UNIVERSIDADE FEDERAL DO RECÔNCAVO DA BAHIA
CENTRO DE CIÊNCIAS AGRÁRIAS, AMBIENTAIS E BIOLÓGICAS
EMBRAPA MANDIOCA E FRUTICULTURA
PROGRAMA DE PÓS-GRADUAÇÃO EM RECURSOS GENÉTICOS VEGETAIS
CURSO DE MESTRADO

INTELIGÊNCIA COMPUTACIONAL PARA ESTUDOS DE DIVERSIDADE GENÉTICA ENTRE GENÓTIPOS DE *Nicotiana tabacum* L.

LUCAS GABRIEL SOUZA SANTOS

CRUZ DAS ALMAS - BA
FEVEREIRO - 2022

INTELIGÊNCIA COMPUTACIONAL PARA ESTUDOS DE DIVERSIDADE GENÉTICA ENTRE GENÓTIPOS DE *Nicotiana tabacum* L.

LUCAS GABRIEL SOUZA SANTOS

Engenheiro Florestal

Universidade Federal do Recôncavo da Bahia, 2018

Dissertação apresentada ao Colegiado do Programa de Pós-Graduação em Recursos Genéticos Vegetais da Universidade Federal do Recôncavo da Bahia, como requisito parcial para a obtenção do Título de Mestre em Recursos Genéticos Vegetais (Área de Concentração: Melhoramento e Biotecnologia Vegetal).

Orientador: Prof. Dr. Ricardo Franco Cunha Moreira

Coorientador: Prof. Dr. Liniker Fernandes da Silva

CRUZ DAS ALMAS - BA

FEVEREIRO - 2022

FICHA CATALOGRÁFICA

UNIVERSIDADE FEDERAL DO RECÔNCAVO DA BAHIA
CENTRO DE CIÊNCIAS AGRÁRIAS, AMBIENTAIS E BIOLÓGICAS
EMBRAPA MANDIOCA E FRUTICULTURA
PROGRAMA DE PÓS-GRADUAÇÃO EM RECURSOS GENÉTICOS VEGETAIS
CURSO DE MESTRADO

INTELIGÊNCIA COMPUTACIONAL PARA ESTUDOS DE DIVERSIDADE GENÉTICA ENTRE GENÓTIPOS DE *Nicotiana tabacum* L.

COMISSÃO EXAMIDADORA DA DEFESA DE DISSERTAÇÃO DE
LUCAS GABRIEL SOUZA SANTOS

Aprovada em 11 de fevereiro de 2022

Prof. Dr. Ricardo Franco Cunha Moreira
Universidade Federal do Recôncavo da Bahia – UFRB
Examinador interno (Orientador)

Prof. Dr. Jair Wyzykowski
Universidade Federal do Recôncavo da Bahia – UFRB
Examinador interno

Profª. Drª. Thamara Moura Lima
Instituto Federal de Educação, Ciência e Tecnologia da Bahia – IFBA
Examinador externo

AGRADECIMENTOS

A Deus por ter me dado saúde e força para superar as dificuldades.

Ao meu orientador, Dr. Ricardo Franco, pela confiança, oportunidade e presteza na orientação.

Ao meu coorientador Dr. Liniker Fernandes, pela orientação, conhecimentos transmitidos e contribuições para aprimoramento do trabalho.

Aos meus colegas do mestrado de Recurso Genéticos Vegetais, por partilharmos da mesma caminhada.

A minha noiva Nara, pelo apoio, carinho e incentivo.

A Universidade Federal do Recôncavo da Bahia, pela infraestrutura e pela iniciativa da realização do curso. e por todos os ensinamentos que foram transmitidos pelos professores.

A Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (CAPES), pela concessão da bolsa.

A todos os que contribuíram direta ou indiretamente para a concretização deste trabalho.

INTELIGÊNCIA COMPUTACIONAL PARA ESTUDOS DE DIVERSIDADE GENÉTICA ENTRE GENÓTIPOS DE *Nicotiana tabacum* L.

Autor: Lucas Gabriel Souza Santos
Orientador: Prof. Dr. Ricardo Franco Cunha Moreira
Coorientador: Prof. Dr. Liniker Fernandes da Silva

RESUMO

O fumo cultivado como fonte de matéria-prima para a indústria é considerado umas das culturas não alimentícias de maior importância do mundo e o Brasil se destaca entre os maiores produtores e importadores mundiais. É uma espécie amplamente utilizada nos estudos de genética e melhoramento vegetal, incluindo a pesquisas em biotecnologia e a crescente utilização de ferramentas tecnológicas para assistir o trabalho dos melhoristas na agricultura digital. Nesse sentido, o objetivo deste trabalho foi investigar a divergência genética entre genótipos de tabaco, por métodos multivariados e verificar a eficiência de algoritmos computacionais, visando à identificação de genótipos promissores. O trabalho foi dividido em dois capítulos. No primeiro capitulo foi feita uma análise de agrupamento hierárquica pelo método UPGMA e pelo método de otimização Tocher, em seguida técnica de rede neural artificial multicamadas. No segundo capítulo foi utilizado o algoritmo computacional de arvore de decisão. Através das análises multivariadas formaram-se 3 grupos divergentes entre si. Uma rede neural artificial com 1 camada oculta contendo 3 neurônios obteve uma acurácia de 0.98, enquanto a arvore de decisão o valor da acurácia foi de 0.96. Estes valores demonstram a quão poderosa são estas ferramentas e reforça o uso delas em estudos que visem a manutenção de banco dos germoplasmas e também em programas de conservação e melhoramento genético de tabaco da região do Recôncavo baiano.

Palavras-chaves: Redes neurais artificiais; Árvore de decisão; Melhoramento vegetal; Fumo.

COMPUTATIONAL INTELLIGENCE FOR GENETIC DIVERSITY STUDIES BETWEEN GENOTYPES OF Nicotiana tabacum L.

ABSTRACT

Tobacco grown as a source of raw material for the industry is considered one of the most important non-food crops in the world and Brazil stands out among the world's largest producers and importers. It is a species widely used in studies of genetics and plant breeding, including research in biotechnology and the increasing use of technological tools to assist the work of breeders in digital agriculture. In this sense, the objective of this work was to investigate the genetic divergence between tobacco genotypes, by multivariate methods and to verify the efficiency of computational algorithms, aiming at the identification of promising genotypes. The work was divided into two chapters. In the first chapter, a hierarchical cluster analysis was performed using the UPGMA method and the Tocher optimization method, followed by the multilayer artificial neural network technique. In the second chapter, the decision tree computational algorithm was used. Through multivariate analyses, 3 groups were formed that diverged from each other. An artificial neural network with 1 hidden layer containing 3 neurons obtained an accuracy of 0.98, while in the decision tree the accuracy value was 0.96. These values demonstrate how powerful these tools are and reinforce their use in studies aimed at maintaining a germplasm bank and also in programs for the conservation and genetic improvement of tobacco in the Recôncavo region of Bahia.

Keywords: Artificial neural networks; Decision tree; Plant breeding; Tobacco.

LISTA DE FIGURAS

CAPÍTULO I

CAPÍTULO II

LISTA DE TABELAS

CAPÍTULO I

CAPÍTULO II

SUMÁRIO

1. INTRODUÇÃO GERAL

O tabaco (*Nicotiana tabacum* L.), ou fumo, como é popularmente conhecido, vem sendo cultivado ao longo de centenas de anos pelo homem. Incialmente empregados pelos índios em rituais religiosos e também com uso medicinal devido ao efeito do alcaloide nicotina, atualmente é considerado a principal cultura não alimentícia explorada por diversos países, entre eles o Brasil, que ao longo dos anos tem se mantido como um dos maiores produtores e exportadores do produto no mundo (LORENCETTI; MALLMANN; SANTOS, 2008; SINDITABACO, 2020).

Considerado a principal matéria prima da indústria do fumo a Bahia é um dos maiores produtores da região nordeste do país. A cultura apresenta grande importância socioeconômica para o recôncavo baiano, região do estado que se destaca o maior produtor de tabaco escuro para charutos no país, havendo ainda o cultivo do produto para fumo em corda no Nordeste do Estado e de tabaco claro para cigarro no Oeste baiano (KIST et al., 2020). Devido, principalmente, ao curto ciclo de vida e as características reprodutiva de suas flores e frutos, a espécie tem sido empregada em diversos estudos de genética quantitativa, sendo utilizada ainda em trabalhos envolvendo cultura de tecido, mutações induzidas dentre outros (GANAPATHI et al., 2004; LORENCETTI; MALLMANN; SANTOS, 2008; PEREIRA, 2014; TROMBIN-SOUZA et al., 2017).

À vista disso, estudos de diversidade genética de cultivares de fumo mostra-se indispensável, tanto para a conservação do germoplasma, quanto na escolha dos genitores para fins de melhoramento ou ampliação das bases genéticas (ZHANG et al., 2008, DAVALIEVA et al., 2010, MALLESHAPPA et al., 2020, YANG et al., 2020). Segundo CRUZ E CARNEIRO (2003) a identificação da diversidade genética pode ser realizada através de analises multivariadas, possibilitando que genótipos promissores sejam identificados, aumentando assim a capacidade da obtenção de híbridos com maior efeito heterótico.

Analises estatísticas multivariadas tem sido frequentemente utilizada por melhoristas em programas de melhoramento genético, duas delas estão entre as mais usuais: analise discriminante e a análise de agrupamento. Entretanto, a aplicação de novas tecnologias na chamada agricultura digital vem agregando novas estratégias que proporcionam vantagens como: a aproximação de funções universais; tolerância a dados com ruídos (outliers) ou incompletos; capacidade de modelar diversas variáveis e suas relações não lineares, assim

como a modelagem com variáveis categóricas e numéricas impulsionando o setor. Nesse sentido, cada vez mais vem sendo empregado de ferramentas adicionais baseadas em inteligência artificial, na resolução de problemas onde a técnica de redes neurais artificiais e de arvore de decisão tem se destacado nas diversas aplicações (HAYKIN, 2001; BARBOSA et al., 2011; HAGAN, 2014).

As redes neurais artificiais são modelos computacionais inspirados nos neurônios biológicos, compostas por uma rede de unidades de processamento (neurônios artificiais) interconectadas possuem a capacidade de reconhecer padrões através de exemplos e de generalizar a informação aprendida, gerando resultados coerentes para dados desconhecidos (BRAGA et al., 2000; HAYKIN, 2001; RUSSELL; NORVIG, 2010). A outra abordagem para problemas de predição é a árvore de decisão (DT). A DT é um modelo estatístico que utiliza treinamento supervisionado podendo ser empregada em problemas de regressão e de classificação assim como a RNA. A grande vantagem desse método em relação as RNA é que requer um custo computacional menor e os resultados obtidos são de fácil interpretação.

A exemplo, na agricultura vêm sendo desenvolvidas pesquisas como: rede neural artificial na predição de produtividade das culturas do trigo (SAFA; SAMARASINGHE; NEJAT, 2015) e soja (ALVES et al., 2018); uso da árvore de decisão e seus refinamentos na predição da resistência à ferrugem do café arábica (SOUSA, 2018); seleção de genótipos de eucalipto (TEIXEIRA, 2018); avaliações sensoriais de qualidade do tabaco (HE et al., 2020); árvore de decisão em classificação de genótipos de feijão (ALMEIDA et al., 2021);, dentre outros.

Assim sendo, o objetivo deste trabalho foi estimar a divergência genética entre 15 genótipos, por métodos multivariados e verificar a eficiência inteligência computacional para estudos de diversidade genética entre genótipos de *nicotiana tabacum* l.

.

2. REVISÃO DE LITERATURA

2.1.Aspectos gerais da cultura do tabaco (Nicotiana tabacum L.)

O gênero *Nicotiana* L. pertence à família Solanaceae, é constituído por mais de 60 espécies conhecidas e está dividido em três subgêneros: *rustica, tabacum* e *petunioides*, originárias da América do Sul, América do Norte, Austrália e Ilhas do Pacífico Sul (GOODSPEED; WHEELER; HUTCHISON, 1954; NARAYAN, 1987).

Entre as muitas e variadas espécies, a que tem maior destaque é o fumo (*Nicotiana tabacum* L.) (REN; TIMKO, 2001; LORENCETTI; MALLMANN; SANTOS, 2008). De acordo com os mesmos autores, a espécie apresenta uma gama de variabilidade genética, sendo esse um dos motivos que a tornou bastante utilizada em programas de melhoramento. Além disso, a espécie é cultivada mundialmente como fonte de matéria-prima para a indústria do fumo, por suas propriedades estimulantes e também utilizada em pesquisas nas áreas de farmácia, fisiologia, virologia e na transgenia, sendo empregada para a produção de proteínas terapêuticas, produtos cosméticos, combustíveis, drogas e vacinas como contra os vírus da hepatite B e do ebola.

Classificada como uma planta herbácea anual, bianual ou perene e autógama, o fumo é um alotetraplóide, com 2n=4x=48, uma das hipóteses sugere que se originou nos vales orientais dos Andes Bolivianos difundindo-se pelo território brasileiro através das migrações indígenas, sobretudo Tupi-Guarani. Possui caule único e ereto com a inflorescência sendo uma panícula terminal e o seu ciclo de vida varia entre 120 e 240 dias. As flores são hermafroditas e apresentam uma variedade principalmente na forma; cor e inserção do estame, apresentando uma morfologia favoreça tanto autofecundação quanto a reprodução cruzada. Outro fato importante, e muito explorado, é que as flores, quando fecundadas, geralmente resultam na produção de centenas e milhares de sementes por frutos (LORENCETTI; MALLMANN; SANTOS, 2008; SINDITABACO, 2019).

O fumo produzido comercialmente pode ser agrupado em diferentes variedades que são influenciadas por alguns fatores como o sistema de produção, o processo de secagem da folha (chamado de cura) e também as características bioquímicas da planta. Os principais tipos são: o Flue-Cured ou Virgínia, Ligh Air - Cured incluindo o Burley e o Maryland, Dark Air-Cured, Fire- Cured, Sun-Cured, Oriental e Cigar Filler (fumos para charuto) (FRICANO et al., 2012), dentre as quais se destacam as variedades Virginia, Burley e Oriental, representando mais de 80% do cultivo mundial (LORENCETTI; MALLMANN; SANTOS, 2008).

No Brasil são plantados os tipos de fumo: Virgínia, Burley, Comum e outros, encontrando-se os fumos para capa de charuto, oriental e fumo em corda, sendo utilizados na fabricação dos cigarros uma mistura contendo 40% de fumo Virgínia, 35% de fumo Burley, 15% de fumo Oriental e 10% de talo picado. A utilização destes tipos de fumo

misturados para a composição do cigarro busca o equilíbrio no sabor e no aroma, tendo em vista, atingir as exigências do mercado consumidor (KIST et al., 2004).

O tabaco também é classificado em grupos, segundo o seu preparo e/ou processo de cura, sendo o tabaco de estufa aqueles que são submetidos à cura (secagem) em estufas, com temperatura e umidade controladas (Flue Cured) e o tabaco de galpão quando submetido à cura (secagem) natural à sombra ou em galpões (Air Cured). Os fumos do tipo estufa compreendem os grupos varietais Virgínia, que possuem colheita de folhas individuais e cura através de calor artificial em estufas apropriadas, sendo empregados para misturas na fabricação de cigarros industrializados e possuem alto teor de açúcares. Os do tipo galpão compreendem os grupos varietais Burley, Comum, Dark e Maryland, cuja colheita é feita pelo corte da planta inteira, estes grupos também são utilizados em misturas na fabricação de cigarros industrializados (MASSOLA et al., 2005).

As variedades com potencial agrícola na Bahia são agrupadas em zonas fisiográficas que atribuem a cada umas delas características únicas, influenciadas pelos microclimas específicos e variações de solos conferem qualidades intrínsecas de cor, sabor e combustibilidade. Na Mata Norte se produz um fumo mais forte. O fumo produzido na Mata de São Gonçalo é mais suave, com características próximas ao da Mata Fina. A Mata Fina é a mais nobre área de produção. Já a Mata Sul produz um fumo suave (OLIVEIRA, 2006). As características do fumo determinarão a usual classificação comercial do produto praticada neste estado, diferenciando preço e determinando o uso da folha para capa (revestimento externo), capote (revestimento intermediário) ou enchimento dos charutos.

2.2. Importância da cultura

O tabaco é uma das culturas não alimentícias de maior relevância no mundo. O Brasil se destaca como um dos principais produtores ao longo dos anos e maior exportador de tabaco no cenário internacional. De acordo com o Sindicato da Indústria do Tabaco (Sinditabaco), é o segundo maior produtor com 603 mil toneladas produzidas em 544 municípios (safra 2019/2020), atrás somente da China (SINDITABACO, 2020). Ainda de conforme os dados do Sinditabaco (2020), o setor do tabaco ocupou a 8ª posição no ranking das exportações do agronegócio brasileiro, representando 0,8% do total dos embarques no ano de 2020.

A produção brasileira está concentrada na região do Sul e Nordeste, sendo que, estes três estados do Sul do Brasil se destacam com 97% da produção e 99% da exportação de folhas claras para cigarros. O Nordeste tem uma participação menor, voltada para a produção fumos escuros para charutos, cigarrilhas e fumo desfiado para palheiro (KIST et al., 2020). Nesta região, se destaca o estado Bahia como um dos principais produtores, com cerca de 36 municípios com produção de fumo distribuídos em suas principais zonas, compreendidas pelas regiões de Feira de Santana, Cruz das Almas e Alagoinhas. O tabaco e seus derivados ficaram em décimo oitavo lugar na pauta das exportações do agronegócio baiano no ano de 2020, participando com 0,42% do valor das exportações, correspondendo a US$ 20,507 milhões (SEI, 2020).

A produção da Bahia no ano de 2020, foi cerca de 12 milhões de charutos. Os plantios ficam concentrados nas regiões do Recôncavo e do Nordeste da Bahia com uma produção de 3 mil toneladas em cerca de 23 municípios. O Recôncavo da Bahia, localizado a aproximadamente 100 km da capital, Salvador, é o maior produtor de tabaco escuro para charutos no país, havendo ainda o cultivo do produto para fumo em corda no Nordeste do Estado e de tabaco claro para cigarro em 7 mil hectares no Oeste baiano (KIST et al., 2020).

O tabaco baiano é considerado um dos melhores do mundo, sendo a maior parte da produção exportada em forma de folha. Entretanto, no Recôncavo, a região mais importante de produção do estado, também chamada de “Mata-Fina”, destaca-se ainda a presença de indústrias tradicionais de charutos e cigarrilhas, que possuem grande importância econômica e social, contribuindo na geração de oportunidades de trabalho, principalmente para as mulheres, com empregos diretos e indiretos, além da participação dos produtores de agricultura familiar da região, envolvendo cerca de 3 mil famílias (OLIVEIRA, 2006; KIST et al., 2020).

2.3.Redes Neurais Artificiais

O conhecimento da diversidade genética presente nas culturas é um fator muito importante em estudos que tenham como objetivos a conservação dos recursos genéticos, a ampliação da base genética e as aplicações práticas em programas de melhoramento. Estudos como esses podem disponibilizar informações acerca do grau de similaridade ou dissimilaridade entre dois ou mais genótipos e fornecem parâmetros que auxiliam na escolha dos melhores genitores (CRUZ; CARNEIRO, 2006). Dessa forma, a análise dos dados e as estimativas obtidas se tornam um desafio diante do qual o uso de métodos que agrupem os

genótipos pode ser apresentado como uma das melhores alternativas para análise e interpretação dos dados.

Um dos objetivos da maioria dos programas de melhoramento genético de plantas tem como foco os mecanismos biológicos para identificar o crescimento da cultura e melhorar seu rendimento. Diante disso a modelagem de safras desempenha um papel significativo na produção agrícola. As previsões de produtividade agrícola é uma das grandes dificuldades da agricultura. Realizar essa tarefa requer a disponibilidade um banco de dados grande, uma vez que a produção agrícola depende de muitos fatores diferentes, como clima, clima, solo, variedade do genótipo, entre outros (XU et al., 2019; KLOMPENBURG; KASSAHUN; CATAL, 2020).

São vários os métodos de previsão de rendimento relatados na literatura, que incluem regressão, simulação, sistemas especialistas, e rede neural artificial (RNA) que ao logo dos anos sido amplamente utilizada. Segundo BARBOSA et al., (2011) o uso de tecnologia de redes neurais artificiais tem se encaixado no contexto da agricultura de diferentes maneiras, auxiliando nas diversas etapas do melhoramento vegetal.

As RNA também são chamadas de conexionismo ou sistemas de processamento paralelo e distribuído, consistem em técnicas computacionais. A estrutura e o funcionamento de uma rede são inspirados nos neurônios biológicos, sendo esta, compostas por uma rede de unidades de processamento (neurônios artificiais) interconectadas responsável por aprender através de exemplos e de generalizar a informação aprendida, ou seja, a capacidade de gerar resultados coerentes para dados desconhecidos (BRAGA et al., 2000; HAYKIN, 2001; RUSSELL; NORVIG, 2010).

O aprendizado de uma rede neural ocorre a partir da adaptação dos parâmetros pelo processo de estimulação no ambiente, após esta etapa os pesos sinápticos são responsáveis por armazenar a informação obtida (HAYKIN, 2001). Devido a esta capacidade de aprender e armazenar o conhecimento adquirido as RNA’s tem sido amplamente utilizada tanto para realizar classificações de padrões, quanto previsões.

Existem vários trabalhos empregando o uso de RNA’s com as mais diversas finalidades nas mais distintas áreas. A exemplo, na agricultura vêm sendo desenvolvidas pesquisas como: predição de produtividade das culturas do trigo (SAFA; SAMARASINGHE; NEJAT, 2015) e soja (ALVES et al., 2018); seleção de genótipos de

eucalipto (TEIXEIRA, 2018); agrupamento de genótipos de mamão (BARBOSA et al., 2011); seleção precoce entre famílias de cana-de-açúcar (PETERNELLI et al., 2017); avaliações sensoriais de qualidade do tabaco (HE et al., 2020); previsões de valores genéticos em cenários simulados (SILVA et al., 2014), dentre outros.

3. Neurônio artificial

O elemento básico de uma RNA são os neurônios, uma unidade de processamento de informação fundamental para operação de uma rede. Os primeiros princípios dos neurônios artificiais, foram propostos por McCulloch e Pitts (1943) e apesar de serem considerados simples, eles serviram como base para o avanço do estudo na área (HAYKIN, 2001). Atualmente a representação de uma modelo básico de um neurônio artificial pode ser feita conforme a figura 1.

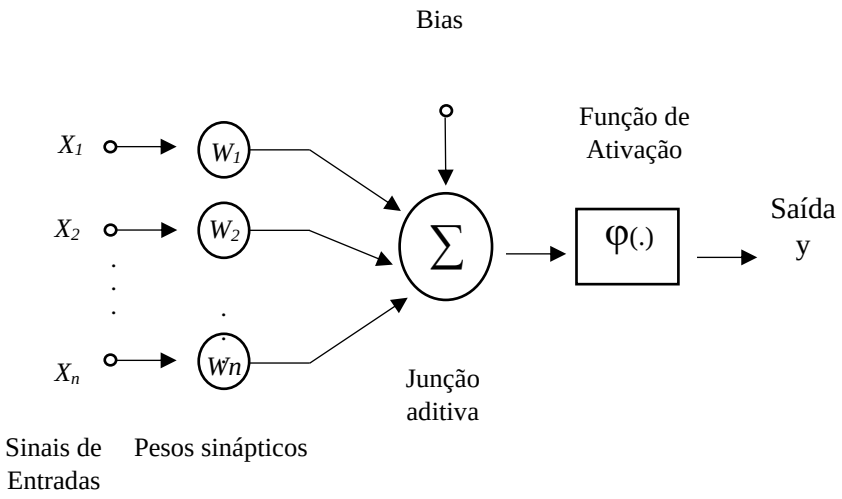

Figura 1- Estrutura do modelo do neurônio artificial
Adaptado de: HAYKIN, 2001

Um modelo típico de um neurônio é composto por alguns elementos básicos, sinais de entrada; conjunto de pesos sinápticos, um somador e uma função de ativação. Onde os sinais são valores externos assumidos pelas variáveis de uma aplicação específica, apresentados à entrada; cada sinal é multiplicado por um peso que representa o seu nível de relevância na saída da unidade; é feita uma soma ponderada dos sinais, a fim de produzir um valor potencial de ativação; caso este nível de atividade exceda certo limite (threshold), então o neurônio produz uma saída. O modelo também inclui um bias b (limiar) que tem o efeito de aumentar ou diminuir a entrada liquida da função de ativação se positivo ou negativo, respectivamente. Geralmente este parâmetro é considerado como um peso sináptico

associado a um novo sinal de entrada fixo em +1 (BRAGA et al., 2000; HAYKIN, 2001; SILVA et al., 2010).

Assim, baseado nessa divisão, as arquiteturas de uma RNA apresentam uma estrutura base, porém apresentam particularidades próprias. Um modelo básico de neurônio artificial pode ser representado matematicamente como:

$$Y = \varphi(V) \tag{1}$$

Em que: Y = saída do neurônio artificial; φ= função de ativação; V = resultado do combinador linear, ou seja:

$$V = \sum_0^m x_m . w_m \tag{2}$$

Em que: V = combinador linear; x_m é o número de entradas; e w_m é o peso para cada entrada de *m*.

Existem alguns tipos básicos de função de ativação como; função degrau, função degrau (bipolar), e função sigmoide são algumas delas (BRAGA et al., 2000; SILVA et al 2010). O neurônio com a função degrau terá saída se o nível de atividade interna do neurônio for um valor positivo ou igual a zero, a saída assumirá o valor 1; caso contrário, assumirá o valor 0 (Figura 2).

$$g(u) = \begin{cases} 1, \ se\ u \geq 0 \\ 0, \ se\ u < 0 \end{cases} \tag{3}$$

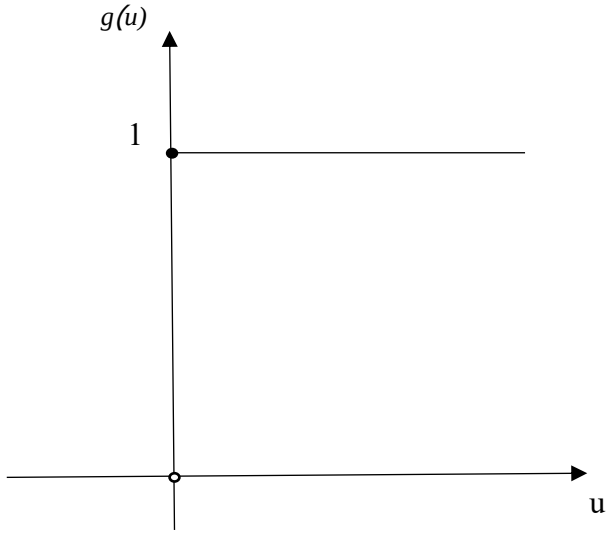

Figura 2 - Representação gráfica função de ativação degrau

A função degrau (bipolar) e sua representação gráfica (Figura 3), da mesma forma que a função de degrau binária, quando o nível de atividade interna do neurônio for positivo ou igual a zero, sua saída assumirá o valor 1; contudo, neste caso, se o nível de atividade do neurônio for um valor negativo, a saída do neurónio assumirá o valor -1.

$$g(u) = \begin{cases} 1, \ se \ u > 0 \\ 0, \ se \ u = 0 \\ -1, \ se \ u < 0 \end{cases} \quad (4)$$

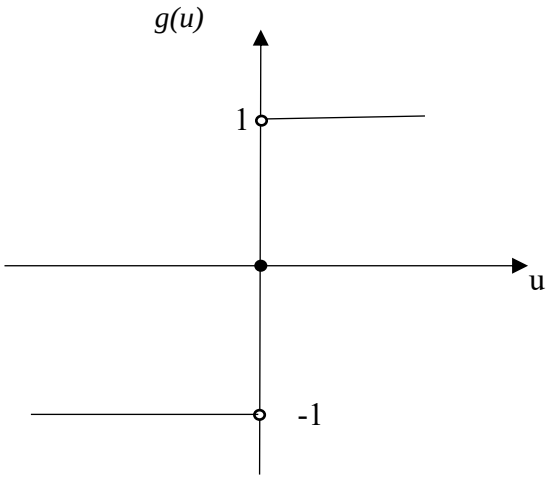

Figura 3 - Representação gráfica função de ativação degrau bipolar

Enquanto que função de ativação sigmoide onde o parâmetro β defina a suavidade ou grau de inclinação da curva da função sigmoide. A saída no neurônio assumirá valores entre 0 e 1, é a mais comum na construção de redes neurais artificiais. Seu gráfico (Figura 4) tem a forma de *s* e um exemplo deste tipo de função é a logística, definida por:

$$g(u) = \frac{1}{1+e^{(-\beta u)}} \quad (5)$$

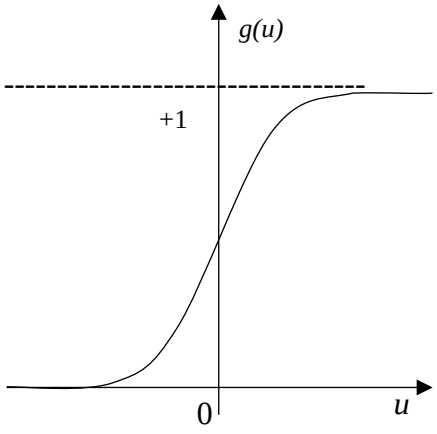

Figura 4 - Representação gráfica função de ativação sigmoide

4. Arquitetura de Redes neurais artificiais

A arquitetura de uma rede neural artificial é definida pela forma como os neurônios estão estruturados entre si. Esta estrutura é um parâmetro de suma importância, pois será um fator restritivo em relação ao tipo de problema tratado pela rede (BRAGA et al., 2000; SILVA et al., 2010). Existem diferentes parâmetros que definem a arquitetura, número de camadas da rede; número de nodos em cada camada; tipo de conexão entre os nodos e a topologia da rede (BRAGA et al., 2000).

A estrutura básica de uma rede neural pode ser dividida em três partes (Figura 5): camada de entrada – onde as informações do banco de dados são inseridas e os padrões são apresentados a rede; intermediarias ou ocultas – onde ocorre quase todo processamento e os ajustes necessários, extraindo as características necessárias a fim de representar a informação fornecida e desempenhar uma determinada tarefa.; camada saída -responsável pela produção e apresentação do resultado obtido, o qual foi processado através de todas as camadas anteriores da rede (SILVA et al., 2010).

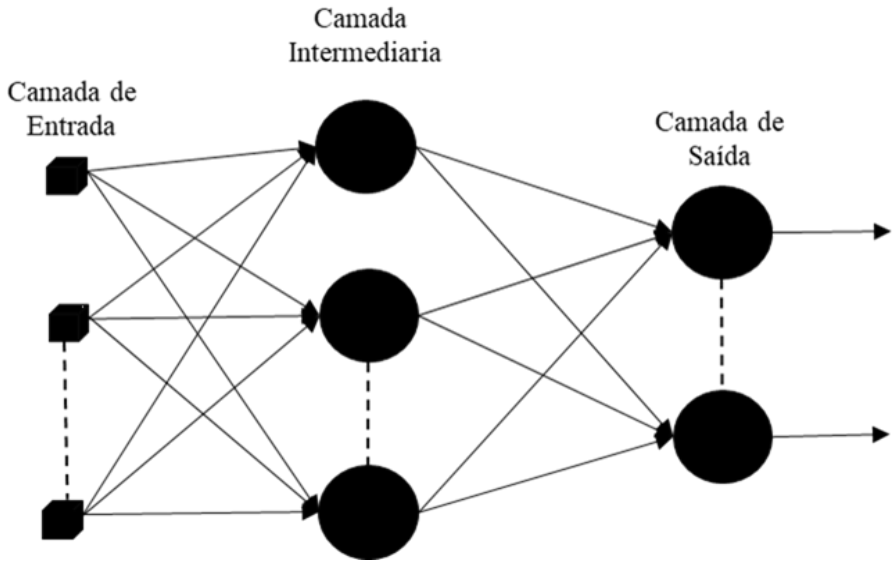

Figura 5 - Representação das camadas de uma rede neural artificial
Fonte: Silva et al., 2010

As principais arquiteturas de redes neurais artificiais, considerando a disposição e interconexão dos seus neurônios, além da constituição de suas camadas, podem ser divididas em: rede Feedforward (alimentação à frente) de camada simples, rede Feedforward de camadas múltiplas e redes recorrentes (SILVA et al., 2010; HAYKIN, 2001). É valido ressaltar que existem outros tipos de redes, tais como: Redes de função de base radial Adaline, redes recorrentes de Hopfield, redes auto organizáveis de Kohonen (SOM), Redes LVQ (Learning Vector Quantization), counter-propagation e redes ART (Adaptive Resonance Theory) (RUSSELL; NORVIG, 2010).

4.1 Redes Feedforward de camada simples

Este tipo de arquitetura possui apenas uma camada de entrada e uma única camada de neurônios que é a própria camada de saída (Figura 6). Utilizada geralmente em reconhecimentos de padrões e em memórias associativas sendo composta por n entradas e m saídas, as redes mais conhecidas desse tipo de arquitetura é a Perceptron e ADALINE (HAYKIN, 2001; SILVA, et al., 2010). A função de ativação utilizada na rede Perceptron geralmente é do tipo degrau ou degrau bipolar.

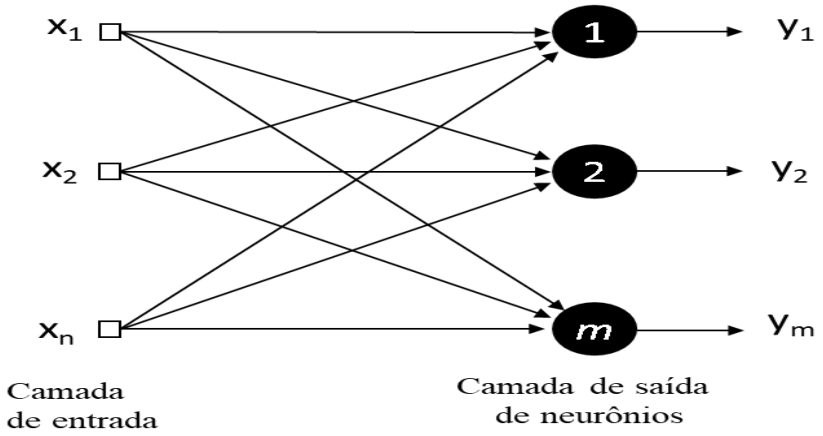

Figura 6 – Rede Feedforward de camada simples
Fonte: Silva et al., 2010

4.2 Redes Feedforward (Multicamadas)

Está arquitetura é constituída de uma ou mais camadas ocultas (Figura 7). As informações são recebidas pela camada de entrada, e processadas de forma sucessiva pelas camadas ocultas, onde as informações são extraídas e codificada por meio dos pesos

sinápticos, formando assim sua representação interna do ambiente externo, geralmente as entradas dos neurônios em cada camada será os sinais da camada anterior, por fim um resultado é apresentada pela camada de saída (HAYKIN, 2001). Tipicamente é aplicada em reconhecimento de padrões e como aproximações de funções, pois é capaz e aproximar funções não lineares, otimizador, identificador de sistemas, etc. (SILVA et al., 2010). A principal rede que utiliza esta arquitetura é o Multilayer Perceptron (MLP).

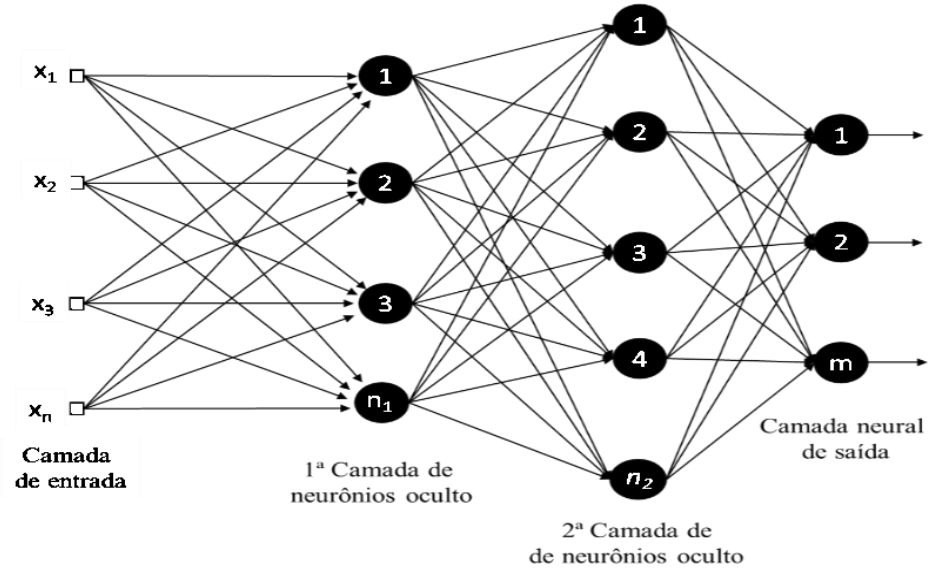

Figura 7 - Feedforward de camadas múltiplas
Fonte: Silva et al., 2010

4.3 Redes Recorrentes

As redes recorrentes (figura 8) têm como características serem realimentadas durante o processo de execução. Isso a difere dos demais tipos de arquitetura e tem grande influência na performance e na capacidade de aprendizado da rede. Esta arquitetura é utilizada principalmente pelas redes Perceptron com realimentação e redes de Hopfield, sendo aplicadas em sistemas dinâmicos, séries temporais, previsões, identificação e controle. As conexões de realimentação se originam dos neurônios da camada de saída e possuem uma memória de atraso que operam sobre todas as entradas x(n) produzindo uma versão atrasada (HAYKIN, 2001).

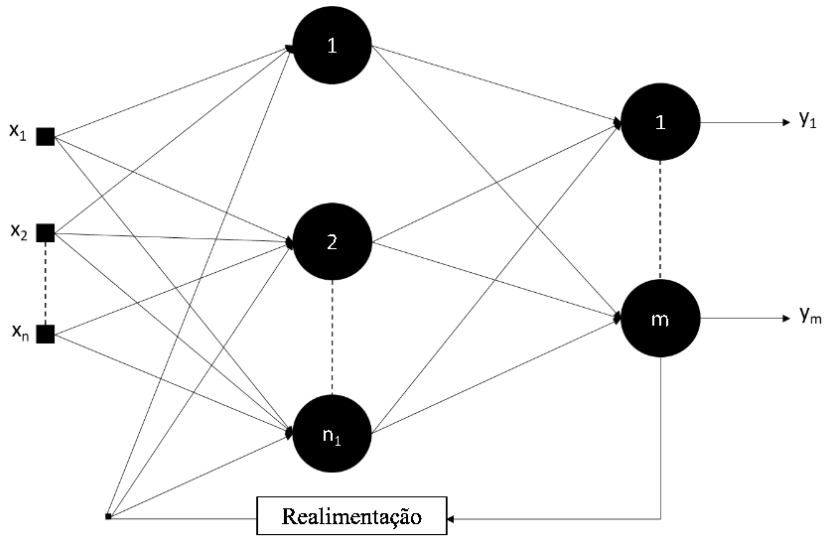

Figura 8 - Rede recorrente
Fonte: Silva et al., 2010

5. Processos de aprendizado

Uma das principais propriedades de uma RNA é a capacidade de adaptação ou aprendizado, é a etapa por meio da qual os parâmetros de uma rede são ajustados através de estímulos fornecidos pelo ambiente de treinamento e se torna capaz de fornecer uma solução generalizada para uma classe de problemas (HAYKIN, 2001; SILVA et al., 2010). O tipo de aprendizagem é determinado pela maneira pela qual os parâmetros são modificados. O treinamento das redes neurais artificiais pode ser dividido em dois modelos, as que utilizam a aprendizagem supervisionada e as que utilizam aprendizagem não-supervisionada (SILVA et al., 2010).

No modelo de aprendizagem supervisionado é necessário fornecer para rede um banco de dados com as entradas e saídas. O processo ocorre da seguinte maneira: os pesos sinápticos e os limiares são continuamente ajustados na etapa de treinamento, conforme o algoritmo de aprendizagem for comparando a saída calculada com a desejada, a diferença encontrada é utilizada no procedimento do ajuste com objetivo de minimizar o erro. A cada pequeno ajuste realizado – se houver solução possível – se torna mais próxima (BRAGA et al., 2000; SILVA et al., 2010). Os algoritmos mais conhecidos para essa abordagem são a Regra Delta e o Algoritmo Backpropagation (ALBANEZ, 2017).

Na aprendizagem não supervisionada não existe uma saída desejada. A rede deve se auto organizar buscando encontrar características similares nos subconjuntos do total de amostras, de forma que os pesos sinápticos e os limiares são ajustados pelo algoritmo de treinamento para criar sua própria representação das entradas (SILVA et al., 2010). Para que este de tipo de aprendizado ocorra é necessário haver redundância nos dados de entrada, cumprindo esse requisito, quaisquer padrões ou características poderão ser encontrados (BRAGA et al., 2000).

6. Árvore de decisão

A árvore de decisão (DT) é um modelo estatístico que utiliza treinamento supervisionado podendo ser empregada em problemas de: regressão, quando a variável resposta é do tipo quantitativa; e de classificação, quando a variável dependente assume valores qualitativos. Aplicando a estratégia de dividir para conquistar, esse método consiste em decompor um problema complexo, simplifica-lo e executa-lo recursivamente, enquanto de forma simultânea, um gráfico de árvore de fácil compreensão associado é gerado (FACELI et al. 2011).

Uma DT é representada na forma de gráfico acíclico direcionado composta pelos nós internos, ramos e nós externos/folhas (Figura 9). Cada nó de decisão contém um teste para algum atributo, cada ramo descendente corresponde a um possível valor deste atributo, e, sendo distintos o conjunto dos ramos, cada folha está associada a uma classe e cada percurso da árvore - da raiz à folha - corresponde a uma regra de classificação.

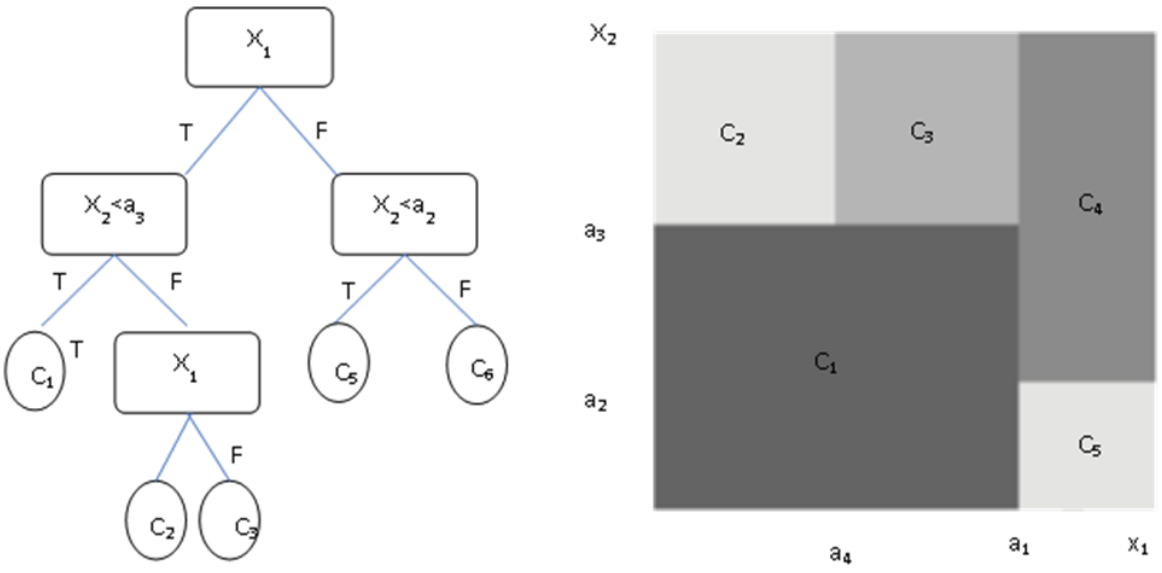

Figura 9 - Arvore de decisão e as regiões de decisão no espaço do objeto
Fonte: James et al., 2013.

6.1.Arvore de regressão

O processo de construção de uma arvore de regressão consiste basicamente em duas etapas, sendo que a primeira é dividir o conjunto de valores possíveis para X1, X2 ...Xp em M regiões distintas R1, R2, ..., RM. Em seguida, para cada região formada é feita uma previsão, que será utilizado para predizer o valor da variável resposta de um novo indivíduo, sendo este valor a média dos valores de resposta para as observações de treinamento pertencente a região utilizada (SOUSA, 2018).

O objetivo então construir regiões R1..., RM que minimiza a Soma de Quadrados dos Resíduos dado por:

$$\sum_{m=1}^{M} \sum_{i \in R_m} \left(y_i - \hat{y}_{R_m}\right)^2, \tag{6}$$

Onde $\hat{y}_{R_m}$ é a média da variável resposta das observações de treinamento dentro da m-éssima região.

O alto custo computacional torna inviável considerar cada partição possível do espaço em M regiões. A solução para contornar esse problema é adotar um procedimento baseado em divisões binárias recursivas, começando no topo da árvore e então dividindo sucessivamente o espaço preditor; na qual objetiva-se, obter a variável X_p e o ponto s, que dívida o espaço em duas regiões (SOUSA, 2018).

$$R_1(p,s) = \{X | X_p \leq s\} \text{ e } R_2(p,s) = \{X | X_p > s\}, \tag{7}$$

tal que o ponto s divida a p-ésima variável em duas regiões que obtenha a menor soma de quadrados dos resíduos, por fim utilizamos a variável que obteve o menor SQR para a primeira divisão, em seguida repetimos o processo para cada região gerada.

6.2.Arvore de classificação

Em uma arvore de classificação a predição é em cada observação pertence à classe de observações de treinamento mais comum na região à qual pertence. Ao construir a arvore de

classificação, se tem como meta obter regiões R1, R2, ..., RM que minimizam um dos três critérios apresentados a seguir (JAMES et al., 2013):

- Taxa de erro aparente $E = 1 - \max_k(\hat{p}_{mk})$ (8)

- Índice de Gini $G = \sum_{k=1}^{K} \hat{p}_{mk}(1 - \hat{p}_{mk})$, (9)

- Deviance $D = -\sum_{k=1}^{K} \hat{p}_{mk} \log \hat{p}_{mk}$, (10)

onde $\hat{p}_{mk}$ representa a proporção de observações na m-ésima região pertencentes a k-ésima classe. Segundo James et al., na construção da árvore de classificação é indicado utilizar o índice Gini ou o Deviance, pois estes apresentam uma maior sensibilidade a pureza. Os índices diminuem de acordo com o crescimento da árvore que ocorre através da divisão binária recursiva. Uma arvore de decisão quando apresenta muitos galhos pode ocasionar em overfitting, a solução mais usual para esse tipo de situação é através dos métodos de "poda".

7. Utilização de algoritmos computacionais na agricultura

As redes Neurais artificias é uma técnica que permite uma abordagem multivariada, possuem uma capacidade de aprendizagem e generalização que possibilita a modelagem de sistemas não lineares onde a relação entre as variáveis estudadas não é conhecida ou são altamente complexas.

A capacidade de aprender a partir do banco de dados e obter resultados precisos é considerada por muitos como a principal vantagem na utilização da RNA. Segundo HAGAN (2014), é preciso encontrar uma rede que generalize bem e adapte aos dados, portanto, o autor recomenda partir de modelos mais simples, pois assim haverá uma redução das possibilidades de erro. Se tratando de redes neurais o modelo considerado mais simples é aquele que contém o menor número de parâmetros livres (pesos e vieses), ou, de forma equivalente, o menor número de neurônios.

A simplicidade na aplicação da RNA é um fator importante, tendo em vista que muitas vezes seu desempenho é superior quando comparados a métodos como os modelos de regressão, pois possibilita a obtenção de resultados efetivos. Dessa forma a rede consegue

realizar aproximação de funções não lineares, e através de treinamento, mapear relações de entrada-saída para qualquer grau de precisão. Permitindo modelar sistemas complexos, apresentando propriedades como: capacidade de aproximação de funções universais, tolerância a dados com ruídos (outliers) ou incompletos; capacidade de modelar diversas variáveis e suas relações não lineares; capacidade de modelagem com variáveis categóricas e numéricas (qualitativas e quantitativas) e analogia neurobiológica (HAYKIN, 2001; HAGAN, 2014).

Esses atributos juntamente com a diversidade de modelos encontrados na literatura, amplia a aplicabilidade das redes neurais levando o uso das RNA's em diversos campos da ciência agrárias, por exemplo: na classificação de diferentes variedades de milho CHEN et al., (2010) utilizaram técnicas de processamento de imagens, análise discriminante, distância de Mahalanobis e redes neurais artificias. As características extraídas através das técnicas de processamento de imagens foram classificadas utilizando a analise discriminante e os grão classificados utilizando a distância de Mahalanobis e redes neurais. Segundo os autores os experimentos obtiveram uma média para a precisão da classificação de até 90% para cinco as variedades de milho. Constaram ainda que o método combinando a distância de Mahalanobis e o classificador de rede neural de retropropagação pode ser empregado com sucesso na identificação de variedades de milho. Ao final do artigo são recomendadas mais investigações para estudar o desempenho do método ao testar o composto.

A classificação das folhas de tabaco é uma etapa de grande importância durante o processo produtivo, entretanto, por ser realizada manualmente torna-se passível de falhas humanas que podem comprometer a qualidade do produto. ZHANG e ZHANG (2011) realizaram um estudo buscando aperfeiçoar o sistema de classificação das folhas de tabaco, a partir da utilização de processamento digital de imagens. Um sistema de classificação com base em técnicas de processamento de imagem foi desenvolvido para inspecionar e classificar automaticamente as folhas de tabaco Flue-Cured. Os resultados experimentais da avaliação da lógica Fuzzy de dois níveis mostraram que a taxa de precisão da classificação encontrada foi cerca de 94% para as folhas de tabaco treinadas e a taxa de precisão das folhas de tabaco não treinadas cerca de 72%. Os autores concluíram que a avaliação abrangente Fuzzy é uma forma viável para a classificação automática e avaliação da qualidade das folhas de tabaco.

Ainda na produção vegetal as RNA's foram utilizadas para prever o rendimento de diversas culturas, como o milho, ADISA et al. (2019) propuseram um modelo de rede neural artificial para prever a produção em quatro províncias sul-africana. As variáveis de entrada foram precipitação, temperatura máxima, temperatura mínima, evapotranspiração potencial, umidade do solo e terra cultivada para milho. Os autores analisaram um conjunto de dados do ano de 1990 até 2007, que foram divididos em dois subconjuntos um de treinamento com 80% e testes com 20% para o treinamento do modelo. Após testar várias combinações as melhores diferentes arquiteturas foram obtidas nas RNA para previsão da produção de milho encontrada para as variáveis. As medidas de desempenho da previsão foram acessadas usando o R^2 ajustado e a comparação entre a produção real e a prevista de milho utilizando os dados de teste indicou a precisão de desempenho ajustada R^2 acima de 0.6 para as diferentes regiões. Segundo os autores o os resultados obtidos com o modelo demonstram a eficiência das redes neurais na previsão, sugerem ainda que, os resultados da pesquisa podem ajudar no planejamento e na tomada de decisão dos agricultores e outros formuladores de políticas.

Na previsão da produção da cultura da cana de açúcar, FERNANDES, EBECKEN e ESQUERDO (2017) utilizaram métricas derivadas de uma série temporal entre os anos de 2003 e 2012 Índice de Vegetação por Diferença Normalizada, (NDVI) com conjunto de redes neurais artificiais, em 60 municípios de São Paulo. Para melhoria do desempenho na previsão, foram retirados todos os Recursos espectrais gerados por meio de séries temporais NDVI redundantes e/ou irrelevantes, aplicados no modelo de aprendizagem da RNA, presentes nos bancos de dados, e através de um modelo de conjunto de empilhamento com RNA o rendimento previsto. Os resultados obtidos apresentaram a raiz do erro quadrático médio (RMSE) de 6,8%, após a retirada das informações dos últimos três meses o novo processamento dos dados apresentou uma melhora na seleção dos recursos, houve um aumento do RMSE para 8%. Com base na média dos rendimentos de todo o estado comparados com os dados oficiais constataram a eficiência do método de empilhamento para a previsão do rendimento da safra três meses antes da colheita, visto que o RMSE foi menor do que a dos dados oficiais.

8. REFERÊNCIAS BIBLIOGRÁFICAS

ABIFUMO - Associação Brasileira da Industria do Fumo. **Perfil da Indústria Brasileira do Tabaco**. Rio de Janeiro. Edição comemorativa 20 anos**.** p. 30, 1999.

ADISA, O.; BOTAI, J.; ADEOLA, A.M.; HASSEN, A.; BOTAI, C.; DARKEY, D.; TESFAMARIAM, E. Application of Artificial Neural Network for Predicting Maize Production in South Africa. **Sustainability**, v. 11, n. 4, p. 1145, 2019.

LUDEMIR, T. B., BRAGA, A. P., & CARVALHO, A. C. P. L. F. Redes neurais artificiais: teoria e aplicações. **Livros Técnicos e Científicos Editora**, 17, 2000.

ALBANEZ, D. D. O. **Redes neurais artificiais aplicadas à segmentação de imagens.** Dissertação (Mestrado em Modelagem e Otimização) - Universidade Federal de Goiás, Catalão, p. 152, 2017.

ALMEIDA, R.D.C.; ASSUNÇÃO NETO, W.V.D.; SILVA, V.B.D.; CARVALHO, L.C.B.; LOPES, Â.C.D.A.; GOMES, R.L.F. Decision tree as a tool in the classification of lima bean accessions**. Revista Caatinga**, v. 34, p. 471-478, 2021.

ALVES, G.R.; TEIXEIRA, I.R.; MELO, F.R.; SOUZA, R.T.; SILVA, A.G. Estimating soybean yields with artificial neural networks. **Acta Scientiarum-agronomy**, v. 40, 35250, 2018.

BARBOSA, C.D.; VIANA, A.P.; QUINTAL, S.S.R.; PEREIRA, M.G. Artificial neural network analysis of genetic diversity in *Carica papaya* L. **Crop Breeding and Applied Biotechnology**, v. 11, p. 224-231, 2011.

CEMEK, B.; ÜNLÜKARA, A.; KURUNÇ, A.; KÜÇÜKTOPCU, E. Leaf area modeling of bell pepper (Capsicum annuum L.) grown under different stress conditions by soft computing approaches. **Computers and Electronics in Agriculture**, v. 174, p. 105514, 2020.

CHEN, X.; XUN, Y.; LI, W.; ZHANG, J. Combining discriminant analysis and neural networks for corn variety identification. **Computers and electronics in agriculture**. Apr 1, v. 71, p. S48-53, 2010.

CRUZ, C.D.; CARNEIRO, P.C.S. **Modelos biométricos aplicados ao melhoramento genético.** Viçosa: Editora UFV, 2. ed. p. 585, 2006.

DAVALIEVA, K.; MALEVA, I.; FILIPOSKI, K.; SPIROSKI, O.; EFREMOV, G.D. Genetic variability of Macedonian tobacco varieties determined by microsatellite marker analysis. **Diversity,** v. 2, n. 4, p. 439-449, 2010.

FACELI, K.; LORENA, A.C.; GAMA, J.; CARVALHO, A.C.P.L.F. de. **Inteligência Artificial: uma abordagem de aprendizado de máquina.** Rio de Janeiro: Livros Técnicos e Científicos Editora, p. 378, 2011.

FAOSTAT - Food and Agriculture Organization of the United Nations Statistical Database. **Crops database**. 2019. Disponível em: <http://faostat3.fao.org/browse/Q/ QC/E>. Acesso em: 1 de fevereiro de 2021.

FERNANDES, J.L.; EBECKEN, N.F.F.; ESQUERDO, J.C.D.M. Sugarcane yield prediction in Brazil using NDVI time series and neural networks ensemble. **International Journal of Remote Sensing**, 38.16: 4631-4644, 2017.

FRICANO, A.; BAKAHER, N.; DELORVO, M.; PIFFANELLI, P.; DONINI, P.; STELLA, A.; POZZI, C. Molecular diversity, population structure, and linkage disequilibrium in a worldwide collection of tobacco (Nicotiana tabacum L.) germplasm. **BMC genetics**, v. 13, n. 1, p. 1-13, 2012.

GANAPATHI, T.R.; SUPRASANNA, P.; RAO, P.S.; BAPAT, V.A. Tobaco (Nicotiana tabacum L.) – A model system for tissue culture interventions and genetc engineering. **Indian Journal of Biotechnology,** v. 3, n. 2, p. 171-184, 2004.

GOODSPEED, T.H.; WHEELER H-M.; HUTCHISON, P.C. Taxonomy of *Nicotiana.* In: GOODSPEED, T. H. **The genus Nicotiana**. Waltham: Chronica Botanica. v. 16, pt. 6, p. 321-492, 1954.

HAGAN, M. T.; DEMUTH, H. B.; BEALE, M. H. **Neural network design.** [S.l.]: Martin Hagan, 2014.

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

HE, C.; CHEN, R.; REN, K.; ZHAO, G.; HE, C.; HU, B.; ZOU, C.; JIANG, Y.; CHEN; Y. A predictive model for the sensory aroma characteristics of flue-cured tobacco based on a back-propagation neural network. **SN Applied Sciences**. v. 2, n. 11, p.1-11, 2020.

KIST, B. B.; DE CARVALHO, C.; FARDIN, I.; GARCIA, P.; BELING, R. **Anuário Brasileiro do Tabaco 2020**. Santa Cruz do Sul: Editora Gazeta Santa Cruz, 135p. 2020.

KLOMPENBURG, T.V.; KASSAHUN, A.; CATAL, C. Crop yield prediction using machine learning: A systematic literature review. **Computers and Electronics in Agriculture**, v. 177, p. 105709, 2020.

LORENCETTI, C.; MALLMANN, I. L.; SANTOS, M. Fumo. In: BARBIERI, R. L.; STUMPF, E. R. T. (Ed.). **Origem e evolução de plantas cultivadas**. Brasília, DF: Embrapa Informação Tecnológica, p. 379-401. 2008.

MALLESHAPPA, C.; SOWMYA, T.M.; KUAMARA, B.N.; MUUTURAJ, M.D. (2020). Genetic variability and heritability studies in flue cured Verginia tobacco (Nicotiana tobaccum L.) germplasm. **Journal of Pharmacognosy and Phytochemistry**, v. 9, n. 5, p. 3171-3173, 2020.

MCCULLOCH, W. S.; PITTS, W. A logical calculus of the ideas immanent in nervous activity. Illinois: **The Bulletin of Mathematical Biophysics,** . v. 5, n. 4, p. 115–133. 1943.

MASSOLA JUNIOR; N.S.; PULCINELLI, C.E.; JESUS JUNIOR; W.C.; GODOY, C.V. Doenças do fumo. In: KIMATI, H.; AMORIM, L.; REZENDE, J.A.M.; BERGAMIN FILHO, A.; CAMARGO, L.E.A. (Ed.). **Manual de Fitopatologia: doenças de plantas cultivadas.** São Paulo: Ceres, v. 2, p. 361-371, 2005.

NARAYAN, R.K. Nuclear DNA changes, genoma differentiation and evolution in *Nicotiana* (Solanaceae). **Plant Systematic and Evolution**, Vienna, v. 157, p. 161- 180, 1987.

OLIVEIRA, J.M.C. A Cultura Do Fumo Na Bahia: refletindo sobre a convenção-quadro. **Revista Bahia Agrícola**, Salvador, v. 7, n. 2.p. 59-65, 2006.

PEREIRA, W. E. L. **Uso de Nicotiana tabacum e Arabidopsis thaliana como plantas modelo para estudo funcional de genes associados à resistência a clorose variegada dos citros**. Dissertação (Mestrado em Genética e Biologia Molecular). Universidade Estadual de Campinas, Campinas, p. 90, 2014.

PETERNELLI, L. A; MOREIRA, É. F. A., NASCIMENTO, M; & CRUZ, C. D. Artificial neural networks and linear discriminant analysis in early selection among sugarcane families. **Crop Breeding and Applied Biotechnology**, 17, 299-305, 2017.

REN, N., e TIMKO, M. P. AFLP analysis of genetic polymorphism and evolutionary relationships among cultivated and wild Nicotiana species. **Genome**, *44*(4), 2001

RUSSELL, S.; NORVIG, P. **Artificial intelligence: A modern Approach.** New Jersey: Pearson Education, Inc; 2010.

SAFA, M; SAMARASINGHE, S; NEJAT, M. Prediction of Wheat Production Using Artificial Neural Networks and Investigating Indirect Factors Affecting It: Case Study in Canterbury Province, New Zealand. **Journal of Agricultural Science and Technology**, v. 17, p. 791-803, 2015.

SEI - Superintendência de Estudos Econômicos e Sociais da Bahia. **Boletim de Comércio Exterior da Bahia - agosto 2020.** 2020. Disponível em: <http://www.sei.ba.gov.br/images/releases_mensais/pdf/bce/bce_ago_2020.pdf>. Acesso em: 07 de fevereiro de 2021.

SILVA, G.N.; TOMAZ, R.S.; SANT'ANNA, I.D.C.; NASCIMENTO, M.; BHERING, L.L.; CRUZ, C.D. Neural networks for predicting breeding values and genetic gains. **Scientia Agricola,** v. 71, n. 6, p. 494-498, 2014.

SILVA, I.N. dac R. **Redes neurais artificiais para engenharia e ciências aplicadas: curso prático**. Artliber Editora Ltda, São Paulo, SP, Brasil, 2010.

SINDITABACO - Sindicato Interestadual da Indústria do Tabaco. **Origem do Tabaco**. 2019. Disponível em: <http://www.sinditabaco.com.br/sobre-o-setor/origem-do-tabaco/>. Acesso em: 02 de fevereiro de 2020.

SINDITABACO – Sindicato Interestadual da Indústria do Tabaco. **Sinditabaco News.** Ed. Janeiro/Abril 2020. Santa Cruz do Sul, p. 6, 2020. Disponível em: <http://www.sinditabaco.com.br/site/wp-content/uploads/2020/02/SindiTabacoNews 37-PT.pdf>. Acesso em: 02 de fevereiro de 2020.

SOUSA, I.C. de. Predição genômica da resistência à ferrugem alaranjada em café arábica via algoritmos de aprendizagem de máquina**. Dissertação (Mestrado em Estatística Aplicada e Biometria)** – Universidade Federal de Viçosa, Viçosa, p. 42, 2018.

TEIXEIRA, V. L. Seleção de genótipos de eucalipto para produção de carvão vegetal utilizando análise multivariada e redes neurais. **(Mestrado em Ciência Florestal)** – Universidade Federal de Viçosa, Viçosa, p. 96, 2018.

TROMBIN-SOUZA, M.; GRZYBOWSKI, C.R.D.S.; OLIVEIRA-CAUDURO, Y.D.; VIEIRA, E.S.N.; PANOBIANCO, M. Osmotic stress on genetically transformed tobacco plant seeds. **Journal of Seed Science**, v. 39, p. 426-432, 2017.

UPOV (Union pour la Protection des Obtentions Variétales). Disponível em http://www.upov.int/tabaco. Acesso em: 20 dez. 2011.

XU, X.; GAO, P.; ZHU, X.; GUO, W.; DING, J.; LI, C.; WU, X. Design of an integrated climatic assessment indicator (ICAI) for wheat production: A case study in Jiangsu Province, China. **Ecological Indicators,** v. 101, p. 943-953, 2019.

YANG, H.; GENG, X.; ZHAO, S.; SHI, H. Genomic diversity analysis and identification of novel SSR markers in four tobacco varieties by high-throughput resequencing. **Plant Physiology and Biochemistry**, v. 150, p. 80-89, 2020.

ZHANG, F.; ZHANG, X. Classification and Quality Evaluation of Tobacco Leaves Based on Image Processing and Fuzzy Comprehensive Evaluation. **Sensors,** v.11, p. 2369-2384, 2011.

CAPITULO I

DIVERGÊNCIA GENÉTICA ENTRE GENÓTIPOS DE *Nicotiana tabacum* L. POR MEIO DE ANÁLISE MULTIVARIADA E REDES NEURAIS ARTIFICIAIS

DIVERGÊNCIA GENÉTICA ENTRE GENÓTIPOS DE Nicotiana tabacum L. POR MEIO DE ANÁLISE MULTIVARIADA E REDES NEURAIS ARTIFICIAIS

RESUMO

O conhecimento da diversidade genética presente em banco de germoplasmas de espécie agrícola é de fundamental importância para conservação e manutenção da variabilidade genética das espécies. Devido a sua relativa simplicidade quando comparadas a outros modelos, as redes neurais artificiais apresentam amplas possibilidades de utilização na biotecnologia vegetal, haja vista a grande capacidade de aprendizagem, adaptação a novos cenários e de reconhecimento de padrões, características essenciais para auxiliar na redução do tempo. Diante do exposto este trabalho foi realizado com o objetivo estimar a divergência genética entre 15 genótipos, por métodos multivariados e verificar a eficiência das redes neurais artificiais visando à identificação de genótipos promissores. A partir de 15 descritores quantitativos foram realizados os agrupamentos com o método hierárquico UPGMA e o método de otimização de Tocher. Para o desenvolvimento da rede neural o banco de dados foi dividido de forma aleatória em dois conjuntos sendo 70% para o treinamento e 30% para validação. Diferentes arquiteturas de redes neurais artificiais foram analisadas. Os métodos de agrupamentos definiram três grupos divergentes. Uma rede neural artificial contendo uma camada oculta com três neurônios foi capaz de encontrar a solução ideal para o problema proposto, com uma acurácia de 0.98 demostrando o potencial do método para uso em estudos classificatórios.

Palavras-chave: Inteligência artificial; aprendizado de máquina; Tabaco tipo Sumatra.

GENETIC DIVERGENCE BETWEEN GENOTYPES OF Nicotiana tabacum L. THROUGH MULTIVARIATE ANALYSIS AND ARTIFICIAL NEURAL NETWORKS

ABSTRACT

The knowledge of the genetic diversity present in the germplasm bank of agricultural species is of fundamental importance for the conservation and maintenance of the genetic variability of the species. Due to their relative simplicity when compared to other models, artificial neural networks present ample possibilities for use in plant biotechnology, given their great capacity for learning, adapting to new scenarios and recognizing patterns, essential characteristics to help reduce time. . In view of the above, this work was carried out with the objective of estimating the genetic divergence between 15 genotypes, by multivariate methods and verifying the efficiency of artificial neural networks in order to identify promising genotypes. From 15 quantitative descriptors, clusters were performed using the hierarchical UPGMA method and the Tocher optimization method. For the development of the neural network, the database was randomly divided into two sets, 70% for training and 30% for validation. Different architectures of artificial neural networks were analyzed. The clustering methods defined three divergent groups. An artificial neural network containing a hidden layer with three neurons was able to find the ideal solution for the proposed problem, with an accuracy of 0.98, demonstrating the potential of the method for use in classificatory studies.

Keywords: Artificial intelligence; machine learning; sumatra tobacco.

1. INTRODUÇÃO

O tabaco (*Nicotiana tabacum* L.) é uma das culturas não alimentícia de maior importância no mundo, devido ao seu impacto socioeconômico. A área ocupada pela cultura é de aproximadamente 3.6 milhões de hectares (FAOSTAT, 2019). No entanto, mais de 50% da produção global é concentrada em quatro países; China, Índia, Brasil e Estados Unidos da América (WENJIE et al., 2011, DASSARI et al., 2018, FAOSTAT, 2019). Dentre estes o Brasil tem se mantido ao longo dos anos, como um dos principais produtores e exportadores no cenário internacional. No ano de 2020 o país produziu cerca de 603 mil toneladas. Dentro do cenário nacional o setor se manteve entre os dez primeiros no ranking das exportações do agronegócio (SINDITABACO, 2020).

O estreitamento da base genética em recursos genéticos, a exemplo do tabaco, é uma preocupação crescente sobre a variabilidade genética remanescente nos bancos de germoplasma (MOON et al., 2009). A avaliação eficiente da variabilidade tem como objetivo identificar acessos com menor redundância possível, com base em características fenotípicas. Diante disto, as atividades de pré-melhoramento vêm a ser ainda mais importantes para caracteres com pouca variabilidade genética em materiais já melhorados ou em germoplasma-elite (FAVERO et al. 2008). Torna-se então imprescindível, o conhecimento da diversidade genética, pois a mesma auxilia na escolha de genótipos, para fins de conservação de ex situ e melhoramento com objetivo de obtenção de híbridos com maior efeito heterótico ou ampliação das bases genéticas (CRUZ e CARNEIRO, 2003, ZHANG et al., 2008, DAVALIEVA et al., 2010, MALLESHAPPA et al., 2020, YANG et al., 2020).

Estudos de diversidade genética têm sido realizados utilizando analises multivariadas, permitindo estimar a dissimilaridade genética existente. Entre as técnicas mais utilizadas por melhoristas em programas de melhoramento genético, estão a analise discriminante e a análise de agrupamento (CRUZ e CARNEIRO, 2003). Entretanto, o uso de novas de metodologias já é uma realidade nas atividades agrícolas, e vem auxiliando na tomada de decisões levando a alcançar melhores resultados. Nesse contexto, cada vez mais vem sendo empregado o uso do aprendizado de maquinas, no qual, se destaca a técnica de redes neurais artificiais como uma ferramenta adicional no processo de tomada de decisão nas diferentes áreas da agricultura e nas diversas etapas do melhoramento vegetal (SILVA

et al. 2014; BARBOSA et al., 2011; PEIXOTO et al. 2015; SANT'ANNA et al 2015; YANG et al. 2019; COSTA et al 2019; SABZI-NOJADEH, et al. 2021).

Modelos de redes neurais artificiais (ANN) são abordagens de modelagem multivariada, relativamente simples em comparação com outros modelos. São amplamente utilizadas, pois possuem uma capacidade de aprendizagem e generalização que possibilita a modelagem de sistemas não lineares onde a relação entre as variáveis estudadas não é conhecida ou são altamente complexas. Outras características que a tornam vantajosas são, tolerância a falhas, adaptabilidade a novas condições, resolução de problemas com base no conhecimento passado e reconhecimento de padrões (HAYKIN, 2001).

No entanto, poucas aplicações de uso de diferentes metodologias na avaliação da diversidade genética foram observadas na cultura do tabaco. Diante do exposto, este trabalho tem por objetivo estimar a divergência genética entre 15 genótipos, por métodos multivariados e verificar a eficiência das redes neurais artificiais visando à identificação de genótipos promissores.

2. MATERIAL E MÉTODOS

2.1. Área de estudo

Os dados utilizados se referem a um experimento conduzido no ano agrícola de 2011, no campo de produção comercial onde foram caracterizados 15 genótipos de tabaco (*Nicotiana tabacum* L.) tipo Sumatra (Tabela 1), cultivado sob tela preta com 30% de sombreamento pela empresa ERMOR TABARAMA TABACOS DO BRASIL Ltda.

O campo de produção onde foi implantado o experimento localiza-se no município de Muritiba – BA, a uma altitude de 220 m em relação ao nível do mar, que apresentam precipitação pluviométrica anual média de 1.220 mm, apresentando temperatura média anual de 24,1 º. O delineamento experimental utilizado foi o de blocos casualizados com quatro repetições. Cada parcela foi constituída de cinco linhas de 10 plantas e cada linha teve 4,5 metros de comprimento com espaçamento de 1,0 metros entre linhas e 0,42 metros entre plantas.

Tabela 1 - Relação dos genótipos de Tabaco provenientes da empresa Ermor Tabarama Tabacos do Brasil.

Código (Nº)	Genótipos	Tipo	Origem
01	ER 03-107	Sumatra	Bahia
02	ER 04-090	Sumatra	Bahia

03	ER 04-095	Sumatra	Bahia
04	ER 05-005	Sumatra	Bahia
05	ER 05-070	Sumatra	Bahia
06	ER 12-040	Sumatra	Bahia
07	ER 13-061	Sumatra	Bahia
08	ER 13-065	Sumatra	Bahia
09	ER 28-027	Sumatra	Bahia
10	ER 33-021	Sumatra	Bahia
11	ER 33-022	Sumatra	Bahia
12	ER 33-023	Sumatra	Bahia
13	109 PD	Sumatra	Bahia
14	125 PD	Sumatra	Bahia
15	221 PD	Sumatra	Bahia

2.2. Conjunto de dados

Foram analisados 15 descritores quantitativos, definidos conforme a Subcomissão de Sementes do SINDIFUMO, com base na descrição recomendada pela International Union for the Protection of Varieties of Plants (UPOV) e Legislações Americana e Italiana (Tabela 2).

Tabela 2 - Relação das variáveis quantitativas de 15 genótipos de tabaco (formatar tamanho e fonte e ver artigos a fim de melhorar este título)

Variáveis quantitativas	**Unidade de medida**
Altura total da planta (ALT)	cm
Nº de folhas (NF)	-
Diâmetro médio do caule (DCM)	mm
Índice cilíndrico (IC) =quociente entre diâmetro médio e base da inflorescência	-
Largura da 3º folha (CFT)	cm
Comprimento da 3º folha (LFT)	cm
Largura da 10º folha (LFD)	cm
Comprimento da 10º folha (CFD)	cm
Largura da base 10º folha (LBD)	cm
Ângulo de inserção 10º folha (AI)	(º)

Comprimento dos internódios (MINT)	cm
Comprimento da flor (CFRL)	cm
Diâmetro do tubo da flor (DFRL)	mm
Engrossamento do tubo da flor (EFRL)	mm
Comprimento da corola (CCRL)	cm

2.3. Medida de dissimilaridade

Através da análise multivariada foram obtidas as matrizes de médias dos descritores e covariâncias residuais. A partir dessas matrizes foi realizada a análise de agrupamento utilizando a distância generalizada de Mahalanobis (D^2_{ij}) como medida de dissimilaridade. Definida pela expressão:

$$D_{IJ}^2 = \left(X_I - X_J\right)' E^{-1} \left(X_I - X_J\right)$$

Onde X_I e X_j são vetores médios associados aos acessos i e j; E^{-1} é a matriz de covariâncias residuais (CRUZ e REGAZZI, 1994).

2.4. Métodos de agrupamentos

2.4.1. Método UPGMA

Para a análise de agrupamento utilizou-se a distância generalizada de Mahalanobis (D2) como medida de dissimilaridade a partir dos dados padronizados. Os agrupamentos hierárquicos foram obtidos pelo método UPGMA - Unweighted Pair Group Method with Arithmetic Mean (SNEATH; SOKAL, 1973). A validação dos agrupamentos foi determinada pelo coeficiente de correlação cofenética de acordo com SOKAL e ROHLF (1962). A significância dos coeficientes de correlação cofenética foi calculada pelo teste de Mantel com 1000 permutações (MANTEL, 1967). O critério para definição do número de grupos foi feito pelo método do pseudo-t2 (MINGOTTI, 2005) utilizando o pacote NbClust pertencente ao programa computacional R (CHARRAD et al., 2013). Toda as análises foram realizadas com auxílio do software R versão 4.0.5 (R CORE TEAM, 2021).

2.4.2. Método de Tocher

O método de otimização de Tocher foi realizado a partir da matriz de distâncias de Mahalanobis, sobre a qual é identificado o par de indivíduos que apresentam maior semelhança a partir desses o primeiro grupo será formado. Em seguida, a probabilidade da

inclusão de novos indivíduos é avaliada, adotando-se o critério de que a distância média intragrupo seja menor que a distância média intergrupo. Sempre que ocorrer a inclusão de um novo indivíduo distância média intragrupo é aumentada (CRUZ et al., 2014). As análises foram realizadas utilizando o pacote MultivariateAnalysis pertencente ao programa computacional R (AZEVEDO et al., 2021).

A decisão de inclui ou não um novo indivíduo k no grupo é, então, feita considerando:

Se $\frac{d_{(grupo)k}}{\text{n}} \leq \theta$, inclui-se o indivíduo k no grupo;

Se $\frac{d_{(grupo)}}{\text{n}} > \theta$, o indivíduo k não é incluído no grupo.

Sendo *n* o número de indivíduos que constitui o grupo original.

Para estabelecimento dos grupos a distância entre os indivíduos k e o grupo formado pelos indivíduos ij é dada por: $d_{(ij)k} = d_{ik} + d_{ij}$

A distância média intragrupo é determinada da seguinte maneira:

$$\frac{\sum id}{C_{n,2}} \text{ ; com } i = 1,2, \ldots, C_{n,2}.$$

A distância média dentro do grupo é a média das distâncias entre cada par de genitores que o constitui. Pelo critério adotado, esta distância é sempre menor que as distâncias médias intergrupos que são obtidas de forma similar, pela soma das distancias de todos os pares possíveis de indivíduos entre dois grupos, dividida pelo número de pares (CRUZ et al., 2014).

2.5. Redes neurais artificiais

A rede utilizada neste trabalho é baseada em uma rede neural artificial de Feedforward multicamadas usando retropropagação. Para desenvolvimento das redes MLP, foi utilizado o algoritmo DeepLearning do pacote H2O (LEDELL et al. 2022) implementado no software R Core Team (2021). O banco de dados foi dividido de forma aleatória em dois conjuntos: treinamento (70%) e validação (30%). Os dados de treinamento são utilizados para determinar os parâmetros do modelo (normalização, pesos e viés da RNA) e os dados de validação para estimar o desempenho alcançado.

O aprendizado das redes foi do tipo supervisionado, com isso foram dados para a rede dois conjuntos de valores: o conjunto de valores de entrada e o conjunto de valores de saída. Desta forma, o treinamento consistiu em um problema de otimização dos parâmetros da rede (seus pesos sinápticos), para que pudessem responder às entradas conforme esperado, até que o erro entre os padrões de saída gerados pela rede alcançasse o valor mínimo desejado.

Diferentes arquiteturas contendo duas camadas ocultas foram analisadas, variando o número de ciclos de 100 a 1000 (com intervalo de 100) e o número de neurônios de 5 a 10 nas duas camadas respectivamente, como apresentados na tabela 3. A função de ativação utilizada foi a unidade linear retificada (ReLu). Como critério de escolha foram selecionadas as redes obtiveram com arquitetura mais simples e com as maiores taxa de acerto na etapa de treinamento e validação respectivamente. Uma taxa de acerto alta, principalmente na validação, indica que a RNA obteve êxito na generalização dos resultados.

Tabela 3 - Configurações utilizadas para realização das etapas de treinamento e validação do modelo da Rede Neural Artificial.

Nº Ciclos	**Arquiteturas**					
	RNA 2	**RNA 3**	**RNA 4**	**RNA 5**	**RNA 6**	**RNA 7**
100-1000	16-5-5-3	16-6-5-3	16-7-5-3	16-8-5-3	16-9-5-3	16-10-5-3
100-1000	16-5-6-3	16-6-6-3	16-7-6-3	16-8-6-3	16-9-6-3	16-10-6-3
100-1000	16-5-7-3	16-6-7-3	16-7-7-3	16-8-7-3	16-9-7-3	16-10-7-3
100-1000	16-5-8-3	16-6-8-3	16-7-8-3	16-8-8-3	16-9-8-3	16-10-8-3
100-1000	16-5-9-3	16-6-9-3	16-7-9-3	16-8-9-3	16-9-9-3	16-10-9-3
100-1000	16-5-10-3	16-6-10-3	16-7-10-3	16-8-10-3	16-9-10-3	16-10-10-3

Nota: o primeiro número representa o nº de neurônios na camada de entrada, o ultimo número repesenta o nº de neurônios na camada de saída, e os números intermendiários representam o nº de neurônios na camada intermediária.

Para avaliação do classificador foi gerada matriz de confusão (Tabela 4) e nível de exatidão ou confiança da classificação (índice Kappa) para cada uma das arquiteturas treinada. A matriz de confusão oferece uma medida efetiva do classificador utilizado, onde cada coluna da matriz representa os resultados reais enquanto que cada linha corresponde aos resultados preditos pelo classificador.

Tabela 4 - Modelo da matriz de confusão com os resultados corretos e incorretos para fins classificação.

	Predição Positiva	**Predição Negativa**
Classe Positiva	Verdadeiro Positivo (VP)	Falso Negativo (FN)
Classe Negativa	Falso Positivo	Verdadeiro Negativo

(FP) (VN)

O valor da acurácia é calculado levando em consideração os acertos da classificação e o índice Kappa considera toda a matriz de contingência no seu cálculo (SARMIENTO et al., 2014).

$$\text{Precisão} = p0 = \frac{\text{VP} + \text{VN}}{\text{VP} + \text{FP} + \text{FN} + \text{VN}}$$

O índice Kappa é uma proporção de acerto depois da eliminação dos casos de acerto por acaso (ROSENFIELD; FITZPATRICLINS, 1986; PANTALEÃO; SCOTFIEL, 2009).

$$p_{Sim} = \frac{\text{VP} + \text{FN}}{\text{VP} + \text{FN} + \text{FP} + \text{FN}} \times \frac{\text{VP} + \text{VN}}{\text{VP} + \text{VN} + \text{FP} + \text{FN}}$$

Kappa:

$$p_{Não} = \frac{\text{FP} + \text{VN}}{\text{VP} + \text{FN} + \text{FP} + \text{FN}} \times \frac{\text{FN} + \text{VN}}{\text{VP} + \text{FN} + \text{FP} + \text{VN}}$$

$$p_e = p_{Sim} + p_{\text{Não}}$$

Definindo estas medidas, o *Kappa* é dado por:

$$Kappa = \frac{p_0 + p_e}{1 + p_e}$$

3. RESULTADOS E DISCUSSÃO

O agrupamento hierárquico UPGMA (Figura 1) apresentou valor para a correlação cofenética (r = 0,95**) considerado altamente significativo, de acordo com BUSSAB et al. (1990). A partir desse resultado é possível concluir que a distorção provocada pelo agrupamento é baixa, assegurando as inferências realizadas através da análise foram definidos três grupos considerados ideais sendo, o grupo um (G1) formado pelos genótipos 109PD e 221PD, o grupo dois (G2) composto por apenas o genótipo 125PD e o grupo três (G3) formado por 12 genótipos (PD ER03-107; ER04-090; ER04-095; ER05-005; ER05-070; ER12-040; ER13-061; ER13-065; ER28-027; ER33-021; ER33-022).

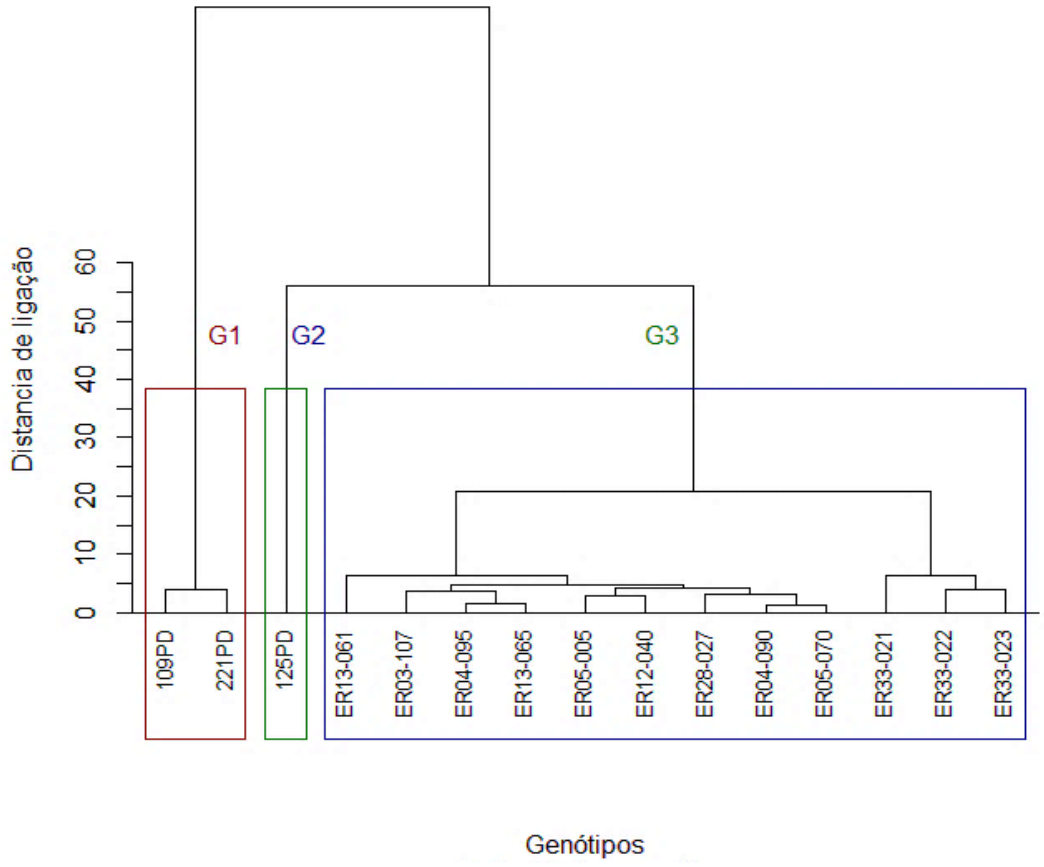

Figura 1 - Dendrograma de dissimilaridade genética entre 15 genótipos de tabaco resultante do agrupamento pelo método UPGMA obtido pela distância de Mahalanobis (D2) estimados em 15 variáveis quantitativas

3.1. Otimização de Tocher

Utilizando o método de otimização de Tocher, foi possível identificar a formação de três grupos (Tabela 5). Os grupos são formados com base na magnitude de suas dissimilaridades, assim os genótipos que fazem parte do mesmo grupo apresentam maior similaridade genética entre si, dessa forma é possível escolher os genitores mais distantes geneticamente. O grupo I foi o que apresentou maior número, contemplando 80% dos genótipos de tabaco avaliados, sendo os genótipos (ER03-107; ER04-090; ER04-095; ER05-005; ER05-070; ER12-040; ER13-061; ER13-065; ER28-027; ER33-021; ER33-022). O grupo 2 foi formado por 13% dos genótipos (109PD e 221PD) o grupo 3 composto por apenas o genótipo 125PD.

Tabela 5 - Formação dos grupos de 15 genótipos de tabaco segundo o método de otimização de Tocher (Original) com a distância generalizada de Mahalanobis.

Grupos	Genótipos					
1	ER03-107	ER04-090	ER04-095	ER05-005	ER05-070	ER12-040
	ER13-061	ER13-065	ER28-027	ER33-021	ER33-022	ER33-023
2	109PD	221PD				
3	125PD					

Fonte: Dados do Autor.

A partir do agrupamento é possível identificar os genótipos que são mais distantes geneticamente. Esse resultado pode ser aproveitado em possíveis cruzamentos para se obter maior heterose na produção de híbridos. Assim, o cruzamento desses genótipos, entre si ou com os demais, deve produzir as populações mais segregantes. É possível identificar nos resultados obtidos que há concordância na formação dos grupos entre os genótipos estudados, tanto pelo método hierárquico UPGMA quanto pelo método de otimização por Tocher.

3.2. Redes Neurais Artificiais

Considerando a classificação dos genótipos quanto aos grupos, entre as redes neurais artificiais (ANN) treinadas, com uma camada oculta contendo 3 neurónios foi o suficiente para solucionar o problema proposto. Verifica-se então a facilidade da ANN ao encontrar uma solução desejável quando se trabalha com um banco de dados robusto. Uma das dificuldades encontrada quando se trabalha com ANN é a falta de um banco com número de observações suficientes para gerar resultados considerados satisfatório. Visto que é preciso realizar a divisão deste em partes desiguais, geralmente acima de 60% para a etapa de treino, e o restante dividido entre o teste e a validação, a uma redução da representatividade da amostra e consequentemente uma limitação na eficiência.

A solução considerada ideal neste estudo foi encontrada em uma configuração de arquitetura de ANN considerada mais simples por conter poucos neurônios na camada oculta. A quantidade de neurônios na camada oculta pode influenciar diretamente na qualidade da rede. Sendo assim, tem-se notado que configurações mais simples, ou seja, menor número de neurônios na camada oculta possível, favorece o processo de busca da melhor configuração para a tarefa designada, além de evitar problemas como o overfitting.

Nas tabelas 6 e 7 respetivamente é possível observar a matriz de confusão gerada para avaliar a eficácia da ANN nas etapas de treinamento e validação. Os valores da diagonal indicam a quantidade de acertos. Portanto, é esperado que em um modelo de classificação

correto, os valores da diagonal da tabela sejam maiores que os valores fora da diagonal. Na Tabela 6 é possível observar que os valores da diagonal seguiram esta premissa, visto que, os valores fora da diagonal foram iguais a 0 demonstrando que a RNA treinada obteve êxito, classificando corretamente as classes G1 G2 e G3 na etapa de treinamento. O mesmo ocorre na etapa de validação (Tabela 7) com ressalva para o G3 onde há uma classificação incorreta, no entanto, ainda é possível afirmar que houve eficiência da rede na generalização dos dados.

Tabela 6 - Matriz de confusão do modelo de treinamento pela Rede Neural Artificial valores da diagonal indicam a classificação correta dos grupos.

Treinamento	G1	G2	G3
G1	193	0	0
G2	0	32	0
G3	0	0	15

Fonte: Dados do Autor.

Tabela 7 - Matriz de confusão do modelo de validação pela Rede Neural Artificial valores da diagonal indicam a classificação correta dos grupos.

Validação	G1	G2	G3
G1	48	0	0
G2	0	8	0
G3	1	0	3

Fonte: Dados do Autor.

As estatísticas indicaram um índice Kappa de 0,9488 na etapa da validação, o que configura a confiabilidade da classificação. Dado que, o índice Kappa é uma medida de exatidão baseada na diferença entre a concordância real na matriz e a concordância por chance, indicada pelo total das linhas e das colunas. Um ponto relevante no nesse índice é o fato de considerar todos os elementos da matriz de confusão no cálculo.

Tabela *8* - - Níveis de precisão (acurácia) e confiança da classificação (Kappa) do modelo de rede neural artificial das etapas de treinamento e validação

Estatística	Treinamento	Validação
Acurácia	1	0,9833
Índice Kappa	1	0,9488

Fonte: Dados do Autor.

Em programas de melhoramento genético e conservação o tempo é crucial, visto que este é influenciado diretamente pela quantidade de recursos disponíveis, sendo um fator limitante no desempenho do programa. Nesse contexto, verifica-se que os métodos baseados em inteligência computacional como a ANN são eficientes e têm contribuído em estudos, sejam

na classificação de populações ou predição, por tratar-se de uma metodologia que é capaz de gerar resultados precisos, mesmo com dados complexos.

De acordo SANT'ANNA et al. (2018) ainda há pouca informação relacionada na escolha das melhores estratégias biométricas que são responsáveis pelo tempo e o êxito na execução do trabalho. Portanto, estudos de ANN com enfoque no melhoramento genético, como o realizado nesse trabalho, são relevantes por contribuírem com informações a respeito da topologia da rede utilizada, visto que é um dos principais fatores relacionados na escolha das melhores estratégias biométricas.

Os trabalhos com tabaco nesta área recentemente têm sido realizados visando classificar a qualidade da folha. Assim como no presente estudo, resultados satisfatórios também foram encontrados por HE et al. 2020. Os autores construíram um modelo utilizando rede neural Backpropagation (BPNN) e regressão múltipla (SRA) com o intuito de disponibilizar metodologias para melhorar qualidade de cura da folha do tabaco Flue-Cured de oito regiões da China, a partir de componentes químicos e dados de qualidade sensorial. A avaliação da precisão dos modelos estudados foi feita através dos índices, erro quadrático médio (MSE); erro padrão médio da regressão (RMSE) e erro absoluto médio (MAE). Os índices MSE, RMSE e MAE para o modelo BPNN da qualidade do aroma melhorou em 17%, 8% e 55%, respectivamente, em comparação com os do modelo SRA. As acurácias dos parâmetros do modelo de BPNN foram melhores do que aqueles do modelo construído usando a análise de SRA. Com base nos resultados afirmaram que o modelo de previsão BPNN é confiável e pode fazer predições com precisão das qualidades sensoriais características do aroma da folha de tabaco Flue-Cured.

Ainda com classificação de folha de tabaco WANG et al., (2018) propuseram um método para controlar o processo de secagem das folhas, por meio de um modelo de rede neural artificial (ANN). Através de imagens, as características de cor e textura das folhas de tabaco, a ANN foi capaz de ajudar na tomada a decisão do ajuste a temperatura e a umidade do galpão de cura em tempo real. A métrica utilizada para avaliar a eficiência do modelo foi o Erro Quadrático Médio (MSE). Os melhores resultados de MSE foram encontrados no ajuste das temperaturas de bulbo seco e úmido 2.03 e 1.65, respectivamente, e na mudança do ponto de ajuste 8.18 quando combinado às características de cor e textura das imagens de folhas. O modelo proposto melhorou significativamente o tempo de trabalho e a estabilidade na produção de folhas de tabaco finas, na cor ideal, de alta qualidade de fragrância no processo de cura do tabaco.

Pesquisas envolvendo outras culturas também vêm sendo realizados. CHEN et al. (2010) combinando a análise de distância de Mahalanobis e rede neural Backpropagation na identificação de cinco variedades de milho permitiram a classificação em três tipos e posteriormente utilizaram uma rede neural com três camadas ocultas para identificação. Segundo os autores os experimentos obtiveram uma média para a precisão da classificação de até 90% de acurácia na classificação dos três tipos de milho. Constaram ainda que o método combinando a distância de Mahalanobis e o classificador de rede neural pode ser empregado com chances de se obter sucessos na identificação de variedades de milho.

4. CONCLUSÃO

A partir do agrupamento e da descrição dos padrões de similaridade genética através das técnicas hierárquica UPGMA e da otimização Tocher, a rede neural artificial foi eficiente ao classificar os 15 genótipos de tabaco tipo Sumatra com base nos caracteres quantitativos.

Os altos valores das estatísticas de validação encontrada garantem a eficácia do modelo aplicado. Ao solucionar o problema proposto utilizando uma topologia considerada mais simples, com uma camada oculta e três neurônios, demostrando o potencial do método para uso em estudos classificatórios.

Sugestão para trabalhos futuros

5. REFERENCIAS BIBLIOGRÁFICAS

AZEVEDO, A. M. MultivariateAnalysis: Pacote Para Analise Multivariada. **R package version 0.4.4,** 2021

BARBOSA, C.D.; VIANA, A.P.; QUINTAL, S.S.R.; PEREIRA, M.G. Artificial neural network analysis of genetic diversity in *Carica papaya* L. **Crop Breeding and Applied Biotechnology**, v. 11, p. 224-231, 2011.

BUSSAB, W. de O.; MIAZAKI, E.S.; ANDRADE, D.F. **Introdução à Análise de Agrupamentos**. In: 9º Simpósio Nacional de Probabilidade e Estatística, São Paulo. Associação Brasileira de Estatística, p. 105,1990.

CHARRAD, M.; GHAZZALI, N.; BOITEAU, V.; NIKNAFS, A. NbClust: An R Package for Determining the Relevant Number of Clusters in a Data Set. **Journal of Statistical Software,** 61(6), 1-36, 2014.

LEDELL, E; GILL, N., AIELLO, S; FU, A; CANDEL, A; CLICK, C; MALOHLAVA, M. **h2o: R Interface for the 'H2O'**, 2022.

CRUZ, C.D., CARNEIRO, P.C.S., REGAZZI, A.J. **Modelos biométricos aplicado ao melhoramento genético**. Viçosa: UFV, 3. ed., v. 2, p. 668, 2014.

CRUZ, C.D.; CARNEIRO, P.C.S. **Modelos biométricos aplicados ao melhoramento genético.** Viçosa: Editora UFV, 2. ed. p. 585, 2006.

DASSARI S.K., CHINTADA K.R., PATRUNI M. Flue-Cured Tobacco Leaves Classification: A Generalized Approach Using Deep Convolutional Neural Networks. In: Cognitive Science and Artificial Intelligence. **SpringerBriefs in Applied Sciences and Technology**. Springer, Singapore. 2018.

DAVALIEVA, K.; MALEVA, I.; FILIPOSKI, K.; SPIROSKI, O.; EFREMOV, G.D. Genetic variability of Macedonian tobacco varieties determined by microsatellite marker analysis. **Diversity,** v. 2, n. 4, p. 439-449, 2010.

FAVERO, A., LOPES, M., & FALEIRO, F. Estado da arte do Pré-melhoramento de espécies vegetais. **Pré-melhoramento, melhoramento e pós-melhoramento: estratégias e desafios.** Brasília: EMBRAPA, p 31-42, 2008.

FAOSTAT - Food and Agriculture Organization of the United Nations Statistical Database. **Crops database**. 2019. Disponível em: <http://faostat3.fao.org/browse/Q/ QC/E>. Acesso em: 1 de fevereiro de 2021.

HAYKIN S. **Redes neurais: princípios e prática**. Bookman, Porto Alegre, p. 900, 2001.HE, C.; CHEN, R.; REN, K.; ZHAO, G.; HE, C.; HU, B.; ZOU, C.; JIANG, Y.; CHEN; Y. A predictive model for the sensory aroma characteristics of flue-cured tobacco based on a back-propagation neural network. **SN Applied Sciences**. v. 2, n. 11, p.1-11, 2020.

HE, C.; CHEN, R.; REN, K.; ZHAO, G.; HE, C.; HU, B.; ZOU, C.; JIANG, Y.; CHEN; Y. A predictive model for the sensory aroma characteristics of flue-cured tobacco based on a back-propagation neural network. **SN Applied Sciences**. v. 2, n. 11, p.1-11, 2020.

MALLESHAPPA, C.; SOWMYA, T.M.; KUAMARA, B.N.; MUUTURAJ, M.D. (2020). Genetic variability and heritability studies in flue cured Verginia tobacco (Nicotiana tobaccum L.) germplasm. **Journal of Pharmacognosy and Phytochemistry**, v. 9, n. 5, p. 3171-3173, 2020.

MANTEL, N. **The detection of disease clustering and generalized regression approach**. *Cancer Research*, Birmingham, v.27, n.2, p.209-220, 1967.

MINGOTI, S.A. Análise de dados através de métodos de estatística multivariada: uma abordagem aplicada. Belo Horizonte: Editora UFMG. p. 297, 2005.

MOON, H.S; NIFONG, J.M; NICHOLSON, J.S; HEINEMAN, A; LION, K; VAN DER HOEVEN, R; HAYES, A.J. AND LEWIS, R.S. Microsatellite-based analysis of tobacco (Nicotiana tabacum L.) genetic resources. **Crop Science**, 49, p. 2149-2159, 2009.

PANTALEÃO, E; SCOFIELD, G. Comparação entre medidas de acurácia de classificação para imagens do satélite ALOS. Anais XIV Simpósio Brasileiro de Sensoriamento Remoto, Natal, Brasil, 25-30 abril 2009, INPE, p. 7039-7046. 2009.

PEIXOTO, L.A.; BHERING, L.L.; CRUZ, C.D. Artificial neural networks reveal efficiency in genetic value prediction. ***Genet. Mol. Res***. *14*, 6796–6807, 2015.

R Core Team. R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. 2021. Disponível em: <https://www.R-project.org/>

ROSENFIELD, G. H.; FITZPATRICK-LINS, K. A coefficient of agreement as a measure of thematic classification accuracy.Photogrammetric engineering and remote sensing, v. 52, n. 2, p. 223-227, 1986.

SABZI-NOJADEH, M.; NIEDBAŁA, G.; YOUNESSI-HAMZEKHANLU, M.; AHARIZAD, S.; ESMAEILPOUR, M.; ABDIPOUR, M.; KUJAWA, S.; NIAZIAN, M. Modeling the Essential Oil and *Trans*-Anethole Yield of Fennel (*Foeniculum vulgare* Mill. var. *vulgare*) by Application Artificial Neural Network and Multiple Linear Regression Methods. **Agriculture**, 11, 1191, 2021.

SANT'ANNA, I.C.; TOMAZ, R.S.; SILVA, G.N.; NASCIMENTO, M.; BHERING, L.L.; CRUZ, C.D. Superiority of artificial neural networks for a genetic classification procedure. **Genet. Mol. Res**, *14*, 9898–9906, 2015.

SARMIENTO, C. M.; RAMIREZ, G. M.; COLTRI, P. P.; LIMA, L. F.; NASSUR, O. A. C.; SOARES, J. F. Comparação de classificadores supervisionados na discriminação de áreas cafeeiras em Campos Gerais-Minas Gerais. **Coffee Science**, v. 9, n. 4, p. 546-557, 2014.

SINDITABACO - Sindicato Interestadual da Indústria do Tabaco. **Origem do Tabaco**. 2019. Disponível em: <http://www.sinditabaco.com.br/sobre-o-setor/origem-do-tabaco/>. Acesso em: 02 de fevereiro de 2020.

SILVA, G.N.; TOMAZ, R.S.; SANT'ANNA, I.D.C.; NASCIMENTO, M.; BHERING, L.L.; CRUZ, C.D. Neural networks for predicting breeding values and genetic gains. **Scientia Agricola,** v. 71, n. 6, p. 494-498, 2014.

SNEATH, P.H.; SOKAL, R.R. **Numerical taxonomy: the principles and practice of numerical classification**. San Francisco: W.H. Freeman, 1973.

SOKAL, R. R. and ROHLF, F. J. The comparison of dendrograms by objective methods. **Taxon**, v.11 p.33-40. 1962.

UPOV (Union pour la Protection des Obtentions Variétales). Disponível em http://www.upov.int/tabaco. : 05 de fevereiro de 2020.

WANG, L., CHENG, B., LI, Z., LIU, T., e LI, J.Intelligent tobacco flue-curing method based on leaf texture feature analysis. **Optik - International Journal for Light and Electron Optics**, 150, p. 117–130. 2017.

WENJIE, P. CHAOYING, J.; YUANJU, T.; YANG, S.X. Tobacco dry weight estimation based on artificial neural network. **Intelligent Automation & Soft Computing**, v. 17, n. 7, p. 997-1007, 2011.

YANG, H.; GENG, X.; ZHAO, S.; SHI, H. Genomic diversity analysis and identification of novel SSR markers in four tobacco varieties by high-throughput resequencing. **Plant Physiology and Biochemistry**, v. 150, p. 80-89, 2020.

Yang, H.-W.; Hsu, H.-C.; Yang, C.-K.; Tsai, M.-J.; Kuo, Y.-F. Differentiating between morphologically similar species in genus Cinnamomum (Lauraceae) using deep convolutional neural networks. *Comput. Electron. Agric. 162*, 739–748, 2019.

ZHANG, F.; ZHANG, X. Classification and Quality Evaluation of Tobacco Leaves Based on Image Processing and Fuzzy Comprehensive Evaluation. **Sensors,** v.11, p. 2369-2384, 2011.

CAPITULO II

DIVERGÊNCIA GENÉTICA ENTRE GENÓTIPOS DE Nicotiana tabacum L. POR MEIO DO ALGORITMO DE ÁRVORE DE DECISÃO

DIVERGÊNCIA GENÉTICA ENTRE GENÓTIPOS DE *Nicotiana tabacum* L. POR MEIO DO ALGORITMO DE ÁRVORE DE DECISÃO

RESUMO

A utilização de inteligência computacional tem se tornado cada vez mais frequente em diversas áreas da ciência, entre elas a genética vegetal. Diferentes modelos quando bem empregados são capazes, por exemplo, de classificar, prever ou agrupar informações com altas taxas de acerto em pequenos intervalos de tempo, fatores fundamentais ao melhoramento genético. Uma das técnicas de aprendizagem supervisionada a árvore de decisão apresenta uma vantagem ao particionar problemas complexos e assim buscar uma solução adequada. Este trabalho tem por objetivo classificar 15 genótipos de tabaco tipo Sumatra, aplicando árvore de decisão por meio de grupos pré-definidos com base em descritores quantitativos. dois conjuntos, treinamento e validação de dados foram criados a fim de avaliar o modelo gerado pelo algoritmo de árvore de decisão. A variável mais importante foi localizada na raiz da arvore definido com a altura total da planta. Ao analisar a matriz de confusão gerada para árvore de classificação obteve-se uma acurácia de 0.97. Os resultados demostram que a metodologia de aprendizado de máquina arvore de classificação é uma técnica promissora, que pode auxiliar em estudos que visem a manutenção de banco dos germoplasmas e também em programas de conservação e melhoramento.

Palavras-chave: Algoritmos de classificação; Aprendizagem supervisionada; Tabaco tipo Sumatra.

GENETIC DIVERGENCE BETWEEN GENOTYPES OF Nicotiana tabacum L. THROUGH THE DECISION TREE ALGORITHM

ABSTRACT

The use of computational intelligence has become increasingly frequent in several areas of science, including plant genetics. Different models, when well used, are capable, for example, of classifying, predicting or grouping information with high accuracy rates in small time intervals, fundamental factors for genetic improvement. One of the supervised learning techniques the decision tree has an advantage when partitioning complex problems and thus seeking an adequate solution. This work aims to classify 15 Sumatra-type tobacco (Nicotiana tabacum L.) genotypes, applying a decision tree through pre-defined groups based on quantitative descriptors. two sets, training and data validation were created in order to evaluate the model generated by the decision tree algorithm. The most important variable was located at the root of the tree defined with the total height of the plant. When analyzing the confusion matrix generated for the classification tree, an accuracy of 0.97 was obtained. The results show that the classification tree machine learning methodology is a promising technique, which can help in studies aimed at maintaining a germplasm bank and also in conservation and improvement programs.

Keywords: Classification algorithms; Supervised learning; Sumatra tobacco.

1. INTRODUÇÃO

Nicotiana tabacum L., popularmente conhecida como fumo é um anfidiplóide e uma das 64 espécies pertencente pertence à família Solanaceae (LORENCETTI; MALLMANN; SANTOS, 2008). Amplamente cultivada em diversos países do mundo, como fonte de matéria prima para indústria do tabaco, é também considerado um dos mais importantes sistemas modelo em biotecnologia vegetal e altamente promissores para farmacologia e cosmetologia, se tornando então uma das principais culturas não alimentícia movimentando bilhões de dólares e gerando milhões de empregos diretos e indiretos (LORENCETTI; MALLMANN; SANTOS, 2008; WENJIE et al., 2011, DASSARI et al., 2018; FAOSTAT, 2019).

A alta variabilidade genética encontrada na espécie tem sido alvo de estudos em programas de melhoramento e conservação. Nesse sentido, as pesquisas de discriminação de populações em cultivares de fumo têm se destacado por possibilitarem a seleção de progenitores adequados, levando consequentemente a otimização dos ganhos seletivos, motivo que confere a grande importância desta etapa para o trabalho do melhorista, e também para fins de conservação de germoplasma e/ou ampliação das bases genéticas (ZHANG et al., 2008, DAVALIEVA et al., 2010, MALLESHAPPA et al., 2020, YANG et al., 2020).

Análises multivariadas têm sido empregadas em populações de programas de melhoramento genético, permitindo estimar a dissimilaridade genética existente (CRUZ e CARNEIRO, 2003). No entanto, o emprego de métodos baseados em inteligência computacional é uma das áreas que tem se destacado ao longo dos anos (FACELI et al., 2011). Em razão da capacidade de aprender e generalizar durante o processo, o modelo é utilizado para classificar, prever ou agrupar novos exemplos a partir da experiência obtida durante o treinamento. Outra vantagem é a tolerância a ruídos que possibilita a modelagem de sistemas não lineares, onde a relação entre as variáveis estudadas não é conhecida ou são altamente complexas (LIAKOS et al., 2018; SANT'ANNA et al., 2018).

Dentre as técnicas de aprendizagem supervisionada, a árvore de decisão (DT)é um método relativamente simples em relação a outros e que vem sendo amplamente utilizada na solução de problemas complexos de classificação e regressão. A vantagem de uma DT em relação a outros métodos é a estratégia utilizada para encontrar a solução desejada. O primeiro passo é a transformação de um problema complexo em subproblemas mais simples

e repetidamente este passo é aplicado a cada subproblema até encontrar a solução ideal. Além disso, outra vantagem é a facilidade de interpretação dos resultados, devido a sua capacidade fornecer representações gráficas simples (ELOUEDI; MELLOULI; SMETS, 2001, FACELI et al., 2011, TRABELSI et al., 2018).

Diante do exposto, este trabalho tem por objetivo aplicar a árvore de decisão na classificação de 15 genótipos de tabaco (*Nicotiana tabacum* L.) tipo Sumatra, por meio de grupos pré-definidos com base em descritores quantitativos.

2. MATERIAL E MÉTODOS

2.1. Área de estudo

Os dados utilizados se referem a um experimento conduzido no ano agrícola de 2011, no campo de produção comercial onde foram caracterizados 15 genótipos de tabaco (*Nicotiana tabacum* L.) tipo Sumatra (Tabela 1), cultivado sob tela preta com 30% de sombreamento pela empresa Ermor Tabarama Tabacos Do Brasil LTDA.

O campo de produção onde foi implantado o experimento localiza-se no município de Muritiba – BA, a uma altitude de 220 m em relação ao nível do mar, que apresentam precipitação pluviométrica anual média de 1.220 mm, apresentando temperatura média anual de 24,1 º C. O delineamento experimental utilizado foi o de blocos casualizados com quatro repetições. Cada parcela foi constituída de cinco linhas de 10 plantas e cada linha teve 4,5 metros de comprimento com espaçamento de 1,0 metros entre linhas e 0,42 metros entre plantas.

Tabela 2 - Relação dos genótipos de Tabaco provenientes da empresa Ermor Tabarama Tabacos do Brasil

Código (Nº)	Genótipos	Tipo	Origem
01	ER 03-107	Sumatra	Bahia
02	ER 04-090	Sumatra	Bahia
03	ER 04-095	Sumatra	Bahia
04	ER 05-005	Sumatra	Bahia
05	ER 05-070	Sumatra	Bahia
06	ER 12-040	Sumatra	Bahia
07	ER 13-061	Sumatra	Bahia
08	ER 13-065	Sumatra	Bahia
09	ER 28-027	Sumatra	Bahia

10	ER 33-021	Sumatra	Bahia
11	ER 33-022	Sumatra	Bahia
12	ER 33-023	Sumatra	Bahia
13	109 PD	Sumatra	Bahia
Código (Nº)	**Genótipos**	**Tipo**	**Origem**
14	125 PD	Sumatra	Bahia
15	221 PD	Sumatra	Bahia

2.2. Conjunto de dados

Foram analisados 15 descritores quantitativos, definidos conforme a Subcomissão de Sementes do SINDIFUMO, com base na descrição recomendada pela International Union for the Protection of Varieties of Plants (UPOV) e Legislações Americana e Italiana (Tabela 2).

Tabela 3 - Relação das variáveis quantitativas de 15 genótipos de tabaco

Variáveis quantitativas	**Unidade de medida**
Altura total da planta (ALT)	cm
Nº de folhas (NF)	-
Diâmetro médio do caule (DCM)	mm
Índice cilíndrico (IC) =quociente entre diâmetro médio e base da inflorescência	-
Largura da 3º folha (CFT)	cm
Comprimento da 3º folha (LFT)	cm
Largura da 10º folha (LFD)	cm
Comprimento da 10º folha (CFD)	cm
Largura da base 10º folha (LBD)	cm
Ângulo de inserção 10º folha (AI)	(º)
Comprimento dos internódios (MINT)	cm
Comprimento da flor (CFRL)	cm
Diâmetro do tubo da flor (DFRL)	mm
Engrossamento do tubo da flor (EFRL)	mm
Comprimento da corola (CCRL)	cm

2.3. Arvore de classificação

A modelagem da classificação proposta tem como intuito criar regras para definir se os genótipos pertencem aos grupos 1,2 ou 3. As analises foram realizadas por meio do pacote rpart "Recursive Partitioning and Regression Trees"(THERNEAU e ATKINSON, 2019) com essa ferramenta é possível construir modelos de classificação ou regressão que podem ser representados como árvores binarias (WILLIAMS G, 2009)

Ao construir a arvore de classificação, se tem como meta obter regiões R1, R2, ..., RM que minimizam um dos três critérios apresentados a seguir (JAMES et al., 2013).

Índice de Gini $$G = \sum_{k=1}^{K} \hat{p}_{mk}(1 - \hat{p}_{mk}),$$

onde $\hat{p}_{mk}$ representa a proporção de observações na m-ésima região pertencentes a k-ésima classe. Segundo JAMES et al., (2013) na construção da árvore de classificação o índice Gini é um dos mais indicado por apresentar uma maior sensibilidade a pureza. O índice diminui de acordo com o crescimento da árvore que ocorre através da divisão binária recursiva. Uma arvore de decisão quando apresenta muitos galhos pode ocasionar em overfitting, a solução mais usual para esse tipo de situação é através dos métodos de "poda".

A fim de avaliar o modelo gerado pelo algoritmo de DT, dois conjuntos de dados foram criados: conjunto de treino de predição e validação. Optou-se por se utilizar 80% dos dados para o treino e 20% para validação. Para avaliação do classificador foi gerada matriz de confusão (Tabela 3) e nível de exatidão ou confiança da classificação (índice Kappa) para cada uma das arquiteturas treinada. A matriz de confusão oferece uma medida efetiva do classificador utilizado, onde cada coluna da matriz representa os resultados reais enquanto que cada linha corresponde aos resultados preditos pelo classificador. Toda as análises foram realizadas com auxílio do software R versão 4.0.5 (R CORE TEAM, 2021).

Tabela 4 - Modelo da matriz de confusão com os resultados corretos e incorretos para fins avaliação do modelo

	Predição Positiva	**Predição Negativa**
Classe Positiva	Verdadeiro Positivo (VP)	Falso Negativo (FN)
Classe Negativa	Falso Positivo (FP)	Verdadeiro Negativo (VN)

O valor da acurácia é calculado levando em consideração os acertos da classificação e o índice Kappa considera toda a matriz de contingência no seu cálculo (SARMIENTO et al., 2014).

$$\text{Precisão} = p0 = \frac{\text{VP} + \text{VN}}{\text{VP} + \text{FP} + \text{FN} + \text{VN}}$$

O índice Kappa é uma proporção de acerto depois da eliminação dos casos de acerto por acaso (ROSENFIELD; FITZPATRICK LINS, 1986; PANTALEÃO; SCOTFIEL, 2009).

$$p_{Sim} = \frac{\text{VP} + \text{FN}}{\text{VP} + \text{FN} + \text{FP} + \text{FN}} \times \frac{\text{VP} + \text{VN}}{\text{VP} + \text{VN} + \text{FP} + \text{FN}}$$

Kappa:

$$p_{Não} = \frac{\text{FP} + \text{VN}}{\text{VP} + \text{FN} + \text{FP} + \text{FN}} \times \frac{\text{FN} + \text{VN}}{\text{VP} + \text{FN} + \text{FP} + \text{VN}}$$

$$p_e = p_{Sim} + p_{\text{Não}}$$

Definindo estas medidas, o *Kappa* é dado por:

$$Kappa = \frac{p_0 + p_e}{1 + p_e}$$

3. RESULTADOS E DISCUSSÃO

A construção da árvore é realizada a partir do topo (nó "raiz") seguido pelos seus ramos, de acordo com os testes nas variáveis explicativas, até se chegar nos nós "folhas" que caracterizam a decisão a ser tomada. Na análise de uma árvore de classificação a interpretação é realizada a partir da previsão da classe pertencente a um nó final e as suas proporções correspondentes da classe entre as observações escolhidas (JAMES et al., 2013). Na figura 1 é possível observar a árvore de decisão utilizada como ferramenta para distinguir melhor a classificação dos genótipos de tabaco quanto aos descritores quantitativos. Dentre todas as variáveis de entrada o modelo preditivo mostra que apenas duas foram significativas na árvore de classificação.

Com base nos valores das variáveis o nó raiz foi definido com a altura total da planta (ETT), os genótipos que apresentaram valores maior ou igual que 2.2 foram classificados como o grupo 1 totalizando 79% dos escolhidos. Do outro lado da árvore para valores menores da ETT está a variável da Largura da base 10º (LBL10) diretamente correlacionada com o fato dos genótipos serem do grupo 1 ou 2. A partir deste ponto foi definido uma nova regra, na qual, os genótipos com LBL10 maior que 8.8 formaram o grupo 2, sendo este o segundo mais representativo, com 13% dos escolhidos. Por fim, na última classificação, os genótipos com ETT e LBL10 com valores menores que o estabelecido nas regras, formaram o grupo 3, compondo 8% dos escolhidos.

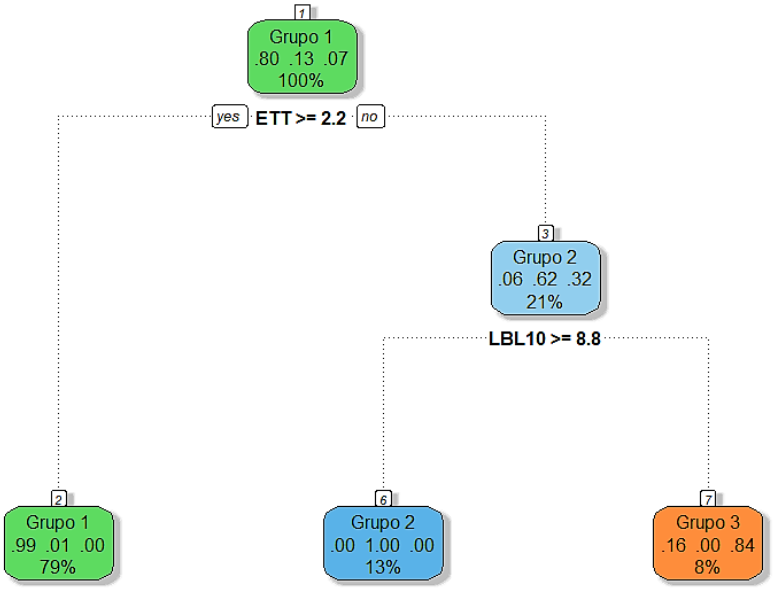

Figura 1 - Árvore de classificação com base no índice de Gini **(**BREIMAN et al., 1984**)** para o conjunto de dados de treinamento dos 15 genótipos de tabaco

Ao analisar a matriz de confusão gerada para árvore de classificação, a pontuação da previsão para acurácia dos dados de treino na Tabela 4 (80% do conjunto de dados) foi de 100% para validar os resultados considerando somente as amostras de validação (20% do conjunto de dados) obteve-se aproximadamente 0.97 de acurácia (Tabela 5).

Tabela 5 - Matriz de confusão do modelo de treinamento pela arvore de decisão valores da diagonal indicam a classificação correta dos grupos.

Treinamento	G1	G2	G3
G1	191	0	0
G2	0	30	0
G3	0	0	20

Tabela 6 - Matriz de confusão do modelo de validação pela arvore de decisão valores da diagonal indicam a classificação correta dos grupos.

Validação	G1	G2	G3
G1	48	2	0
G2	0	6	0
G3	0	0	4

O índice kappa foi de 0.89 para a validação demonstrando a confiabilidade dos resultados. Estes resultados comprovam como as árvores de classificação são intuitivas, pois constroem as regras que levam até grupos pré-definidos em relação à variável estudada. De acordo com MCHUGH (2012) um coeficiente Kappa de Cohen superior a 0,4 pode ser considerado uma estimativa boa a excelente de concordância entre as metodologias.

Tabela 7 - Níveis de precisão (acurácia) e confiança da classificação (Kappa) do algoritmo da árvore de classificação das etapas de treinamento e validação.

Estatística	Treinamento	Validação
Acurácia	1	0,9667
Índice Kappa	1	0,8944

Estudos de divergência genética em cultivares de fumo são de suma importância, por se tratar de um tema que envolve diretamente a conservação de recursos genéticos, além de possibilitar a identificação de duplicatas em um banco de germoplasma e no melhoramento genético quando o objetivo é estabelecer grupos heterótico, identificando combinações com maiores expressões do vigor híbrido.

Ao avaliar os resultados da classificação do tabaco no presente estudo, fica evidente que a arvore de decisão (DT) permite compreender melhor o sentido por trás da diferenciação entre os genótipos com base nos descritores quantitativos. Em razão disso, a árvore de decisão é uma das ferramentas de aprendizado supervisionado de fácil interpretação dos resultados, que pode ser utilizada para resolver problemas de classificação complexos (FACELI et al., 2011; PILTAVER et al., 2016).

Aplicações com sucesso do algoritmo foram também obtidas por ALMEIDA et al. (2021) em um estudo visando classificar genótipos de feijão-fava quanto aos centros de

domesticação e estado biológico. Os autores utilizaram as técnicas de análise discriminante de componentes principais (DAPC) e árvore de decisão (DT) com base de descritores quantitativos. A variável nó raiz foi a área da semente, e a variável mais importante para a classificação foi o peso da semente. Os resultados obtidos demonstraram que a DT foi eficiente na identificação de padrões de classificação nos acessos de feijão-fava.

As abordagens de aprendizado de máquina supervisionado na agricultura têm sido utilizadas como forma de evitar o problema de não linearidade dos dados, por exemplo, SOUSA et al. (2020) avaliaram o desempenho árvore de decisão em comparação com redes neurais e com a tradicional Regressão Linear Generalizada Bayesiana (GBLASSO) na previsão da resistência à ferrugem da folha em diferentes genótipos de café arábica e na identificação da importância de marcadores relacionados à característica de interesse. Os resultados mostraram que as metodologias de aprendizado de máquina ANN e os refinamentos da árvore de decisão (Poda, Bagging, RF e Boosting) foram mais eficazes, visto que, seus valores de precisão foram maiores que o método tradicional, além disso, o melhor marcador foi identificado pela DT.

Embora os estudos de aprendizado de máquinas estejam amplamente difundidos algumas culturas ainda carecem de pesquisas envolvendo o método. Assim, TORABI-GIGLOU et al. (2020) estimaram a estrutura genética com marcadores microssatélites SSR e avaliaram cinco algoritmos supervisionados (Naive Bayesian, regressão logística, máquina de vetor suporte (SVM), Randon Forest e árvore de decisão) na classificação de populações de batatas selvagem e comercial, como também identificaram os melhores primers SSR para distinguir cada população de batata. A precisão geral mais baixa foi de 50% para árvore de decisão e a de maior valor foi 90% para SVM, além disso, identificaram que o primer SS110 foi o melhor para a análise da população de batata. Segundo os autores, os métodos utilizados podem ser aplicados de forma eficiente como ferramentas adequadas para categorizar as populações de batata, sendo ressaltada ainda à importância do estudo, por ser pioneiro na classificação de batatas utilizando bioinformática e aprendizado de máquinas.

4. CONCLUSÃO

A partir do agrupamento previamente estabelecido através das técnicas hierárquica UPGMA e da otimização Tocher, a utilização da árvore de decisão na classificação do tabaco tipo Sumatra, com base em descritores quantitativos, demonstrou-se uma técnica promissora, que pode auxiliar na tomada de decisões, em estudos que visem a manutenção

de banco dos germoplasmas e também em programas de conservação e melhoramento genético de tabaco da região do Recôncavo baiano.

5. REFERENCIAS BIBLIOGRÁFICAS

ALMEIDA, R.D.C.; ASSUNÇÃO NETO, W.V.D.; SILVA, V.B.D.; CARVALHO, L.C.B.; LOPES, Â.C.D.A.; GOMES, R.L.F. Decision tree as a tool in the classification of lima bean accessions**. Revista Caatinga**, v. 34, p. 471-478, 2021.

BREIMAN, L., FRIEDMAN, J. H., OLSHEN, R. A., & STONE, C. J. Classification and Regression Trees; Wadsworth and Brooks/Cole: Monterey, CA, USA. 426p, 1984.

DASARI, S.K.; CHINTADA, K.R.; PATRUNI, M. Flue-cured tobacco leaves classification: A generalized approach using deep convolutional neural networks. In: **Cognitive Science and Artificial Intelligence.** Springer, Singapore, p. 13-21, 2018.

ELOUEDI, Z.; MELLOULI, K.; SMETS, P. Belief decision trees: theoretical foundations. **International Journal of Approximate Reasoning,** v. 28, n. 2-3, p. 91-124, 2001.

FACELI, K.; LORENA, A.C.; GAMA, J.; CARVALHO, A.C.P.L.F. de. **Inteligência Artificial: uma abordagem de aprendizado de máquina.** Rio de Janeiro: Livros Técnicos e Científicos Editora, p. 378, 2011.

FAOSTAT - Food and Agriculture Organization of the United Nations Statistical Database. **Crops database**. 2019. Disponível em: <http://faostat3.fao.org/browse/Q/QC/E>. Acesso em: 1 de fevereiro de 2021.

INTERNATIONAL UNION FOR THE PROTECTION OF NEW VARIETIES OF PLANTS - UPOV. Disponível em: <http:// www.upov.int/tabaco>

SOUSA, I. C. D; NASCIMENTO, M; SILVA, G. N; NASCIMENTO, A. C. C; CRUZ, C. D; ALMEIDA, D. P. D; CAIXETA, E. T. Genomic prediction of leaf rust resistance to Arabica coffee using machine learning algorithms. **Scientia Agricola**, p. 78. 2020.

JAMES, G.; WITTEN, D.; HASTIE, T.; TIBSHIRANI, R. **An Introduction to Statistical Learning with Applications in R**. 1st ed. New York, NY: Springer, 2013.

LIAKOS, K. G; BUSATO, P; MOSHOU, D; PEARSON, S e BOCHTIS, D. Machine learning in agriculture: A review. **Sensors,** p. *18*(8), 2674, 2018.

LORENCETTI, C.; MALLMANN, I. L.; SANTOS, M. Fumo. In: BARBIERI, R. L.; STUMPF, E. R. T. **Origem e evolução de plantas cultivadas**. Brasília, DF: Embrapa Informação Tecnológica, p. 379-401. 2008.

MCHUGH, M. L. Interrater reliability: the kappa statistic. **Biochemia medica: Biochemia medica**, 22(3), p. 276-282, 2012.

MALLESHAPPA, C.; SOWMYA, T.M.; KUAMARA, B.N.; MUUTURAJ, M.D. (2020). Genetic variability and heritability studies in flue cured Verginia tobacco (Nicotiana tobaccum L.) germplasm. **Journal of Pharmacognosy and Phytochemistry**, v. 9, n. 5, p. 3171-3173, 2020.

PANTALEÃO, E; SCOFIELD, G. Comparação entre medidas de acurácia de classificação para imagens do satélite ALOS. **Anais** XIV Simpósio Brasileiro de Sensoriamento Remoto, Natal, Brasil, 25-30 abril 2009, INPE, p. 7039-7046. 2009.

PILTAVER, R. et al. What makes classification trees comprehensible? **Expert Systems With Applications,** 62: 333-346, 2016.

R CORE TEAM. R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. 2021. Disponível em: URL https://www.R-project.org/

ROSENFIELD, G. H.; FITZPATRICK-LINS, K. A coefficient of agreement as a measure of thematic classification accuracy**. Photogrammetric engineering and remote sensing**, v. 52, n. 2, p. 223-227, 1986.

THERNEAU, T. E ATKINSON B. rpart: Recursive Partitioning and Regression Trees. **R package version 4.1-15**, 2019.

SANT'ANNA, I. C. et al. RNA - Aplicações em estudos classificatórios. In: CRUZ, Cosme Damião; NASCIMENTO, Moysés. (Eds.). **Inteligência computacional aplicada ao melhoramento genético** 1. ed. Viçosa: UFV, 2018. cap. 7. p. 189-214.

SARMIENTO, C. M.; RAMIREZ, G. M.; COLTRI, P. P.; LIMA, L. F.; NASSUR, O. A. C.; SOARES, J. F. Comparação de classificadores supervisionados na discriminação de áreas cafeeiras em Campos Gerais-Minas Gerais. **Coffee Science**, v. 9, n. 4, p. 546-557, 2014.

TORABI-GIGLOU, M; MOHARRAMNEJAD, S; PANAHANDEH, J; EBADI-SEGHERLOO, A; e GHASEMI, E. Machine Learning for Detecting Potato Populations Using SSR Markers. **Iranian Journal of Science and Technology, Transactions A: Science**, v. 44, p. 911-918, 2020.

TRABELSI, A.; ELOUEDI, Z.; LEFEVRE, E. Decision tree classifiers for evidential attribute values and class labels. **Fuzzy Sets and Systems,** v. 366, p. 46-62, 2019.

WENJIE, P. CHAOYING, J.; YUANJU, T.; YANG, S.X. Tobacco dry weight estimation based on artificial neural network. **Intelligent Automation & Soft Computing**, v. 17, n. 7, p. 997-1007, 2011.

WILLIAMS, G. Rattle: A Data Mining GUI for R , Graham J Williams, **The R Journal,** 1 (2): 45-55, 2009.

ZHANG HY, LIU XZ, HE CS, YANG YM. Genetic diversity among flue-cured tobacco cultivars based on RAPD and AFLP markers**. Braz Arch Biol Techno**; 51(6):1097-1101, l. 2008.

CONSIDERAÇÕES FINAIS

O uso de ferramentas baseadas em inteligência artificial tem sido amplamente estudado no meio cientifico. Este trabalho demonstrou que, utilizando-se método convencional e algoritmos computacionais é possível realizar o estudo da diversidade genética em genótipos de tabaco. No que tange o uso das redes neurais artificiais na cultura do tabaco têm sido utilizadas para classificação e avaliação de qualidades de folhas e métodos para automatizar os processos de cura de folhas.

Por fim, vale ressaltar que com a experiência adquirida com este trabalho, os dados apresentados são potencialmente úteis em estudos como de pré-melhoramento e também aumentam a quantidade de dados disponíveis favorecendo o crescimento do setor principalmente na região do Recôncavo da Bahia. Como sugestões para trabalhos futuros, fica sugerido mais estudos quanto a utilização das técnicas com inclusão de dados moleculares; e também outros métodos de classificação se cita como exemplo: método de mapa auto organizáveis de kohonen.