medidas de dissimilaridade entre fenótipos (Resende, 2007).

As medidas de dissimilaridade mais utilizadas nos estudos genéticos segundo Cruz *et al.* (2011) são: a distância euclidiana, o quadrado da distância euclidiana, a distância euclidiana média e a distância generalizada de Mahalanobis.

Considerando $Y_{ij}$ a observação no i-ésimo genótipo para a j-ésima característica, define-se a distância euclidiana entre o par de genótipos i e i' por meio da expressão:

$$D_E(i, i') = [\sum_{j=1}^{v}(Y_{ij} - Y_{i'j})^2]^{1/2}$$

Sendo o número de características estudadas: $1 \leq j \leq v$.

Como a distância euclidiana sempre aumenta com o acréscimo do número de características consideradas na análise, tem sido usada, de forma alternativa, a distância euclidiana média, dada por:

$$D_{EM}(i,i') = [\sum_{j=1}^{v}(Y_{ij} - Y_{i'j})^2/v]^{1/2}$$

Outra forma de expressar a dissimilaridade, algumas vezes preferida por manter a relação com a soma de quadrado de desvio, é por meio do quadrado da distância euclidiana média, dada por:

$$D_{QEM}(i,i') = (\sum_{j=1}^{v}(Y_{ij} - Y_{i'j})^2)/v$$

Essas estimativas de dissimilaridade, porém, apenas informam sobre o grau de semelhança ou de diferença entre dois quaisquer genótipos, tornando impraticável o reconhecimento de grupos homogêneos pelo simples exame visual quando o número de estimativas é relativamente grande. Nestes casos, é necessário a utilização de métodos de agrupamento ou de projeção de distâncias em gráficos bidimensionais, sendo, dos métodos de agrupamento, os mais utilizados os métodos hierárquicos e os de otimização ou mutuamente exclusivos (Cruz *et al*., 2014).

Algoritmos de agrupamentos baseados no método hierárquico organizam um conjunto de dados em uma estrutura hierárquica de acordo com a proximidade entre os indivíduos, resultando na formação de uma árvore binária ou um dendograma. A altura do dendograma expressa a distância entre um par de indivíduos ou entre um par de grupos ou ainda entre um indivíduo e um grupo. O resultado do agrupamento pode ser obtido cortando-se o dendograma em diferentes níveis (Cruz*,* 2006a).

Vários estudos utilizaram análises de agrupamento na avaliação da diversidade genética, com caracteres morfológicos e agronômicos, como no caso do guaraná (Nascimento Filho *et al*., 2001), a melancia (Souza *et al*., 2005), a cana-de-açúcar (Silva *et al*., 2011), dentre outras.

Entre os principais métodos de agrupamento hierárquico está o método da variância mínima de Ward (Cruz et al., 2011), sendo um método hierárquico e aglomerativo (Ward, 1963) que consiste na formação de grupos pela maximização da homogeneidade dentro dos grupos, obtida pela minimização da soma de quadrados

dentro deste. Esse método tende a resultar em agrupamentos de tamanhos aproximadamente iguais devido a sua minimização de variação interna (Hair *et al*., 2005). Este método tem sido testado e utilizado com bons resultados na avaliação da diversidade genética de plantas (Silva *et al*., 2017; Pessoa *et al*., 2019).

## Aprendizagem de máquina

A aprendizagem de máquinas (AM) é a área da Inteligência Artificial responsável por estudar formas de transferir o conhecimento às máquinas, ou seja, dotar as máquinas do elemento essencial para um comportamento inteligente, a capacidade de aprendizagem (Coppin, 2010). A aprendizagem possibilita que o sistema faça a mesma tarefa ou tarefas sobre uma mesma população de uma maneira mais eficaz a cada execução. O campo da aprendizagem de máquina é concebido pela questão de como construir programas, que automaticamente melhoram com a sua experiência, conforme definição de Mitchell (1997).

Algoritmos de AM têm sido amplamente utilizados em diversas tarefas, que podem ser organizadas de acordo com diferentes critérios. De acordo com esse critério, as tarefas de aprendizagem podem ser divididas, segundo Faceli *et al.* (2011), em preditiva e descritiva.

Em tarefas de previsão, a meta é encontrar, a partir de dados de treinamento com atributos de entrada e de saída conhecidos, uma função (ou modelo) que possa ser utilizada para prever um rótulo ou valor que caracterize um novo exemplo, com base nos valores de seus atributos de entrada. Os algoritmos ou métodos de AM utilizados nessa tarefa seguem o paradigma de aprendizagem supervisionada. O termo supervisionado vem da simulação da presença de um “supervisor externo”, que conhece a saída (rótulo) desejada para cada exemplo (Faceli *et al.*, 2011).

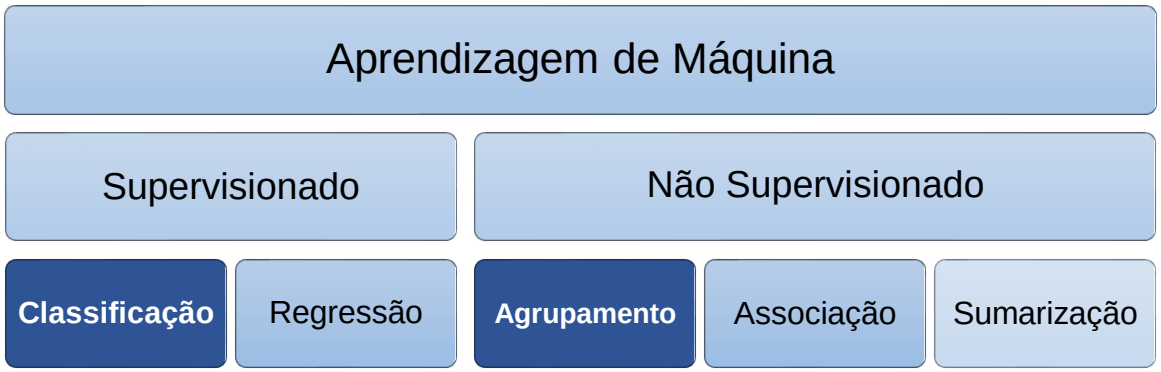
Em tarefas de descrição, a meta é explorar ou descrever um conjunto de dados, e neste caso os algoritmos de AM utilizados nessas tarefas não fazem uso do atributo de saída e, por isso, seguem o paradigma de aprendizagem não supervisionada.

Como a existência de dados com rotulação prévia são escassos, há ainda um modelo intermediário que é o semi-supervisionado, onde o princípio é então utilizar os exemplos rotulados existentes para se obter informações sobre o problema e utiliza-

las para guiar o processo de aprendizado a partir de exemplos não rotulados (Bruce, 2001).

Segundo Russell e Norvig (1995), há ainda o paradigma de aprendizagem por reforço, em que um agente aprendiz procura maximizar uma medida de desempenho baseada nas recompensas ou punições que recebe ao interagir com um ambiente desconhecido. Aprendizado por reforço se distingue do aprendizado supervisionado no sentindo em que pares de entrada/saída corretos nunca são apresentados.

Na Figura 1 pode-se observar a hierarquia de acordo com os paradigmas de aprendizagem, com destaque dos tipos avaliados no presente estudo.




Adaptado de (Faceli *et al.*, 2011).

**Figura 1.** Hierarquia de aprendizagem dos paradigmas de AM utilizados neste estudo.

As tarefas supervisionadas se distinguem pelo tipo dos dados da variável dependente: discreto, no caso de classificação; contínuo, no caso de regressão. As tarefas não supervisionadas são genericamente divididas em: agrupamento, em que os dados são agrupados de acordo com sua similaridade; sumarização, cujo objetivo é encontrar uma descrição simples e compacta para um conjunto de dados; e associação, que consiste em encontrar padrões frequentes de associações entre os atributos de um conjunto de dados (Faceli *et al.*, 2011).

## Aprendizagem supervisionada

Um algoritmo de AM supervisionado é uma função que, dado um conjunto de exemplos rotulados, constrói um estimador. O rótulo ou etiqueta toma valores num domínio conhecido. Se esse domínio for um conjunto de valores nominais, tem-se um

problema de classificação, também conhecido como aprendizado de conceitos, e o estimador gerado é um classificador. Se o domínio for um conjunto infinito e ordenado de valores, tem-se um problema de regressão, que induz um regressor (Dietterich, 1998).

Os principais métodos supervisionados com sua classificação segundo Faceli *et al.* (2011) são: *Naive Bayes* - métodos probabilísticos; Árvore de decisão e Floresta Aleatória - métodos baseados em procura ou em árvores; K-vizinhos mais próximos - métodos baseados em distância; e Máquina de Vetor de Suporte e Rede Neural Perceptron Multicamadas - métodos baseados em otimização.

O método *Naive Bayes* (John e Langley, 1995) é um classificador probabilístico simples baseado no teorema de Bayes. Esse método é denominado ingênuo (*naive*) porque ele assume que os atributos contribuem de forma independente para formar probabilidades estimativas da ocorrência de uma determinada classe do problema durante o processo de classificação.

O algoritmo de Árvore de decisão (Faceli et al., 2011) usa a estratégia dividir para conquistar para resolver um problema de decisão. Um problema complexo é dividido em problemas mais simples, aos quais recursivamente é aplicada a mesma estratégia. As soluções dos subproblemas podem ser combinadas, na forma de uma árvore, para produzir uma solução ao do problema complexo. A força dessa proposta vem da capacidade de dividir o espaço de instâncias em subespaços e cada subespaço é ajustado usando diferentes modelos.

A Floresta aleatória (R*andom forest*) (Breiman, 2001) é uma combinação de árvores de decisão, em que cada árvore depende dos valores de vetores aleatórios amostrados de forma independente e distribuídos igualmente para todas as árvores na floresta. Nesse método, depois que um determinado número de árvores é gerado, cada uma lança um voto para uma classe do problema, considerando um vetor de entrada. Então, a classe mais votada será escolhida na predição do classificador.

O K-vizinhos mais próximos (K *nearest neighboors* – KNN) é um método não-paramétrico usado para classificação e regressão e foi proposto por Fukunaga e Narendra (1975). O princípio do método é determinar o rótulo de classificação de uma amostra baseado nas amostras vizinhas provenientes de um conjunto de treinamento.

Máquinas de vetores de suporte (*Support Vector Machines* - SVM) (Haykin, 2009) é um método de aprendizagem de máquina que pode ser usado para problemas de classificação e regressão e outras tarefas de aprendizagem. Elas foram

conceitualmente implementadas seguindo a ideia de que vetores de entrada são não-linearmente mapeados para um espaço de atributos de alta dimensão. Nesse espaço, é construída uma superfície de decisão que permite distinguir as classes dos exemplos de entrada. Para conseguir separar dados linearmente ou não-linearmente separáveis, um dos principais elementos usados pelo método SVM é uma função de *kernel*, pela qual constrói uma superfície de decisão que é não-linear no espaço de entrada, mas é linear no espaço de atributos (Haykin, 2009). As principais funções de kernel que podem ser utilizadas no SVM são: linear, função de base radial, polinomial e sigmoidal.

A Rede neural artificial do tipo perceptron multicamadas (Haykin, 2009) é uma rede que possui um conjunto de unidades sensoriais que formam, conforme esquematizado na Figura 2, a camada de entrada, uma ou mais camadas intermediárias de neurônios computacionais e uma camada de saída.

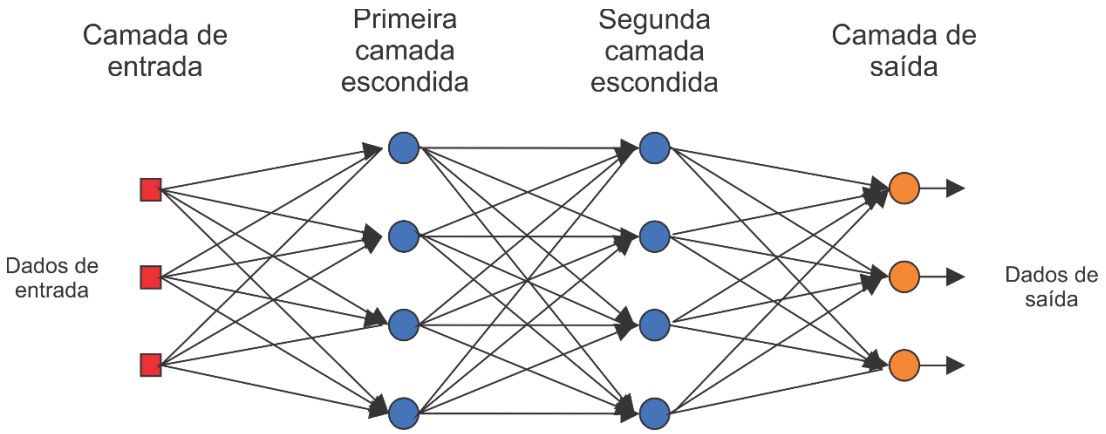



**Figura 2**. Exemplo simplificado de uma rede neural multicamadas.

## Aprendizagem não supervisionada

As tarefas da aprendizagem descritiva, ou não supervisionada, se referem a identificação de informações relevantes nos dados sem a presença de um elemento externo para guiar a aprendizagem. Essencialmente, a aprendizagem reside na identificação de propriedades intrínsecas aos dados de entrada, de maneira a construir representações desses dados que possam servir a diversos propósitos como auxílio a tomada de decisões ou descoberta de conhecimento.

Essas técnicas são utilizadas principalmente quando o objetivo da aprendizagem é encontrar padrões ou tendências que auxiliem no entendimento dos dados (Souto *et al*., 2003). Os principais métodos desse tipo de aprendizagem são descritos a seguir.

O método k-Means é um procedimento de otimização com inicialização aleatória que garante a convergência para um mínimo local da soma dos erros quadráticos (SEQ). O algoritmo procura uma partição que minimize a SEQ entre os objetos de um conjunto de dados e o centroide dos seus respectivos grupos (Jain, 2010). Seja $\mu_k$ a média do cluster $c_k$, o erro quadrático entre $\mu_k$ e os pontos no cluster $c_k$ é definido como:

$$J(c_k) = \sum_{x_i \in c_k} \|x_i - \mu_k\|^2$$

O objetivo do k-Means é minimizar a soma do erro quadrático sobre todos os clusters K,

$$J(c) = \sum_{k=1}^{K} \sum_{x_i \in c_k} \|x_i - \mu_k\|^2$$

O método *Density-Based Spatial Clustering of Applications with Noise,* conhecido pela sigla DBSCAN, busca por grupos definidos, como regiões com alta densidade de objetos, separados por regiões de baixa densidade. Uma das principais vantagens desse algoritmo advém do fato de não ser necessário informar previamente o número desejado de grupos. Para isso, baseia na classificação de cada objeto do conjunto de dados em uma dentre 3 categorias: objeto central - todo objeto $x_i$ que somado ao número de objetos dentro de um raio máximo igual a *eps,* totaliza uma quantidade maior ou igual a um parâmetro *MinPts*; objeto de borda - todo objeto que não satisfaz as condições para objeto central, mas que pertence à vizinhança de um objeto central; ruído - todo objeto que não pertence a nenhuma das duas categorias anteriores (Ester *et al*. 1996).

Os mapas auto-organizáveis de Kohonen (SOM - *Self Organizing Maps*) são um tipo de rede neural não supervisionada (Kohonen, 1990) e realizam uma redução de dimensionalidade, mostrando em um mapa bidimensional as informações de similaridade existentes entre amostras compostas de diversos atributos, sendo utilizado na classificação e análise de dados (Vesanto e Alhoniemi, 2000).

O *Affinity Propagation Clustering*, ou agrupamento por propagação de afinidade, é um algoritmo relativamente recente introduzido por Frey e Dueck (2007). O método, segundo os autores, toma como entrada medidas de similaridade entre pares de pontos de dados e identifica pontos que sejam representativos, “exemplares” dos conjuntos de dados, no entorno dos quais forma os agrupamentos. Ele opera considerando simultaneamente todos os pontos de dados como possíveis exemplos e realiza a troca de “mensagens” entre os pontos de dados até que um bom conjunto de exemplos e clusters de alta qualidade surja.

## Aprendizagem de máquina no melhoramento de plantas

Na área agrícola como um todo, já existe um número muito expressivo de pesquisas e aplicações do AM, especialmente redes neurais artificiais (RNA), para praticamente todos os produtos agrícolas. Contudo, Cruz e Nascimento (2018) afirmam que a utilização das RNA no melhoramento de plantas é um fato recente, mas tem se observado que ela pode ser utilizada de forma concomitante ou mesmo substituir as metodologias tradicionais nesta área, portanto apresenta um grande potencial de aplicação.

Abaixo elencamos os trabalhos encontrados referente à utilização de métodos de AM aplicados a alguns procedimentos da atividade do melhoramento de plantas.

Classificação de genótipos:

- Barbosa *et al.* (2011) obtiveram bons resultados, acurácia de 91,90%, com o uso de RNA para classificar trinta e sete acessos de mamão, tomando como base a média de três épocas de cultivo e oito características quantitativas;

- Oliveira *et al*. (2013) também testaram o uso de RNA, do tipo *Multi Layer Perceptron*, para classificação de 114 autotetraploides de bananeira por meio de dados de correlação entre a massa fresca de discos foliares e o conteúdo de DNA, proveniente de outra pesquisa realizada previamente, e obteve 90,90% de acurácia na classificação;

- Ornella e Tapia (2010) avaliaram o desempenho de vários algoritmos de aprendizagem supervisionada, dentre eles *Naive Bayes* e máquina de vetor de suporte, para classificação de três conjuntos de dados de marcadores moleculares

representando um amplo espectro de padrões heteróticos de milho, e embora os resultados obtidos tenham sido considerados relativamente ruins, os autores afirmam que há uma forte evidência de que o uso de dados com mais instâncias de treinamento pode gerar classificadores bem-sucedidos;

- Fuentes *et al*. (2018) avaliaram o uso de RNA na classificação de 16 cultivares de videira obtiveram 94% de acurácia na classificação utilizando dados morfo-colorimétricos das folhas, e 92% de acurácia utilizando dados de espectrometria de infravermelho próximo (NIR) também das folhas;

- Plotze (2004), em sua dissertação de mestrado, avaliou a eficácia de RNA na classificação de 20 espécies de maracujás nativos do gênero *Passiflora,* utilizando dados extraídos das folhas, e obteve 85% de acurácia na classificação;

- Sant'anna *et al.* (2015) utilizaram dados simulados para comparar a acurácia entre metodologias estatísticas convencionais (análises discriminantes de Fisher e Anderson) e RNA na classificação de genótipos conforme sua distância genética. Os autores obtiveram valores acima de 80% de acurácia com as RNAs nas condições de maior dificuldade de classificação, ou seja, quando os genótipos possuem alta similaridade genética, enquanto que metodologias convencionais obtiveram entre 18% e 20% de acurácia nestas condições.

Classificação de genótipos obtida por meio de metodologias de adaptabilidade e estabilidade:

- Nascimento *et al*. (2013) propuseram uma metodologia para classificação de genótipos usando RNA com base na análise da adaptabilidade e estabilidade fenotípica de genótipos de alfafa, considerando a metodologia de Eberhart e Russell (1966). Foram utilizados dados provenientes de um experimento sobre produção de matéria seca de 92 genótipos, e observaram que a RNA foi capaz de classificar satisfatoriamente os genótipos e que a análise apresentou altas taxas de concordância, em comparação com os resultados obtidos pela metodologia de Eberhart e Russell.

- Barroso *et al.* (2013), utilizando a metodologia baseada em RNA proposta por Nascimento *et al*. (2013), e os mesmos dados de 92 genótipos de alfafa utilizada por esses autores, compararam os resultados desta com os resultados obtidos pela metodologia de Eberhart e Russell (1966) e com os resultados obtidos pela análise discriminante, e verificaram que a RNA apresentou índices de concordância mais

elevados do que a análise discriminante com relação aos resultados obtidos pela metodologia de Eberhart e Russell.

- Teodoro *et al.* (2015) objetivaram verificar a concordância entre as RNAs e o método de Eberhart e Russel na identificação de genótipos de feijão-caupi com alta adaptabilidade e estabilidade fenotípicas. Para isso, utilizaram no trabalho dados de quatro ensaios de valor de cultivo e uso, com 18 linhagens experimentais e duas cultivares de feijão-caupi. Os autores também encontraram elevada concordância entre os métodos avaliados, incluindo as RNAs, quanto à discriminação da adaptabilidade fenotípica dos genótipos.

Seleção e predição de valores genéticos:

- Brasileiro *et al.* (2015) avaliaram as RNAs aplicadas ao processo de seleção de plantas individuais de cana-de-açúcar dentro das melhores famílias. Foram utilizados dados de 128 famílias de meios-irmãos derivadas de cruzamentos realizados na estação experimental da Universidade Federal de Alagoas. Os autores testaram dois modelos de RNA, e o melhor modelo conseguiu classificar todos os genótipos corretamente, ou seja, a rede fez a mesma escolha seletiva que o melhorista durante a simulação do melhor preditor imparcial linear individual (BLUPIS), sendo que o segundo modelo apresentou apenas 5,1% de diferença em relação ao uso do BLUPIS.

- Silva *et al*. (2016) propuseram o uso de RNA como uma alternativa aos métodos de predição de valores genéticos convencionais, utilizando dados simulados referentes a oito cenários e, para fins de previsão de valor genético, sete parâmetros estatísticos, além da média fenotípica. Avaliando diferentes configurações de rede, os resultados demonstraram a superioridade das redes neurais em comparação com os procedimentos de estimativa baseados em modelos lineares e indicaram alta precisão preditiva e eficiência da RNA, coincidindo com trabalho similar realizado anteriormente por Silva *et al.* (2014).

- Peixoto *et al.* (2015) objetivaram avaliar a eficácia de RNAs na predição de valor genético, comparando a correlação do valor da rede e do valor fenotípico com o valor genético. Utilizaram dados simulados de dezesseis cenários com diferentes valores de herdabilidade, coeficiente de variação e número de genótipos por bloco. Concluíram que as RNAs foram eficientes em prever o valor genético com um ganho

de 0,64 a 10,3% em comparação com o valor fenotípico, independentemente do tamanho da população simulada, herdabilidade ou coeficiente de variação.

Predição de valores genéticos por meio da associação com análise genômica:

- Gianola *et al.* (2011) testaram a capacidade preditiva de RNAs para valores genéticos, utilizando dados fenotípicos de rendimento médio de grãos de 599 linhagens de trigo, e dados moleculares de cada linhagem genotipadas com 1279 marcadores DArT (Diversity Array Technology). Tiveram como resultado o bom desempenho das RNAs não lineares, superando o desempenho de modelos lineares de referência em capacidade preditiva.

Observando as referências apresentadas, podemos notar que a maioria dos trabalhos publicados utilizaram as Redes Neurais Artificiais como modelo de aprendizado de máquina aplicado ao melhoramento de plantas. Assim, podemos estender a afirmação que o uso de RNA é um fato recente (Cruz; Nascimento, 2018) para as outras técnicas de aprendizado de máquina também, ou seja, o uso de diversas outras técnicas e modelos desta área no melhoramento genético, de forma mais abrangente, é também muito recentemente, o que caracteriza a importância de trabalhos nesta área.

## Uso de dados simulados

Uma problemática encontrada para a execução de pesquisas que se utilizem de dados genotípicos ou fenotípicos de uma ou várias populações, é a necessidade de realização de estudos ou ensaios experimentais, os quais demandam muitas vezes de longo tempo de execução e altos custos, tornando-se, por vezes, inviável (Silva, 2014). Uma das grandes contribuições da informática, nesse sentido, é viabilizar o estudo de fenômenos, por meio da simulação de uma situação complexa, em que são estabelecidos parâmetros e restrições, de forma que os efeitos de certos fatores controláveis possam ser convenientemente estudados (Cruz, 2006a).

A simulação computacional demanda dos geneticistas o desenvolvimento de modelos biológicos que retratem, da melhor maneira possível, os fenômenos de interesse, e dos programadores as rotinas para o processamento adequado, apesar

de impor restrições, para que a influência de certos fatores possa ser avaliada (Cruz; Sant'Anna, 2017). Destaca-se também que, em certos estudos da área de melhoramento genético, é prudente que cada característica simulada tenha a propriedade de descrever uma variável com a média, herdabilidade e precisão experimental estabelecidas pelo pesquisador, pois tais informações são rotineiramente trabalhadas pelo pesquisador e conseguem captar o potencial genético do material estudado, sua variabilidade e a influência ambiental (Cruz, 2006a).

Um exemplo de aplicação da simulação dado por Silva (2014) é a simulação de ensaios segundo o delineamento de blocos casualizados com o objetivo principal simular planilhas de dados experimentais que possam representar um conjunto de genótipos e acompanhar ou predizer a sua diversidade, a precisão experimental, a distribuição de erros, a estrutura das matrizes de covariâncias, dentre outros.

De acordo com Sant'anna (2018), a simulação de dados tem sido empregada nos estudos genéticos sob vários contextos, como o de populações, do indivíduo ou do próprio genoma, como pode-se observar nos trabalhos de: Sant'Anna (2015), com análise discriminante em estruturas de populações; Price *et al.* (2006), com estudos de genoma; Coelho e Barbin (2006), na comparação de diferentes métodos de estimação dos componentes de variância e coeficientes de herdabilidade; e Peixoto (2015), nos estudos acerca do processo de aprendizagem de redes neurais artificiais e sua utilização em processo de predição de valores genéticos.

## Referências

BARBOSA, C. D. *et al* . Artificial neural network analysis of genetic diversity in *Carica papaya* L. **Crop Breed. Appl. Biotechnol. (Online)**, Viçosa, v. 11, n. 3, p. 224-231, Sept. 2011.

BARROSO, L. M. A.; NASCIMENTO, M.; NASCIMENTO, A. C. C.; Silva, F. F.; Ferreira, R. D. P. Uso do Método de Eberhart e Russell como informação a priori para aplicação de redes neurais artificiais e análise discriminante visando a classificação de genótipos de alfafa quanto à adaptabilidade e estabilidade fenotípica. **Revista Brasileira de Biometria**, Lavras, v. 31, n. 2, p. 176-188, 2013.

BORÉM, A.; MIRANDA, G. V.; FRITSCHE-NETO, R. **Melhoramento de plantas**. 7. ed. Viçosa: Editora UFV, 2017.

BRASILEIRO, B. P.; MARINHO, C. D.; COSTA, P. M. de A.; *et al*. Selection in sugarcane families with artificial neural networks. **Crop Breeding and Applied Biotechnology**, v. 15, n. 2, p. 72–78, 2015.

BREIMAN, L. Random forests. **Machine learning**, v. 45, n. 1, p. 5-32, 2001.

BRUCE, R. F. A Bayesian Approach to Semi-Supervised Learning. In: **NLPRS**. 2001. p. 57-64.

CARVALHO, V. P. de. **Discriminação de população por meio de inteligência computacional**. 2016. Dissertação (Mestrado em Estatística Aplicada a Biometria) – Universidade Federal de Viçosa, Viçosa, 2016.
COPPIN, B. **Inteligência artificial**. Rio de Janeiro: LTC, 2010.

COSTA, C. C.; FREITAS, R E. Contribuição do Melhoramento Genético para a Redução de Preço dos Alimentos. Instituto de Pesquisa Econômica Aplicada - IPEA, **Discussion Papers**. 2006.

CRUZ, C. D. **Programa GENES: Análise multivariada e simulação**. Viçosa: Editora UFV, 2006a.

CRUZ, C. D. **Programa GENES: Biometria**. Viçosa: Editora UFV, 2006b.

CRUZ, C. D.; FERREIRA, F. M.; PESSONI, L. A. **Biometria aplicada ao estudo da diversidade genética**. Visconde do Rio Branco: Suprema, 2011.

CRUZ, C. D.; NASCIMENTO, M. (ed.). **Inteligência computacional aplicada ao melhoramento genético**. Viçosa: Editora UFV, 2018.

CRUZ, C. D.; REGAZZI, A. J.; CARNEIRO, P. C. S. **Modelos biométricos aplicados ao melhoramento de plantas**. Vol. 2. 2014.

CRUZ, C. D.; SANT'ANNA, I. C. Bioinformática e os avanços computacionais nas análises biométricas aplicadas ao melhoramento. *In*: LUDKE W. H. *et al*. (ed.) **Desafios biométricos aplicados ao melhoramento genético**. Viçosa: GenMelhor, 2017. p. 148-166.

COELHO, A. M.; BARBIN, Décio. Simulação de dados visando à estimação de componentes de variância e coeficientes de herdabilidade. **Rev. Mat. Estat**, v. 24, n. 2, p. 103-120, 2006.

DIETTERICH, T. G. Approximate statistical tests for comparing supervised classification learning algorithms. **Neural Computation**, v. 10, p. 1895–1923, 1998.

EBERHART, S. A.; RUSSELL, W. A. Stability parameters for comparing varieties. **Crop Science**, v.6, p.36-40, 1966.

ESTER, M.; KRIEGEL, H-P.; SANDER, J.; et al. A Density-Based Algorithm for Discovering Clusters in Large Spatial Databases with Noise. In: **Kdd**, v. 96, p. 226–231, 1996.

FACELI, K. *et al*. **Inteligência Artificial - Uma Abordagem de Aprendizado de Máquina.** Rio de Janeiro: LTC, 2011.

FISHER, R. A. The use of multiple measurements in taxonomic problems. **Annals of Eugenics**, v.7, p.179-188, 1936.

FREY, B. J.; DUECK, D. Clustering by Passing Messages Between Data Points. **Science**, v. 315, n. 5814, p. 972–976, 2007.

FRITSCHE-NETO, R.; BORÉM, A. (Ed.). **Phenomics: How next-generation phenotyping is revolutionizing plant breeding**. Springer, 2015.

FUENTES, S; HERNÁNDEZ-MONTES, E; ESCALONA, J M; *et al*. Automated grapevine cultivar classification based on machine learning using leaf morpho-colorimetry, fractal dimension and near-infrared spectroscopy parameters. **Computers and Electronics in Agriculture**, v. 151, p. 311–318, 2018.

FUKUNAGA, K.; NARENDRA, P. M. A branch and bound algorithm for computing k-nearest neighbors. **IEEE Transactions on Computers**, v. 100, n. 7, p. 750–753, 1975.

GIANOLA, D.; OKUT, H.; WEIGEL, K. A.; *et al*. Predicting complex quantitative traits with Bayesian neural networks: a case study with Jersey cows and wheat. **BMC Genetics**, v. 12, n. 1, p. 87, 2011.

HAYKIN, S. **Neural networks and learning machines.** 3. ed. New York: Prentice Hall, 2009. 936 p.

HAIR, J. F.; et al. **Análise multivariada de dados**. Trad. Adonai S. Sant'Anna e Anselmo C. Neto. 5 ed. Porto Alegre: Bookman, 2005.

JAIN, A. K. Data clustering: 50 years beyond K-means. **Pattern Recognition Letters**, v. 31, n. 8, p. 651–666, 2010.

JOHNSON, R. A.; WICHERN, D. W. **Applied Multivariate Statistical Analysis.** 6. Ed. Upper Saddle River:Prentice Hall, 2007.

JOHN, G. H.; LANGLEY, P. Estimating Continuous Distributions in Bayesian Classifiers. In: **Proceedings of the Eleventh conference on Uncertainty in artificial intelligence (UAI'95)**. San Francisco, CA: Morgan Kaufmann Publishers Inc. p. 338–345, 1995.

KOHONEN, T. The self-organizing map. **Proceedings of the IEEE**, v. 78, n. 9, p. 1464–1480, 1990.

LOPES, J. F.; CARVALHO, S. I. C. de. A **Variabilidade Genética e o Pré-melhoramento**. In: FALEIRO, F. G.; FARIAS NETO, A. L. de; RIBEIRO JUNIOR, W. Q. Pré-melhoramento, melhoramento e pós-melhoramento: estratégias e desafios. 2008.

MITCHELL, T. M. **Machine learning**. 1. ed. New York, NY, USA: McGraw-Hill, Inc., 1997.

NASCIMENTO FILHO, F. J. do; ATROCH, A. L.; SOUSA, N. R. de; *et al*. Divergência genética entre clones de guaranazeiro. **Pesquisa Agropecuaria Brasileira**, v. 36, n. 3, p. 501–506, 2001.

NASCIMENTO, M.; PETERNELLI, L. A.; CRUZ, C. D.; NASCIMENTO, A. C. C.; FERREIRA, R. D. P.; BHERING, L. L.; SALGADO, C. C. Artificial neural networks for adaptability and stability evaluation in alfalfa genotypes. **Crop Breeding and Applied Biotechnology**, Viçosa, v. 13, n. 2, p. 152-156, 2013.

NASS, L. L.; PATERNIANI, E. Pre-breeding: a link between genetic resources and maize breeding. **Scientia Agricola**, v. 57, n. 3, p. 581–587, 2000.

OLIVEIRA, A. C. L. de; PASQUAL, M.; PIO, L. A. S.; *et al*. Utilização da modelagem matemática (redes neurais artificiais) na classificação de autotetraploides de bananeira (*Musa acuminata Colla*). **Bioscience Journal**, v. 29, n. 3, p. 617–622, 2013.

ORNELLA, L.; TAPIA, E. Supervised machine learning and heterotic classification of maize (Zea mays L.) using molecular marker data. **Computers and Electronics in Agriculture**, v. 74, n. 2, p. 250-257, 2010.

PEIXOTO, L. A.; BHERING, L. L.; CRUZ, C. D. Artificial neural networks reveal efficiency in genetic value prediction. **Genetics and Molecular Research**, v. 14, n. 2, p. 6796–6807, 2015.

PESSOA, A. M. S.; RÊGO, E. R.; SILVA, A. P. G.; et al. Genetic diversity in F3 population of ornamental peppers (Capsicum annuum L.). **Revista Ceres**, v. 66, n. 6, p. 442–450, 2019.

PLOTZE, R. O. **Identificação de espécies vegetais através da análise da forma interna de órgãos foliares**. 2004. Dissertação (Mestrado em Ciência de Computação e Matemática Computacional) - Universidade de São Paulo, São Carlos, 2004.

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.

RAMALHO, M. A. P.; DIAS, L. A. dos S.; CARVALHO, B. L. Contributions of plant breeding in Brazil: progress and perspectives. **Crop Breeding and Applied Biotechnology**, v. 12, n. spe, p. 111–120, 2012.

RESENDE, M. D. V. de. **Matemática e estatística na análise de experimentos e no melhoramento genético**. Colombo: Embrapa Florestas, 2007.

RUSSELL, Stuart; NORVIG, Peter. **Artificial Intelligence: A Modern Approach**. [s.l.]: Prentice Hall, 1995.

SANT'ANNA, I. de C. **Redes neurais artificiais para predição genômica na presença de interações epistáticas**. 2018. Tese (Doutorado em Estatística aplicada a biometria) - Universidade Federal de Viçosa, Viçosa, 2018.

SANT'ANNA, I. C. *et al*. Superiority of artificial neural networks for a genetic classification procedure. **Genetics And Molecular Research**. Ribeirao Preto: Funpec-editora, v. 14, n. 3, p. 9898-9906, 2015.

SILVA, A. R.; SILVA, S. A.; SANTOS, L. A.; et al. Genetic divergence among castor bean lines and parental strains using ward's method based on morpho-agronomic descriptors. **Acta Scientiarum. Agronomy**, v. 39, n. 3, p. 307, 2017.
SILVA, G. N. **Redes Neurais Artificiais: Novo Paradigma para a Predição de Valores Genéticos**. 2014. Dissertação (Mestrado em Estatística aplicada a biometria) - Universidade Federal de Viçosa, Viçosa, 2014.

SILVA, G. N.; TOMAZ, R. S.; SANT'ANNA, I. de C.; *et al*. Neural networks for predicting breeding values and genetic gains. **Scientia Agricola**, v. 71, n. 6, p. 494–498, 2014.

SILVA, G. N.; TOMAZ, R. S.; SANT'ANNA, I. C.; *et al*. Evaluation of the efficiency of artificial neural networks for genetic value prediction. **Genetics and Molecular Research**, v. 15, n. 1, p. 1–11, 2016.

SILVA, G. C.; OLIVEIRA, F. J.; ANUNCIAÇÃO FILHO, C. J.; *et al*. Divergência genética entre genótipos de cana-de-açúcar. **Revista Brasileira de Ciências Agrárias**, v. 6, n. 1, p. 52–58, 2011.

SQUILASSI, M. G. Melhoramento de plantas e a produção de alimentos. Embrapa Tabuleiros Costeiros. **Documentos**, n. 56, 2003.

SOUTO, M. C. P. ; LORENA, A. C. ; DELBEM, A. C. B. ; CARVALHO, A. C. P. L. F. . Técnicas de Aprendizado de Máquinas para Problemas em Biologia Molecular. In: Sociedade Brasileira de Computação. (Org.). **Anais da III Jornadas de Mini-Cursos de Inteligência Artificial**. Sociedade Brasileira de Computação, 2003, v. VIII, p. 103-152.

SOUZA, F. de F.; QUEIRÓZ, M. A. de; DIAS, R. de C. S. Divergência genética em linhagens de melancia. **Horticultura Brasileira**, v. 23, n. 2, p. 179–183, 2005.

TEODORO, P. E.; BARROSO, L. M. A.; NASCIMENTO, M.; *et al*. Redes neurais artificiais para identificar genótipos de feijão-caupi semiprostrado com alta adaptabilidade e estabilidade fenotípicas. **Pesquisa Agropecuária Brasileira**, v. 50, n. 11, p. 1054–1060, 2015.

VESANTO, J.; ALHONIEMI, E. Clustering of the self-organizing map. **IEEE Transactions on Neural Networks**, v. 11, n. 3, p. 586–600, 2000.

WARD, J. H. Hierarquical grouping to optimize an objective function. **Journal of the American Statistical Association**, v. 58, p. 236–244. Mar. 1963.

# CAPÍTULO 1 - Algoritmos de aprendizagem supervisionada na classificação de populações de plantas com diferentes graus de parentesco


## Resumo

A discriminação de populações e classificação de indivíduos tem grande importância para fins de melhoramento genético, nos estudos populacionais e na conservação da diversidade genética, sendo as abordagens multivariadas utilizadas com frequência, especialmente as funções discriminantes de Fisher e de Anderson. Contudo, novas metodologias baseadas em aprendizagem de máquina (AM) têm-se mostrado promissoras para tais procedimentos, porém ainda carecem de aprofundamento na avaliação e comparação destes métodos. Assim, neste estudo propomos avaliar a eficácia de algoritmos de AM supervisionada na classificação de populações com diferentes graus de similaridade, comparando-os com as técnicas de análise discriminante propostas por Anderson e por Fisher. Os métodos de AM supervisionada testados foram: *Naive Bayes*, Árvore de decisão, K-vizinhos mais próximos (kNN), Floresta Aleatória (RF), Máquina de Vetor de Suporte (SVM) e Rede Neural Perceptron Multicamadas (RNA/MLP). Para a comparação dos métodos de classificação foram utilizados dados fenotípicos de populações com diferentes graus de similaridade genética, resultantes da simulação de informações genotípicas para diferentes populações submetidas ao esquema de retrocruzamento. As médias de acurácia de 30 repetições de cada método de classificação foram comparadas pelos testes de Friedman e Nemenyi com nível de confiança de 95%. Os métodos de classificação baseados em algoritmos de AM se mostraram superiores às funções discriminantes de Fisher e de Anderson, permitindo obter alta acurácia em condições de maior similaridade entre as populações, sendo os algoritmos kNN, Floresta Aleatória, SVM e *Naive Bayes* aqueles que apresentaram maior acurácia, superando o algoritmo de Árvore de Decisão e até mesmo a RNA/MLP, que perderam acurácia na condição de 96,88% de similaridade entre as populações.



**Palavras-chave**: Aprendizagem de máquina; melhoramento genético; métodos de classificação; similaridade entre populações.

# Introdução

A discriminação de populações e a classificação de indivíduos têm sido de grande importância para fins de melhoramento genético, nos estudos populacionais e no processo de conservação da diversidade genética, sendo utilizadas frequentemente em bancos de germoplasma. Para isso as abordagens multivariadas têm sido utilizadas com frequência e, dentre as metodologias mais usuais as funções discriminantes de Fisher e as funções discriminantes de Anderson se destacam (Fonseca *et al*., 2004; Carvalho *et al*., 2018b). Com estes métodos procura-se obter funções que permitam classificar um elemento amostral a partir das informações de um conjunto de características mensuradas em uma dentre várias populações conhecidas, buscando minimizar a probabilidade de classificação incorreta (Cruz, 2011).

Nas últimas décadas, o paradigma da inteligência artificial, amplamente consolidada nas áreas computacionais, tem apresentado resultados promissores e vem se tornando uma ferramenta com crescente aplicação na área de melhoramento de plantas, especialmente na classificação de genótipos (Ornella e Tapia, 2010; Barbosa *et al*., 2011; Oliveira *et al*., 2013; Fuentes *et al*., 2018), na predição de parâmetros genéticos e ganhos no melhoramento (Silva *et al*., 2016) e na avaliação de adaptabilidade e estabilidade de genótipos (Nascimento *et al*., 2013; Barroso *et al*., 2013). Uma das áreas da inteligência artificial que pode ser utilizada para resolver diversos problemas da estatística são as técnicas de aprendizagem de máquina, podendo ser utilizadas para agrupamento de indivíduos, previsão de séries temporais e, em especial, nos problemas de classificação.

A aprendizagem de máquina é a área da inteligência artificial responsável por estudar formas de transferir o conhecimento às máquinas (Coppin, 2010). Os seus algoritmos possuem a vantagem de serem, na sua maioria, não-paramétricos, não necessitarem de informações detalhadas sobre os processos físicos do sistema a ser modelado, tolerar a perda de dados e capazes de solucionar problemas de grande complexidade, especialmente ao utilizar as Redes Neurais Artificiais (RNAs), que tem apresentado potencial para aplicações em análises no melhoramento genético (Sant'Anna, 2015).

As RNAs, especialmente aquelas mais complexa como as do tipo Perceptron Multicamadas (MLP), demandam maior poder computacional e exigem o

conhecimento da melhor configuração de arquitetura e topologia para a resolução de cada problema. No entanto, outros algoritmos de aprendizagem de máquinas também têm mostrado eficiência nos processos de classificação e possuem potencial de utilização no melhoramento genético, como os algoritmos *Naive Bayes*, Árvore de decisão, k-Vizinhos mais próximos (kNN), Floresta Aleatória e Máquina de Vetor de Suporte (SVM) (Ornella e Tapia, 2010; Carvalho, 2018a; Naik *et al*., 2017), porém ainda são escassos os testes com estes algoritmos, geralmente de forma isolada, sem uma comparação mais ampla entre diferentes métodos, concentrando em um pequeno número de métodos de AM (Singh *et al*., 2016).

Outra limitação dos trabalhos já realizados é a carência de aplicação de procedimentos experimentais adequados, como validação cruzada e o emprego de repetição na validação dos algoritmos. Assim, torna-se necessário o aprofundamento das avaliações destes métodos para aplicação no melhoramento de plantas, e a comparação destes com as metodologias estatísticas tradicionais.

Desta forma, o presente estudo objetivou avaliar a eficácia de algoritmos de aprendizagem de máquina supervisionada na classificação e discriminação de populações com diferentes graus de parentesco, comparando-os também com as técnicas de análise discriminante tradicionais, propostas por Anderson e por Fisher.

## Material e Métodos

Para comparação dos métodos de classificação foram utilizados dados fenotípicos de populações com diferentes graus de similaridade genética, resultantes da simulação de informações genotípicas para diferentes populações submetidas ao esquema de retrocruzamento *in silico*.

A simulação de dados tem sido de grande importância no melhoramento vegetal e vem sendo cada vez mais utilizada para avaliação e testes de metodologias, tanto com dados quantitativos (Vasconcelos *et al*., 2007) como moleculares (Odong *et al*., 2011), uma vez que nem sempre há disponibilidade de dados reais e sua obtenção demanda muito trabalho e custo significativo. Outra vantagem da simulação é a possibilidade de controle na geração dos dados, o que permite o melhor conhecimento do seu comportamento, sendo fator favorável para a avaliação de desempenho de novas metodologias (Maurer *et al*., 2008).

O delineamento de retrocruzamentos gera populações com proporção média de similaridade previamente conhecidas entre a população de retrocruzamento e a população do genitor recorrente (Borém, 2017), sendo esta informação fundamental para compor um conjunto de dados com dificuldade de discriminação conhecida. A geração dos dados utilizados nestas comparações segue a metodologia utilizada por Sant'Anna *et al*. (2015), com a utilização do Software Genes (Cruz, 2013).

**Simulação dos dados Genotípicos**

Foram simulados dados genotípicos para dez populações, em equilíbrio de Hardy-Weinberg, contendo 100 indivíduos cada. Para cada indivíduo foi gerada informações de 50 locos com dois alelos codominantes cada.

A partir desses dados genéticos foram calculadas as medidas de dissimilaridade genotípica de Nei (1972), as quais foram projetadas graficamente em plano bidimensional para visualização da diversidade genética destas populações (Figura 1). Foram selecionadas as duas populações mais divergentes, representadas pelas populações com maior distância gráfica entre elas, para serem utilizadas como genitores no sistema de retrocruzamentos, neste caso, as populações 4 e 10, que foram utilizadas como genitores A e B, respectivamente.

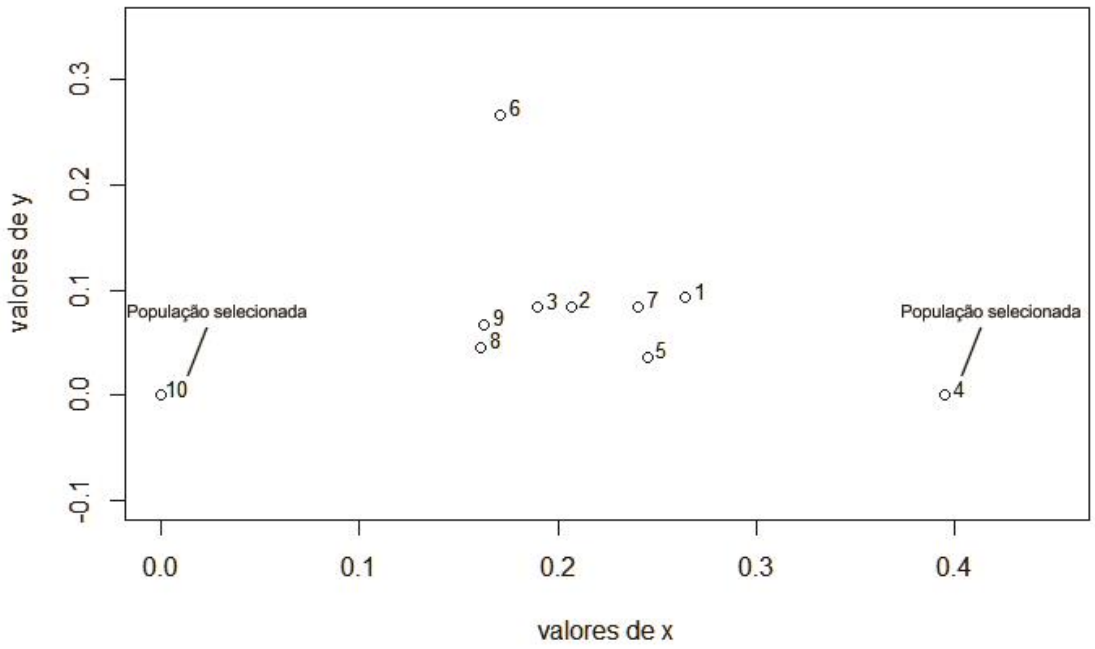



**Figura 1.** Projeção gráfica das medidas de dissimilaridade de Nei, com base nos dados genéticos, das dez populações simuladas, indicando as duas populações selecionadas.

As demais populações foram geradas conforme delineamento mostrado na Figura 2, pela opção “Cruzamentos” em “Simulação Genotípica” do Software Genes, com 100 indivíduos cada, a partir de cruzamentos planta-a-planta.

Buscando proximidade com os cenários reais, onde a genotipagem nem sempre é possível e, portanto, se busca discriminar indivíduos pertencentes a populações distintas por meio de características fenotípicas, foram gerados os dados fenotípicos destas populações por meio dos dados genotípicos simulados.

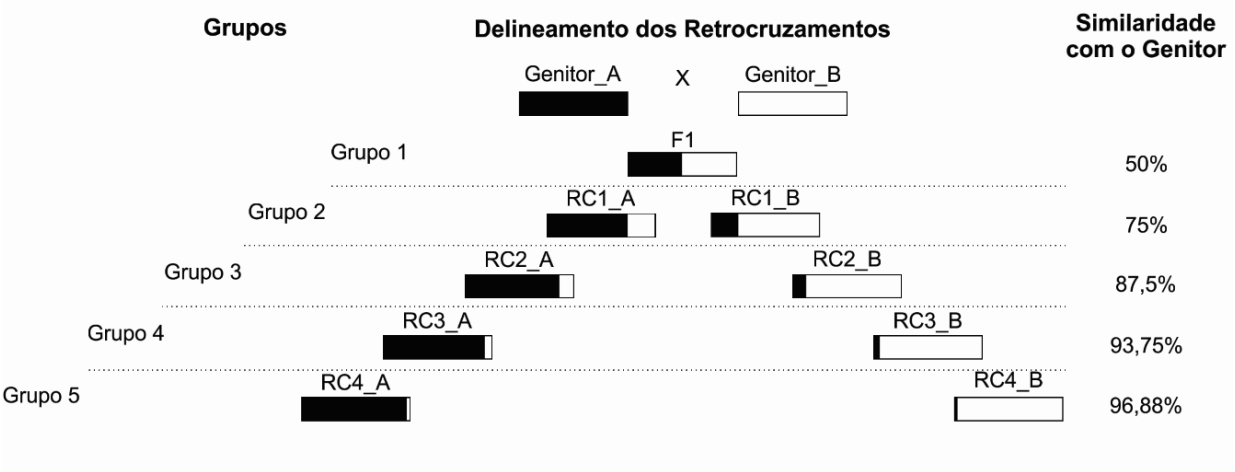



**Figura 2**. Delineamento dos cruzamentos entre os genitores selecionados e suas gerações de retrocruzamentos, similaridade esperada das populações de retrocruzamento com os genitores, e composição dos grupos utilizados nos testes dos métodos de discriminação.

**Geração dos valores fenotípicos e dos grupos de populações**

Para cada população foram gerados valores fenotípicos para seis variáveis quantitativas, contínuas, com distribuição normal e com valores de média, herdabilidades e coeficientes de variação previamente estabelecidos. Foi considerado o valor único de 12% do coeficiente de variação experimental para todas as variáveis e populações, correspondendo ao valor médio aceitável encontrado em experimentos de campo.

Os valores foram gerados a partir da ação de alelos de 20 locos, tomados ao acaso entre os 50 previamente simulados, com efeito aditivo diferencial e com pesos da importância do loco sobre a variabilidade genotípica total do caráter, estabelecidos a partir de uma distribuição binomial e grau médio de dominância nulo.

O delineamento experimental adotado foi inteiramente casualizado, utilizando o modelo estatístico determinante do valor fenotípico que segue:

$$Y_{ij} = \mu_j + G_i + \varepsilon_{ij}$$

Em que:
$Y_{ij}$: observação simulada de uma dada característica para o i-ésimo indivíduo pertencente à j-ésima população;
$\mu_j$: média geral da característica especificada previamente para a j-ésima população;
$G_i$: efeito genotípico do i-ésimo indivíduo;
$\varepsilon_{ij}$: erro aleatório

Foram escolhidos os valores de herdabilidade de 30, 40, 50, 60, 70 e 80%, respectivamente, para cada uma das seis variáveis simuladas, valores de média semelhante ao valor da herdabilidade para o Genitor A e 50% dos valores de média do Genitor A para o Genitor B. As médias das demais populações foram calculadas tendo em vista o percentual médio esperado dos genótipos parentais em cada geração (Tabela 1).

**Tabela 1.** Médias paramétricas e herdabilidade das características simuladas para cada população.

| Populações | Variáveis quantitativas | | | | | |
|---|---|---|---|---|---|---|
| | V1 | V2 | V3 | V4 | V5 | V6 |
| 1 - Genitor_A (P_A) | 30,00 | 40,00 | 50,00 | 60,00 | 70,00 | 80,00 |
| 2 - Genitor_B (P_B) | 15,00 | 20,00 | 25,00 | 30,00 | 35,00 | 40,00 |
| 3 - F1 | 22,50 | 30,00 | 37,50 | 45,00 | 52,50 | 60,00 |
| 4 - RC1_A | 26,25 | 35,00 | 43,75 | 52,50 | 61,25 | 70,00 |
| 5 - RC1_B | 18,75 | 25,00 | 31,25 | 37,50 | 43,75 | 50,00 |
| 6 - RC2_A | 28,13 | 37,50 | 46,88 | 56,25 | 65,62 | 75,00 |
| 7 - RC2_B | 16,88 | 22,50 | 28,13 | 33,75 | 39,38 | 45,00 |
| 8 - RC3_A | 29,06 | 38,75 | 48,44 | 58,13 | 67,81 | 77,50 |
| 9 - RC3_B | 15,94 | 21,25 | 26,56 | 31,88 | 37,19 | 42,50 |
| 10 - RC4_A | 29.53 | 39.38 | 49.22 | 59.06 | 68.91 | 78.75 |
| 11 - RC4_B | 15.47 | 20.63 | 25.78 | 30.94 | 36.09 | 41.25 |
| Herdabilidade (***h²***) | 30,00 | 40,00 | 50,00 | 60,00 | 70,00 | 80,00 |

Os dados fenotípicos para cada indivíduo das 11 populações foram gerados com auxílio do Software Genes (Cruz, 2013) a partir do procedimento: “Simulação” > “Simulação de delineamentos” > “Fenotipagem de populações segregantes”.

As distâncias de Mahalanobis entre as populações podem ser observadas na Figura 3, demonstrando a gradativa aproximação das populações de retrocruzamento com seu genitor recorrente com o avanço das gerações de retrocruzamento.

Para testar o nível de capacidade de discriminação dos indivíduos pelos métodos, foram formados cinco grupos de populações com distintos níveis de similaridade, sendo o grupo 1 constituído pela população do Genitor A, Genitor B e o Híbrido F1, e a cada grupo formado a partir deste foram acrescentadas as populações da geração de retrocruzamentos seguinte, constituindo assim grupos com nível crescente de similaridade entre as populações (Figura 2).

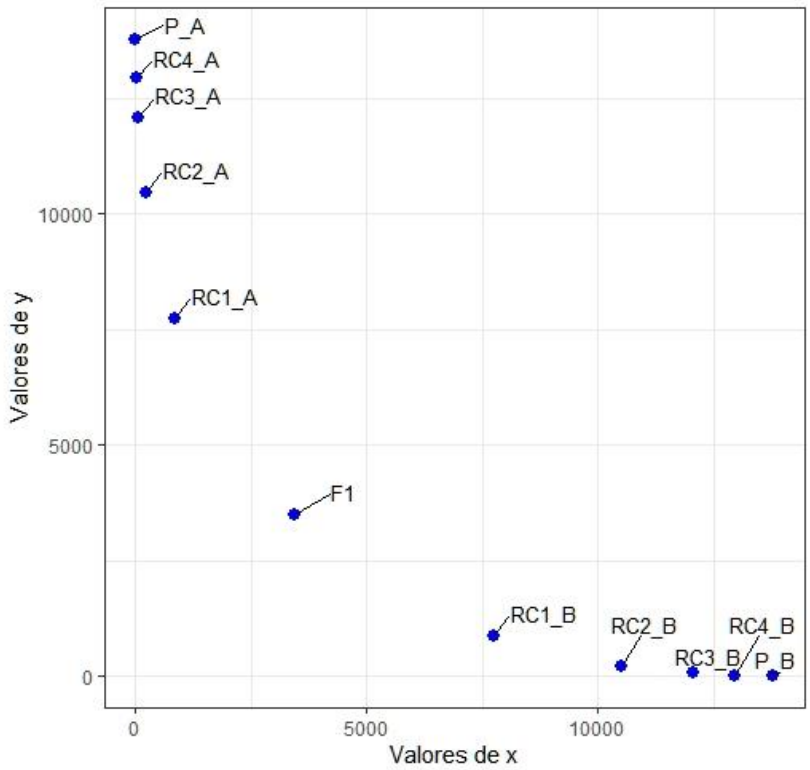



**Figura 3**. Projeção gráfica das distâncias de Mahalanobis entre as onze populações simuladas, utilizando dados fenotípicos.

**Métodos de Classificação**

Os métodos de classificação supervisionada testados foram: Naive Bayes, Árvore de decisão, kNN, Floresta Aleatória, Máquina de Vetor de Suporte e Rede Neural MLP. Estes também foram comparados com os métodos estatísticos clássicos: Função discriminante de Fisher e Função discriminante de Anderson. Os métodos de análise discriminante são aplicados a populações que possuem uma partição definida *a priori*, descritas por diversas variáveis explicativas. O objetivo é discriminar as classes da partição, a partir das características definidas pelas variáveis explicativas. Pretende-se, então, construir uma regra de decisão que permita, no futuro, alocar novos indivíduos, minimizando os erros de alocação.

As análises discriminantes foram realizadas no Software Genes (Cruz, 2013), já o treinamento e classificação por meio dos métodos de aprendizagem supervisionada foram realizados por meio do Software R (R Core Team 2018).

Análise Discriminante de Fisher:

A análise discriminante de Fisher é uma técnica de análise multivariada utilizada para diferenciar ou discriminar populações e classificar ou alocar indivíduos em populações pré-definidas. Para a discriminação são estabelecidas funções das variáveis observadas que sejam responsáveis ou possam explicar as diferenças entre populações. Para a alocação ou classificação, são determinadas as funções que além de separar as populações, sejam capazes de alocar ou classificar novos indivíduos em uma das populações. Segundo Cruz *et al*. (2011) a função discriminante possibilita alocar um determinado indivíduo, com vetor de observações $\tilde{x}$, em uma população *i*, ou *i'*, com máxima probabilidade de acerto, e é dada pela expressão:

$$D_{ii},(\tilde{x}) = \alpha'\tilde{x} = (\mu_i - \mu_{i'})' \sum^{-1} \tilde{x}$$

Análise discriminante de Anderson:

Na análise discriminante de Anderson, inicialmente são consideradas as informações previamente conhecidas do pertencimento de indivíduos a determinadas populações. Então são geradas funções que constituem combinações lineares das características avaliadas e tem por finalidade obter a melhor discriminação entre os indivíduos , alocando-os em suas devidas populações. Uma vez estimadas estas funções, permitem classificar novos genótipos, de comportamento desconhecido, em populações já conhecidas (Cruz 2006).

A função discriminante de Anderson é dada por:

$$D_j(\tilde{x}) = \ln(p_j) + \left(\tilde{x} - \frac{1}{2}\mu_j\right) \sum^{-1} \mu_j$$

Métodos de Classificação Supervisionado:

Neste estudo foram testados os algoritmos: *Naive Bayes*, Árvore de Decisão, Floresta Aleatória, kNN, Máquina de Vetores de Suporte (SVM) e Rede Neural Artificial - Perceptron Multicamada (RNA-MLP). A construção dos modelos de classificação para os diferentes algoritmos foi gerada a partir do Software R (R Core Team 2018) utilizando os pacotes, métodos e configurações descritos na Tabela 2.

A arquitetura de Rede Neural utilizada foi a Perceptron Multicamadas, com três camadas, contendo 15, 35 e 40 neurônios, respectivamente para cada camada. Para os demais algoritmos foram empregadas as configurações padrão de cada pacote do Software R.

**Tabela 2.** Algoritmos, pacotes utilizados no Software R, e configurações utilizadas na geração dos modelos.

| Algoritmos | Pacotes | Método/Configurações |
|---|---|---|
| *Naive Bayes* | klaR | nb |
| Árvore de decisão | Rpart | rpart |
| Floresta Aleatória | randomForest | rf |
| kNN | Knn | knn |
| SVM | Kernlab | svmRadial |
| RNA-MLP | RSNNS | mlpML / layer 1: 15, layer 2: 35, layer 3: 40 |

**Avaliação das metodologias de classificação**

Inicialmente, para evitar vieses no processo de treinamento dos algoritmos, foi utilizada a metodologia da validação cruzada k-fold. Neste procedimento dividimos o conjunto de dados em k subconjuntos, então cada um dos k subconjuntos se revezam como conjunto de validação, sendo o modelo treinado no resto dos k-1 grupos (Zheng 2015). Neste estudo foi utilizado o valor de k igual a 5.

Para cada algoritmo testado, o processo de treinamento e classificação foi repetido 30 vezes e a medida do desempenho da classificação de cada método foi realizada pela métrica de acurácia ou precisão média, que mede a frequência com que o classificador faz a previsão correta. A classificação de novos indivíduos é realizada em populações previamente conhecidas, o que torna possível identificar o número de previsões corretas. Então, o percentual de acurácia foi calculado como segue:

$$\%Acurácia = \frac{nº\ de\ previsões\ corretas}{nº\ total\ de\ indivíduos\ classificados}$$

A média de acurácia das 30 repetições de cada método de classificação supervisionada juntamente com a acurácia dos métodos estatísticos clássicos foram comparadas pelos testes de Friedman e Nemenyi com nível de confiança de 95,0%.

## Resultados e Discussão

Os valores médios de acurácia obtidos de 30 repetições dos métodos de classificação aplicados aos diferentes grupos de similaridade são apresentados na Tabela 3. Pode-se notar que as funções discriminantes de Fisher e Anderson apresentaram desempenho de acurácia idênticas, resultado também obtido por Pereira (2009), e foram satisfatórias para discriminar os conjuntos de populações do Grupo 1 e 2, apresentando 99,0 e 90,8% de acurácia, para os níveis de similaridade de 50 e 75%, respectivamente. Contudo, com o aumento da similaridade entre as populações estes métodos se tornaram pouco eficazes, com acurácia de apenas 35,82% para o grupo 5, de maior similaridade.

**Tabela 3.** Acurácia média dos métodos de classificação para os diferentes grupos de populações com diferentes níveis de similaridade.

| Grupos de Similaridade | Métodos de Classificação | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| | Fisher | Anderson | *Naive Bayes* | Árvore de Decisão | *Random Forest* | kNN | SVM | RNA |
| Grupo 1 | 99,00 | 99,00 | 100,00 | 100,00 | 100,00 | 100,00 | 100,00 | 100,00 |
| Grupo 2 | 90,80 | 90,80 | 100,00 | 100,00 | 100,00 | 100,00 | 100,00 | 100,00 |
| Grupo 3 | 63,57 | 63,57 | 100,00 | 99,44 | 100,00 | 100,00 | 100,00 | 100,00 |
| Grupo 4 | 47,33 | 47,33 | 100,00 | 98,92 | 99,84 | 100,00 | 100,00 | 99,18 |
| Grupo 5 | 35,82 | 35,82 | 97,62 | 84,52 | 97,31 | 97,26 | 97,57 | 86,73 |

Com base nos valores de acurácia foram obtidos os valores médios do ranking dos métodos de classificação para os diferentes grupos de populações, contidos na Tabela 4, que mostraram diferenças estatísticas entre si, conforme teste de Friedman. Estes dados foram utilizados para a comparação de desempenho entre os métodos, tendo como referência a distância crítica segundo Nemenyi (Tabela 5).

**Tabela 4.** Ranking de desempenho médio dos métodos de classificação para os diferentes grupos de populações com diferentes níveis de similaridade.

| Grupos de Similaridade | Métodos de Classificação | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| | Fisher | Anderson | *Naive Bayes* | Árvore de Decisão | *Random Forest* | kNN | SVM | RNA |
| Grupo 1 | 7.50 | 7.50 | 3.50 | 3.50 | 3.50 | 3.50 | 3.50 | 3.50 |
| Grupo 2 | 7.50 | 7.50 | 3.50 | 3.50 | 3.50 | 3.50 | 3.50 | 3.50 |
| Grupo 3 | 7.50 | 7.50 | 3.00 | 6.00 | 3.00 | 3.00 | 3.00 | 3.00 |
| Grupo 4 | 7.50 | 7.50 | 2.32 | 5.73 | 4.50 | 2.32 | 2.32 | 3.82 |
| Grupo 5 | 7.50 | 7.50 | 1.65 | 5.70 | 3.17 | 3.20 | 1.98 | 5.30 |

Friedman: 0,00. Distância crítica segundo Nemenyi, com 95,0% de confiança: 1.917

**Tabela 5.** Comparação entre os métodos de classificação baseados na acurácia, para os grupos de populações com diferentes níveis de similaridade.

| | Fisher | | | | | Anderson | | | | | *Naive Bayes* | | | | | Árvore de Decisão | | | | | Floresta Aleatória | | | | | kNN | | | | | SVM | | | | | RNA-MLP | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | Grupos | | | | | Grupos | | | | | Grupos | | | | | Grupos | | | | | Grupos | | | | | Grupos | | | | | Grupos | | | | | Grupos | | | | |
| | 1 | 2 | 3 | 4 | 5 | 1 | 2 | 3 | 4 | 5 | 1 | 2 | 3 | 4 | 5 | 1 | 2 | 3 | 4 | 5 | 1 | 2 | 3 | 4 | 5 | 1 | 2 | 3 | 4 | 5 | 1 | 2 | 3 | 4 | 5 | 1 | 2 | 3 | 4 | 5 |
| Fisher | | | | | | ○ | ○ | ○ | ○ | ○ | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - |
| Anderson | ○ | ○ | ○ | ○ | ○ | | | | | | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - |
| *Naive Bayes* | + | + | + | + | + | + | + | + | + | + | | | | | | ○ | ○ | ○ | + | + | ○ | ○ | ○ | ○ | ○ | ○ | ○ | ○ | ○ | ○ | ○ | ○ | ○ | ○ | ○ | ○ | ○ | ○ | ○ | + |
| Árvore de Decisão | + | + | + | + | + | + | + | + | + | + | ○ | ○ | ○ | - | - | | | | | | ○ | ○ | ○ | - | - | ○ | ○ | ○ | - | - | ○ | ○ | ○ | - | - | ○ | ○ | ○ | - | ○ |
| *Random Forest* | + | + | + | + | + | + | + | + | + | + | ○ | ○ | ○ | ○ | ○ | ○ | ○ | ○ | + | + | | | | | | ○ | ○ | ○ | ○ | ○ | ○ | ○ | ○ | ○ | ○ | ○ | ○ | ○ | ○ | + |
| kNN | + | + | + | + | + | + | + | + | + | + | ○ | ○ | ○ | ○ | ○ | ○ | ○ | ○ | + | + | ○ | ○ | ○ | ○ | ○ | | | | | | ○ | ○ | ○ | ○ | ○ | ○ | ○ | ○ | ○ | + |
| SVM | + | + | + | + | + | + | + | + | + | + | ○ | ○ | ○ | ○ | ○ | ○ | ○ | ○ | + | + | ○ | ○ | ○ | ○ | ○ | ○ | ○ | ○ | ○ | ○ | | | | | | ○ | ○ | ○ | ○ | + |
| RNA | + | + | + | + | + | + | + | + | + | + | ○ | ○ | ○ | ○ | - | ○ | ○ | ○ | + | ○ | ○ | ○ | ○ | ○ | - | ○ | ○ | ○ | ○ | - | ○ | ○ | ○ | ○ | - | | | | | |

| | |
|---|---|
| ○ | Diferença não significativa entre os métodos de classificação na linha em comparação com a coluna. |
| + | Métodos de classificação na linha com desempenho significativamente superior em comparação com a coluna. |
| - | Métodos de classificação na linha com desempenho significativamente inferior em comparação com a coluna. |

Na comparação da acurácia entre os métodos, pode-se constatar a superioridade dos algoritmos de aprendizagem de máquina sobre as funções discriminantes clássicas (Fisher e Anderson), praticamente em todas as condições de similaridade entre as populações (Tabela 5), mas especialmente naqueles casos de maior similaridade entre as populações. Li *et al*. (2006), comparando o uso da análise discriminante linear de Fisher, SVM e *Naive Bayes* para classificação multi-classe, mostraram que a precisão da função de Fisher é comparável às outras, porém o trabalho não avalia o nível de similaridade entre as classes. No trabalho de Sant'Anna *et al*. (2015), porém, esta perda de acurácia pelas funções discriminantes, tanto de Fisher como de Anderson, para classes ou grupos com maior dificuldade de discriminação, é também observada.

Apesar de diversos trabalhos terem focado nos algoritmos de Redes Neurais com aplicação na classificação em melhoramento de plantas, pode-se observar na Tabela 5 que nas comparações realizadas com o Grupo 5, com similaridade entre as populações chegando até 96,88%, os métodos de Árvore de Decisão e RNA-MLP perderam acurácia, se diferenciando estatisticamente do grupo de algoritmos que manteve alta acurácia nesta condição, composto por: kNN, Floresta Aleatória, SVM e *Naive Bayes*.

No presente trabalho foi utilizada apenas uma configuração de RNA, mas há possibilidade de outras configurações melhorarem a acurácia do método para as condições de maior similaridade entre as populações, porém são escassas as informações que orientam a definição da estrutura das RNA para esta finalidade. Nesse sentido, Sant'Anna *et al*. (2015) testando 64.800 diferentes arquiteturas diferentes de RNA, também constatam perda de acurácia nas condições de maior similaridade entre as populações.

Na análise da matriz de confusão destes métodos (Figura 4), pode-se verificar que nos métodos Árvore de Decisão e RNA-MLP a perda de acurácia se deve à dificuldade de classificação das populações que apresentaram maior similaridade. Corroborado pela Figura 3, pode-se identificar a proximidade entre as populações 1, 10 e 8, e de outro lado a proximidade entre as populações 2, 11 e 9, justamente aquelas onde se encontra os erros de classificação, como por exemplo, no caso do método Árvore de decisão, onde a matriz de confusão mostrou que as populações 10 e 11 foram erroneamente classificadas como populações 1 e 2, com 82,4 e 49,7% de erro, respectivamente.

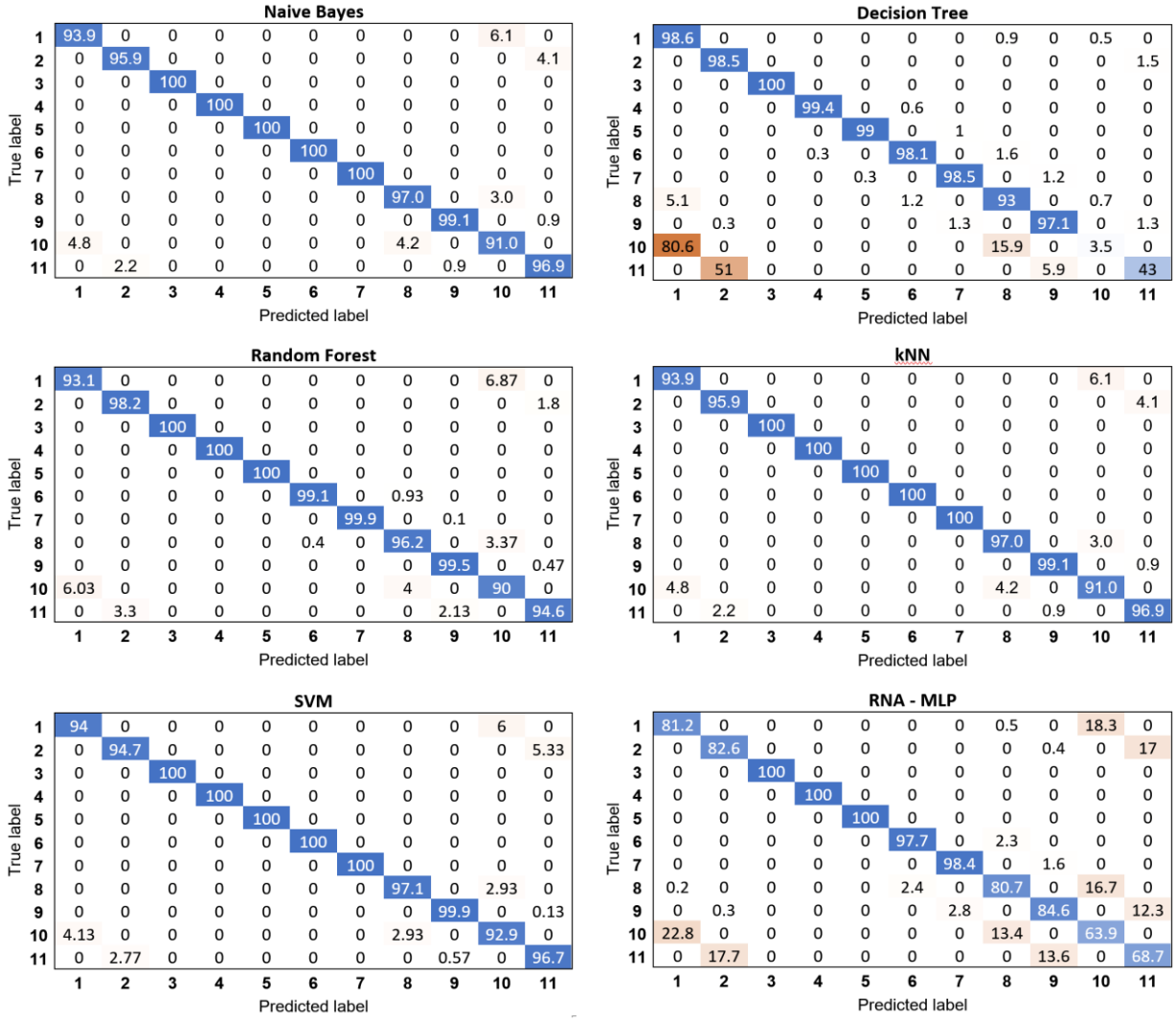



**Figura 4**. Matrizes de confusão dos métodos de classificação de AM para o Grupo 5 de similaridade.

Em concordância com os resultados de testes com dados simulados de Carvalho *et al*. (2018b), utilizando cenários com distintas distribuições (vetores de dados e heterogeneidade de matrizes de covariância), o algoritmo SVM mostrou resultados consistentes de acurácia nas diferentes condições, sendo que RNA e Fisher tiveram bom desempenho somente em cenários específicos. Árvore de Decisão apresentou acurácia inferior à SVM, mas acima de 75%. Em contraposição, em outro trabalho também avaliando SVM, RNA e Fisher em populações com distintas similaridades, os autores não obtiveram resultados de acurácia satisfatório, com valores abaixo de 50% em condições de 87,5% de similaridade entre as populações (Carvalho *et al*., 2018a).

Outra vantagem dos algoritmos que apresentaram melhor desempenho em condições de maior dificuldade de distinguibilidade, é que demandam menor poder computacional do que as RNAs, e que pode constituir um ganho significativo, especialmente quando se trata de classificar um grande número de indivíduos usando muitas variáveis de entrada.

## Conclusões

Os métodos de classificação baseados em algoritmos de aprendizagem de máquina se mostraram superiores às funções discriminantes de Fisher e Anderson, permitindo obter alta acurácia em condições de alta similaridade entre as populações.

Em condições de máxima similaridade avaliada neste trabalho (96,88%), e nas configurações aplicadas, os algoritmos kNN, Random Forest, SVM e *Naive Bayes*, foram os que apresentaram maior eficácia, superando o algoritmo de Árvore de Decisão e até mesmo a RNA-MLP utilizada.

## Referências

BARBOSA, C. D. *et al* . Artificial neural network analysis of genetic diversity in Carica papaya L. **Crop Breed. Appl. Biotechnol. (Online)**, Viçosa, v. 11, n. 3, p. 224-231, Sept. 2011.

BARROSO, L. M. A.; NASCIMENTO, M.; NASCIMENTO, A. C. C.; Silva, F. F.; Ferreira, R. D. P. Uso do Método de Eberhart e Russell como informação a priori para aplicação de redes neurais artificiais e análise discriminante visando a classificação de genótipos de alfafa quanto à adaptabilidade e estabilidade fenotípica. **Revista Brasileira de Biometria**, Lavras – MG, Brasil, v. 31, n. 2, p. 176-188, 2013.

BORÉM, A.; MIRANDA, G. V.; FRITSCHE-NETO, R. **Melhoramento de plantas**. 2017.

CARVALHO, V. P.; SANT'ANNA, I.C.; NASCIMENTO, M.; *et al*. Support vector machines applied to the genetic classification problem of hybrid populations with high degrees of similarity. **Genetics and Molecular Research**, v. 17, n. 4, p. 1–10, 2018a.

CARVALHO, V. P.; DE SOUSA, I. C.; NASCIMENTO, M.; *et al*. Discrimination of populations under covariance matrix heterogeneity and non-normal random vectors in genetic diversity studies. **Científica**, v. 46, n. 4, p. 344, 2018b.

COPPIN, B**. Inteligência artificial**. Rio de Janeiro: LTC, 2010.

CRUZ, C. D. GENES - a software package for analysis in experimental statistics and quantitative genetics. **Acta Scientiarum. Agronomy**, v. 35, n. 3, p. 271–276, 2013.

CRUZ, C. D. **Programa genes: Análise multivariada e simulação**. Viçosa: UFV, 2006. 175 p.

CRUZ, C. D.; FERREIRA, F. M.; PESSONI, L. A. **Biometria aplicada ao estudo da diversidade genética**. Viçosa: Suprema, 2011. p. 2-28.

FONSECA, A.F.A.; SEDIYAMA, T.; CRUZ, C.D.; *et al*. Discriminant analysis for the classification and clustering of robusta coffee genotypes. **Cropp Breeding and Applied Biotechnology**, v. 4, n. 3, p. 285–289, 2004.

FUENTES, S; HERNÁNDEZ-MONTES, E; ESCALONA, J M; *et al*. Automated grapevine cultivar classification based on machine learning using leaf morpho-colorimetry, fractal dimension and near-infrared spectroscopy parameters. **Computers and Electronics in Agriculture**, v. 151, p. 311–318, 2018.

HAYKIN, S. **Neural networks and learning machines.** 3. ed. New York: Prentice Hall, 2009. 936 p.

LI, T.; ZHU, S.; OGIHARA, M. Using discriminant analysis for multi-class classification: an experimental investigation. **Knowledge and Information Systems**, v. 10, n. 4, p. 453–472, 2006.

MAURER, H. P.; MELCHINGER, A. E.; FRISCH, M. Population genetic simulation and data analysis with Plabsoft. **Euphytica**, v. 161, n. 1–2, p. 133–139, 2008.

NAIK, H. S.; ZHANG, J.; LOFQUIST, A.; et al. A real-time phenotyping framework using machine learning for plant stress severity rating in soybean. **Plant Methods**, v. 13, n. 1, p. 23, 2017.

NASCIMENTO, M.; PETERNELLI, L. A.; CRUZ, C. D.; NASCIMENTO, A. C. C.; FERREIRA, R. D. P.; BHERING, L. L.; SALGADO, C. C. Artificial neural networks for adaptability and stability evaluation in alfalfa genotypes. **Crop Breeding and Applied Biotechnology**, Viçosa - MG - Brasil, v. 13, n. 2, p. 152-156, 2013.

NEI, M. Genetic Distance between Populations. **The American Naturalist**, v. 106, n. 949, p. 283–292, 1972.

ODONG, T. L.; VAN HEERWAARDEN, J.; JANSEN, J.; *et al*. Determination of genetic structure of germplasm collections: are traditional hierarchical clustering methods appropriate for molecular marker data? **Theoretical and Applied Genetics**, v. 123, n. 2, p. 195–205, 2011.

OLIVEIRA, A. C. L.; PASQUAL, M.; PIO, L. A. S.; *et al*. Utilização da modelagem matemática (redes neurais artificiais) na classificação de autotetraploides de bananeira (*Musa acuminata Colla*). **Bioscience Journal**, v. 29, n. 3, p. 617–622, 2013.

ORNELLA, L.; TAPIA, E. Supervised machine learning and heterotic classification of maize (*Zea mays* L.) using molecular marker data. **Computers and Electronics in Agriculture**, v. 74, n. 2, p. 250-257, 2010.

R Core Team. *R:* **A language and environment for statistical computing**. R Foundation for Statistical Computing, Vienna, Austria. 2018. URL https://www.R-project.org/.

SANT'ANNA, I. C. *et al*. Superiority of artificial neural networks for a genetic classification procedure. **Genetics And Molecular Research**. Ribeirao Preto: Funpec-editora, v. 14, n. 3, p. 9898-9906, 2015.

SILVA, G.N.; TOMAZ, R.S.; SANT'ANNA, I.C.; *et al*. Evaluation of the efficiency of artificial neural networks for genetic value prediction. **Genetics and Molecular Research**, v. 15, n. 1, p. 1–11, 2016.

SILVA, G. N.; TOMAZ, R. S.; SANT'ANNA, I. C.; *et al*. Neural networks for predicting breeding values and genetic gains. **Scientia Agricola**, v.71, p.494–498, 2014.

SINGH, A.; GANAPATHYSUBRAMANIAN, B.; SINGH, A. K.; *et al*. Machine Learning for High-Throughput Stress Phenotyping in Plants. **Trends in Plant Science**, v. 21, n. 2, p. 110–124, 2016. doi:10.1016/j.tplants.2015.10.015.

VASCONCELOS, E. S.; CRUZ, C. D.; BHERING, L. L.; *et al*. Método alternativo para análise de agrupamento. **Pesquisa Agropecuaria Brasileira**, v. 42, n. 10, p. 1421–1428, 2007.

ZHENG, A. **Evaluating Machine Learning Algorithms: A Beginner's Guide to Key Concepts and Pitfalls**. First Edit. Sebastopol, CA: O'Reilly Media, Inc., 2015.

# CAPÍTULO 2 - Algoritmos de aprendizagem não supervisionada no agrupamento de populações de plantas com diferentes graus de similaridade


## Resumo

O presente estudo objetivou avaliar técnica de agrupamento tradicional e algoritmos de aprendizagem de máquina não supervisionada no agrupamento de populações de plantas com diferentes graus de similaridade. Os métodos de agrupamento hierárquico pelo método de Ward, *k-Means*, *Density-Based Spatial Clustering of Applications with Noise* (DBSCAN), *Self-Organizing Maps* (SOM) e *Affinity Propagation Clustering* (APC), foram testados para o agrupamento de 11 populações de plantas com dados simulados de seis variáveis quantitativas, em dois cenários: com procedimento convencional de definição do número de grupos e com a indução da formação de 11 grupos. Os resultados foram avaliados por meio das matrizes de confusão e por índices internos (Dunn, Calinski Harabasz, Davies Bouldin e Silhouette) e externos (Rand ajustado, Jaccard, Precisão, Recall). No primeiro cenário, os métodos de aprendizagem de máquina não supervisionados mostraram-se superiores ao método Ward. No segundo cenário os métodos k-Means e Ward mostraram-se melhores, e permitiu concluir que a utilização de outros métodos de estimação do número de grupos pode levar a resultados mais acertados, como o índice de Calinski-Harabasz que se mostrou um bom indicador. DBSCAN apresentou baixo índice de precisão nestas condições**,** com populações de alta similaridade, não conseguindo diferenciar os materiais mais próximos geneticamente. O método APC foi consistente nos dois cenários, apresentando bons índices, sendo considerado promissor por não necessitar de definição prévia do número de grupos, facilitando seu uso.



**Palavras-chave**: Aprendizagem de máquina; melhoramento de plantas; métodos de agrupamento; diversidade genética.

## Introdução

O conhecimento do grau de variabilidade genética a partir de estudos de divergência é essencial no processo de identificação de novas fontes de genes de interesse para o melhoramento genético de plantas (Falconer e Mackay, 1996), seja entre e dentro de populações encontradas em suas condições naturais, em bancos de germoplasma ou desenvolvidas nos programas de melhoramento genético. As análises de diversidade desses diferentes genótipos têm permitido estudos da biodiversidade, no sentido de orientar a utilização de diferentes genitores nos programas de melhoramento (Cruz et al., 2011).

As abordagens multivariadas têm sido utilizadas com frequência com este propósito de avaliar a diversidade, e dentre as metodologias mais usuais destacam-se os métodos de agrupamento, especialmente as técnicas hierárquicas, que são as mais amplamente difundidas (Siegmund et al., 2004). Entre os principais métodos de agrupamento hierárquico está o método da variância mínima de Ward (Cruz et al., 2011), que tem sido testado e utilizado com bons resultados na avaliação da diversidade genética de plantas (Silva *et al*., 2017; Pessoa *et al*., 2019).

Contudo, o paradigma da inteligência artificial, amplamente consolidado nas áreas computacionais, tem apresentado resultados promissores e vem se tornando uma ferramenta com crescente aplicação na área de melhoramento de plantas, especialmente na classificação de genótipos (Ornella e Tapia, 2010; Barbosa *et al*., 2011; Oliveira *et al*., 2013; Fuentes *et al*., 2018), na predição de parâmetros genéticos e ganhos no melhoramento (Silva *et al*., 2016) e na avaliação de adaptabilidade e estabilidade de genótipos (Nascimento *et al*., 2013; Barroso *et al*., 2013). Uma das áreas da inteligência artificial que pode ser utilizada para resolver diversos problemas da estatística são as técnicas de aprendizagem de máquina (AM), podendo ser utilizadas nos problemas de classificação, previsão de séries temporais e, em especial, para agrupamento de indivíduos.

Os trabalhos de avaliação comparativa, especialmente com as metodologias estatísticas tradicionais, e de desempenho dos diferentes métodos de AM não supervisionado, aplicados ao melhoramento de plantas ainda são escassos, sendo necessária a ampliação dos estudos nesta área.

Desta forma, o presente trabalho objetivou avaliar algoritmos de aprendizagem de máquina não supervisionada no agrupamento de populações de plantas com

diferentes graus de similaridade, usando técnica de agrupamento tradicional, em dois cenários de uso: com procedimento convencional de definição do número de grupos e com a indução da formação dos 11 grupos.

## Material e Métodos

Na comparação dos métodos de agrupamento foram utilizados dados fenotípicos quantitativos de populações com diferentes graus de similaridade genética, resultantes da simulação de informações genotípicas para diferentes populações submetidas ao esquema de retrocruzamento *in silico*.

O delineamento de retrocruzamentos gera populações com proporção média de similaridade previamente conhecidas (Borém, 2017), sendo uma informação fundamental para compor um conjunto de dados com nível de dificuldade de agrupamento conhecido. Para a geração dos dados empregados nestas comparações foi utilizado o Software Genes (Cruz, 2013).

**Simulação dos dados genotípicos**

Foram simulados dados genotípicos para dez populações com 100 indivíduos cada, em equilíbrio de Hardy-Weinberg, e para cada indivíduo foi gerada informações de 50 locos com dois alelos codominantes cada. Com estes dados foram calculadas as medidas de dissimilaridade genotípica de Nei (1972), selecionando as duas populações mais divergentes para serem utilizadas como genitores parentais (P_A e P_B).

A partir de cruzamentos *in silico* planta-a-planta dos genitores selecionados no sistema de retrocruzamentos, foram geradas mais nove populações com 100 indivíduos cada, no delineamento contido na Figura 1, onde pode-se observar também o percentual de similaridade de cada geração com os genitores, originando assim as 11 populações utilizadas nos testes de agrupamento.

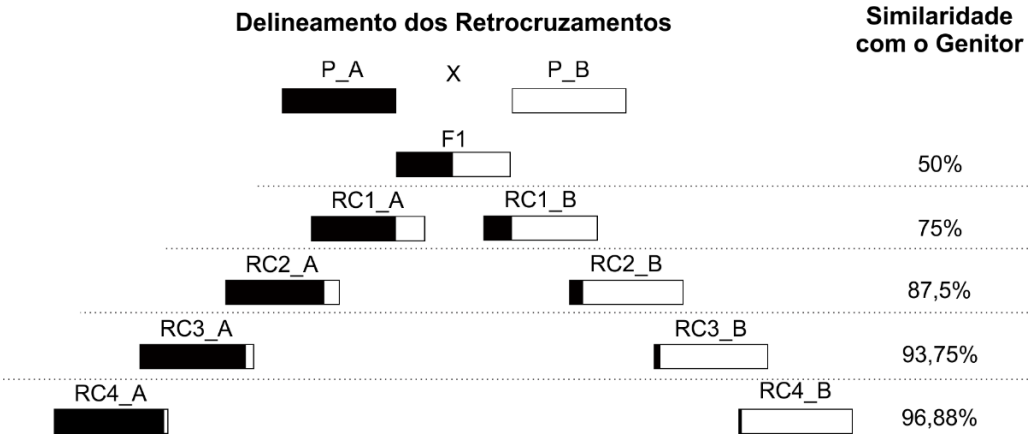



**Figura 1.** Delineamento dos cruzamentos entre os genitores selecionados e suas gerações de retrocruzamentos e similaridade esperada das populações com os genitores.

**Geração dos valores fenotípicos**

Para cada uma das 11 populações foram gerados valores fenotípicos para seis variáveis quantitativas, contínuas, com distribuição normal e com valores de média, herdabilidades e coeficientes de variação previamente estabelecidos. Para o coeficiente de variação experimental foi considerado o valor de 12% em todas as populações.

Os valores foram gerados a partir da ação de alelos de 20 locos, tomados ao acaso entre os 50 previamente simulados, com efeito aditivo diferencial e com pesos da importância do loco sobre a variabilidade genotípica total do caráter, estabelecidos a partir de uma distribuição binomial e grau médio de dominância nulo.

O delineamento experimental adotado foi inteiramente casualizado, utilizando o modelo estatístico determinante do valor fenotípico que segue:

$$Y_{ij} = \mu_j + G_i + \varepsilon_{ij}$$

Em que:
$Y_{ij}$: observação simulada de uma dada característica para o i-ésimo indivíduo pertencente à j-ésima população;
$\mu_j$: média geral da característica especificada previamente para a j-ésima população;
$G_i$: efeito genotípico do i-ésimo indivíduo;
$\varepsilon_{ij}$: erro aleatório

Foram escolhidos os valores de herdabilidade de 30, 40, 50, 60, 70 e 80%, respectivamente, para cada uma das seis variáveis simuladas, com valores de média semelhante ao valor da herdabilidade para o Genitor A e 50% dos valores de média

do Genitor A para o Genitor B. As médias das demais populações foram calculadas tendo em vista o percentual médio esperado dos genótipos parentais em cada geração (Tabela 1).

**Tabela 1.** Médias paramétricas e herdabilidades das características simuladas para cada população.

| Populações | Variáveis quantitativas | | | | | |
|---|---|---|---|---|---|---|
| | V1 | V2 | V3 | V4 | V5 | V6 |
| 1 - P_A | 30,00 | 40,00 | 50,00 | 60,00 | 70,00 | 80,00 |
| 2 - P_B | 15,00 | 20,00 | 25,00 | 30,00 | 35,00 | 40,00 |
| 3 - F1 | 22,50 | 30,00 | 37,50 | 45,00 | 52,50 | 60,00 |
| 4 - RC1_A | 26,25 | 35,00 | 43,75 | 52,50 | 61,25 | 70,00 |
| 5 - RC2_A | 28,13 | 37,50 | 46,88 | 56,25 | 65,62 | 75,00 |
| 6 - RC3_A | 29,06 | 38,75 | 48,44 | 58,13 | 67,81 | 77,50 |
| 7 - RC4_A | 29.53 | 39.38 | 49.22 | 59.06 | 68.91 | 78.75 |
| 8 - RC1_B | 18,75 | 25,00 | 31,25 | 37,50 | 43,75 | 50,00 |
| 9 - RC2_B | 16,88 | 22,50 | 28,13 | 33,75 | 39,38 | 45,00 |
| 10 - RC3_B | 15,94 | 21,25 | 26,56 | 31,88 | 37,19 | 42,50 |
| 11 - RC4_B | 15.47 | 20.63 | 25.78 | 30.94 | 36.09 | 41.25 |
| Herdabilidade ($h^2$) | 30,00 | 40,00 | 50,00 | 60,00 | 70,00 | 80,00 |

Com os valores fenotípicos foram calculadas as distâncias médias de Mahalanobis entre as populações, gerando os agrupamentos pelo método de Ward (Figura 2).

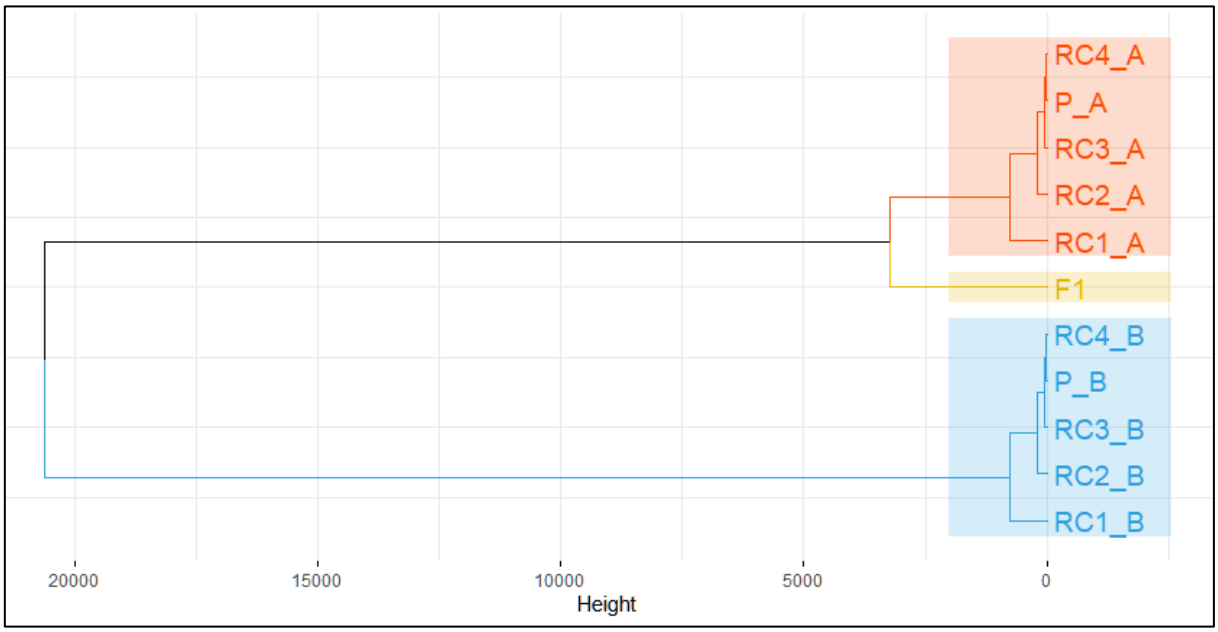



**Figura 2**. Dendrograma do agrupamento das onze populações simuladas pelo método de Ward, utilizando as distâncias de Mahalanobis.

O agrupamento é um indicativo da tendência de como as populações deverão ser agrupadas pelos métodos testados neste trabalho, sendo possível a distinção de pelo menos três grupos nas populações na Figura 2. Pode-se observar ainda que há uma grande similaridade das populações de retrocruzamento com seu genitor recorrente, o que pode torna o processo de distinção entre as populações mais difícil.

**Métodos de agrupamento**

Utilizamos cinco métodos de agrupamento, sendo um método estatístico clássico, o Agrupamento Hierárquico pelo método de Ward, que serviu de referência para comparação com os métodos de agrupamento não supervisionados, a saber: *k-Means*, *Density-Based Spatial Clustering of Applications with Noise* (DBSCAN), *Self-Organizing Maps* (SOM) e *Affinity Propagation Clustering* (APC).

Ward é um método hierárquico e aglomerativo (Ward, 1963) que consiste na formação de grupos pela maximização da homogeneidade dentro dos grupos (Mingoti, 2005), obtida pela minimização da soma de quadrados dentro deste. Esse método tende a resultar em agrupamentos de tamanhos aproximadamente iguais devido a sua minimização de variação interna (Hair et al., 2005).

k-Means procura uma partição que minimize a soma dos erros quadráticos (SEQ) entre os objetos de um conjunto de dados e o centróide dos seus respectivos grupos (Jain, 2010). O algoritmo k-means é um procedimento de otimização com inicialização aleatória que garante a convergência para um mínimo local da SEQ.

DBSCAN busca por grupos definidos, como regiões com alta densidade de objetos, separados por regiões de baixa densidade. Uma das principais vantagens desse algoritmo advém do fato de não ser necessário informar previamente o número desejado de grupos. Para isso, baseia na classificação de cada objeto do conjunto de dados em uma dentre 3 categorias: objeto central - todo objeto $x_i$ que somado ao número de objetos dentro de um raio máximo igual a *eps,* totaliza uma quantidade maior ou igual a um parâmetro *MinPts*; objeto de borda - todo objeto que não satisfaz as condições para objeto central, mas que pertence à vizinhança de um objeto central; ruído - todo objeto que não pertence a nenhuma das duas categorias anteriores (Ester *et al*. 1996).

A Figura 3 ilustra um exemplo do agrupamento de um conjunto de objetos considerando MinPts = 3. Nela, os objetos azuis são centrais, o verde de borda e o vermelho é ruído. Os círculos em torno de cada objeto denotam o raio que define as vizinhanças, ajustado pelo parâmetro *eps*.

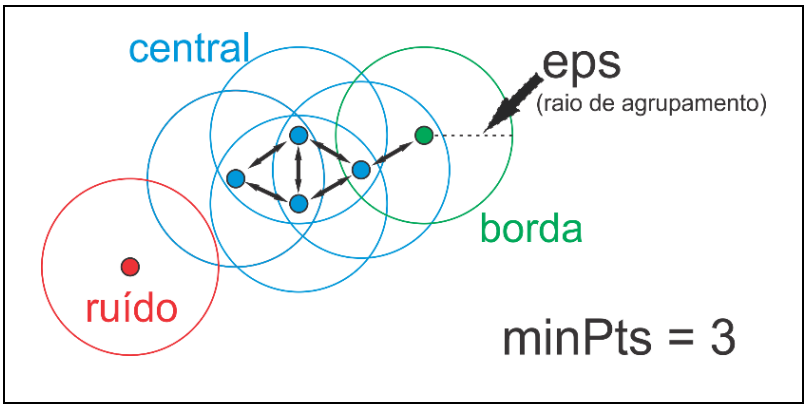



**Figura 3**. Exemplo de agrupamento de objetos feito pelo DBSCAN.

SOM refere-se aos mapas auto-organizáveis que constitui um tipo de rede neural não supervisionada desenvolvida por Kohonen (Kohonen, 1990). Esta técnica realiza uma redução de dimensionalidade, mostrando em um mapa bidimensional as informações de similaridade existentes entre amostras compostas de diversos atributos, sendo utilizado na classificação e análise de dados (Vesanto e Alhoniemi, 2000). A quantidade de células utilizadas foi definida a partir da regra empírica $5\sqrt{N}$, onde N é o número de pontos para análise.

APC refere-se ao agrupamento por propagação de afinidade. É um algoritmo introduzido por Frey e Dueck (2007). O método toma como entrada medidas de similaridade entre pares de pontos de dados e identifica pontos que sejam representativos, “exemplares” dos conjuntos de dados, no entorno dos quais forma os agrupamentos. Ele opera considerando simultaneamente todos os pontos de dados como possíveis exemplos e realiza a troca de “mensagens” entre os pontos de dados até que um bom conjunto de exemplos e grupos de alta qualidade surja.

Os agrupamentos pelos métodos testados foram realizados através do Software R (R Core Team 2018) por meio de pacotes específicos e configurações descritas na Tabela 2.

Os dados de entrada utilizados em cada método, proveniente da simulação, foram normalizados entre valores zero e um, para evitar que um atributo seja dominante nas medidas de distância.

**Tabela 2.** Algoritmos, pacotes e configurações utilizados no Software R para os agrupamentos.

| Algoritmos | Pacotes | Método / Configurações |
|---|---|---|
| Hierárquico | stats | hclust / distance = ‘euclidean’, method = ‘ward.D’ |
| K-Means | ClusterR | KMeans_rcpp / num_init = *n*, max_iters = 100, initializer = 'kmeans++' |
| DBSCAN | dbscan | dbscan / eps = 0.14, minPts = 3 (7 grupos)<br>eps = 0.0955, minPts = 3 (11 grupos) |
| SOM | kohonen | som / grid (13 x 13, hexagonal), rlen=10000, alpha=c(0.05,0.01) |
| APC | apcluster | apcluster / negDistMat(r = 2) |

Com exceção dos métodos DBSCAN e APC, os demais necessitam da definição prévia do número de grupos a serem formados. Para tanto, foram utilizadas metodologias de definição de grupos comumente indicadas. Para o método de Ward e k-Means foi utilizada metodologia proposta por Kassambara (2017) que, por meio da função ‘*NbClust’* do Software R, calcula o número de grupos ideais por meio de 30 índices diferentes, sendo recomendada a quantidade de grupos com maior frequência de indicação. Para o método SOM foi utilizando o “método cotovelo” (Kassambara, 2017), baseado na métrica da soma dos quadrados dentro dos grupos (WCSS). A cada aumento no número de grupos o WCSS diminui, sendo que o número estimado de grupos ideal é definido quando ocorre a última alteração substancial da métrica e as alterações restantes são insignificantes.

Os métodos foram avaliados em dois cenários diferentes. Como tem-se a informação prévia de que os dados estão divididos em 11 populações distintas, além do procedimento comumente indicado para formação dos grupos pelos diferentes métodos testados, descrito anteriormente e que passam a ser denominados de procedimento convencional, ainda aplicou-se os mesmos métodos de forma a induzir a formação de 11 grupos, com a finalidade de identificar o desempenho dos métodos nestas condições.

Nos dois cenários de aplicação dos métodos, com procedimento convencional e formando 11 grupos, os métodos foram avaliados por índices internos ou externos descritos como segue.

**Avaliação dos métodos de agrupamento**

Para avaliar a qualidade dos agrupamentos encontrados, inicialmente foram geradas matrizes de confusão para cada método, que permitem avaliar visualmente a quantidade e a qualidade dos grupos formados, bem como calculadas algumas medidas de avaliação ou validação. As medidas ou critérios de validação podem ser relativos, internos ou externos.

Critérios de validação interna medem o grau em que uma solução de agrupamento é justificada com base apenas no conjunto de dados original ou em uma matriz de similaridades ou dissimilaridades calculadas a partir do mesmo. Assim, um índice de validação interna pode ser visto como o grau de concordância entre um agrupamento encontrado por um algoritmo de agrupamento e o próprio conjunto de dados (Giancarlo e Utro, 2019).

Critérios de validação externa avaliam o grau de concordância entre duas soluções de agrupamento de dados. Em muitos casos, uma das partições comparadas consistirá em uma solução obtida por algum algoritmo, enquanto a outra partição representará uma solução de referência para o conjunto de dados estudado (Giancarlo e Utro, 2019).

Para avaliar os métodos foram utilizados, neste trabalho, critérios internos e, uma vez que se dispõe de uma partição de referência que são as populações previamente conhecidas dos dados simulados, utilizou-se também critérios externos para validação dos agrupamentos, descritos na sequência.

Critérios de validação interna

Índice de Dunn: é um critério de validação relativa baseado na ideia de compactação intragrupo e separação intergrupos. É apropriado para identificar agrupamentos que contém grupos cujos objetos estão próximos entre si e distantes de objetos contidos em outros grupos (Vendramin *et al*., 2010). Portanto, valores altos desse índice indicam soluções que obedecem a esta condição.

Índice Calinski-Harabasz: método estatístico utilizado para determinação da distribuição ótima de objetos em grupos, indicado pelo maior valor do índice (Maulik e Bandyopadhyay, 2002).

Índice de Davies-Bouldin: utiliza características do próprio conjunto de dados para avaliar o número de divisões pela relação entre a distância intra-classe e inter-classes, com base em uma medida de dispersão de um cluster *i* e uma medida de dissimilaridade entre dois clusters (*i* e *j)*. Quanto menor o valor do índice, melhor, pois significa baixas medidas de dispersão intragrupo e grandes distâncias intergrupo (Halkidi *et al*., 2001).

Largura de silhueta (Silhouette): assim como o índice de Dunn, a largura de silhueta baseia-se nos conceitos de compactação intragrupo e separação intergrupos, e da mesma forma os valores altos expressam boas soluções de agrupamento.

Critérios de validação externa

Índice Rand ajustado: calcula a proporção de concordância no agrupamento de pares de objetos entre duas partições (a obtida e a de referência) em relação ao total de pares possíveis de objetos. Esta medida retorna valores no intervalo (-∞;1], onde valores positivos indicam que a similaridade entre as partições é maior do que o valor esperado ao comparar agrupamentos gerados aleatoriamente (Hubert e Arabie, 1985). Valores mais altos indicando uma maior similaridade entre duas partições.

Índice Jaccard: calcula a proporção de pares de objetos agrupados conjuntamente na partição obtida e na partição de referência em relação à quantidade de pares de objetos em um mesmo grupo na partição obtida ou na partição de referência. O índice está contido no intervalo [0;1] com valores mais altos apontando uma maior concordância entre as partições (Jaccard, 1908).

Precisão: definido como a proporção de pontos agrupados corretamente na partição obtida, de acordo com a partição de referência, em relação ao total de pontos agrupados. Assim, este índice indica a qualidade de separação intergrupos.

*Recall*: refere-se ao coeficiente de revocação como a proporção de pontos agrupados na partição de referência e que são efetivamente também agrupados na partição obtida, indicando, portanto a eficácia na compactação intragrupos.

Todos os índices e coeficientes utilizados nas comparações dos métodos nos diferentes cenários foram gerados a partir do Software R por meio do pacote *'clusterCrit'* (Desgraupes, 2018).

# Resultados e Discussão

## Cálculo do número de grupos

Os resultados obtidos pela função '*NbClust*' de frequência de indicação do número de grupos (Figura 4), sugerem que para o método Ward forme-se três grupos. Para k-Means temos a indicação mais frequente de 2 grupos, porém, como se trata de índices que apontam apenas estimativas e tomando em conta que para o melhoramento busca-se a maior diversidade possível, foi escolhida a quantidade de sete grupos para este método, destacando-se que esta quantidade também teve uma frequência alta.

Para o SOM optou-se por cinco grupos, com base no método do cotovelo, indicado pelo ponto vermelho no gráfico, onde há redução significativa na diferença do valor da soma de quadrado de um número de grupo para outro (Figura 4c).

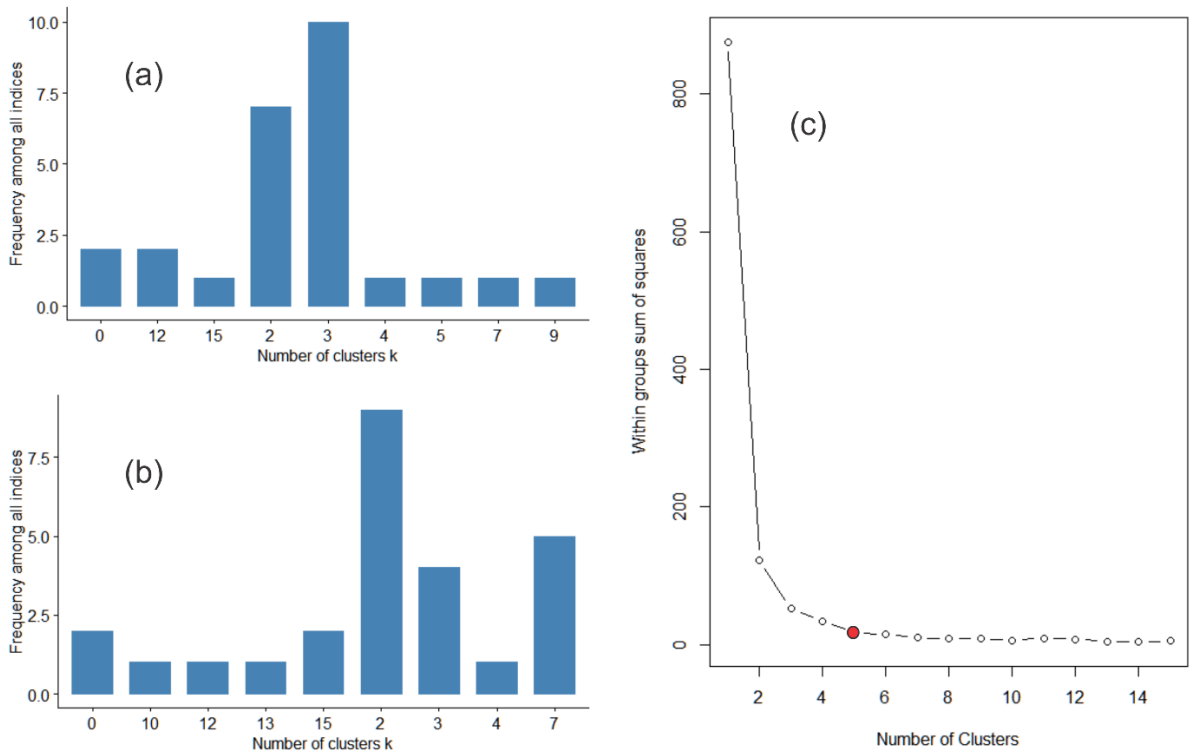



**Figura 4**. (a) e (b) Frequência de indicação do número de grupos segundo 30 índices, para o método Hierárquico de Ward e k-Means respectivamente, e (c) Método cotovelo para indicação do número de grupos para o método SOM.

Outra forma indicada para quantificar o número de grupos no SOM é pela matriz-U, que mostra a distância entre os neurônios vizinhos, buscando dessa forma regiões no mapa com maior proximidade e que são divididas por áreas de maior distância. Porém se for analisada a Figura 5, pode-se identificar visualmente apenas dois grupos.

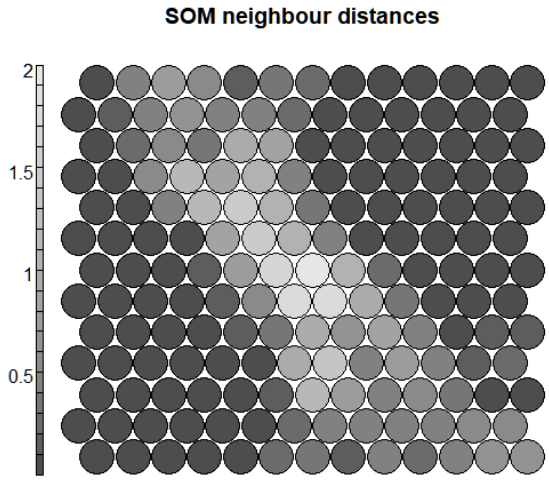



**Figura 5**. Matriz-U do método SOM, com as distâncias entre os neurônios vizinhos.

O ajuste dos parâmetros do método DBSCAN para a melhor qualidade de agrupamento resultou na formação de sete grupos, e apenas um ponto que foi classificado como ruído, indicado como grupo zero.

O método APC, que define automaticamente a partição em um número de grupos mais adequada aos dados, gerou a maior quantidade de grupos dentre os métodos testados, formando nove grupos.

**Resultados com procedimento convencional**

Observa-se nas matrizes de confusão (Figura 6) que houve coerência em todos os métodos na alocação de grupos de maior similaridade em partições comuns. Assim, como os métodos criaram menor número de partições em relação as populações existentes, estes agruparam corretamente as últimas gerações de retrocruzamento com seus respectivos genitores recorrentes, conforme indicado previamente na Figura 2.

Ainda com base nas matrizes de confusão, constata-se que há grande coesão intragrupo para todos os métodos, ou seja, os indivíduos pertencentes a uma determinada população foram agrupados em grande parte numa mesma partição,

com exceção do método k-Means que teve poucos indivíduos, apenas 20 da terceira geração de retrocruzamento, agrupados em partições diferentes da maioria do restante da população, e do APC que dividiu praticamente ao meio as últimas populações de retrocruzamento. Isto é esperado devido à alta similaridade destas com os genitores P_A e P_B (96,88%), assim como com a terceira geração de retrocruzamento (96,87%), sendo condizente no agrupamento com o esperado.

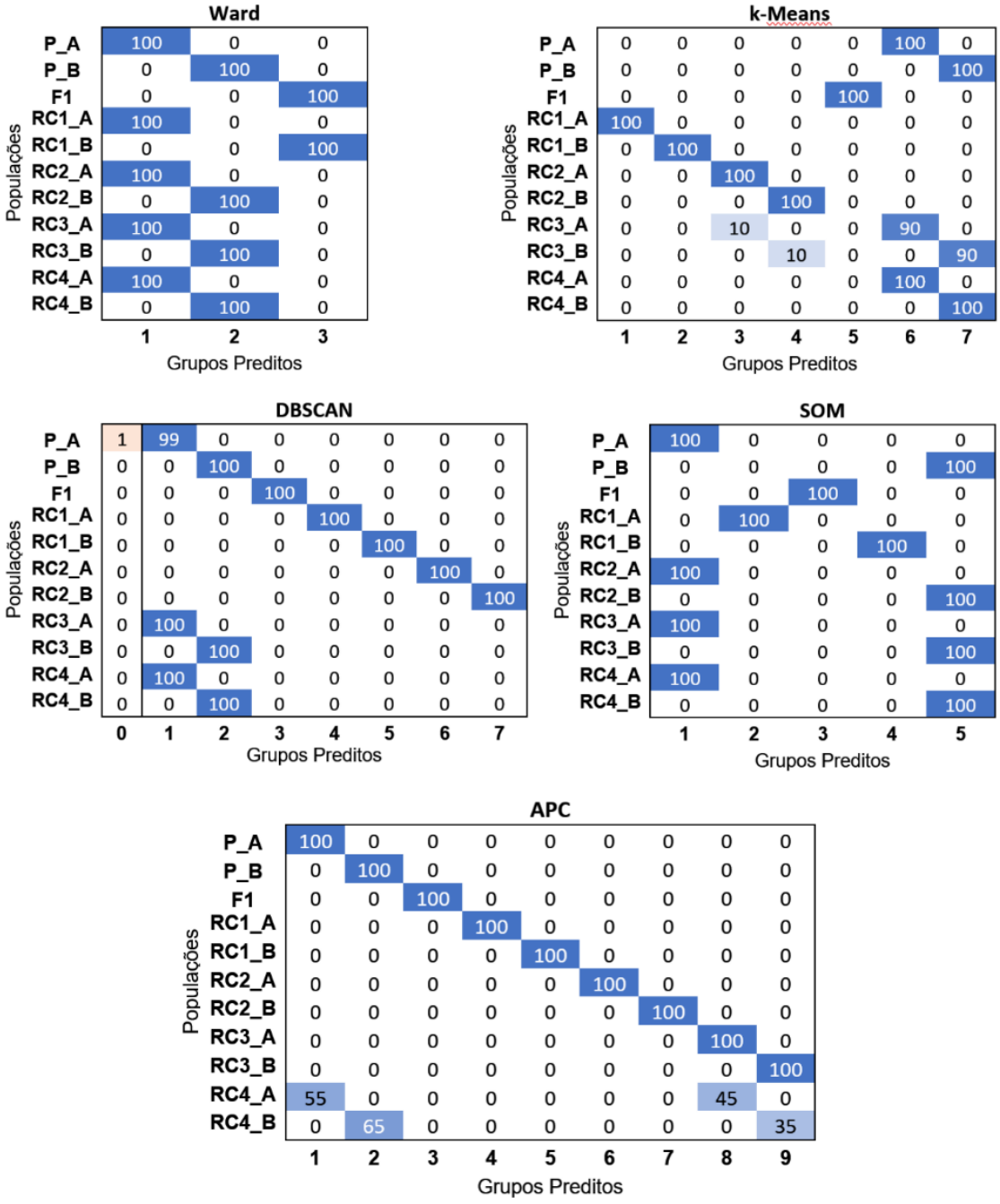



**Figura 6**. Matrizes de confusão dos resultados dos métodos de agrupamento obtidos com base no procedimento convencional.

Considerando a expectativa da capacidade de detecção do maior número possível das 11 populações, o método APC se destaca pela formação de quantidade maior de partições, e com a manutenção de um alto nível de agrupamento intragrupos. Isso é corroborado por alguns índices calculados (Tabela 3), especialmente aqueles externos, que comparam os agrupamentos gerados com o agrupamento original de referência, que são as 11 populações.

**Tabela 3.** Índices internos e externos dos métodos de agrupamento para o procedimento convencional.

| Métodos | Índices Internos | | | | Índices Externos | | | |
|---|---|---|---|---|---|---|---|---|
| | Dunn | Calinski Harabasz | Davies Bouldin | Silhouette | Rand ajustado | Jaccard | Precisão | Recall |
| Ward | 0,2537 | 10661,1 | 0,48723 | 0,6847 | 0,2871 | 0,2426 | 0,2426 | **1,0000** |
| k-Means | 0,0625 | 37069,9 | 0,43057 | 0,6874 | 0,5881 | 0,4682 | 0,4758 | 0,9669 |
| DBSCAN | 0,1986 | 31581,8 | 0,41150 | 0,7055 | 0,5954 | 0,4757 | 0,4761 | 0,9982 |
| SOM | **0,4158** | 20842,1 | **0,32069** | **0,7720** | 0,3923 | 0,3121 | 0,3121 | **1,0000** |
| APC | 0,0521 | **45691,4** | 0,59756 | 0,5941 | **0,7782** | **0,6676** | **0,7131** | 0,9128 |

Os índices de Rand e Jaccard mostraram comportamento similar a Precisão, tanto na Tabela 3 como na Tabela 4, sendo essa última mais fácil de ser interpretada por mostrar diretamente a proporção de acerto na formação das partições e alocação dos grupos, portanto focou-se na análise com esta métrica em detrimento das outras. Este índice aponta 71,31% de acerto para APC, seguido de DBSCAN e k-Means com 47,61% e 47,58% de acerto**,** respectivamente.

O método hierárquico de Ward, que é uma das metodologias clássicas recomendadas no melhoramento de plantas, apresentou o pior desempenho em precisão (24,26%) seguida do SOM (31,21%). Da mesma forma, Santos *et al.* (2018) encontraram respostas similares dos dois métodos na ordenação dos genótipos, comparando os métodos SOM e agrupamento hierárquico UPGMA no estudo da diversidade genética de arroz irrigado.

Diferentemente dos resultados obtidos neste trabalho, Campos *et al.* (2016), em um trabalho com goiaba, concluíram que a metodologia SOM é eficiente para detectar a divergência genética e a formação de grupos heteróticos para a espécie, com consistência do agrupamento de 86%. Porém**,** este valor foi determinado por comparação com a Análise discriminante linear, também conhecida como função

discriminante de Fisher, metodologia que se mostrou ineficiente quando há alta similaridade entre as populações (Sant'anna *et al.,* 2015).

Recall é uma métrica que se diferencia dos demais índices externos avaliados, pois avalia a acurácia na aglomeração intragrupo, e neste aspecto os métodos de Ward e SOM tiveram o máximo desempenho, pois agruparam todos os indivíduos que deveriam ser agrupados segundo os dados de referência. Porém, os demais métodos também apresentaram valores muito altos, acima de 90%.

**Resultados com a indução da formação de 11 grupos**

Quando induzimos a formação dos 11 grupos pelos métodos testados, notamos nas matrizes de confusão (Figura 7) que tanto DBSCAN como SOM, apesar de criarem 11 partições, alocaram poucos ou nenhum indivíduo em quatro e três partições, respectivamente, sendo que no caso de DBSCAN ainda houveram 52 indivíduos alocados como ruído (grupo 0, em vermelho). Mas apesar dos dois métodos terem baixa Precisão (Tabela 4), SOM apresentou o maior valor de Recall, o que demonstra consistência do método, mantendo a agregação intragrupo, e formando oito grupos distintos. Adicionalmente, o método SOM tem a vantagem de permitir uma avaliação mais detalhada sobre a influência e importância de cada variável (Wehrens e Buydens, 2007).

Apesar de perder Precisão e valor de Recall, o método APC ainda manteve valores significativos para estes índices, 80,04% e 89,80%, respectivamente, sendo o melhor avaliado conforme os índices de Calinski-Harabasz e Dunn.

k-Means foi o melhor método com 11 partições, com o maior valor de Precisão (93,69%) e alto valor de Recall (93,88%). Porém, com valores próximos, destaca-se também o método hierárquico de Ward, com a pior classificação no primeiro teste, e agora apresentou Precisão de 91,71% e Recall de 92,78%.

Para os métodos que necessitam da definição prévia do número de grupos a serem formados, observamos que as formas de estimativas comumente indicadas podem estar subestimando estes valores, indicando a formação de um número menor de grupos do que realmente os métodos de agrupamento tem potencial de separar, como é o caso de k-Means e Ward, o que confirma a afirmação de Mingoti (2005) de que estas estimativas são subjetivas e não fornecem respostas exatas.

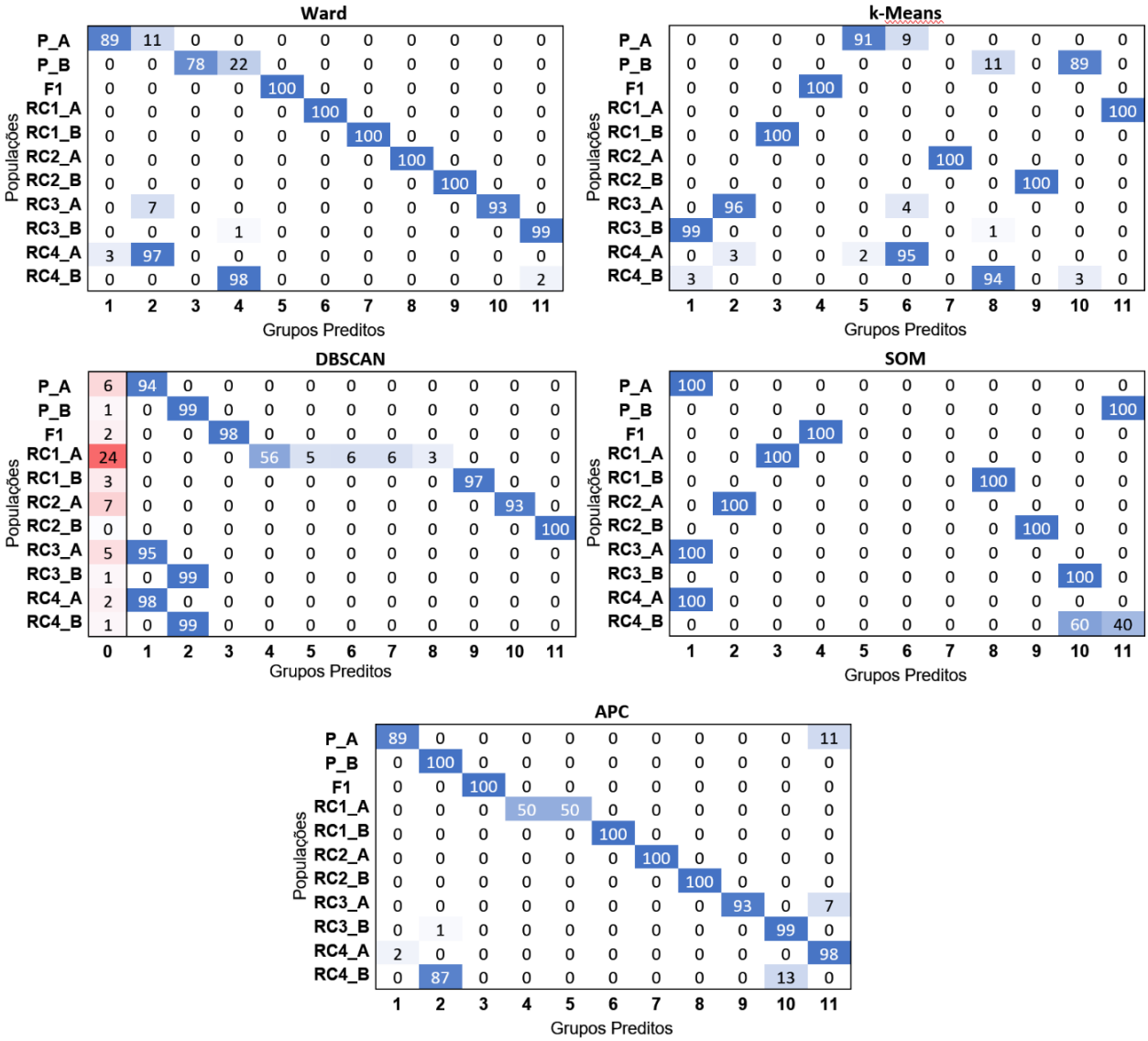



**Figura 7**. Matrizes de confusão dos resultados dos métodos de agrupamento obtidos com a indução da formação de 11 grupos.

**Tabela 4.** Índices internos e externos dos métodos de agrupamento para o procedimento com a indução da formação de 11 grupos.

| Métodos | Índices Internos | | | | Índices Externos | | | |
|---|---|---|---|---|---|---|---|---|
| | Dunn | Calinski Harabasz | Davies Bouldin | Silhouette | Rand ajustado | Jaccard | Precisão | Recall |
| Ward | 0,0766 | 42872,2 | 0,7522 | 0,5165 | 0,9147 | 0,8560 | 0,9171 | 0,9278 |
| k-Means | 0,0637 | 43674,1 | 0,7394 | 0,5205 | **0,9316** | **0,8829** | **0,9369** | 0,9388 |
| DBSCAN | 0,0152 | 4188,6 | 5,1006 | 0,4026 | 0,5514 | 0,4335 | 0,4570 | 0,8942 |
| SOM | 0,0504 | 38912,4 | **0,5161** | **0,6417** | 0,6736 | 0,5511 | 0,5655 | **0,9559** |
| APC | **0,0788** | **44163,1** | 0,7447 | 0,5137 | 0,8302 | 0,7337 | 0,8004 | 0,8980 |

Nesse sentido, para esta condição de teste que o presente trabalho foi desenvolvido, o índice interno Calinski-Harabasz parece ser o mais indicado para prever a quantidade de grupos mais adequada, aproximando-se da classificação dos índices externos em conjunto. Como os índices Davies-Boudin e Silhouette novamente mostraram uma avaliação centrada na agregação intragrupo, não são recomendados para esse fim, uma vez que no melhoramento busca-se a identificação da maior diversidade possível, o que se dá no particionamento intergrupo, como mencionado anteriormente.

## Conclusões

Nas condições em que são utilizados os estimadores de números de grupos recomendados usualmente, os métodos de agrupamentos baseados em algoritmos de aprendizagem de máquina não supervisionado testados são superiores quando comparados ao método estatístico clássico de agrupamento hierárquico de Ward.

Os melhores métodos de agrupamento na condição de indução da formação de um número de grupos maior do que o obtido pelos métodos comumente indicados de estimação foram k-Means e Ward, conseguindo níveis significativos de diferenciação intergrupos e alta agregação intragrupos. O que leva a concluir também que a utilização de outros métodos para definição do número de grupos pode levar a resultados mais acertados, como o índice de Calinski-Harabasz que se mostrou um bom indicador neste estudo.

É importante a continuidade das pesquisas, especialmente no que diz respeito aos métodos de estimação da quantidade de grupos a serem formados, buscando índices que propiciem uma maior precisão nessa estimativa.

Dentre os métodos testados, DBSCAN não se mostrou bom para agrupamento de populações com alta similaridade, por não conseguir diferenciar os genótipos mais próximos geneticamente e apresentando baixo índice de precisão.

O método SOM apresenta desempenho intermediário, mas ainda deve ser considerado quando é necessária a avaliação individual da importância de cada variável.

O método APC é consistente nas duas condições de teste, apresentando bons índices, especialmente os internos, e se mostra promissor por não necessitar da definição prévia do número de grupos, tornando-se prático em seu uso.

# Referências


BARBOSA, C. D. *et al* . Artificial neural network analysis of genetic diversity in Carica papaya L. **Crop Breed. Appl. Biotechnol. (Online)**, Viçosa, v. 11, n. 3, p. 224-231, Sept. 2011.

BARROSO, L. M. A.; NASCIMENTO, M.; NASCIMENTO, A. C. C.; Silva, F. F.; Ferreira, R. D. P. Uso do Método de Eberhart e Russell como informação a priori para aplicação de redes neurais artificiais e análise discriminante visando a classificação de genótipos de alfafa quanto à adaptabilidade e estabilidade fenotípica. **Revista Brasileira de Biometria**, Lavras – MG, Brasil, v. 31, n. 2, p. 176-188, 2013.

BORÉM, A.; MIRANDA, G. V.; FRITSCHE-NETO, R. **Melhoramento de plantas**. 2017.

CAMPOS, B. M.; VIANA, A. P.; QUINTAL, S. S. R.; *et al.* Heterotic group formation in psidium guajava l. by artificial neural network and discriminant analysis. **Revista Brasileira de Fruticultura**, v. 38, n. 1, p. 151–157, 2016. doi: 10.1590/0100-2945-258/14.

CRUZ, C. D. GENES - a software package for analysis in experimental statistics and quantitative genetics. **Acta Scientiarum. Agronomy**, v. 35, n. 3, p. 271–276, 2013.

CRUZ, C. D.; FERREIRA, F. M.; PESSONI, L. A. **Biometria aplicada ao estudo da diversidade genética**. Viçosa: Suprema, 2011. p. 2-28.

DESGRAUPES, B. **Package "clusterCrit"**. [s.l.: s.n.], 2018. Disponível em: <https://cran.r-project.org/web/packages/clusterCrit/clusterCrit.pdf>. Acesso em: 20 out. 2019.

ESTER, M.; KRIEGEL, H. P.; SANDER, J.; et al. A Density-Based Algorithm for Discovering Clusters in Large Spatial Databases with Noise. In: **Kdd**, v. 96, p. 226–231, 1996.

FALCONER, D. S.; MACKAY, T. F. C. **Introduction to quantitative genetics, Longman**. Essex, England, 1996.

FREY, B. J.; DUECK, D. Clustering by Passing Messages Between Data Points. **Science**, v. 315, n. 5814, p. 972–976, 2007. doi: 10.1126/science.1136800.

FUENTES, S; HERNÁNDEZ-MONTES, E; ESCALONA, J M; *et al*. Automated grapevine cultivar classification based on machine learning using leaf morpho-colorimetry, fractal dimension and near-infrared spectroscopy parameters. **Computers and Electronics in Agriculture**, v. 151, p. 311–318, 2018.

GIANCARLO, R.; UTRO, F. Computation Cluster Validation in the Big Data Era. In: **Encyclopedia of Bioinformatics and Computational Biology**. [s.l.]: Elsevier, 2019, p.449–455. doi:10.1016/b978-0-12-809633-8.20385-3.

HAIR, J. F.; et al. **Análise multivariada de dados**. Trad. Adonai S. Sant'Anna e Anselmo C. Neto. 5 ed. Porto Alegre: Bookman, 2005.

HALKIDI, M.; BATISTAKIS, Y.; VAZIRGIANNIS, M. On clustering validation techniques. **Journal of Intelligent Information Systems**, v. 17, n. 2–3, p. 107–145, 2001. doi: 10.1023/A:1012801612483

HUBERT, L.; ARABIE, P. Comparing partitions. **Journal of Classification**, v.2, n.1, p.193–218, 1985. doi: 10.1007/BF01908075.

JACCARD, P. Nouvelles Recherches Sur la Distribution Florale. **Bulletin de la Societe Vaudoise des Sciences Naturelles**. v.44. p.223-270, 1908. doi: 10.5169/seals-268384.

JAIN, A. K. Data clustering: 50 years beyond K-means. **Pattern Recognition Letters**, v. 31, n. 8, p. 651–666, 2010. doi: 10.1016/j.patrec.2009.09.011.

KASSAMBARA, A. **Practical guide to cluster analysis in R: Unsupervised machine learning**. STHDA, 2017.

KOHONEN, T. The self-organizing map. **Proceedings of the IEEE**, v. 78, n. 9, p. 1464–1480, 1990. doi: 10.1109/5.58325.

MAULIK, U.; BANDYOPADHYAY, S. Performance evaluation of some clustering algorithms and validity indices. **IEEE Transactions on Pattern Analysis and Machine Intelligence**, v. 24, n. 12, p. 1650–1654, 2002. doi: 10.1109/TPAMI.2002.1114856.

MINGOTI, S. A. **Análise de dados através de métodos de estatística multivariada: uma abordagem aplicada**. Belo Horizonte: Editora UFMG, 2005.

NASCIMENTO, M.; PETERNELLI, L. A.; CRUZ, C. D.; NASCIMENTO, A. C. C.; FERREIRA, R. D. P.; BHERING, L. L.; SALGADO, C. C. Artificial neural networks for adaptability and stability evaluation in alfalfa genotypes. **Crop Breeding and Applied Biotechnology**, Viçosa - MG - Brasil, v. 13, n. 2, p. 152-156, 2013.

NEI, M. Genetic Distance between Populations. **The American Naturalist**, v. 106, n. 949, p. 283–292, 1972.

OLIVEIRA, A. C. L.; PASQUAL, M.; PIO, L. A. S.; *et al*. Utilização da modelagem matemática (redes neurais artificiais) na classificação de autotetraploides de bananeira (*Musa acuminata Colla*). **Bioscience Journal**, v. 29, n. 3, p. 617–622, 2013.

ORNELLA, L.; TAPIA, E. Supervised machine learning and heterotic classification of maize (*Zea mays* L.) using molecular marker data. **Computers and Electronics in Agriculture**, v. 74, n. 2, p. 250-257, 2010.

PESSOA, A. M. S.; RÊGO, E. R.; SILVA, A. P. G.; et al. Genetic diversity in F3 population of ornamental peppers (*Capsicum annuum* L.). **Revista Ceres**, v. 66, n. 6, p. 442–450, 2019. doi: 10.1590/0034-737x201966060005.

R Core Team. *R:* **A language and environment for statistical computing**. R Foundation for Statistical Computing, Vienna, Austria. 2018. URL https://www.R-project.org/.

SANT'ANNA, I.C.; TOMAZ, R.S.; SILVA, G.N.; et al. Superiority of artificial neural networks for a genetic classification procedure. **Genetics and Molecular Research**, v. 14, n. 3, p. 9898–9906, 2015. doi: 10.4238/2015.August.19.24.

SANTOS, I. G.; CARNEIRO, V. Q.; SILVA JUNIOR, A. C.; et al. Self-organizing maps in the study of genetic diversity among irrigated rice genotypes. **Acta Scientiarum. Agronomy**, v. 41, n. 1, p. 39803, 2018. doi: 10.4025/actasciagron.v41i1.39803.

SILVA, A. R.; SILVA, S. A.; SANTOS, L. A.; et al. Genetic divergence among castor bean lines and parental strains using ward's method based on morpho-agronomic descriptors. **Acta Scientiarum. Agronomy**, v. 39, n. 3, p. 307, 2017. doi: 10.4025/actasciagron.v39i3.32504.

SILVA, G. N.; TOMAZ, R.S.; SANT'ANNA, I. C.; et al. Evaluation of the efficiency of artificial neural networks for genetic value prediction. **Genetics and Molecular Research**, v. 15, n. 1, p. 1–11, 2016. doi: 10.4238/gmr.15017676.

SIEGMUND, K. D.; LAIRD, P. W.; LAIRD-OFFRINGA, I. A. A comparison of cluster analysis methods using DNA methylation data. **Bioinformatics**, v. 20, n. 12, p. 1896–1904, 2004. doi: 10.1093/bioinformatics/bth176.

VENDRAMIN, L.; CAMPELLO, R. J. G. B.; HRUSCHKA, E. R. Relative clustering validity criteria: A comparative overview. **Statistical Analysis and Data Mining**, v.3, n.4, p. 209–235, 2010. doi: 10.1002/sam.10080.

VESANTO, J.; ALHONIEMI, E. Clustering of the self-organizing map. **IEEE Transactions on Neural Networks**, v. 11, n. 3, p. 586–600, 2000. doi: 10.1109/72.846731.

WARD, J. H. Hierarquical grouping to optimize an objective function. **Journal of the American Statistical Association**, v. 58, p. 236–244. Mar. 1963.

WEHRENS, R.; BUYDENS, L. M. C. Self- and Super-organizing Maps in R: The kohonen Package. **Journal of Statistical Software**, v. 21, n. 5, 2007. Disponível em: <http://www.jstatsoft.org/>.

# CAPÍTULO 3 - ICMGen: software de inteligência computacional aplicada ao melhoramento genético de plantas

## Resumo

ICMGen é um software para avaliação de métodos de aprendizagem de máquina com aplicação nas atividades de classificação e agrupamento de genótipos de plantas. O objetivo de disponibilizar o software foi permitir que pesquisadores da área do melhoramento genético, não familiarizados com a programação e com a linguagem R possam utilizar desta ferramenta para comparar e encontrar métodos mais eficientes para os procedimentos desejados. O ICMGen é gratuito para sistema operacional Windows, no idioma português, disponível para download no site http://www.agroeco.com.br. Possibilita utilizar e testar os algoritmos supervisionados: Naive Bayes, Decision Tree, k-Nearest Neighbors (kNN), Random Forest, Suport Vector Machine (SVM), Artificial Neural Networks (ANN) e, não supervisionados: Self-Organizing Maps (SOM), K-Means, Affinity Propagation Clustering (APC) e Density-based spatial clustering of applications with noise (DBSCAN). O software gera médias de acurácia dos algoritmos selecionados, gráficos comparativos, matrizes de confusão, e índices internos e externos, para subsidiar a tomada de decisão relacionada ao melhor método de aprendizagem de máquina.

**Palavras-chave**: aprendizagem de máquina, inteligência artificial, teste de algoritmos, métodos supervisionados, métodos não supervisionados.

## Introdução

A inteligência artificial, em especial o Aprendizagem de Máquina (AM), vem se tornando uma ferramenta com crescente aplicação na área de melhoramento de plantas, como na classificação de genótipos (Ornella e Tapia, 2010; Barbosa *et al*., 2011; Oliveira *et al*., 2013; Fuentes *et al*., 2018), na predição de parâmetros genéticos e ganhos no melhoramento (Silva *et al*., 2016) e na avaliação de adaptabilidade e estabilidade de genótipos (Nascimento *et al*., 2013; Barroso *et al*., 2013).

Os algoritmos de AM possuem a vantagem de serem não-paramétricos, não necessitarem de informações detalhadas sobre os processos físicos do sistema a ser modelado, tolerar a perda de dados, e são capazes de solucionar problemas de grande complexidade, especialmente ao utilizar as Redes Neurais Artificiais (RNAs), que tem sido estudada para aplicações no melhoramento genético (Sant'Anna, 2015), assim como outros algoritmos, como Naive Bayes, Decision Tree, k-Nearest Neighbors (kNN), Random Forest e Suport Vector Machine (SVM) que têm mostrado eficiência nos processos de classificação e possuem potencial para serem utilizados no melhoramento (Ornella e Tapia, 2010; Carvalho, 2018;).

Ainda são escassos os testes com estes algoritmos na área do melhoramento genético, sendo necessária a realização de trabalhos de comparação entre os métodos, com a aplicação de procedimentos experimentais adequados, como a validação cruzada e o emprego de repetição na validação dos algoritmos.

O Software R (R Core Team, 2018), a partir de pacotes específicos, permite realizar esse procedimento de validação dos métodos de AM, porém demanda conhecimento da linguagem e tempo para programação. Dessa forma, o desenvolvimento do ICMGen (**I**nteligência **C**omputacional aplicada ao **M**elhoramento **Gen**ético) objetivou facilitar este processo e permitir que pesquisadores da área do melhoramento genético não familiarizados com a programação e linguagem R possam utilizar esta ferramenta para comparar e encontrar métodos mais eficientes para os procedimentos desejados, como a classificação e agrupamento de genótipos.

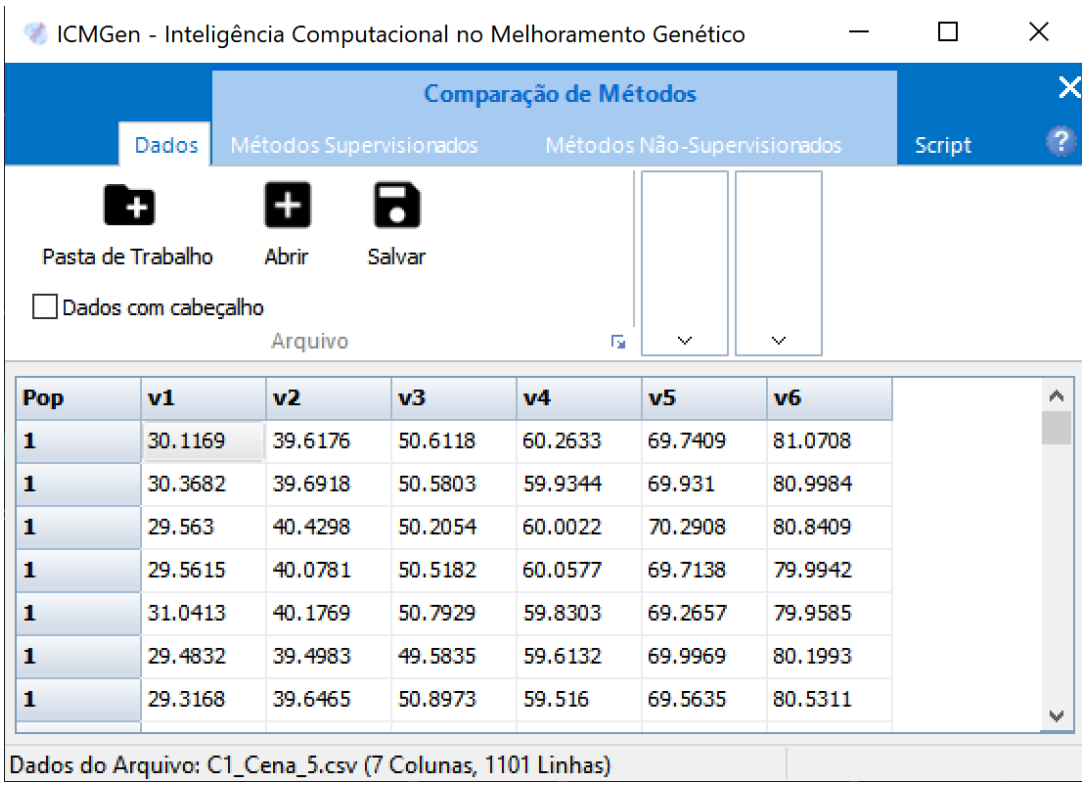
## Descrição

O Software ICMGen foi desenvolvido na linguagem de programação Delphi, no idioma português, para sistema operacional Windows, está patenteado no Brasil

(Costa et al., 2019) porém é gratuito e disponível para download no site http://www.agroeco.com.br. A maioria dos procedimentos é realizada por meio da execução de scripts no Software R, o mesmo deve estar instalado no computador para seu funcionamento. O ICMGen não necessita instalação, sendo utilizado pela execução direta do arquivo executável que pode ser baixado do site http://www.agroeco.com.br/. O Software R também pode ser obtido de forma gratuita no link https://www.r-project.org/.

Destina-se basicamente a teste de algoritmos de aprendizagem de máquina, tanto supervisionado como não supervisionado, utilizados no processo de classificação e agrupamento. O software produz médias de acurácia dos algoritmos selecionados, gerando automaticamente gráficos comparativos, além de matrizes de confusão médias para melhor avaliação dos métodos de aprendizagem de máquina.

A entrada dos dados para análise é feita diretamente pela abertura de arquivo com extensão .xls, .xlsx, .csv, .txt ou .prn, contendo a identificação prévia das classes, ou não, e os valores das variáveis para cada genótipo (Figura 1).




| Pop | v1 | v2 | v3 | v4 | v5 | v6 |
|---|---|---|---|---|---|---|
| 1 | 30.1169 | 39.6176 | 50.6118 | 60.2633 | 69.7409 | 81.0708 |
| 1 | 30.3682 | 39.6918 | 50.5803 | 59.9344 | 69.931 | 80.9984 |
| 1 | 29.563 | 40.4298 | 50.2054 | 60.0022 | 70.2908 | 80.8409 |
| 1 | 29.5615 | 40.0781 | 50.5182 | 60.0577 | 69.7138 | 79.9942 |
| 1 | 31.0413 | 40.1769 | 50.7929 | 59.8303 | 69.2657 | 79.9585 |
| 1 | 29.4832 | 39.4983 | 49.5835 | 59.6132 | 69.9969 | 80.1993 |
| 1 | 29.3168 | 39.6465 | 50.8973 | 59.516 | 69.5635 | 80.5311 |

**Figura 1.** Interface de entrada de dados.

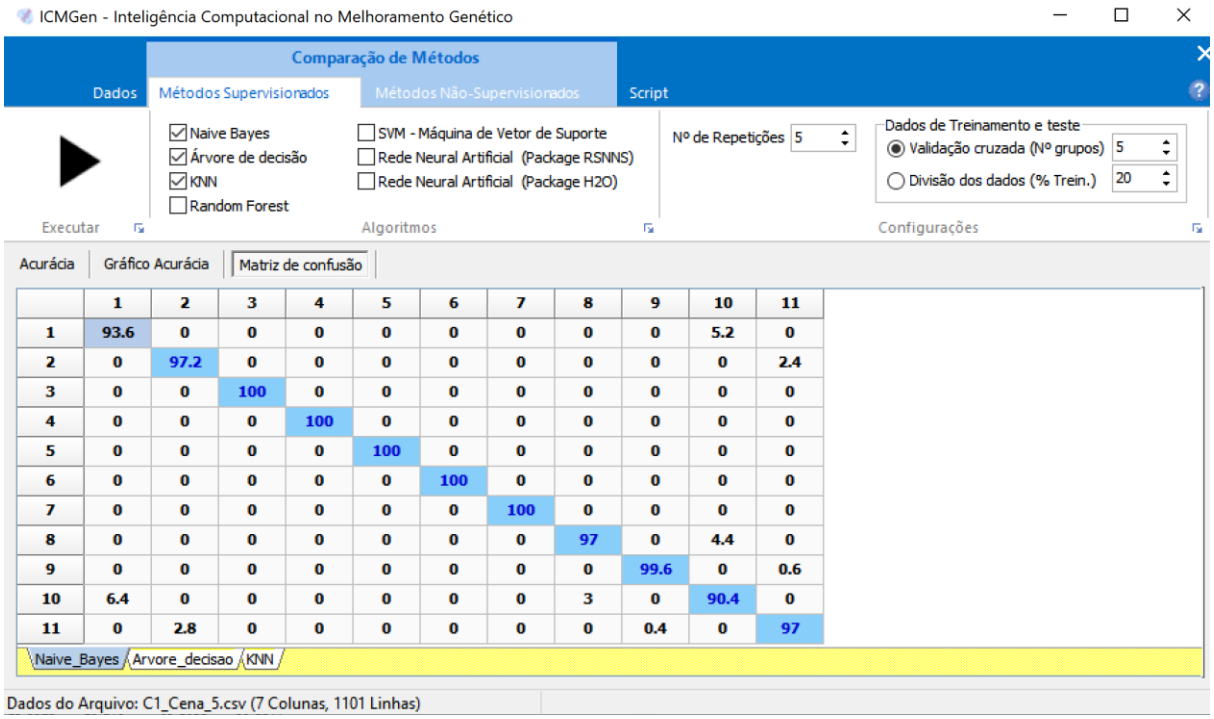
## Procedimentos disponíveis

### Métodos supervisionados

Nesta aba podem ser aplicados os seguintes algoritmos: Naive Bayes, Decision Tree, k-Nearest Neighbors (kNN), Random Forest, Suport Vector Machine (SVM) e Artificial Neural Networks (ANN).

Permite-se escolher o número de repetições e a quantidade de dados a serem utilizados na validação cruzada, por meio da definição do número de grupos ou percentual dos dados.

São mostrados os valores de acurácia para cada repetição e o valor médio, que é utilizado para gerar o gráfico de comparação caso seja escolhido mais de um algoritmo para comparação. Em aba específica são mostradas as matrizes de confusão de cada método com valores médios para cada classe (Figura 2).




| | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 |
|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 | 93.6 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 5.2 | 0 |
| 2 | 0 | 97.2 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 2.4 |
| 3 | 0 | 0 | 100 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 4 | 0 | 0 | 0 | 100 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 5 | 0 | 0 | 0 | 0 | 100 | 0 | 0 | 0 | 0 | 0 | 0 |
| 6 | 0 | 0 | 0 | 0 | 0 | 100 | 0 | 0 | 0 | 0 | 0 |
| 7 | 0 | 0 | 0 | 0 | 0 | 0 | 100 | 0 | 0 | 0 | 0 |
| 8 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 97 | 0 | 4.4 | 0 |
| 9 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 99.6 | 0 | 0.6 |
| 10 | 6.4 | 0 | 0 | 0 | 0 | 0 | 0 | 3 | 0 | 90.4 | 0 |
| 11 | 0 | 2.8 | 0 | 0 | 0 | 0 | 0 | 0 | 0.4 | 0 | 97 |

**Figura 2**. Interface para aplicação de métodos supervisionados de aprendizagem de máquina, com a obtenção de acurácia e matriz de confusão.

### Métodos não supervisionados

Para a abordagem não supervisionada estão disponíveis os seguintes algoritmos: Self Organizing Maps (SOM), que é um tipo de rede neural artificial, k-Means, Affinity Propagation Clustering (APC) e Density-based spatial clustering of applications with noise (DBSCAN).

Neste caso são gerados índices de validação internos, e quando se dispões de uma classificação prévia dos materiais analisados pode-se gerar índices externos de validação, bem como matrizes de confusão.

## Conclusão

O ICMGen é um software que facilita as pesquisas com métodos de aprendizagem de máquina para classificação e agrupamento, com abordagem supervisionada ou não supervisionada, aplicados no melhoramento de plantas. Não exigindo do pesquisador maiores conhecimentos de programação ou domínio de uma linguagem específica, possibilitando focar no que é essencial para a avaliação dos genótipos. O software gera médias de acurácia dos algoritmos selecionados, gráficos comparativos, matrizes de confusão e índices internos e externos para subsidiar a tomada de decisão relacionada ao melhor métodos supervisionados.

## Referências

BARBOSA, C. D. *et al* . Artificial neural network analysis of genetic diversity in *Carica papaya* L. **Crop Breed. Appl. Biotechnol. (Online)**, Viçosa, v. 11, n. 3, p. 224-231, Sept. 2011.

BARROSO, L. M. A.; NASCIMENTO, M.; NASCIMENTO, A. C. C.; Silva, F. F.; Ferreira, R. D. P. Uso do Método de Eberhart e Russell como informação a priori para aplicação de redes neurais artificiais e análise discriminante visando a classificação de genótipos de alfafa quanto à adaptabilidade e estabilidade fenotípica. **Revista Brasileira de Biometria**, Lavras – MG, Brasil, v. 31, n. 2, p. 176-188, 2013.

CARVALHO, V. P.; SANT'ANNA, I.C.; NASCIMENTO, M.; *et al*. Support vector machines applied to the genetic classification problem of hybrid populations with high degrees of similarity. **Genetics and Molecular Research**, v. 17, n. 4, p. 1–10, 2018.

COSTA, R. B.; COSTA, C. S.; OLIVEIRA, M. A. C.; MORAES, P. M.; SKOWRONSKI, L.; NOGUEIRA, M. L.; CARVALHO, M. de A. ICMGen - Inteligência Computacional no Melhoramento Genético. Depositante: Missão Salesiana de Mato Grosso. BR 512019001093-0, Criação: 20 Fev. 2019, Expedição: 4 Jun 2019.

FUENTES, S; HERNÁNDEZ-MONTES, E; ESCALONA, J M; *et al*. Automated grapevine cultivar classification based on machine learning using leaf morpho-colorimetry, fractal dimension and near-infrared spectroscopy parameters. **Computers and Electronics in Agriculture**, v. 151, p. 311–318, 2018.

NASCIMENTO, M.; PETERNELLI, L. A.; CRUZ, C. D.; NASCIMENTO, A. C. C.; FERREIRA, R. D. P.; BHERING, L. L.; SALGADO, C. C. Artificial neural networks for

adaptability and stability evaluation in alfalfa genotypes. **Crop Breeding and Applied Biotechnology**, Viçosa - MG - Brasil, v. 13, n. 2, p. 152-156, 2013.

OLIVEIRA, A. C. L.; PASQUAL, M.; PIO, L. A. S.; *et al*. Utilização da modelagem matemática (redes neurais artificiais) na classificação de autotetraploides de bananeira (*Musa acuminata Colla*). **Bioscience Journal**, v. 29, n. 3, p. 617–622, 2013.

ORNELLA, L.; TAPIA, E. Supervised machine learning and heterotic classification of maize (*Zea mays* L.) using molecular marker data. **Computers and Electronics in Agriculture**, v. 74, n. 2, p. 250-257, 2010.

R Core Team. *R:* **A language and environment for statistical computing**. R Foundation for Statistical Computing, Vienna, Austria. 2018. URL https://www.R-project.org/.

SANT'ANNA, I. C. *et al*. Superiority of artificial neural networks for a genetic classification procedure. **Genetics And Molecular Research**. Ribeirao Preto: Funpec-editora, v. 14, n. 3, p. 9898-9906, 2015.

SILVA, G.N.; TOMAZ, R.S.; SANT'ANNA, I.C.; *et al*. Evaluation of the efficiency of artificial neural networks for genetic value prediction. **Genetics and Molecular Research**, v. 15, n. 1, p. 1–11, 2016.