UNIVERSIDADE ESTADUAL PAULISTA – UNESP
CÂMPUS DE JABOTICABAL

RUMO À AGRICULTURA INTELIGENTE: PREVISÃO DE PRODUTIVIDADE AGRÍCOLA COM DADOS AGROMETEOROLÓGICOS USANDO MACHINE LEARNING

Kamila Cunha de Meneses
Engenheira Agrônoma

2021

UNIVERSIDADE ESTADUAL PAULISTA – UNESP

CÂMPUS DE JABOTICABAL

RUMO À AGRICULTURA INTELIGENTE: PREVISÃO DE PRODUTIVIDADE AGRÍCOLA COM DADOS AGROMETEOROLÓGICOS USANDO MACHINE LEARNING

Kamila Cunha de Meneses

Orientador: Prof. Dr. Glauco de Souza Rolim

Coorientador: Prof. Dr. Newton La Scala Júnior

Tese apresentada à Faculdade de Ciências Agrárias e Veterinárias – Unesp, Câmpus de Jaboticabal, como parte das exigências para a obtenção do título de Doutora em Agronomia (Ciência do Solo).

2021

Meneses, Kamila Cunha de
M543r Rumo à agricultura inteligente: previsão de produtividade agrícola com dados agrometeorológicos usando Machine Learning / Kamila Cunha de Meneses. – – Jaboticabal, 2021
97 p.

Tese (doutorado) - Universidade Estadual Paulista (Unesp), Faculdade de Ciências Agrárias e Veterinárias, Jaboticabal
Orientador: Glauco de Souza Rolim
Coorientador: Newton La Scala Junior

1. Análise de riscos. 2. Agrometeorologia. 3. Inteligência artificial. I. Título.

Sistema de geração automática de fichas catalográficas da Unesp. Biblioteca da Faculdade de Ciências Agrárias e Veterinárias, Jaboticabal. Dados fornecidos pelo autor(a).

UNIVERSIDADE ESTADUAL PAULISTA

Câmpus de Jaboticabal

CERTIFICADO DE APROVAÇÃO

TÍTULO DA TESE: RUMO À AGRICULTURA INTELIGENTE: PREVISÃO DE PRODUTIVIDADE AGRÍCOLA COM DADOS AGROMETEOROLÓGICOS USANDO MACHINE LEARNING

AUTORA: KAMILA CUNHA DE MENESES
ORIENTADOR: GLAUCO DE SOUZA ROLIM
COORIENTADOR: NEWTON LA SCALA JUNIOR

Aprovada como parte das exigências para obtenção do Título de Doutora em AGRONOMIA (CIÊNCIA DO SOLO), pela Comissão Examinadora:

Assinado de forma digital por Glauco de Souza Rolim Dados: 2021.11.21 09:21:13 -03'00'

Prof. Dr. GLAUCO DE SOUZA ROLIM (Participaçao Virtual)
Departamento de Engenharia e Ciências Exatas (DECEx) / FCAV / UNESP - Jaboticabal

Assinado de forma digital por Glauco de Souza Rolim Dados: 2021.11.21 09:21:30 -03'00'

Prof. Dr. GENER TADEU PEREIRA (Participaçao Virtual)
Departamento de Engenharia e Ciências Exatas (DECEx) / FCAV / UNESP - Jaboticabal

Prof. Dr. LUCAS EDUARDO DE OLIVEIRA APARECIDO (Participaçao Virtual)
Instituto Federal de Educação, Ciência e Tecnologia do Sul de Minas Gerais - Campus Muzambinho / Muzambinho/MG

Assinado de forma digital por Glauco de Souza Rolim Dados: 2021.11.21 09:21:49 -03'00'

Assinado de forma digital por Glauco de Souza Rolim Dados: 2021.11.21 09:22:06 -03'00'

Profa. Dra. AMANDA LIZ PACÍFICO MANFRIM PERTICARRARI (Participaçao Virtual)
Departamento de Engenharia e Ciências Exatas (DECEx) / FCAV / UNESP - Jaboticabal

Assinado de forma digital por Glauco de Souza Rolim Dados: 2021.11.21 09:22:20 -03'00'

Prof.ª Dr.ª MARYZÉLIA FURTADO DE FARIAS (Participaçao Virtual)
Universidade Federal do Maranhão-UFMA / São Luís/MA

Jaboticabal, 12 de novembro de 2021

DADOS CURRICULARES DA AUTORA

Kamila Cunha de Meneses – Filha de Francisco Feitosa de Meneses e Maria dos Milagres dos Santos Cunha, nasceu em Alcântara, Maranhão, no dia 06 de novembro de 1991. Cursou Agronomia pela Universidade Federal do Maranhão – Câmpus IV, de Chapadinha – MA, de 2009 a 2013. Durante a graduação, foi bolsista de extensão da PROEX-UFMA, no período de 2011 a 2013. Foi voluntária de projetos de iniciação científica e monitora das disciplinas Hidráulica Agrícola e Irrigação e Drenagem. Em fevereiro de 2018, obteve o título de Mestra em Agronomia pelo Programa de Pós-graduação em Agronomia (Ciência do Solo), na Universidade Estadual Paulista "Júlio de Mesquita Filho" – Unesp, Câmpus de Jaboticabal, sob orientação do Prof. Dr. Glauco de Souza Rolim. Em março de 2018, iniciou o curso de Doutorado pelo mesmo programa de Pós-graduação na Unesp/FCAV, desenvolvendo pesquisas sobre modelagem agrícola, sensoriamento remoto e agrometeorologia. Durante o doutorado também atuou como docente da disciplina de Processamento de Dados no curso de graduação de Engenharia Agronômica da Unesp/FCAV pelo Programa Institucional de Aperfeiçoamento e Apoio à Docência no Ensino Superior – PAADES, coorientou alunos da Universidade Federal do Maranhão - UFMA, participou de diversas bancas de defesa de trabalho de conclusão de curso e publicou artigos em periódicos de alto impacto. É Integrante dos grupos de pesquisa: i) "Group of Agrometeorological Studies" (GAS); ii) Emissão de CO_2 do Solo e balanço de gases de efeito estufa em sistemas agrícolas, iii) Caracterização do Solo para Fins de Manejo Específico (CSME) da Unesp – Câmpus de Jaboticabal e iv) Manejo Sustentável de Sistemas Agropecuários da UFMA – Câmpus IV. Atualmente, é professora substituta das disciplinas de solos, economia rural e administração rural da Universidade Federal do Maranhão e cofundadora da empresa AGRODIMENSÃO. Em 12 de novembro de 2021, submeteu-se à banca para a defesa de Tese, sendo aprovada como Doutora em Agronomia.

"Se fiz descobertas valiosas, foi mais por ter paciência do que qualquer outro talento."

Isaac Newton

"A ciência não conhece os países, porque o conhecimento pertence à humanidade e é a tocha que ilumina o mundo. A ciência é a alma da prosperidade das nações e a fonte de todo progresso"

Louis Pasteur

Dedico ao meu colega e amigo Washigton Bruno Silva Pereira (*in memoriam*) que planejava defender sua dissertação no mesmo dia da minha defesa, mas teve sua vida ceifada pela COVID-19.

AGRADECIMENTOS

A maior parte do meu doutorado foi no período da pandemia COVID-19, na qual passamos por vários sentimentos e incertezas, mas a ciência acendeu uma esperança de dias melhores com o surgimento da vacina. Viva a ciência e o SUS!

Agradeço a Deus por ter mostrado em todos os momentos que a fé e sua presença foram essenciais para contemplar essa fase tão importante na minha vida. Toda honra e glória a ti!

A minha família que estive ao meu lado durante todo este tempo e mesmo de longe sempre conseguiu me dar amor e apoio. Tudo por vocês!

Aos meus amigos que deixaram esse momento mais alegre e divertido.

Ao Departamento de Ciências Exatas e Engenharia pela oportunidade de conviver com pessoas que muito me ensinaram e contribuíram para realização deste trabalho. Conviver com estas pessoas todos os dias é uma alegria muito grande.

Ao meu orientador Glauco de Souza Rolim e coorientador Newton La Scala Junior pelos ensinamentos, paciência, oportunidades e exemplo de profissionalismo. Gratidão por tudo!

Professores da banca de qualificação: professora Dra. Teresa Cristina Tarle Pissarra e professor Dr. Lucas Eduardo de Oliveira Aparecido, pelas sugestões e atenção.

Aos membros da banca de defesa: professor Dr. Gener Tadeu Pereira, professor Dr. Lucas Eduardo de Oliveira Aparecido, professora Dra. Maryzélia Furtado de Farias e professora Dra. Amanda Liz Pacífico Manfrim Perticarrari pelas sugestões que contribuíram na conclusão deste estudo.

À Universidade Estadual Paulista “Júlio de Mesquita Filho”, campus Jaboticabal, especialmente a Pós-Graduação em Agronomia (Ciência do Solo), pela oportunidade oferecida.

O presente trabalho foi realizado com apoio da Coordenação de Aperfeiçoamento de Pessoal de Nível Superior - Brasil (CAPES) - Código de Financiamento 001.

A todos que contribuíram diretamente e indiretamente na realização deste trabalho.

Muito obrigada!

SUMÁRIO

RUMO À AGRICULTURA INTELIGENTE: PREVISÃO DE PRODUTIVIDADE AGRÍCOLA COM DADOS AGROMETEOROLÓGICOS USANDO MACHINE LEARNING

RESUMO – Sistemas agrícolas baseados em tecnologias digitais podem contribuir para um aumento da segurança alimentar global como parte dos esforços de mitigação e adaptação às mudanças climáticas. Um modelo acurado para previsão da produtividade beneficia muitos aspectos do gerenciamento de áreas produtivas. Portanto, nossa hipótese é a possibilidade do uso de algoritmos de machine learning para a previsão da produtividade em regiões do Brasil. Neste contexto, foram utilizadas séries históricas de produtividade de dois principais cultivos do Brasil. No primeiro trabalho, o objetivo foi prever a produtividade da cana-de-açúcar com seis meses de antecedência da colheita com acurácia em várias regiões produtoras do Brasil, utilizando modelos de machine learning. No segundo trabalho, o objetivo foi prever a produtividade do algodão usando algoritmos de machine learning baseado em elementos climáticos. Para cana-de-açúcar, nós utilizamos dados de produtividade da cana-de-açúcar de 62 localidades do Brasil. Foram utilizados também dados meteorológicos diários de temperatura média do ar, temperatura mínima, temperatura máxima, precipitação, velocidade do vento a 2 metros de altura, umidade relativa, irradiância solar no topo da atmosfera, irradiância solar global coletados na plataforma NASA/POWER. Foi utilizado um método de modelagem tradicional de regressão linear múltipla (RLM) e seis métodos de aprendizado de máquina (ML): support vector machine (SVM), random forest approach (RF), Artificial neural network (ANN), Least Absolute Shrinkage and Selection Operator (LASSO) regression, RIDGE regression e eXtreme Gradient Boosting (XGBoost) para prever a produtividade da cana-de-açúcar com seis meses de antecedência. Para o algodão, foi realizada a previsão da produtividade do algodão em função dos elementos climáticos por meio de algoritmos de aprendizado de máquina com quatro parâmetros ajustados por mínimos quadrados ordinários. Para cana-de-açúcar foram separados 4 grupos de localidades conforme a produtividade a partir de análise de cluster. Nos grupos 1 e 2 ocorrem os maiores valores de deficiência hídrica nas localidades produtoras de cana-de-açúcar. No teste dos modelos de ML, os valores de MAPE foram acima de 20%. No teste dos modelos pelo XGBOOST apresentou MAPEs de 29.71%, 26.79%, 43.5% e 33.36% para os grupos 1, 2, 3 e 4, respectivamente. Os modelos XBOOST e MLP apresentam os melhores desempenhos para a previsão de produtividade da cana-de-açúcar. Os modelos mostram que a produtividade do algodão apresenta tendência sigmóide devido ao acúmulo de precipitação, evapotranspiração potencial, armazenamento de água no solo e excedente hídrico durante o ciclo. É possível prever a produtividade do algodão para as principais regiões produtoras do Brasil usando algoritmos de aprendizado de máquina. Modelos de regressores Extra-trees tiveram melhor desempenho na previsão da produtividade do algodão usando dados climáticos do plantio à floração.

Palavras-chaves: análise de riscos, agrometeorologia, inteligência artificial

TOWARDS SMART AGRICULTURE: FORECASTING AGRICULTURAL PRODUCTIVITY WITH AGROMETEOROLOGICAL DATA USING MACHINE LEARNING

ABSTRACT – Agricultural systems based on digital technologies can contribute to an increase in global food security as part of climate change mitigation and adaptation efforts. An accurate model for forecasting productivity benefits many aspects of productive area management. Therefore, our hypothesis is the possibility of using machine learning algorithms to predict productivity in regions of Brazil. In this context, historical series of productivity of two main crops in Brazil were used. In the first work, the objective was to accurately predict the productivity of sugarcane six months before harvesting in several producing regions in Brazil, using machine learning models. In the second work, the objective was to predict cotton productivity using machine learning algorithms based on climatic elements. For sugarcane, we used sugarcane productivity data from 62 locations in Brazil. Daily meteorological data of mean air temperature, minimum temperature, maximum temperature, precipitation, wind speed at 2 meters high, relative humidity, solar irradiance at the top of the atmosphere, global solar irradiance collected on the NASA/POWER platform were also used. A traditional multiple linear regression (RLM) modeling method and six machine learning (ML) methods were used: support vector machine (SVM), random forest approach (RF), Artificial neural network (ANN), Least Absolute Shrinkage and Selection Operator (LASSO) regression, RIDGE regression and eXtreme Gradient Boosting (XGBoost) to predict sugarcane productivity six months in advance. For cotton, cotton yield prediction was performed as a function of climatic elements through machine learning algorithms with four parameters adjusted by ordinary least squares. For sugarcane, 4 groups of locations were separated according to productivity from cluster analysis. In groups 1 and 2, the highest values of water deficit occur in sugarcane-producing localities. In the test of the ML models, the MAPE values were above 20%. In the test of the models by XGBOOST they presented MAPEs of 29.71%, 26.79%, 43.5% and 33.36% for groups 1, 2, 3 and 4, respectively. The XBOOST and MLP models present the best performances for predicting sugarcane productivity. The models show that cotton yield has a sigmoid trend due to the accumulation of P, PET, STO and EXC throughout the cycle. It is possible to predict cotton productivity for the main producing regions in Brazil using machine learning algorithms. Extra-trees regressor models performed better in predicting cotton yield using climatic data from planting to flowering. With this, it is possible to have an average anticipation of around 80 days, allowing the producer time to plan their activities such as harvesting and sales strategies.

Keywords: risk analysis, agrometeorology, artificial intelligence

CAPÍTULO 1 – CONSIDERAÇÕES GERAIS

1 INTRODUÇÃO E JUSTIFICATIVA

1.1 Agricultura inteligente para o clima

A expansão de área e a intensificação rápida do uso do solo a partir de 1961 contribuíram para o aumento da produção total de alimentos em 240%, devido ao aumento da produtividade e da área de uso do solo. No entanto, a produção de alimentos deve dobrar até 2050 para alimentar a população mundial que deverá chegar a 9,7 bilhões, gerando um aumento de até 70% na demanda por alimentos, sem considerar a complexidade da mitigação das mudanças climáticas (FAO, 2013).

A temperatura média do ar na superfície terrestre aumentou 1,53 °C e a temperatura média da superfície 0,87 °C no período de 1850 a 2015. Em 2080, a produtividade agrícola global diminuirá de 3 a 16% (FAO, 2011), pois os sistemas agrícolas são altamente sensíveis às condições climáticas voláteis. Para que ocorra a manutenção e o fortalecimento da segurança alimentar, os sistemas de produção precisam se tornar mais robustos, para serem capazes de ter um bom desempenho em face de condições vitais de estresses e acidentes agrícolas. Além disso, a mudança nesses sistemas também pode levar ao aumento dos sumidouros de carbono, a benefícios significativos de mitigação e redução nas emissões por unidade de produto agrícola (Azadi et al., 2011).

A agricultura inteligente para o clima (Climate-smart Agriculture, em inglês) foi desenvolvida pela Organização para Agricultura e Alimentação (FAO) como uma abordagem unificada para enfrentar os desafios das mudanças climáticas. O conceito de Climate-smart Agriculture (CSA) foi lançado pela FAO em 2010 em um documento de referência preparado para a Conferência de Haia sobre Agricultura, Segurança Alimentar e Mudança Climática (FAO, 2019). A CSA é definida como uma abordagem que visa transformar, reorientar e desenvolver sistemas agrícolas baseados em tecnologias digitais, com o objetivo de contribuir para um aumento na segurança alimentar global como parte dos esforços de mitigação e adaptação às mudanças climáticas (Zecca, 2019).

Um pré-requisito essencial da agricultura inteligente é, definitivamente, a adoção da Tecnologia da Informação e Comunicação (TIC), que é promovida por formuladores de políticas em todo o mundo (Sørensen et al., 2019). A TIC pode incluir, de forma indicativa, sistemas de informação de gestão agrícola, sensores de umidade e solo, acelerômetros, redes de sensores sem fio, câmeras, drones, satélites de baixo custo, serviços online e veículos guiados automatizados (Benos et al., 2021).

Os produtores inovadores se esforçam para reduzir a quebra da produtividade entre a potencial e real usando tecnologias avançadas, uma vez que, a produção agrícola é dependente dos fatores ambientais como radiação solar, água, temperatura, manejo cultural e tipo de solo. Assim, compreender o impacto desses fatores é essencial para quantificar a causa e a magnitude da variação, principalmente em sistemas de cultivos de sequeiro (Al-Shammari et al., 2021).

1.2 Machine learning no sistema agrícola

A produtividade dos cultivos será aumentada com a incorporação de novas tecnologias. Na agricultura, as técnicas de aprendizado de máquina (ML) são consideradas a melhor escolha, pois conseguem prever a produção da safra do ano futuro (Najeeb e Kamalakkannan, 2022). ML surgiu junto com tecnologias de big data e computação de alto desempenho para criar oportunidades para desvendar, quantificar e compreender processos intensivos de dados em ambientes operacionais agrícolas (Liakos et al., 2018).

A definição de aprendizado de máquina é o campo científico que dá às máquinas a capacidade de aprender sem serem estritamente programadas, permitindo que grandes problemas não lineares sejam resolvidos de forma autônoma (Chlingaryan et al., 2018), por exemplo, bioinformática, bioquímica, medicina, meteorologia, ciências econômicas, robótica, aquicultura, segurança alimentar e climatologia.

As regressões lineares não conseguem capturar normalmente as interações complexas entre os fatores climáticos e a produtividade, portanto, os

modelos de aprendizado de máquina vêm demonstrando um desempenho poderoso em vários aplicativos orientados a dados, incluindo na área da agricultura, como é o caso da previsão de cultivos. Há diversos algoritmos de machine learning disponível para regressão, por exemplo, random forest, support vector machine e redes neurais. Além desses, há métodos de aprendizagem profunda que empregam várias camadas de computação em redes neurais que podem ser usados para explorar informações heterogêneas (ou seja, sensoriamento remoto e dados meteorológicos) e encontrar relações não lineares complexas com a produtividade de cultivo (Meroni et al., 2021). A maioria dos modelos tradicionais faz previsões da produtividade atual, mas à medida que a tecnologia avança, os modelos atuais de produtividade não serão confiáveis porque o potencial de produtividade continuará a mudar (Roell et al., 2020).

Tian et al. (2018) observaram que o modelo de support vector machine (SVR) com índices climáticos pode melhorar a precisão da previsão da seca agrícola em comparação com aquele que usa apenas o índice de seca. O modelo SVR mostrou-se eficaz e flexível na previsão de índices de seca. A vantagem do SVR é que o modelo pode transferir um problema não linear para um problema linear usando a função kernel e ser eficaz na resolução de um problema de dimensão elevada. Jian et al. (2018) concluíram que o algoritmo de regressão SVM tem grande potencial para estimar a salinidade do solo usando dados de sensoriamento remoto de múltiplas fontes.

De acordo com Roell et al. (2020), o uso de algoritmos de aprendizado de máquina com adição dados climáticos foram os principais aspectos que ajudaram a aumentar a variação explicada nas previsões de produção do trigo em comparação com o modelo estatístico realizado anteriormente.

Os méritos significativos do acoplamento de ML e modelos de simulação de cultura mostrados por Shahhosseini et al. (2021), fizeram os autores levantar a questão que os modelos de ML podem se beneficiar ainda mais com a adição de mais recursos de entrada de outras fontes. Portanto, uma possível extensão do seu estudo poderia ser a inclusão de dados de sensoriamento remoto na tarefa de previsão de ML e investigar o nível de importância que cada fonte de dados pode exibir.

Oliveira et al. (2021) avaliaram a possibilidade de aplicação de agrotóxicos em taxa variável, utilizando o princípio do volume por fileiras de árvores e os autores observaram que, por meio de sensoriamento remoto e redes neurais artificiais (MLP), é possível estimar o volume do cafeeiro com razoável precisão. Isso pode ser feito usando um modelo perceptron multicamadas para estimar a altura e o diâmetro do cafeeiro usando os índices de vegetação de diferentes partes da planta como dados de entrada.

Aparecido et al. (2019) observaram que random forest (RFT) foi mais preciso na previsão da ferrugem do café, cercospora, bicho-mineiro e broca-do-café utilizando dados climáticos. Além disso, o RFT apresentou maior acurácia nas previsões para o Cerrado Mineiro em anos de alta e baixa produtividade e para todas as doenças. Enquanto, os valores RMSE variaram de 0,227 a 0,853 para café de alta produtividade e 0,147 e 0,827 para café de baixa produtividade na previsão da broca do café.

Embora os métodos de regressão utilizados no trabalho de Shafiee et al. (2021) tenham mostrado uma boa capacidade de previsão da produtividade de grãos, os autores observaram que o regressor LASSO provou ser mais acessível e econômico em termos de tempo.

No zoneamento agrícola de risco climático para o girassol em diferentes épocas de semeadura, Aparecido et al. (2019) concluíram que a rede neural é uma ferramenta eficiente e pode ser usada na espacialização de variáveis climáticas de forma rápida e precisa. Os autores encontraram que a semeadura de girassol na primavera e no verão são as que proporcionam as maiores áreas aptas no sudeste do Brasil, com 58,1 e 64,46%, respectivamente.

Kernel Ridge Regression (KRR) são considerados muito proeminentes e têm amplas aplicações em previsões. KRR é uma abordagem de regressão não linear onde uma função de kernel não linear é aplicada no espaço original para definir um produto interno em um espaço transformado de dimensão superior para fornecer desempenho de generalização com base na solução de mínimos quadrados de regularização (Naik et al., 2018).

XGBoost é um método baseado em árvore de boost que, por sua vez, é baseado em árvores de decisão. Considerando que a combinação linear para múltiplas árvores capacidades que podem ajustar bem os dados de

treinamento e descrever a complexa relação não linear entre os dados de entrada e saída, torna este método considerado um dos melhores métodos de aprendizagem estatística (Cardoso et al., 2020).

1.3 Fatores meteorológicos importantes nos modelos

Os sistemas agrícolas são altamente dependentes das condições climáticas. A variabilidade climática representa um fator de risco para à produção agrícola em muitas regiões do mundo (IPCC, 2019). O aumento da temperatura do ar, a mudança nos padrões de precipitação e a maior frequência de secas que induzem o estresse hídrico da cultura, bem como as mudanças no estresse de geada, calor ou frio, afetam o crescimento, o desenvolvimento e a produtividade da cultura (Žydelis et al., 2021). As variáveis climáticas podem explicar entre 20 e 49% da variabilidade de diferentes anomalias de produtividade dos cultivos em escala global, em que a força do impacto da mudança climática na produção agrícola varia entre regiões e culturas (Vogel et al., 2019).

Conjuntos de dados meteorológicos em grid desempenham um papel importante no monitoramento de secas, modelagem agrometeorológica e gestão de recursos hídricos (Wang et al., 2020). A temperatura do ar é uma variável meteorológica importante para compreensão da física de muitos processos da superfície terrestre. É comumente utilizada para monitorar o estresse hídrico da vegetação, avaliar o balanço de energia de superfície, estimar a precipitação e a produtividade dos cultivos e derivar a evaporação diária e a distribuição da umidade do solo em macroescala (Hadria et al 2018). Os três componentes da temperatura do ar comumente usados no ciclo de energia e água do sistema terra-atmosfera são máxima, mínima e média.

Xiangyu e Vico (2021) destacaram que as precipitações menos frequentes e mais intensas causaram teores de água no solo mais variáveis, levando a temperaturas do dossel mais altas e mais variáveis, e uma fração maior de dias em que o limite de temperatura para dano potencial por estresse térmico foi excedido.

Rolim et al. (2020) observaram que as faixas de valores mensais para cada elemento meteorológico indicaram que temperaturas mais altas, menor

precipitação, menor armazenamento de água no solo e maiores déficits hídricos são necessários para a produção de café de alta qualidade. Os melhores modelos agrometeorológicos desenvolvidos pelos autores foram com a precipitação, a evapotranspiração real e o déficit hídrico, armazenamento de água distribuídos conforme a região estudada.

Já as variáveis meteorológicas mais importantes para os modelos agrometeorológicos para previsão da produção de açaí (*Euterpe oleracea* Mart.) foram temperatura do ar, radiação solar e déficit de pressão de vapor para sistema irrigado e para de sequeiro, o estresse hídrico teve o maior efeito (Moraes et al., 2020).

Vários estudos têm mostrado que o efeito da chuva na produtividade depende de vários fatores, relacionados à sua quantidade total e distribuição dentro da estação de crescimento, características do solo, em particular as condições iniciais de água do solo, reserva do solo e profundidade de enraizamento do solo (Schlenker e Lobell, 2010; Ray et al., 2015; Lobell e Asseng, 2017a,b).

As chuvas são de grande importância para as lavouras, pois disponibilizam água para o solo. A água atua em diversos processos do metabolismo vegetal (Sanches et al., 2019). O estresse hídrico induz o fechamento estomático, reduz a taxa fotossintética e acelera a senescência (Silva et al. 2009), assim sendo um fator limitante para várias culturas (Taiz et al. 2017; Anda et al. 2020). Duarte e Sentelhas (2019a) usaram o déficit hídrico gerado pelo método de Thornthwaite e Mather (1955) como uma variável para prever a produtividade de milho atingível.

O balanço hídrico climatológico é a contabilização da entrada e saída de água do solo, determinando períodos de deficiência e excedente hídrico (Aparecido et al., 2020). O balanço hídrico climatológico é uma importante ferramenta para o planejamento agrícola, pois, é possível, classificar o clima, realizar zoneamento agroclimático e manejo adequado da irrigação, além de identificar melhores datas de plantio (Singh et al., 2019). Silveira et al. (2020) utilizaram o balanço hídrico de Thornthwaite e Mather (1955) para estudar a eficiência do uso da água na laranja Pêra (*Citrus sinensis L. Osbeck*) na região sudoeste do estado de São Paulo.

Com base em nosso conhecimento atual, os efeitos sobre a produtividade das safras resultantes das variações das condições climáticas são complexos e difíceis de prever, especialmente por meio da combinação de efeitos e interações com o ambiente determinado. Por outro lado, a previsão da produtividade de cultivo com base em séries temporais meteorológicas e de produtividade de longo prazo utilizando modelos de machine learning têm potencial, pois as tendências nas safras futuras são determinadas pelas anteriores.

1.3 OBJETIVO GERAL

Nosso principal objetivo foi prever a produtividade de cultivos agrícolas com antecedência à colheita em regiões do Brasil, utilizando modelos de machine learning.

REFERÊNCIAS

Ahmed, G. Najeeb; Kamalakkannan, S. Developing an IoT-Based Data Analytics System for Predicting Soil Nutrient Degradation Level. In: **Expert Clouds and Applications**. Springer, Singapore, 2022. p. 125-137.

Al-Shammari, D.; Whelan, B. M.; Wang, C.; Bramley, R. G.; Fajardo, M.; Bishop, T. F.. Impact of spatial resolution on the quality of crop yield predictions for site-specific crop management. **Agricultural and Forest Meteorology**, v. 310, p. 108622, 2021.

Aparecido, L. E.O.; Lorençone, P. A.; Lorençone, J. A.; Meneses, K. C.; Moraes, J. R. S. C.. Climate changes and their influences in water balance of Pantanal biome. **Theoretical and Applied Climatology**, v. 143, n. 1, p. 659-674, 2021.

Azadi, H.; Moghaddam, S. M.; Burkart, S.; Mahmoudi, H.; Van Passel, S.; Kurban, A.; Lopez-Carr, D. Rethinking resilient agriculture: From Climate-Smart Agriculture to Vulnerable-Smart Agriculture. **Journal of Cleaner Production**, p. 128602, 2021.

Cardoso, J.; Glória, André; Sebastião, Pedro. Improve Irrigation Timing Decision for Agriculture using Real Time Data and Machine Learning. In: **2020 International Conference on Data Analytics for Business and Industry: Way Towards a Sustainable Economy (ICDABI)**. IEEE, 2020. p. 1-5.

Chlingaryan, Anna; Sukkarieh, Salah; Whelan, Brett. Machine learning approaches for crop yield prediction and nitrogen status estimation in precision agriculture: A review. **Computers and electronics in agriculture**, v. 151, p. 61-69, 2018.

Moraes, J. R. S. C.; Rolim, G. S.; Martorano, L. G.; Aparecido, L. E. O.; Padilha, M. D. S.O.; Farias Neto, J. T. Agrometeorological models to forecast açaí (Euterpe oleracea Mart.) yield in the Eastern Amazon. **Journal of the Science of Food and Agriculture**, v. 100, n. 4, p. 1558-1569, 2020.

FAO (2019). Food and agriculture organization, climate-smart agriculture history. http://www.fao.org/climate-smart-agriculture/overview/faqs/history/en.

FAO, 2011. Energy-Smart Food for People and Climate Food and Agriculture Organization of the United Nations (ONU), Rome, Italy (2011), p. 66. Disponível em: http://www.fao.org/docrep/014/i2454e/i2454e00.pdf.

FAO, 2013. Climate smart agriculture source book. Disponível em: http://www.fao.org/docrep/018/i3325e/i3325e.pdf (2013), Accesso em 23 Aug 2021.

Hadria, R.; Benabdelouahab, T.; Mahyou, H.; Balaghi, R.; Bydekerke, L.; El Hairech, T.; Ceccato, P. Relationships between the three components of air temperature and remotely sensed land surface temperature of agricultural areas in Morocco. **International Journal of Remote Sensing**, v. 39, n. 2, p. 356-373, 2018.

IPCC, 2019: Summary for policymakers. In: Climate Change and Land: An IPCC Special Report on Climate Change, Desertification, Land Degradation, Sustainable Land Management, Food Security, and Greenhouse Gas Fluxes in Terrestrial Ecosystems [P.R. Shukla, J. Skea, E. Calvo Buendia, V. Masson-Delmotte, H.-O. Pörtner, D. C. Roberts, P. Zhai, R. Slade, S. Connors, R. van Diemen, M. Ferrat, E. Haughey, S. Luz, S. Neogi, M. Pathak, J. Petzold, J. Portugal Pereira, P. Vyas, E. Huntley, K. Kissick, M. Belkacemi, J. Malley, (eds.)]. In press.

Jiang, H.; Rusuli, Y.; Amuti, T.; He, Q.. Quantitative assessment of soil salinity using multi-source remote sensing data based on the support vector machine and artificial neural network. **International journal of remote sensing**, v. 40, n. 1, p. 284-306, 2019.

Liakos, K. G.; Busato, P.; Moshou, D.; Pearson, S.; Bochtis, D.. Machine learning in agriculture: A review. **Sensors**, v. 18, n. 8, p. 2674, 2018.

Lobell, D. B.; Asseng, S. Comparing estimates of climate change impacts from process-based and statistical crop models. **Environmental Research Letters**, v. 12, n. 1, p. 015001, 2017.

Naik, J.; Bisoi, R.; Dash, P. K. Prediction interval forecasting of wind speed and wind power using modes decomposition based low rank multi-kernel ridge regression. **Renewable energy**, v. 129, p. 357-383, 2018.

Ray, D. K.; Gerber, J. S.; MacDonald, G. K.; West, P. C. . Climate variation explains a third of global crop yield variability. **Nature communications**, v. 6, n. 1, p. 1-9, 2015.

Roell, Y. E.; Beucher, A.; Møller, P. G.; Greve, M. B.; Greve, M. H.. Comparing a Random Forest based prediction of winter wheat yield to historical yield potential. **Agronomy**, v. 10, n. 3, p. 395, 2020.

Schlenker, W.; Lobell, D. B. Robust negative impacts of climate change on African agriculture. **Environmental Research Letters**, v. 5, n. 1, p. 014010, 2010.

Shafiee, S.; Lied, L. M.; Burud, I.; Dieseth, J. A.; Alsheikh, M.; Lillemo, M. Sequential forward selection and support vector regression in comparison to LASSO regression for spring wheat yield prediction based on UAV imagery. **Computers and Electronics in Agriculture**, v. 183, p. 106036, 2021.

Tian, Y.; Xu, Y.; Wang, G.. Agricultural drought prediction using climate indices based on Support Vector Regression in Xiangjiang River basin. **Science of the Total Environment**, v. 622, p. 710-720, 2018.

Vogel, E.; Donat, M. G.; Alexander, L. V.; Meinshausen, M.; Ray, D. K.; Karoly, D.; Meinshausen, N.; Frieler, K. The effects of climate extremes on global agricultural yields. **Environmental Research Letters**, v. 14, n. 5, p. 054010, 2019.

Wang, Q.; Li, W.; Xiao, C.; Ai, W.. Evaluation of High-Resolution Crop Model Meteorological Forcing Datasets at Regional Scale: Air Temperature and Precipitation over Major Land Areas of China. **Atmosphere**, v. 11, n. 9, p. 1011, 2020.

Žydelis, R.; Weihermüller, L.; Herbst, M. Future climate change will accelerate maize phenological development and increase yield in the Nemoral climate. **Science of The Total Environment**, v. 784, p. 147175, 2021.

CAPÍTULO 2 – PREVISÃO DA PRODUTIVIDADE DE CANA-DE-AÇÚCAR NO BRASIL A PARTIR DE MODELOS DE MACHINE LEARNING

RESUMO – A previsão da produtividade da cana-de-açúcar com seis meses de antecedência à colheita fornece informações significativas aos gerentes, autoridades e produtores para decisões estratégicas e tomada de decisões assertivas durante o período de cultivo. Portanto, nosso principal objetivo foi prever a produtividade da cana-de-açúcar com seis meses de antecedência da colheita com acurácia em várias regiões produtoras do Brasil, utilizando modelos de machine learning. Neste estudo, nós utilizamos dados de produtividade da cana-de-açúcar de 62 localidades do Brasil. Para cada local, os dados de produtividade foram coletados no sistema IBGE-SIDRA, durante o período de 1985 a 2020. Foram utilizados também dados meteorológicos diários de temperatura média do ar, temperatura mínima, temperatura máxima, precipitação, velocidade do vento a 2 metros de altura, umidade relativa, irradiância solar no topo da atmosfera, irradiância solar global coletados na plataforma NASA/POWER. Nós utilizamos um método de modelagem tradicional de regressão linear múltipla (RLM) e seis métodos de aprendizado de máquina (ML): support vector machine (SVM), random forest approach (RF), Artificial neural network (ANN), Least Absolute Shrinkage and Selection Operator (LASSO) regression, RIDGE regression e eXtreme Gradient Boosting (XGBoost) para prever a produtividade da cana-de-açúcar com seis meses de antecedência. As otimizações dos parâmetros para os algoritmos foram ajustadas usando GridSearchCV no Scikit-Learn em linguagem python. As localidades foram separadas em 4 grupos com climas semelhantes por análise de cluster. A produtividade dos grupos 1, 2, 3 e 4 das localidades produtoras de cana-de-açúcar foi de 65.689 kg ha^{-1}, 66.050 kg ha^{-1}, 61.667 kg ha^{-1} e 60.930 kg ha^{-1}, respectivamente. Nos grupos 1 e 2 ocorrem os maiores valores de deficiência hídrica nas localidades produtoras de cana-de-açúcar. No teste dos modelos de ML, os valores de MAPE foram acima de 20%. No teste dos modelos pelo XGBOOST apresentou MAPEs de 29.71%, 26.79%, 43.5% e 33.36% para os grupos 1, 2, 3 e 4, respectivamente. O RLM apresentou melhor desempenho no teste nos grupos. Para o grupo 2, nós observamos que MLP apresentou maior MAPE (113.92%) no teste. E finalmente os modelos XBOOST e MLP apresentam os melhores desempenhos para a previsão de produtividade da cana-de-açúcar.

Palavras-chaves: *Saccharum officinarum*; modelos agrometeorológicos, RIDGE, LASSO, XBGBOOST

PREDICTION OF SUGARCANE YIELD IN BRAZIL FROM MACHINE LEARNING MODELS

ABSTRACT - The prediction of sugarcane yield six months before harvesting provides significant information to managers, authorities and producers for strategic decisions and assertive decision-making during the cultivation period. Therefore, our main objective was to accurately predict the sugarcane yield six months in advance of harvesting in several producing regions in Brazil, using machine learning models. In this study, we used sugarcane yield data from 62 locations in Brazil. For each location, yield data were collected in the IBGE-SIDRA system, during the period from 1985 to 2020. Daily meteorological data of average air temperature, minimum temperature, maximum temperature, precipitation, wind speed at 2 meters were also used height, relative humidity, top atmospheric solar irradiance, global solar irradiance collected on NASA/POWER platform. We use a traditional multiple linear regression (RLM) modeling method and six machine learning (ML) methods: support vector machine (SVM), random forest approach (RF), Artificial neural network (ANN), Least Absolute Shrinkage and Selection Operator (LASSO) regression, RIDGE regression and eXtreme Gradient Boosting (XGBoost) to predict sugarcane yield six months in advance. The parameter optimizations for the algorithms were adjusted using GridSearchCV in Scikit-Learn in Python language. The locations were separated into 4 groups with similar climates by cluster analysis. The yield of groups 1, 2, 3 and 4 of the sugarcane producing localities was 65,689 kg ha^{-1}, 66,050 kg ha^{-1}, 61,667 kg ha^{-1} and 60,930 kg ha^{-1}, respectively. In groups 1 and 2, the highest values of water deficit occur in sugarcane-producing localities. In the test of the ML models, the MAPE values were above 20%. In the model test by XGBOOST, MAPEs were 29.71%, 26.79%, 43.5% and 33.36% for groups 1, 2, 3 and 4, respectively. The RLM performed better in the test in the groups. For group 2, we observed that MLP had higher MAPE (113.92%) in the test. Finally, the XBOOST and MLP models present the best performances for predicting sugarcane yield.

Keywords: *Saccharum officinarum*; agrometeorological models, RIDGE, LASSO, XBGBOOST

2.1 INTRODUÇÃO

A cana-de-açúcar é produzida em 100 países, com mais de 1,91 bilhões de toneladas comercializadas em 2018 (FAOSTAT, 2020) em 26,3 milhões de ha. O Brasil é o maior produtor de cana-de-açúcar com área colhida de 8,6 Mha e produção de 620 Mt (FAOSTAT, 2020). Os maiores estados produtores são São Paulo, Goiás e Minas Gerais (CONAB, 2020). A produção de cana-de-açúcar concentra-se 90% na região centro-sul do Brasil. A expansão da cana-de-açúcar no território brasileiro aumentou desde 2014/2015 devido ao estímulo do governo nacional ao uso de biocombustível (Vera *et al.*, 2020; Zilli *et al.*, 2020), principalmente em áreas com solos arenosos, altas temperaturas do ar e precipitação pluviométrica irregular (Guo *et al.*, 2021; Walter et al., 2014; Canisares et al., 2020).

O clima e o solo afetam a produtividade da cana-de-açúcar variando bastante entre as regiões, pois causam mudanças em suas características bioquímicas, fisiológicas e morfológicas, resultando em diferentes taxas de crescimento (Paixão et al., 2020). A previsão da produtividade da cana-de-açúcar em áreas maiores é um desafio devido à relação entre as variáveis climáticas, a fenologia e suas distribuições espaciais incertas ligadas à logística (Yu et al., 2020). Portanto, é essencial melhorar a acurácia das previsões da produtividade da cana-de-açúcar para enfrentar o desafio das mudanças climáticas globais, garantir a segurança alimentar e maior sustentabilidade ecológica (Guo et al., 2021, Webber et al., 2020).

A integração de dados espacializados, análises estatísticas e modelos de cultivos vem sendo bastante usados para interpretar a heterogeneidade espacial e melhorar a acurácia na determinação da biomassa (Shawon et al., 2020),

estimativas (Singla et al., 2020; Yu et al., 2020; Chen e Tao, 2020; Ashapure et al., 2020) e previsões de produtividade da cana-de-açúcar (Feng et al., 2020; Tedesco-Oliveira et al., 2020; Shendryk et al., 2021) são encontradas na literatura.

A previsão da produtividade da cana-de-açúcar com seis meses de antecedência à colheita fornece informações significativas aos gerentes, autoridades e produtores para uma tomada de decisão rápida durante período de cultivo.

As previsões de produtividades também são úteis em relação ao comércio, políticas de desenvolvimento e assistência humanitária ligada à segurança alimentar (JRC, 2018). No entanto, os parâmetros de entrada, calibração, procedimentos de avaliação e validação e os métodos de simulação de respostas da cultura a vários fatores ambientais, por exemplo, sob mudanças climáticas extremas, como calor e frio extremos, são limitações inerentes aos modelos baseados em processos, assim levando a incertezas na previsão das produtividades dos cultivos e identificação de medidas adequadas para a adaptação (Webber et al., 2020). Já as limitações dos modelos estatísticos são a quantidade e qualidade dos dados de entrada, e podem gerar grandes incertezas (Roberts et al., 2017). Os métodos estatísticos tradicionais geralmente consideram as correlações lineares entre a produtividade do cultivo e as variáveis independentes, assim não refletem com acurácia as relações não lineares inerentes ao sistema agrícola (Haqiqi et al., 2019).

Nos últimos anos, os modelos de machine learning (ML) são uma ferramenta poderosa para avaliar os efeitos das variáveis climáticas sobre a produção agrícola (Aparecido et al., 2020; Elavarasan et al., 2018; Schwalbert et al., 2020). Os modelos de ML capturam relações não lineares, lidam com as interações

entre preditores e não assumem nenhum pré-requisito para utilização (Chlingaryan et al., 2018).

Um modelo acurado para prever a produtividade da cana-de-açúcar beneficia muitos aspectos do gerenciamento de áreas produtivas de cana-de-açúcar. Portanto, nossa hipótese é a possibilidade do uso de algoritmos de ML para a previsão da produtividade da cana-de-açúcar em regiões do Brasil. Portanto, nossos principais objetivos foram: 1- prever a produtividade da cana-de-açúcar com seis meses de antecedência da colheita, 2 - avaliar o desempenho de diferentes modelos de machine learning na previsão da produtividade da cana-de-açúcar em diferentes regiões do Brasil.

2.2MATERIAL E MÉTODOS

2..2.1 Área de estudo e dados

Neste estudo, nós utilizamos dados de produtividade da cana-de-açúcar de 62 localidades do Brasil (Figura 1). Para cada local, os dados de produtividade de cana-de-açúcar foram coletados no Sistema IBGE de Recuperação Automática - SIDRA (IBGE, 2020) durante o período de 1985 a 2020.

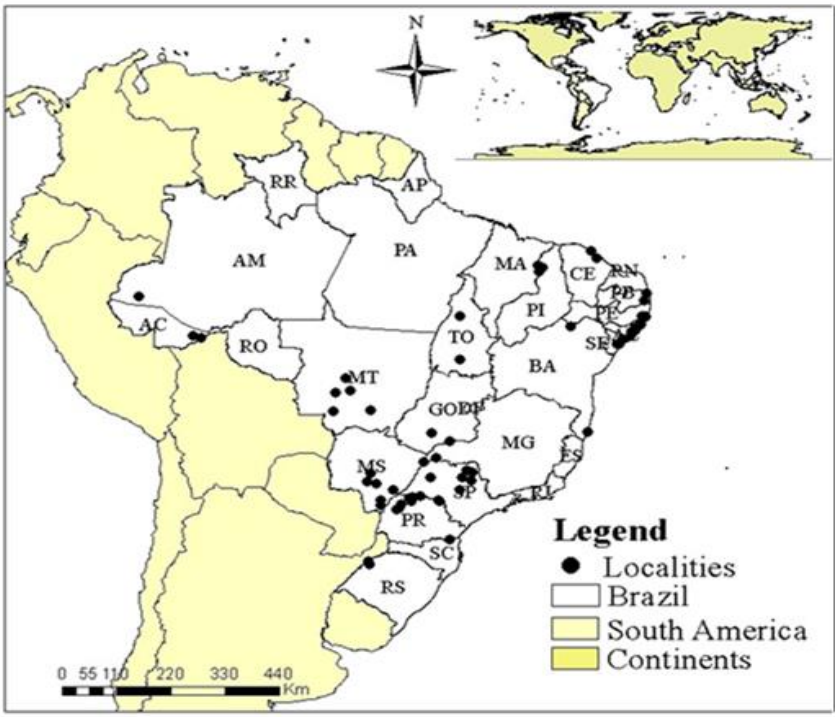

Figura 1. Localização da área de estudo.

Essas localidades foram classificadas em quatro grupos por meio da análise de agrupamento por cluster, levando em consideração os dados climáticos (Figura 1). A matriz de similaridade foi construída com a distância euclidiana, e os grupos foram ligados usando o método de Ward (Hair et al., 2005).

As séries históricas dos dados climáticos diários de 1984 a 2020 das localidades continham temperatura média do ar (Tmean, °C), temperatura mínima (Tmin, °C), temperatura máxima (Tmax, °C), precipitação (P, mm), velocidade do vento a 2 metros de altura (WS2M, m s^{-1}), umidade relativa (%), irradiância solar no topo da atmosfera (MJ m2dia), irradiância solar global (Qg, MJ m^{-2} day^{-1}). Os dados climáticos foram coletados na plataforma National Aeronautics and Space Administration/Prediction of Worldwide Energy Resources NASA/POWER (Sparks, 2018). Esta plataforma de dados foi desenvolvida para fornecer informações meteorológicas derivadas em grades com uma resolução de 0.5°

(latitude-longitude, equivalente a aproximadamente 56 km). Em seguida, nós organizamos os dados climáticos em período de 10 dias.

2.2.2 Cálculo da evapotranspiração potencial

Nós calculamos a evapotranspiração potencial (PET) a partir dos dados climáticos usando o método de Penman-Monteith (PM) (Allen et al., 1998) em escala diária e após estratificado na escala decendial, conforme as Equações 1 a 8.

$$PET = \left(\frac{0.48\times\Delta\times(Rn-G)+\left(\frac{900}{T+273}\right)\times U2(es-e)}{\Delta+\gamma\times(1+0.34\times U2)}\right) \quad (1)$$

$$\Delta = \frac{4098\times es}{(T+273)^2} \quad (2)$$

$$es = 0.6108 \times e^{\frac{17.27\times T}{237.+T}} \quad (3)$$

$$ea = \frac{RH\times es}{100} \quad (4)$$

$$Rn = BOC - BOL \quad (5)$$

$$BOC = (1-\alpha)\times \mathrm{Rs} \quad (6)$$

$$BOL = -\left[4.903\times 10^{-9}\left[\frac{{T_{max}}^4+{T_{min}}^4}{2}\right]\left(0.34-0.14\sqrt{ea}\right)\left(1.35\frac{Rs}{Rso}-0.35\right)\right] \quad (7)$$

$$Rso = (0.75+2\times 10^{-5}\,11)Ra \quad (8)$$

em que: PET – evapotranspiração potencial (mm dia^{1}); Rn – radiação líquida (MJ m^{-2} dia^{-1}); G – fluxo de calor no solo (MJ m^{-2} dia^{-1}); T – temperatura média do ar diária (ºC); γ – constante psicometrico (0.063 kPa ºC^{-1}); Δ- curva de pressão de umidade declínio na temperatura do ar (kPa ºC^{-1}); U2 - Velocidade média diária do vento a 2 metros (m s^{-1}); ea - pressão parcial de umidade (kPa); es - pressão

de saturação de umidade, média diária (kPa); Tmax - Temperatura máxima (°C); Tmin - Temperatura mínima (°C); RH – umidade relativa (%); BOC - balanço de radiação de ondas curtas ($MJm^{-2}dia^{-1}$) e BOL é o balanço de radiação de onda longa ($MJ\ m^{-2}dia^{-1}$); Rs é a radiação solar incidente ($MJm^{-2}dia^{-1}$); α é o coeficiente de reflexão da vegetação; Rso é a radiação solar incidente na ausência de nuvens ($MJm^{-2}dia^{-1}$), e Ra = radiação solar no topo da atmosfera ($MJm^{-2}dia^{-1}$).

2.2.3 Balanço Hídrico

A deficiência hídrica (DEF, em mm), excedente hídrico (EXC, em mm) e o armazenamento água no solo (STO, em mm) foram calculados para todas as localidades estudadas conforme o método de Thornthwaite e Mather (1955) na escala decendial. Nós utilizamos a capacidade disponível de água (CAD) no solo igual a 100 mm. A CAD de 100 mm foi utilizada para todas as localidades, pois é um valor padrão para fins climáticos e para caracterização da disponibilidade hídrica em escala regional (BRASIL, 1981; Aparecido et al., 2020), descrito na Figura 2.

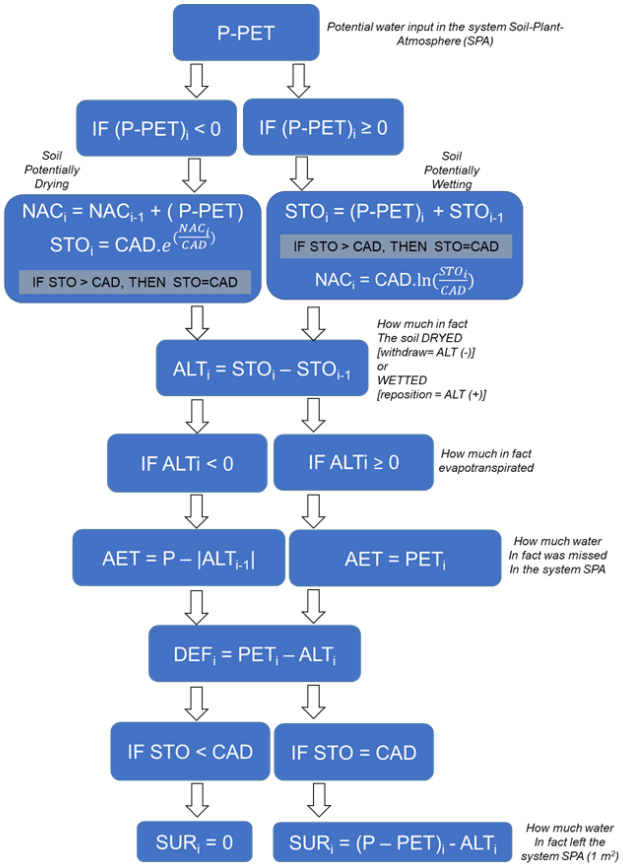

Figura 2. Fluxograma do modelo de balanço hídrico modificado a partir de Thornthwaite e Mather (1955). Legenda: P - precipitação (mm), PET evapotranspiração potencial, AET – evapotranspiração atual (mm), STO - armazenamento de água do solo (mm), CAD – capacidade de água do solo (mm), NAC - negativo acumulado (mm), significando o potencial de secagem do solo, ALT – alteração da STO, SUR - excesso de água no sistema solo-planta-atmosfera (mm), DEF - déficit hídrico do sistema solo-planta-atmosfera (mm), e i - determinado período, i−1 - período anterior. Fonte: Rolim et al. (2020).

2.2.4 Época de colheita

Neste trabalho, nós utilizamos a fenologia média da cana planta que é composta por quatro fases fenológicas (brotação, perfilhamento, desenvolvimento e

maturação), além da colheita e plantio (Marcari et al., 2015), conforme a Figura 3. Esses períodos foram utilizados como entrada (Features) nos modelos de machine learning.

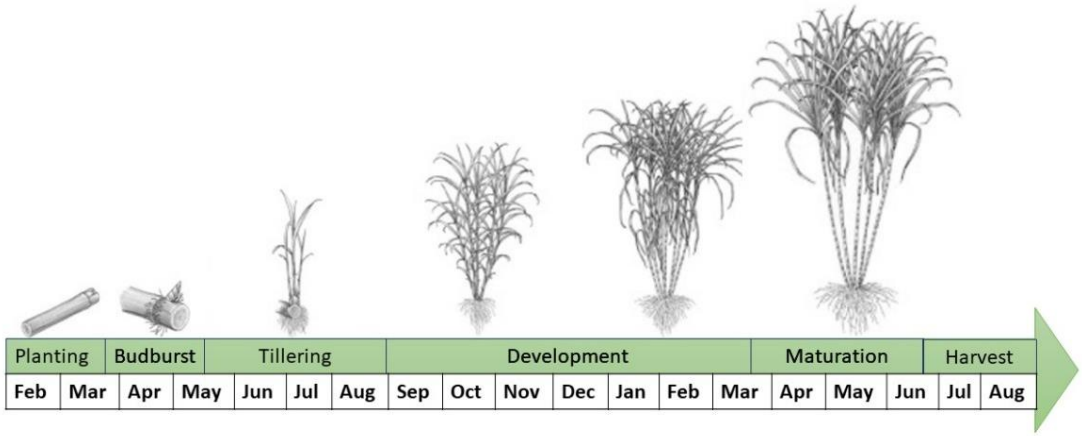

Figura 3. Esquema da fenologia no ano de desenvolvimento (1º ano) e produção (2º ano) para cana-de-açúcar.

2.2.5 Modelos de previsão

Nós utilizamos um método de regressão linear tradicional: regressão linear múltipla (RLM) e seis métodos de aprendizado de máquina (ML): support vector machine (SVM), random forest (RF), Artificial neural network (ANN), Least Absolute Shrinkage and Selection Operator (LASSO) regression, RIDGE regression e eXtreme Gradient Boosting (XGBoost) para prever a produtividade da cana-de-açúcar com seis meses de antecedência. As variáveis independentes foram os valores decendiais das variáveis climáticas: temperaturas do ar mínima, média e máxima, precipitação e radiação global, evapotranspiração potencial, e dos componentes do balanço hídrico (armazenamento de água do solo, deficiência hídrica e excedente hídrico) com antecedência de seis meses da colheita (Figura 3 e 4). Antes da realização das

previsões de produtividade, os dados foram padronizados e divididos em duas partes: treinamento (70 %) e teste (30%) conforme Zang et al. (2020). Os modelos preditivos foram desenvolvidos em uma plataforma Python IDE (ambiente de desenvolvimento integrado) Jupyter Lab (versão 3.7.4 para Windows) usando funções disponíveis na biblioteca scikit-learn, que é uma ferramenta eficiente para modelagem científica (Pedregosa et al., 2011). As otimizações dos parâmetros para os algoritmos foram ajustadas usando GridSearch-CV no Scikit-Learn (Pedregosa et al., 2011).

	Planting					**Budburst**						**Tillering**	
	Feb1	Feb2	Feb3	Mar1	Mar2	Mar3	Apr1	Apr2	Apr3	May1	May2	May3	Jun1
Tmean	Tmean_Feb1	Tmean_Feb2	Tmean_Feb3	Tmean_Mar1	Tmean_Mar2	Tmean_Mar3	Tmean_Apr1	Tmean_Apr2	Tmean_Apr3	Tmean_May1	Tmean_May2	Tmean_May3	Tmean_Jun1
Tmin	Tmin_Feb1	Tmin_Feb2	Tmin_Feb3	Tmin_Mar1	Tmin_Mar2	Tmin_Mar3	Tmin_Apr1	Tmin_Apr2	Tmin_Apr3	Tmin_May1	Tmin_May2	Tmin_May3	Tmin_Jun1
Tmax	Tmax_Feb1	Tmax_Feb2	Tmax_Feb3	Tmax_Mar1	Tmax_Mar2	Tmax_Mar3	Tmax_Apr1	Tmax_Apr2	Tmax_Apr3	Tmax_May1	Tmax_May2	Tmax_May3	Tmax_Jun1
Prec	Prec_Feb1	Prec_Feb2	Prec_Feb3	Prec_Mar1	Prec_Mar2	Prec_Mar3	Prec_Apr1	Prec_Apr2	Prec_Apr3	Prec_May1	Prec_May2	Prec_May3	Prec_Jun1
WS2M	WS2M_Feb1	WS2M_Feb2	WS2M_Feb3	WS2M_Mar1	WS2M_Mar2	WS2M_Mar3	WS2M_Apr1	WS2M_Apr2	WS2M_Apr3	WS2M_May1	WS2M_May2	WS2M_May3	WS2M_Jun1
UR	UR_Feb1	UR_Feb2	UR_Feb3	UR_Mar1	UR_Mar2	UR_Mar3	UR_Apr1	UR_Apr2	UR_Apr3	UR_May1	UR_May2	UR_May3	UR_Jun1
Qo	Qo_Feb1	Qo_Feb2	Qo_Feb3	Qo_Mar1	Qo_Mar2	Qo_Mar3	Qo_Apr1	Qo_Apr2	Qo_Apr3	Qo_May1	Qo_May2	Qo_May3	Qo_Jun1
Qg	Qg_Feb1	Qg_Feb2	Qg_Feb3	Qg_Mar1	Qg_Mar2	Qg_Mar3	Qg_Apr1	Qg_Apr2	Qg_Apr3	Qg_May1	Qg_May2	Qg_May3	Qg_Jun1
PET	PET_Feb1	PET_Feb2	PET_Feb3	PET_Mar1	PET_Mar2	PET_Mar3	PET_Apr1	PET_Apr2	PET_Apr3	PET_May1	PET_May2	PET_May3	PET_Jun1
STO	STO_Feb1	STO_Feb2	STO_Feb3	STO_Mar1	STO_Mar2	STO_Mar3	STO_Apr1	STO_Apr2	STO_Apr3	STO_May1	STO_May2	STO_May3	STO_Jun1
DEF	DEF_Feb1	DEF_Feb2	DEF_Feb3	DEF_Mar1	DEF_Mar2	DEF_Mar3	DEF_Apr1	DEF_Apr2	DEF_Apr3	DEF_May1	DEF_May2	DEF_May3	DEF_Jun1
SUR	SUR_Feb1	SUR_Feb2	SUR_Feb3	SUR_Mar1	SUR_Mar2	SUR_Mar3	SUR_Apr1	SUR_Apr2	SUR_Apr3	SUR_May1	SUR_May2	SUR_May3	SUR_Jun1

	Tillering								**Development**				
	Jun2	Jun3	Jul1	Jul2	Jul3	Aug1	Aug2	Aug3	Sep1	Sep2	Sep3	Oct1	Oct2
Tmean	Tmean_Jun2	Tmean_Jun3	Tmean_Jul1	Tmean_Jul2	Tmean_Jul3	Tmean_Aug1	Tmean_Aug2	Tmean_Aug3	Tmean_Sep1	Tmean_Sep2	Tmean_Sep3	Tmean_Oct1	Tmean_Oct2
Tmin	Tmin_Jun2	Tmin_Jun3	Tmin_Jul1	Tmin_Jul2	Tmin_Jul3	Tmin_Aug1	Tmin_Aug2	Tmin_Aug3	Tmin_Sep1	Tmin_Sep2	Tmin_Sep3	Tmin_Oct1	Tmin_Oct2
Tmax	Tmax_Jun2	Tmax_Jun3	Tmax_Jul1	Tmax_Jul2	Tmax_Jul3	Tmax_Aug1	Tmax_Aug2	Tmax_Aug3	Tmax_Sep1	Tmax_Sep2	Tmax_Sep3	Tmax_Oct1	Tmax_Oct2
Prec	Prec_Jun2	Prec_Jun3	Prec_Jul1	Prec_Jul2	Prec_Jul3	Prec_Aug1	Prec_Aug2	Prec_Aug3	Prec_Sep1	Prec_Sep2	Prec_Sep3	Prec_Oct1	Prec_Oct2
WS2M	WS2M_Jun2	WS2M_Jun3	WS2M_Jul1	WS2M_Jul2	WS2M_Jul3	WS2M_Aug1	WS2M_Aug2	WS2M_Aug3	WS2M_Sep1	WS2M_Sep2	WS2M_Sep3	WS2M_Oct1	WS2M_Oct2
UR	UR_Jun2	UR_Jun3	UR_Jul1	UR_Jul2	UR_Jul3	UR_Aug1	UR_Aug2	UR_Aug3	UR_Sep1	UR_Sep2	UR_Sep3	UR_Oct1	UR_Oct2
Qo	Qo_Jun2	Qo_Jun3	Qo_Jul1	Qo_Jul2	Qo_Jul3	Qo_Aug1	Qo_Aug2	Qo_Aug3	Qo_Sep1	Qo_Sep2	Qo_Sep3	Qo_Oct1	Qo_Oct2
Qg	Qg_Jun2	Qg_Jun3	Qg_Jul1	Qg_Jul2	Qg_Jul3	Qg_Aug1	Qg_Aug2	Qg_Aug3	Qg_Sep1	Qg_Sep2	Qg_Sep3	Qg_Oct1	Qg_Oct2
PET	PET_Jun2	PET_Jun3	PET_Jul1	PET_Jul2	PET_Jul3	PET_Aug1	PET_Aug2	PET_Aug3	PET_Sep1	PET_Sep2	PET_Sep3	PET_Oct1	PET_Oct2
STO	STO_Jun2	STO_Jun3	STO_Jul1	STO_Jul2	STO_Jul3	STO_Aug1	STO_Aug2	STO_Aug3	STO_Sep1	STO_Sep2	STO_Sep3	STO_Oct1	STO_Oct2
DEF	DEF_Jun2	DEF_Jun3	DEF_Jul1	DEF_Jul2	DEF_Jul3	DEF_Aug1	DEF_Aug2	DEF_Aug3	DEF_Sep1	DEF_Sep2	DEF_Sep3	DEF_Oct1	DEF_Oct2
SUR	SUR_Jun2	SUR_Jun3	SUR_Jul1	SUR_Jul2	SUR_Jul3	SUR_Aug1	SUR_Aug2	SUR_Aug3	SUR_Sep1	SUR_Sep2	SUR_Sep3	SUR_Oct1	SUR_Oct2

	Development			
	Oct3	Nov1	Nov2	Nov3
Tmean	Tmean_Oct3	Tmean_Nov1	Tmean_Nov2	Tmean_Nov3
Tmin	Tmin_Oct3	Tmin_Nov1	Tmin_Nov2	Tmin_Nov3
Tmax	Tmax_Oct3	Tmax_Nov1	Tmax_Nov2	Tmax_Nov3
Prec	Prec_Oct3	Prec_Nov1	Prec_Nov2	Prec_Nov3
WS2M	WS2M_Oct3	WS2M_Nov1	WS2M_Nov2	WS2M_Nov3
UR	UR_Oct3	UR_Nov1	UR_Nov2	UR_Nov3
Qo	Qo_Oct3	Qo_Nov1	Qo_Nov2	Qo_Nov3
Qg	Qg_Oct3	Qg_Nov1	Qg_Nov2	Qg_Nov3
PET	PET_Oct3	PET_Nov1	PET_Nov2	PET_Nov3
STO	STO_Oct3	STO_Nov1	STO_Nov2	STO_Nov3
DEF	DEF_Oct3	DEF_Nov1	DEF_Nov2	DEF_Nov3
SUR	SUR_Oct3	SUR_Nov1	SUR_Nov2	SUR_Nov3

Figura 4. Descrição detalhada das variáveis de entrada (Features) para os modelos utilizados.

2.2.5.1 Multiple linear regression (MLR)

A regressão linear múltipla vem sendo usada em vários trabalhos de modelos preditivos de cultivo que envolve mais de uma variável independente (Obsie *et al.*, 2020; Guo *et al.*, 2021). A MLR sempre foi mais realista que a regressão linear simples, devido essa usar várias variáveis independentes, aumentando a chance de previsões com maiores acurácias (Aiken et al., 2012). Para a regressão linear múltipla, nós realizamos o método stepwise (SW) para selecionar as variáveis para serem adicionadas e/ou removidas para a análise de regressão (Abraham et al., 2017) e foi realizada também uma análise de correlação de Pearson entre as variáveis independentes adicionadas pelo SW e a produtividade da cana-de-açúcar que apresentavam nível mínimo de significância (α <5%) para o entendimento de quanto e quando essas variáveis afetam a produtividade. A variável resposta y, no caso deste trabalho a produtividade da cana-de-açúcar, é modelada em função de mais de uma variável independente. Os modelos de regressão linear múltipla são descritos na Equação 11.

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 \ldots + \beta_k X_k + \epsilon \quad (11)$$

em que: Y = variável dependente (produtividade), X_k = variáveis independentes (ambientais), β_k = coeficientes das variáveis independentes e ϵ = erro aleatório.

2.2.5.2 Support Vector Machine (SVM)

Nós utilizamos três kernels do Support Vector Machine (SVM), os quais foram linear, polinomial e radial basis function (RBF). O modelo SVM baseado em kernel foi estabelecido por Cortes e Vapnik (1995). O SVM era utilizado apenas

para métodos de classificação, mas ao passar do tempo foi introduzido o seu uso para métodos de regressão mantendo todas as principais características do algoritmo (Vapnik, 1999, Chen et al., 2013, Benimam et al., 2020a). O SVR se origina da teoria de aprendizagem estatística, que é aplicada para determinar como regular a generalização e descobrir o equilíbrio ideal entre a complexidade da estrutura do modelo e o risco empírico (Ma et al., 2020), com diferentes funções de kernel, como linear, polinômios, splines, redes de função de base radial, esses foram embutidos e podem ser aplicadas seletivamente (Bennett e Demiriz, 1999; Suykens e Vandewalle, 1999, Zhang et al., 2004). Por isso é usado com frequência para resolver problemas complexos**,** principalmente, para previsão da produtividade dos cultivos (Cauwenberghs e Poggio, 2001; Tong e Chang, 2001).

Os valores testados dos parâmetros pelo GridsearchCV estão discriminados na Tabela 1.

2.2.5.3 Random Forest (RF)

A random forest é um modelo não linear e não paramétrico (Ho, 1995). RF é comumente utilizado para classificação e previsão da produtividade dos cultivos (Tsagkrasoulis e Montana, 2018; Quiroz et al., 2018, Aparecido et al., 2020). Os modelos de regressão e classificação ajustam um conjunto de modelos de árvore de decisão a um conjunto de dados. Nas árvores, os dados são divididos repetidamente em mais unidades homogêneos conhecido como nós, com a finalidade de melhorar a previsibilidade da variável de resposta. Os pontos de divisão são baseados em valores de variáveis preditoras. Assim, as variáveis usadas para dividir os dados são consideradas variáveis explicativas importantes. O valor previsto de uma resposta contínua é a resposta média

ajustada de todos as árvores individuais que resultaram de cada amostra “bootstraped” (Everingham et al., 2016).

2.2.5.4 Multi-layer perceptron artificial neural network (ANN)

Redes neurais artificiais (ANN) são sistemas de computação inspirados nas redes neurais biológicas (Haykin, 1999). Cada uma das unidades de processamento é chamada de neurônio. As redes neurais é uma técnica de machine learning bastante utilizada na área agrícola (Meneses et al., 2020, Aparecido et al., 2019), sendo o tipo de redes neurais mais utilizado na previsão da produtividade a Multi-layer Perceptron (MLP). A MLP é o tipo de redes neurais feed forward que apresenta no mínimo três camadas de nós, que produz um modelo preditivo para uma ou mais variáveis dependentes com base nos valores das variáveis preditoras e é capaz de distinguir dados que não são linearmente separáveis (Khan et al., 2019). O erro calculado durante a etapa de treinamento é distribuído pela rede e ajusta os pesos de conexão entre os neurônios (Haykin,1999), sendo que neste estudo foi utilizado o algoritmo de propagação reversa .

2.2.5.5 Least Absolute Shrinkage and Selection Operator (LASSO)

O menor operador absoluto de redução e seleção (Lasso; Tibshirani 1996) é um método de regressão bem conhecido e poderoso para regularização e seleção de variável para minimizar o erro de predição (Ahmad et al., 2020). O LASSO calcula o coeficiente de regressão por meio de mínimos quadrados penalizados por norma ℓ_1, adicionando uma penalidade aos coeficientes do modelo (Saporta

e Niang, 2009). O Lasso deve ser usado com cuidado no caso de conjuntos de variáveis altamente correlacionadas, uma vez que tende a selecionar arbitrariamente uma variável e ignorar o resto (Friedman et al. 2010).

2.2.5.6 RIDGE

A regressão RIDGE (Gunst & Mason, 1977; Hoerl & Kennard, 1970) (também conhecida como Regularização de Tikhonov ou queda de peso) é uma variante de problemas de mínimos quadrados regularizados, onde a escolha da função de penalidade é a L2. Maior α diminui mais os coeficientes, levando a um aumento do viés e redução da variância e vice-versa (Hastie, 2020).

2.2.5.7 XGBOOST

O XGBoost foi proposto em 2016 com novas funcionalidades, como manipulação de dados esparsos e uso de algoritmo de aproximação para um melhor tempo de processamento (Chen e Guestrin, 2016). Muitos pesquisadores agrícolas têm usado o XGBoost para prever a produtividade das safras (Shahhosseini et al., 2021). XGBoost é uma biblioteca de código aberto que implementa árvores de decisão com gradiente ampliado que são eficientes e altamente otimizadas. No aumento da árvore de gradiente, os modelos não são treinados isoladamente uns dos outros, mas sim em sucessão, onde cada modelo reduz iterativamente os erros cometidos pelos modelos anteriores. Em vez de atribuir pesos diferentes aos classificadores após cada iteração, este método ajusta um novo modelo a novos resíduos da previsão anterior e, em seguida, minimiza a perda ao adicionar a última previsão (Chen e Guestrin, 2016).

Tabela 1. Parâmetros dos modelos

Modelo	Parâmetro	Valores testados	Valores Ajustados			
			Grupo 1	Grupo 2	Grupo 3	Grupo 4
MLR	coeficientes angulares ou pesos (seleção de variáveis por stepwise)		*	*	*	*
SVM-Linear	gamma	auto; scale	auto	auto	auto	auto
	c	1;50;100	100	100	100	100
SVM-Polinomial	degree	2;3;4	2	2	2	2
	gamma	auto; scale	auto	auto	auto	auto
	c	1;50;100	100	100	100	100
SVM- RBF	gamma	auto; scale	auto	auto	auto	auto
	c	1;50;100	100	100	100	100
RF	max_depth	2;4;5;6;7;9;10;12				
	n_estimators	2;3;4;5;6;8				
	min_impurity_decrease	0;1				
ANN	camadas	31;16;20	16	16	16	16
	random_state	11	11	11	11	11
	max_iter	400	400	400	400	400
LASSO	alphas	¯4;-0,5;30				
	max_iter	10000				
RIDGE	alpha	1;0,1;0,01;0,001;0,0001;0				
XGBoost	nthread	4				
	learning_rate	0.03, 0.05, .07				
	max_depth	5;6;7				
	min_child_weight	4				
	silent	1				
	subsample	0,7				
	colsample_bytree	0,7				
	n_estimators	500				

* verificar no texto

2.2.6 Avaliação dos modelos

A previsão da produtividade da cana-de-açúcar foi realizada com seis meses de antecedência à colheita, portanto, nós utilizamos os dados climáticos decendiais desde a data de plantio até seis meses antes da colheita. Nós avaliamos o desempenho dos modelos a partir dos índices estatísticos: 1) Correlação de Pearson (r), 2) Mean absolute percentage error (MAPE), 3) Coeficiente de

determinação ajustado (R^2), e 4) Root Mean Square Error (RMSE) (Equações 12 a 15).

$$r = \frac{\sum_{i=1}^{n}(Y_{OBS_i} - \bar{Y}_{OBS}) \times (Y_{EST_i} - \bar{Y}_{EST})}{\sqrt{\sum_{i=1}^{n}(Y_{OBS_i} - \bar{Y}_{OBS})^2} \times \sqrt{\sum_{i=1}^{n}(Y_{EST_i} - \bar{Y}_{EST})^2}} \quad (12)$$

$$MAPE(\%) = \frac{\sum_{i=1}^{n}\left(\left|\frac{Y_{EST\,i} - Y_{OBS\,i}}{Y_{OBS\,i}}\right| x100\right)}{N} \quad (13)$$

$$R^2 adjusted = \left[1 - \frac{(1 - R^2)x(n-1)}{N-k-1}\right] \quad (14)$$

$$RMSE = \sqrt{\frac{\sum_{i=r}^{N}(Y_{OBS\,i} - Y_{EST\,i}}{N}} \quad (15)$$

em que $Yest_i$: Variável estimada por RNA; Yobsi: Variável observada; N: Número de dados; k: Número de variáveis independentes na regressão.

2.3 RESULTADOS E DISCUSSÃO

A análise de agrupamento por cluster, a partir das condições meteorológicas, dividiu as localidades produtoras de cana-de-açúcar do Brasil em quatro grupos (Figura 5 e Tabela 1). Os estados brasileiros apresentaram heterogeneidade de grupos nas localidades produtoras de cana em relação as condições climáticas (Figura 6).

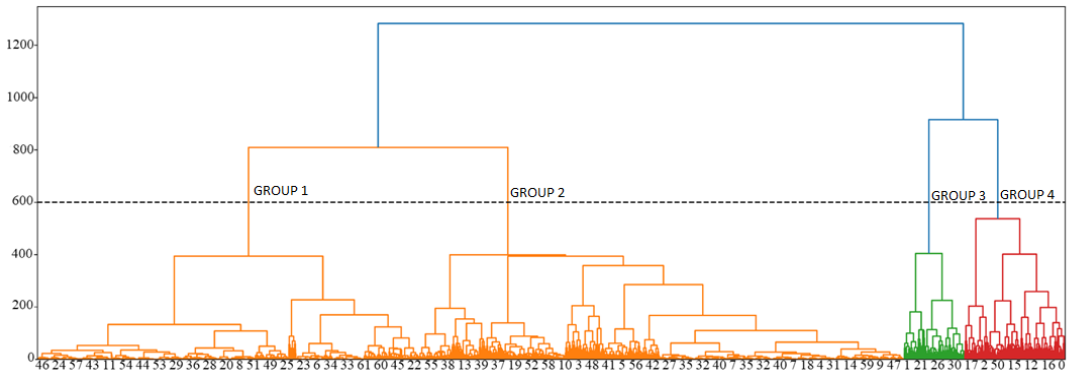

Figura 5. Dendrograma das localidades produtoras de cana-de-açúcar do Brasil em função as variáveis climáticas.

Tabela 1. Localidades produtoras de cana-de-açúcar estudadas e seus respectivos grupos conforme a análise de agrupamento.

N	Localidade	Estado	Grupo
0	Senador Guiomard	Acre	4
1	Plácido de Castro	Acre	3
2	Atalaia	Alagoas	4
3	Coruripe	Alagoas	2
4	Marechal Deodoro	Alagoas	3
5	Matriz de Camaragibe	Alagoas	3
6	Passo de Camaragibe	Alagoas	1
7	Rio Largo	Alagoas	3
8	São Luís do Quitunde	Alagoas	1
9	Campo Alegre	Alagoas	3
10	São Miguel dos Campos	Alagoas	2
11	Ipixuna	Amazonas	1
12	Caravelas	Bahia	4
13	Juazeiro	Bahia	2
14	Paracuru	Ceará	3
15	Presidente Kennedy	Espírito Santo	4
16	Itumbiara	Goiás	4
17	Santa Helena de Goiás	Goiás	4
18	Timon	Maranhão	3
19	Iturama	Minas Gerais	2
20	Maracaju	Mato Grosso do Sul	1
21	Aparecida do Taboado	Mato Grosso do Sul	4
22	Naviraí	Mato Grosso do Sul	2
23	Sidrolândia	Mato Grosso do Sul	1
24	Nova Andradina	Mato Grosso do Sul	1

25	Itaquiraí	Mato Grosso do Sul	1
26	Rio Brilhante	Mato Grosso do Sul	4
27	Diamantino	Mato Grosso	3
28	Cáceres	Mato Grosso	1
29	São José do Rio Claro	Mato Grosso	1
30	Tangará da Serra	Mato Grosso	4
31	Jaciara	Mato Grosso	3
32	Juripiranga	Paraíba	3
33	Rio Tinto	Paraíba	2
34	Gameleira	Pernambuco	2
35	Rio Formoso	Pernambuco	3
36	Sirinhaém	Pernambuco	1
37	José de Freitas	Piauí	2
38	Teresina	Piauí	2
39	União	Piauí	2
40	Cambará	Paraná	3
41	Colorado	Paraná	3
42	Cruzeiro do Oeste	Paraná	3
43	Jacarezinho	Paraná	1
44	Paranacity	Paraná	1
45	Porecatu	Paraná	2
46	Rondon	Paraná	1
47	Tapejara	Paraná	3
48	Porto Xavier	Rio Grande do Sul	3
49	Roque Gonzales	Rio Grande do Sul	1
50	Japoatã	Sergipe	4
51	Laranjeiras	Sergipe	1
52	Pacatuba	Sergipe	2
53	Rosário do Catete	Sergipe	1
54	Santo Amaro das Brotas	Sergipe	1
55	Luís Antônio	São Paulo	2
56	Guararapes	São Paulo	3
57	Batatais	São Paulo	1
58	Jaboticabal	São Paulo	2
59	Jaú	São Paulo	3
60	Morro Agudo	São Paulo	2
61	Peixe	Tocantins	2

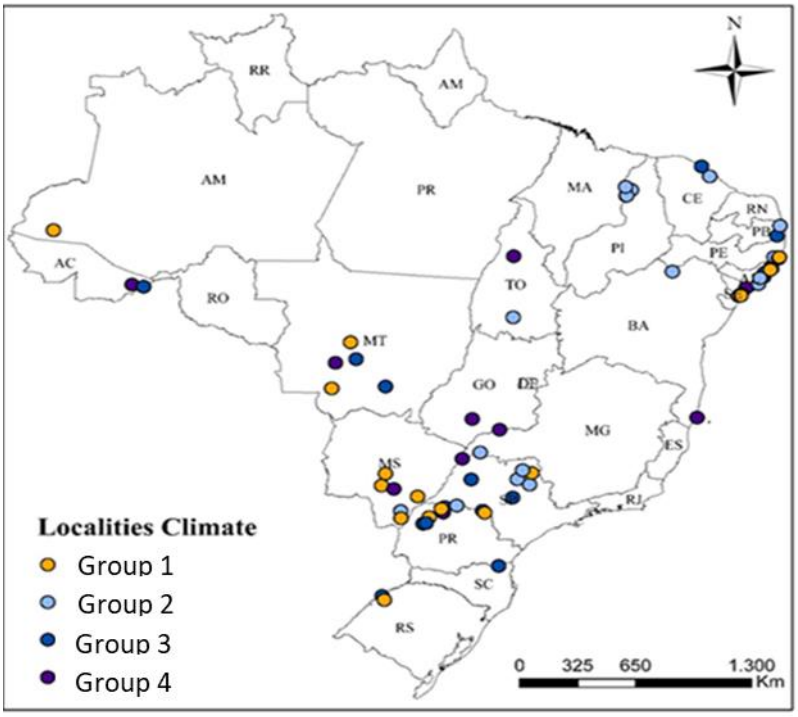

Figura 6. Distribuição espacial das localidades estudadas.

A Precipitação entre os grupos de locais foram muito semelhantes (Figura 7), entretanto no verão (dezembro-março) apresentou maior variabilidade.

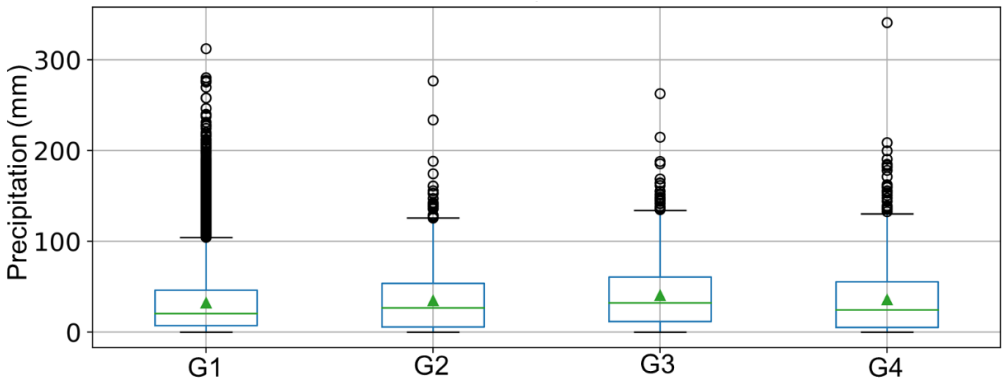

Figura 7. Precipitação dos grupos das localidades produtoras de cana-de-açúcar.

Alta variabilidade entre grupos e localidades (Apêndice Figura 1). A diversidade agrometeorológica dentro do país requer uma compreensão da influência da variabilidade climática na produtividade em escalas regionais (Zachariah et al., 2020).

Os valores máximos decendiais de precipitação no verão para os grupos 1, 2, 3 e 4 foram 310 mm, 280 mm, 270 mm e 350 mm, respectivamente (Figura 8). As localidades com maiores precipitações médias foram Matriz de Camaragibe (AL), Crato (CE) e Cachoeiro de Itapemirim (ES) com 60,19 mm decendial^{-1}, 52,95 mm decendial^{-1} e 49,36 mm decendial^{-1}, respectivamente. O grupo 1 apresentou os menores valores médios de precipitação, principalmente em Benjamin Constant (AM), Linhares (ES) e Pinheiros (ES) com 13,19 mm decendial^{-1}, 27,55 mm decendial^{-1} e 27,55 mm decendial^{-1}, respectivamente. Em geral, os valores de precipitação estão dentro ou perto dos limites prescritos para o cultivo.

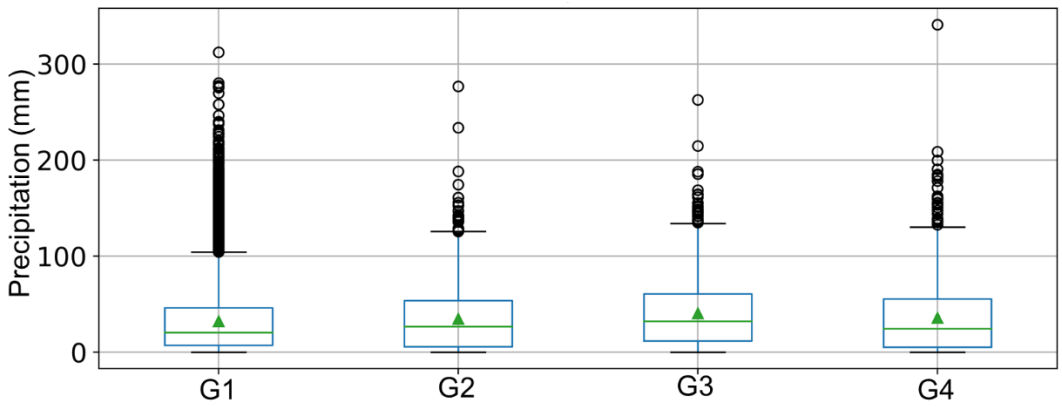

Figura 8. Total de Precipitação decendial dos grupos das localidades produtoras de cana-de-açúcar.

As tendências de precipitação e temperatura do ar são espacialmente não uniformes. O Brasil possui uma extensa área territorial com predomínio do clima

tropical, portanto, há uma alta variação climática nas diferentes regiões brasileiras (Aparecido et al., 2020).

Houve alta variabilidade na temperatura média do ar no grupo 3 (Figura 9), sendo que todos os apresentaram dados com distribuição normal para a temperatura média do ar.

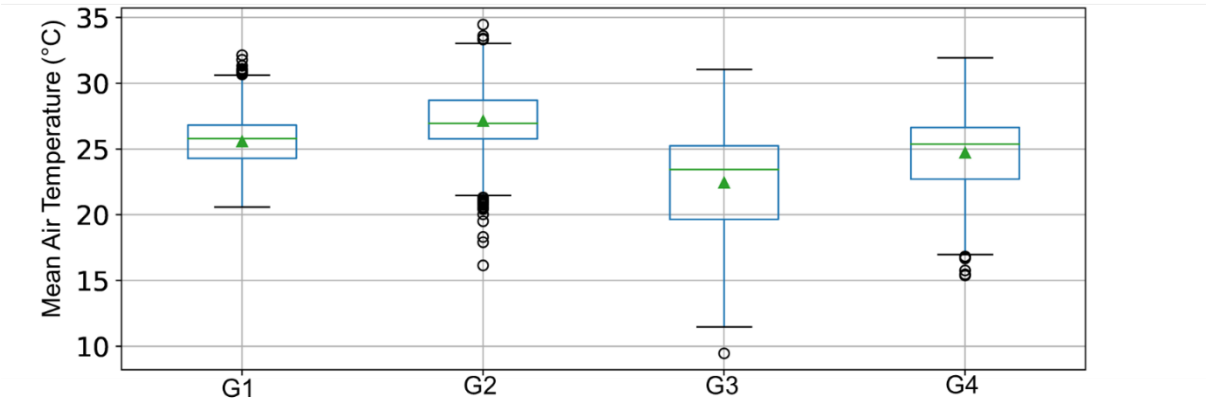

Figura 9. Temperatura média decendial do ar dos grupos das localidades produtoras de cana-de-açúcar.

Nos grupos 1 e 2 apresentaram temperaturas média decendiais acima de 33 °C. Em relação aos grupos, as temperaturas médias decendiais no grupo 3 apresentaram os menores valores no inverno.

As maiores temperaturas médias do ar ocorreram no grupo 2 com valor de 28 °C para as localidades Varjão (GO), Santa Helena de Goiás (GO) e Envira (AM) (Apêndice Figura 2). Enquanto no grupo 3, houve os menores valores médios de temperatura do ar em Rio Branco (AC), Itapemirim (ES) e Paracuru (CE) com 18,2 °C, 20,3 °C e 20,5 °C, respectivamente. O grupo 2 está na faixa adequada de temperatura do ar para a cana-de-açúcar. As temperaturas do ar favoráveis para desenvolvimento da cana-de-açúcar variam entre 28 e 38 °C (Bacchi,

1983). De acordo com Doorenbos e Kassam (1979), a cana-de-açúcar não tem um desenvolvimento viável com temperaturas do ar inferiores a 20 °C. A produtividade da cana-de-açúcar também diminui em temperaturas acima da faixa ideal, pois causa uma redução no enriquecimento de sacarose (FAO, 2018).

Houve alta variabilidade na evapotranspiração potencial entre os grupos, principalmente nos grupos 2 e 3 (Figura 10).

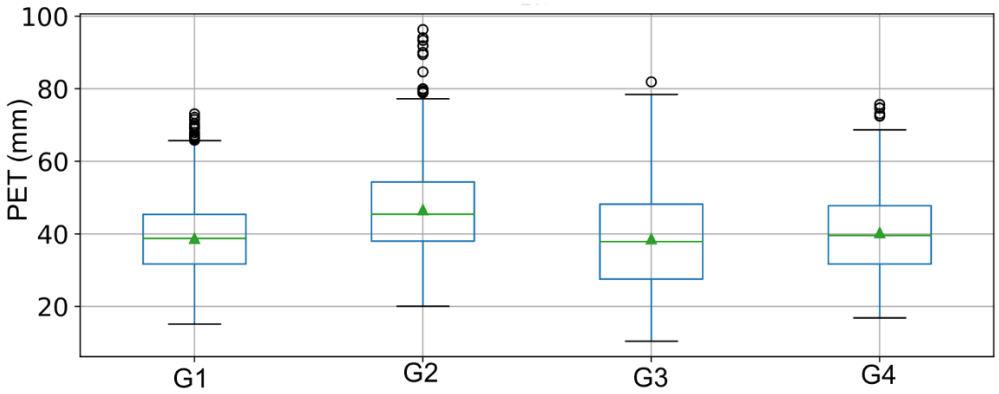

Figura 10. Evapotranspiração potencial decendial dos grupos das localidades produtoras de cana-de-açúcar.

No verão, os valores decendiais de PET foram acima 63 mm, sendo no grupo 2 os maiores valores de PET no verão. Já no inverno, os menores foram no grupo 3.

Os maiores valores médios de PET foram de 56,71 mm decendial^{-1}, 56,43 mm decendial^{-1} e 55,75 mm decendial^{-1} em Viçosa do Ceará (CE), Envira (AM) e Varjão (GO) no grupo 2, respectivamente (Apêndice Figura 4). Os menores valores médios de PETs foram 36,70 mm decendial^{-1}, 37,05 mm decendial^{-1} e 37,14 mm decendial^{-1} em Coruripe (AL), Tianguá (CE) e Itapaci (GO), respectivamente, no grupo 1.

A produção de cana-de-açúcar geralmente ocorre em regiões com alta evapotranspiração potencial, portanto, os valores de PET para as localidades são adequados para a produção de cana-de-açúcar. A necessidade de água da cana-de-açúcar com base na evapotranspiração é inferior a 2000 mm, no entanto, a água aplicada por irrigação no cultivo é em média de 3000 a 4000 mm (Shrivastava et al., 2011; Dingre e Gorantiwar, 2020). A irrigação nos cultivos de cana-de-açúcar no Brasil ainda é uma técnica recente. No entanto, com a expansão do cultivo no país vem aumentando o seu uso, com 10% da área de cultivo (Marin et al., 2020).

O excedente hídrico médio decendial para todas as localidades foram ao redor de 15 mm dia^{-1} (Figura 11).

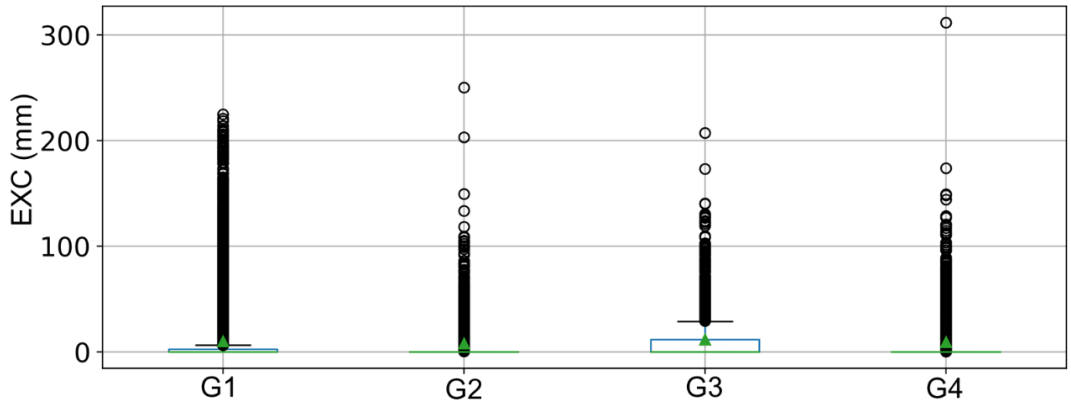

Figura 11. Excedente hídrico dos grupos das localidades produtoras de cana-de-açúcar.

Os maiores valores de EXC foram Matriz de Camaragibe (AL), Guaraciaba do Norte (CE) e Crato (CE) com 24,43 mm decendial^{-1}, 20,94 mm decendial^{-1}, 20,19 mm decendial^{-1} no grupo 2 (Apêndice Figura 5). Os menores valores de EXC foram no grupo 1 nas localidades Benjamin Constant (AM), Pinheiros (ES) e

Linhares (ES) com 1,16 mm decendial^{-1}, 5,75 mm decendial^{-1} e 5,76 mm decendial^{-1}.

Zachariah et al. (2020) observaram que a produtividade de cana-de-açúcar tem alta sensibilidade a chuva e seus marcadores como número de dias chuvosos, umidade do solo e irriação.

A distribuição dos dados de deficiência hídrica do grupo 1 foi parecida no grupo 2, sendo que no grupo 3 ocorreu os menores valores de DEF (Figura 11).

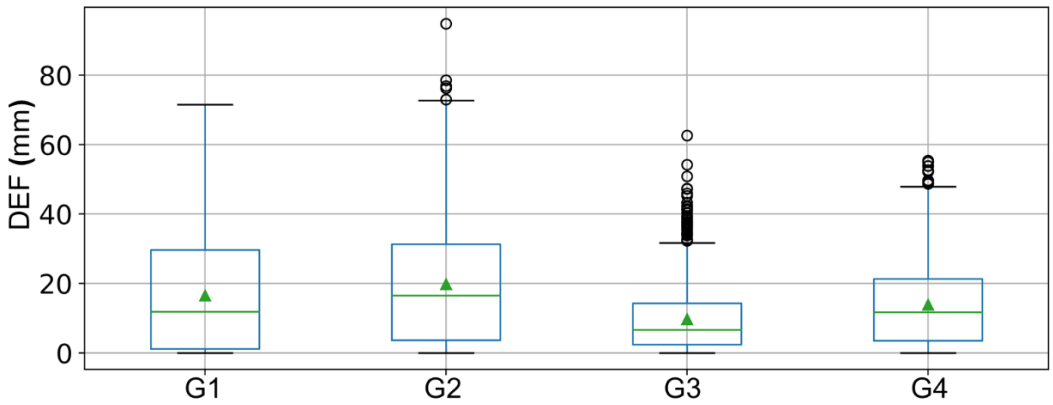

Figura 11. Deficiência hídrica dos grupos das localidades produtoras de cana-de-açúcar.

Os maiores valores médios de DEF ocorreram em Viçosa do Ceará (CE), Varjão (GO) e Santa Helena de Goiás (GO) com 32,26 mm decendial^{-1}, 32,18 mm decendial^{-1} e 32,02 mm decendial^{-1} (Apêndice Figura 6). No grupo 3, os menores valores médios de DEF nesse grupo foram 3,49 mm decendial^{-1}, 8,56 mm decendial^{-1} e 9,07 mm decendial^{-1} em Rio Branco (AC), Mâncio Lima (AC) e Lajedão (BA), respectivamente. Aparecido et al. (2020) concluíram que as regiões adequadas ao cultivo da cana-de-açúcar apresentaram déficit hídrico variando de 0 mm ano^{-1} a 550 mm ano^{-1}. O fechamento estomático é o principal mecanismo fisiológico da cana-de-açúcar resistir período de deficiência hídrica

(Taiz, 2013), consequentemente, aumenta a temperatura foliar e reduz a transpiração da planta, podendo a temperatura da folha aumentar gradualmente até 4 a 6º C acima da temperatura do ar (López et al., 2009).

Os valores médios de produtividade dos grupos 1, 2, 3 e 4 das localidades produtoras de cana-de-açúcar foram de 65,7 t ha^{-1}, 66,0 t ha^{-1}, 61,7 t ha^{-1} e 60,9 t ha^{-1}, respectivamente (Figura 12).

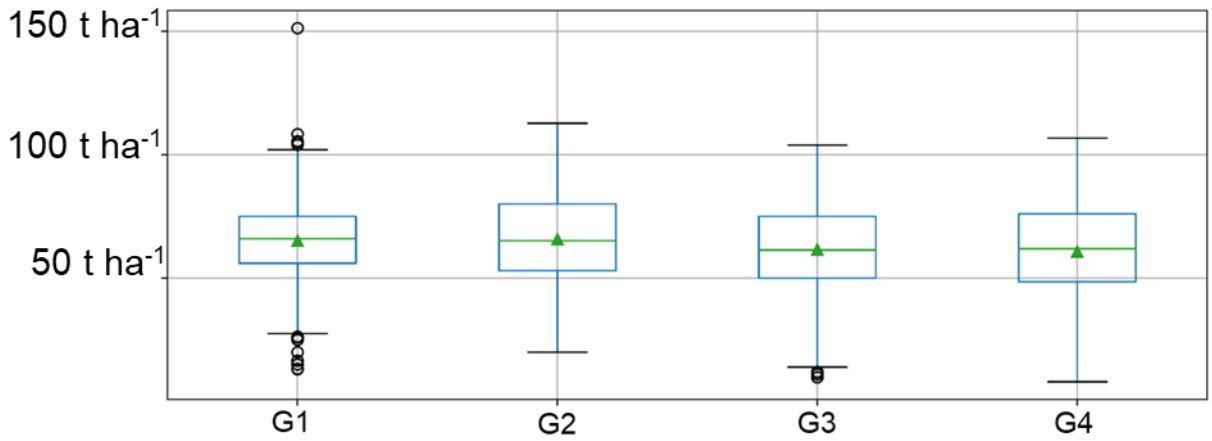

Figura 12**.** Produtividade dos grupos das localidades produtoras de cana-de-açúcar no Brasil.

No grupo 1, a localidade com a maior produtividade média foi Tianguá (CE) com 84,7t ha^{-1} (Apêndice Figura 5). Já Coruripe (AL) apresentou a menor produtividade média com 32,5 t ha^{-1}. Nas localidades do grupo 2, Matriz de Camaragibe (AL) apresentou os maiores valores médios de produtividade de 87,7 t ha^{-1}, enquanto, Santa Helena de Goiás (GO) foi a localidade com a menor produtividade média, de 41,1 t ha^{-1}. A localidade com maior produtividade média no grupo 3 foi Ibiapina (CE) com 85,9 t ha^{-1} e a localidade com menor produtividade média nesse grupo foi Cruzeiro do Sul (AC) com 25,9 t ha^{-1}. No grupo 4, a localidade com maior produtividade média foi São Luís do Quitunde

(AL), com 32,3 t ha^{-1}. Enquanto, Brasiléia (AC) foi a localidade com a menor produtividade do grupo1, com 32,3 t ha^{-1}.

As correlações entre a produtividade e as variáveis climáticas foram variáveis (Figura 13) para cada grupo de variáveis.

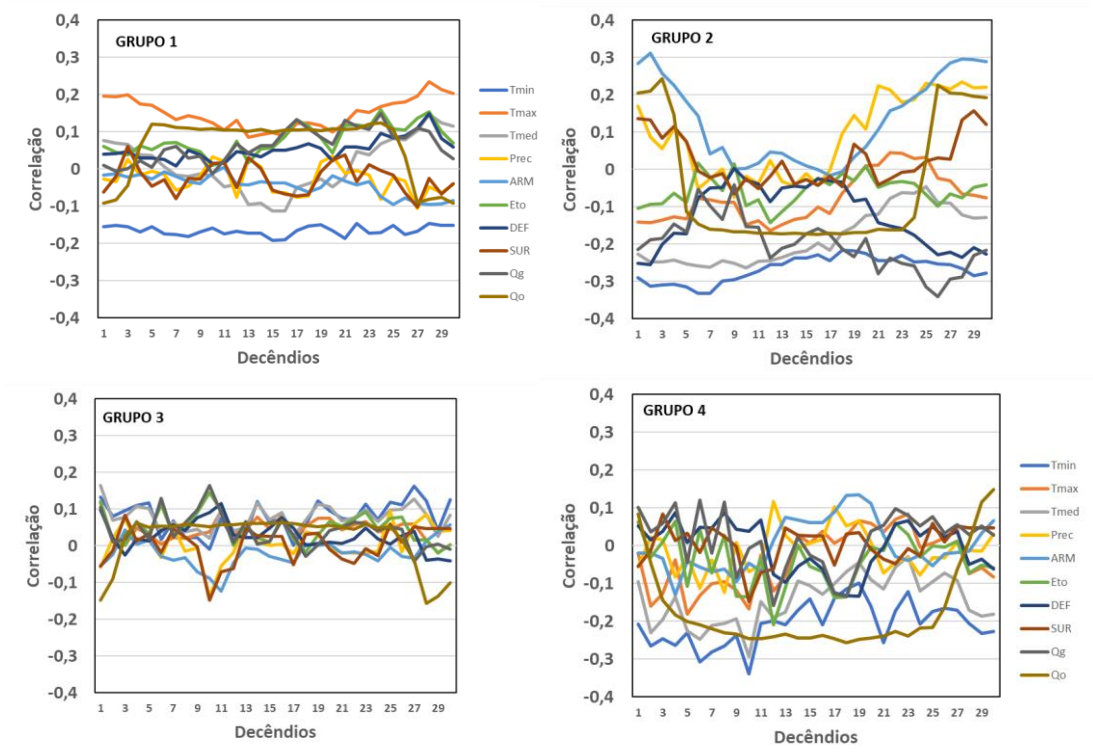

Figura 13. Correlação Linear de Pearson de variáveis climáticas decendiais e a produtividade de cana-de-açúcar para diferentes grupos de localidades.

O grupo 3 apresentou as correlações entre as variáveis climáticas e a produtividade da cana-de-açúcar mais fracas. No grupo 1, as temperaturas mínima e máxima do ar apresentaram maiores correlação com a produtividade. Houve alta variabilidade nas correlações entre as variáveis climáticas e a produtividade no grupo 3 durante o desenvolvimento da cana-de-açúcar.

Nós comparamos o desempenho de modelos de algoritmos de ML em quatro grupos de localidades no Brasil para a previsão de seis meses de antecedência da colheita da cana-de-açúcar. Os resultados estatísticos dos modelos de para a previsão da produtividade da cana-de-açúcar nos grupos de localidades divididos nesse estudo durante o treinamento e teste (Tabelas 2 e 3) foram variáveis indicando que cada modelo tem uma forma particular de caracterizar a relação entre as features e a produtividade em cada grupo de localidades.

Os modelos que apresentaram os melhores desempenhos no treinamento dos grupos foram XGBOOST e MLP (Figura 13 e Tabela 1). Considerando os modelos de ML, alguns dos modelos desenvolvidos forneceram previsões com valores de MAPE tão pequenos quanto 8%. Isso indicou que os modelos desenvolvidos superaram os modelos de previsão de produtividade de cana-de-açúcar desenvolvidos na literatura. Os modelos com os piores desempenhos no treinamento foram RLM e RBF Kernel SVM, contrariando resultados encontrados por Fan et al. (2018) indicando que a variabilidade climática interfere na produtividade de forma diversa.

Já no teste dos modelos de ML, os valores de MAPE foram acima de 20%. No teste dos modelos pelo XGBOOST apresentou MAPEs de 29.71%, 26.79%, 43.5% e 33.36% para os grupos 1, 2, 3 e 4, respectivamente. O RLM apresentou melhor desempenho no teste nos grupos.

O tipo de variáveis de entrada exerce um papel significativo na precisão da estimativa dos modelos de aprendizado de máquina. Alguns trabalhos propuseram, além dos dados climáticos, a utilização de dados de solo (mineralogia e fertilidade) e índice de área foliar (Junliang et al., 2021).

Tabela 2. Estatísticas de avaliação do desempenho na previsão de produtividade no período de treinamento.

	RLM				**SVM-LINEAR KERNEL**			
	G1	G2	G3	G4	G1	G2	G3	G4
MAPE(%)	26.08	20.53	30.54	25.59	23.46	17.32	25.2	23.46
RMSE	16356.36	14757.32	16063.76	14432.95	15885.17	13626.26	15038.89	13870.64
R^2adjust	0.06	0.29	0.23	0.43	0.16	0.39	0.26	0.47
R	0.24	0.54	0.48	0.66	0.41	0.63	0.51	0.68
	SVM-POLYNOMIAL KERNEL				**SVM- RBF KERNEL**			
	G1	G2	G3	G4	G1	G2	G3	G4
MAPE(%)	25.75	21.58	29.54	33.91	26.75	25.02	30.45	34.59
RMSE	16674.55	15500.78	16592.25	17990.16	17078.85	17079.18	17072.18	18490.67
R^2 adjust	0.11	0.24	0.14	0.27	0.24	0.53	0.54	0.68
R	0.34	0.49	0.37	0.52	0.49	0.73	0.73	0.82
	LASSO				**MLP**			
	G1	G2	G3	G4	G1	G2	G3	G4
MAPE(%)	17.79	12	15.62	4.18	7.3	0.42	0.77	0
RMSE	12455.36	9397.26	9337.08	3060.84	8055.27	882.82	2006.12	0
R^2 adjust	0.49	0.71	0.71	0.97	0.78	1	0.99	1
R	0.7	0.84	0.85	0.99	0.89	1	0.99	1
	RF				**RIDGE**			
	G1	G2	G3	G4	G1	G2	G3	G4
MAPE(%)	14.05	11.55	16.1	10.8	21.83	15.84	22.62	16.6
RMSE	10098.96	9327.58	9909.47	6952.8	14428.41	11787.02	12994.66	10074.57
R^2 adjust	0.68	0.73	0.71	0.73	0.32	0.55	0.46	0.73
R	0.82	0.85	0.84	0.85	0.57	0.74	0.68	0.85
	XGBOOST							
	G1	G2	G3	G4				
MAPE(%)	2.53	0.09	0.15	0				
RMSE	2687.63	75.11	108.08	0.06				
R^2 adjust	0.98	1	1	1				
R	0.99	1	1	1				

Tabela 3. Estatísticas de avaliação do desempenho na previsão de produtividade no período de teste.

	RLM				**LINEAR KERNEL**			
	G1	G2	G3	G4	G1	G2	G3	G4
MAPE(%)	23.43	20.6	30.67	32	24.49	22.19	40.14	38.01
RMSE	17071.74	15283.52	17140.88	15820.24	17219.8	15869.56	20464.97	18679.48
R^2adjust	0.05	0.26	0.12	0.42	0.01	0.24	0.01	0.21
R	0.23	0.51	0.34	0.65	0.08	0.49	0.11	0.46
	POLYNOMIAL KERNEL				**RBF KERNEL**			
	G1	G2	G3	G4	G1	G2	G3	G4
MAPE(%)	23.28	23.94	38.92	44.2	23.86	32.61	40.83	64.34
RMSE	16151.52	16894.55	19595.83	20185.64	16373.95	22858.93	24990.91	38288.9
R^2adjust	0.02	0.13	0.03	0.099	0.03	0.09	0.08	0.08
R	0.14	0.36	0.17	0.31	0.17	0.31	0.28	0.28
	LASSO				**MLP**			
	G1	G2	G3	G4	G1	G2	G3	G4

MAPE(%)	36.46	32.61	40.83	64.34	46.73	113.92	44.42	40.39
RMSE	26106.89	22858.93	24990.91	38288.9	36470.22	98512.99	25778.73	22796.69
R^2adjust	0	0.09	0.08	0.08	0.02	0	0.13	0.18
R	0.05	0.31	0.28	0.28	0.15	0.03	0.36	0.42
	RF				**RIDGE**			
	G1	G2	G3	G4	G1	G2	G3	G4
MAPE(%)	28.26	26.24	42.09	39.09	25.99	23.63	39.5	37.39
RMSE	20276.78	18892.73	21632.58	18735.34	18582.81	16878.05	20265.96	18984.09
R^2adjust	0.01	0.04	0	0.17	0	0.19	0.04	0.24
R	0.12	0.21	0.05	0.41	0.02	0.44	0.21	0.49

	XGBOOST			
	G1	G2	G3	G4
MAPE(%)	29.71	26.79	43.5	33.36
RMSE	21639.8	20252.5	22679.37	17039.65
R^2adjust	0	0.04	0	0.3
R	0.01	0.19	0.07	0.55

No RBF kernel, a previsão da produtividade apresentou baixa dispersão tanto treinamento quanto no teste.

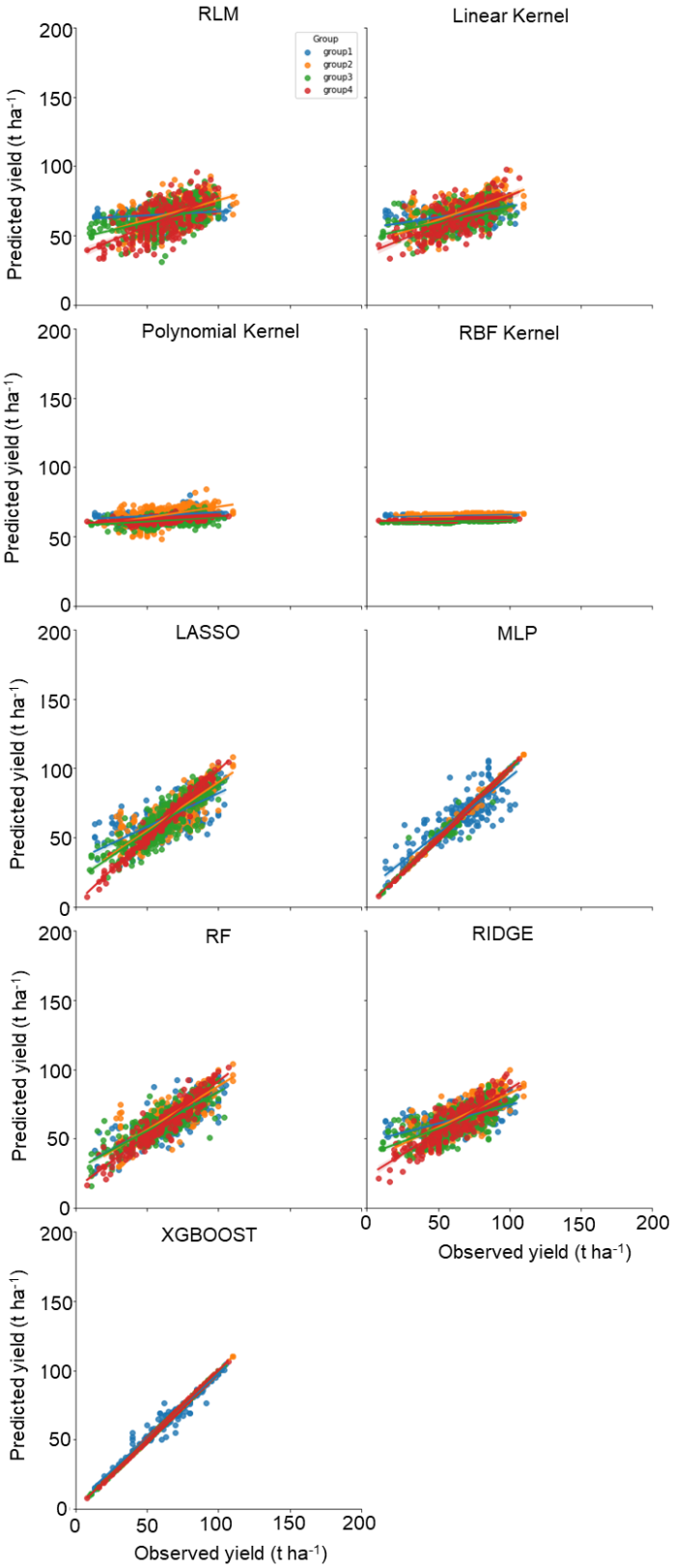

Figura 13. Treinamento da previsão da produtividade dos grupos das localidades produtoras de cana-de-açúcar dos modelos RLM, SVM linear, SVM polinomial, SVM RBF, LASSO, MLP, RF, RIDGE e XGBOOST.

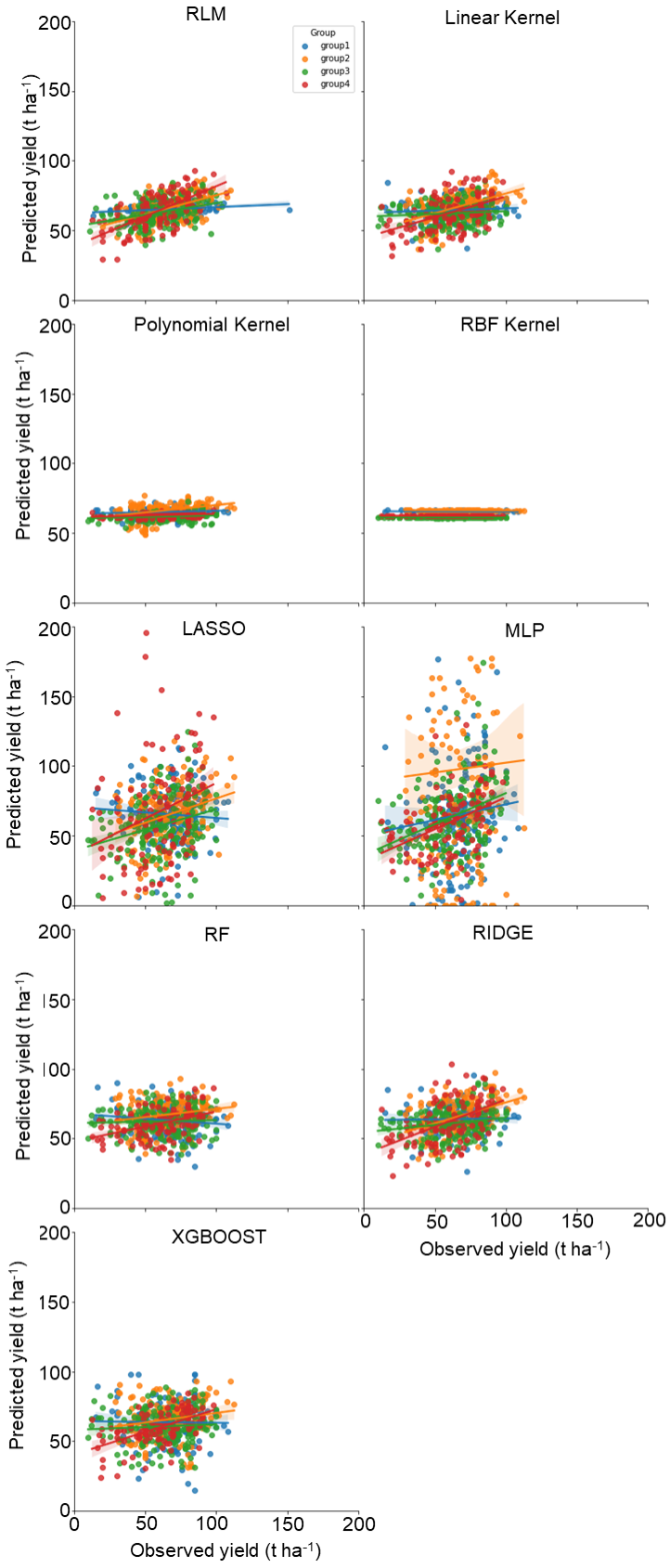
RLM
Linear Kernel
Polynomial Kernel
RBF Kernel
LASSO
MLP
RF
RIDGE
XGBOOST
Group
group1
group2
group3
group4
Predicted yield (t ha-1)
Observed yield (t ha-1)
0
50
100
150
200

Figura 14. Teste da previsão da produtividade dos grupos das localidades produtoras de cana-de-açúcar dos modelos RLM, SVM linear, SVM polinomial, SVM RBF, LASSO, MLP, RF, RIDGE e XGBOOST.

2.4 CONCLUSÕES

Os modelos XBOOST e MLP apresentam os melhores desempenhos para a previsão de produtividade da cana-de-açúcar.

Nos grupos 1 e 2 ocorrem os maiores valores de deficiência hídrica nas localidades produtoras de cana-de-açúcar.

.

REFERÊNCIAS

Abraham, S., Raisee, M., Ghorbaniasl, G., Contino, F., & Lacor, C. (2017). A robust and efficient stepwise regression method for building sparse polynomial chaos expansions. Journal of Computational Physics, 332, 461-474.

AHMAD, Ishfaq et al. Remote sensing-based framework to predict and assess the interannual variability of maize yields in Pakistan using Landsat imagery. **Computers and Electronics in Agriculture**, v. 178, p. 105732, 2020.

Aiken, L. S., West, S. G., Pitts, S. C., Baraldi, A. N., & Wurpts, I. C. (2012). Multiple linear regression. *Handbook of Psychology, Second Edition*, *2*.

ALLEN, Richard G. et al. A recommendation on standardized surface resistance for hourly calculation of reference ETo by the FAO56 Penman-Monteith method. **Agricultural Water Management**, v. 81, n. 1-2, p. 1-22, 2006.

ASHAPURE, Akash et al. Developing a machine learning based cotton yield estimation framework using multi-temporal UAS data. **ISPRS Journal of Photogrammetry and Remote Sensing**, v. 169, p. 180-194, 2020.

BENIMAM, Hania et al. Dragonfly-support vector machine for regression modeling of the activity coefficient at infinite dilution of solutes in imidazolium ionic liquids using σ-profile descriptors. **Journal of Chemical & Engineering Data**, v. 65, n. 6, p. 3161-3172, 2020.

BENNETT, Kristin et al. Semi-supervised support vector machines. **Advances in Neural Information processing systems**, p. 368-374, 1999.

CANISARES, Lucas Pecci et al. Soil microstructure alterations induced by land use change for sugarcane expansion in Brazil. **Soil Use and Management**, v. 36, n. 2, p. 189-199, 2020.

CAUWENBERGHS, Gert; POGGIO, Tomaso. Incremental and decremental support vector machine learning. **Advances in neural information processing systems**, p. 409-415, 2001.

CHEN, Ji-Long; LI, Guo-Sheng; WU, Sheng-Jun. Assessing the potential of support vector machine for estimating daily solar radiation using sunshine duration. **Energy conversion and management**, v. 75, p. 311-318, 2013.

CHEN, Tianqi; GUESTRIN, Carlos. Xgboost: A scalable tree boosting system. In: **Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining**. 2016. p. 785-794.

CHEN, Yi; TAO, Fulu. Improving the practicability of remote sensing data-assimilation-based crop yield estimations over a large area using a spatial assimilation algorithm and ensemble assimilation strategies. **Agricultural and Forest Meteorology**, v. 291, p. 108082, 2020.

CHLINGARYAN, Anna; SUKKARIEH, Salah; WHELAN, Brett. Machine learning approaches for crop yield prediction and nitrogen status estimation in precision agriculture: A review. **Computers and electronics in agriculture**, v. 151, p. 61-69, 2018.

CONAB. Companhia Nacional de Abastecimento (2020) Acompanhamento da safra brasileira de cana-de-açúcar. Available online at https://www.conab.gov.br/info-agro/safras/cana. Accessed in: August 21th, 2020 (in Portuguese).

CORTES, Corinna; VAPNIK, Vladimir. Support vector machine. Machine learning, v. 20, n. 3, p. 273-297, 1995.

Souza Rolim, G., de Oliveira Aparecido, L. E., de Souza, P. S., Lamparelli, R. A. C., & dos Santos, É. R. (2020). Climate and natural quality of Coffea arabica L. drink. Theoretical and Applied Climatology, 1-12.

ELAVARASAN, Dhivya et al. Forecasting yield by integrating agrarian factors and machine learning models: A survey. **Computers and Electronics in Agriculture**, v. 155, p. 257-282, 2018.

EVERINGHAM, Yvette et al. Accurate prediction of sugarcane yield using a random forest algorithm. **Agronomy for sustainable development**, v. 36, n. 2, p. 27, 2016.

FAN, Junliang et al. Effects of earlywood and latewood on sap flux density-based transpiration estimates in conifers. **Agricultural and Forest Meteorology**, v. 249, p. 264-274, 2018.

Fao 2020. http://www.fao.org/faostat/en/#data/QC.

FENG, Puyu et al. Dynamic wheat yield forecasts are improved by a hybrid approach using a biophysical model and machine learning technique. **Agricultural and Forest Meteorology**, v. 285, p. 107922, 2020.

FENG, Puyu et al. Dynamic wheat yield forecasts are improved by a hybrid approach using a biophysical model and machine learning technique. **Agricultural and Forest Meteorology**, v. 285, p. 107922, 2020.

FRIEDMAN, Jerome; HASTIE, Trevor; TIBSHIRANI, Robert. **Applications of the lasso and grouped lasso to the estimation of sparse graphical models**. Technical report, Stanford University, 2010.

Gunst, R. F., & Mason, R. L. (1977). Advantages of examining multicollinearities in regression analysis. Biometrics, pp. 249–260.

GUO, Miao et al. Multi-level system modelling of the resource-food-bioenergy nexus in the global south. **Energy**, v. 197, p. 117196, 2020.

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

HAQIQI, Iman et al. Predicting Crop Yields Using Soil Moisture and Heat: An Extension to Schlenker and Roberts (2009). 2019.

HASTIE, Trevor. Ridge regularization: An essential concept in data science. **Technometrics**, v. 62, n. 4, p. 426-433, 2020.

HO, Tin Kam. Random decision forests. In: **Proceedings of 3rd international conference on document analysis and recognition**. IEEE, 1995. p. 278-282.

HOERL, Arthur E.; KENNARD, Robert W. Ridge regression: Biased estimation for nonorthogonal problems. **Technometrics**, v. 12, n. 1, p. 55-67, 1970.

HYKIN, S. Neural networks: a comprehensive foundation. Printice-hall. **Inc., New Jersey**, p. 120-134, 1999.

MA, Yunqian; GUO, Guodong (Ed.). **Support vector machines applications**. New York, NY, USA:: Springer, 2014.

MARCARI, Marcos Antônio; DE SOUZA ROLIM, Glauco; DE OLIVEIRA APARECIDO, Lucas Eduardo. Agrometeorological models for forecasting yield and quality of sugarcane. Australian Journal of Crop Science, v. 9, n. 11, p. 1049-1056, 2015.

MENESES, Klara Cunha de et al. Estimating Potential Evapotranspiration in Maranhão State Using Artificial Neural Networks. **Revista Brasileira de Meteorologia**, v. 35, n. 4, p. 675-682, 2020.

OBSIE, Efrem Yohannes; QU, Hongchun; DRUMMOND, Francis. Wild blueberry yield prediction using a combination of computer simulation and machine learning algorithms. **Computers and Electronics in Agriculture**, v. 178, p. 105778, 2020.

OLIVEIRA APARECIDO, Lucas Eduardo et al. Machine learning algorithms for forecasting the incidence of Coffea arabica pests and diseases. **International journal of biometeorology**, p. 1-18, 2020.

PAIXÃO, Carla SS et al. Statistical Process Control Applied to Monitor Losses in the Mechanized Sugarcane Harvesting. **Engenharia Agrícola**, v. 40, n. 4, p. 473-480, 2020.

PEDREGOSA, Fabian et al. Scikit-learn: Machine learning in Python. **the Journal of machine Learning research**, v. 12, p. 2825-2830, 2011.

Quiroz, J.C., Mariun, N., Mehrjou, M.R., Izadi, M., Misron, N., and Mohd Radzi, M.A. 2018. “Fault detection of broken rotor bar in LS-PMSM using random forests.” Measurement, Vol. 116: pp. 273–280.

ROBERTS, Michael J. et al. Comparing and combining process-based crop models and statistical models with some implications for climate change. **Environmental Research Letters**, v. 12, n. 9, p. 095010, 2017.

Saporta, Gand N. Niang. 2009. Principal component analysis: application to statistical process control. Data Anal.Vol: 1-23-29.

SCHWALBERT, Raí A. et al. Satellite-based soybean yield forecast: Integrating machine learning and weather data for improving crop yield prediction in southern Brazil. **Agricultural and Forest Meteorology**, v. 284, p. 107886, 2020.

SHAWON, Ashifur Rahman et al. Assessment of a Proximal Sensing-integrated Crop Model for Simulation of Soybean Growth and Yield. **Remote Sensing**, v. 12, n. 3, p. 410, 2020.

SHENDRYK, Yuri; DAVY, Robert; THORBURN, Peter. Integrating satellite imagery and environmental data to predict field-level cane and sugar yields in Australia using machine learning. **Field Crops Research**, v. 260, p. 107984, 2021.

SIDRA, IBGE. Sistema IBGE de recuperação automática–SIDRA. **Censo Agropecuário**, 2020.

Singla, S.K., Garg, R.D., Dubey, O.P. Ensemble machine learning methods to estimate the sugarcane yield based on remote sensing information. Revue d'Intelligence Artificielle. Volume 34, Issue 6, 31 December 2020, Pages 731-743.

SPARKS, Adam H. nasapower: a NASA POWER global meteorology, surface solar energy and climatology data client for R. 2018.

SUYKENS, Johan AK; VANDEWALLE, Joos. Least squares support vector machine classifiers. **Neural processing letters**, v. 9, n. 3, p. 293-300, 1999.

TEDESCO-OLIVEIRA, Danilo et al. Convolutional neural networks in predicting cotton yield from images of commercial fields. **Computers and Electronics in Agriculture**, v. 171, p. 105307, 2020.

THORNTHWAITE, C.W.; MATHER, J.R. The water balance. Centerton, NJ: Drexel Institute of Technology - Laboratory of Climatology, 1955. 104p. (Publications in Climatology, vol. VIII, n.1).

TIBSHIRANI, Robert. Regression shrinkage and selection via the lasso. **Journal of the Royal Statistical Society: Series B (Methodological)**, v. 58, n. 1, p. 267-288, 1996.

TONG, Simon; CHANG, Edward. Support vector machine active learning for image retrieval. In: **Proceedings of the ninth ACM international conference on Multimedia**. 2001. p. 107-118.

TSAGKRASOULIS, Dimosthenis; MONTANA, Giovanni. Random forest regression for manifold-valued responses. **Pattern Recognition Letters**, v. 101, p. 6-13, 2018.

VAPNIK, Vladimir N. An overview of statistical learning theory. IEEE transactions on neural networks, v. 10, n. 5, p. 988-999, 1999.

VERA, Ivan; WICKE, Birka; HILST, Floor van der. Spatial Variation in Environmental Impacts of Sugarcane Expansion in Brazil. Land, v. 9, n. 10, p. 397, 2020.

Walter A, Galdos MV, Scarpare FV, Leal MRLV, Seabra JEA, Cunha MP et al (2014) Brazilian sugarcane ethanol: developments so far and challenges for the future: Brazilian sugarcane ethanol. Wiley Interdisciplinary Reviews: Energy and Environment 3:70–92.

WEBBER, Heidi et al. No perfect storm for crop yield failure in Germany. **Environmental Research Letters**, v. 15, n. 10, p. 104012, 2020.

ZACHARIAH, Mariam et al. On the role of rainfall deficits and cropping choices in loss of agricultural yield in Marathwada, India. **Environmental Research Letters**, v. 15, n. 9, p. 094029, 2020.

ZHANG, Li; ZHOU, Weida; JIAO, Licheng. Wavelet support vector machine. **IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics)**, v. 34, n. 1, p. 34-39, 2004.

ZILLI, Marcia et al. The impact of climate change on Brazil's agriculture. **Science of the Total Environment**, v. 740, p. 139384, 2020.

Apêndice

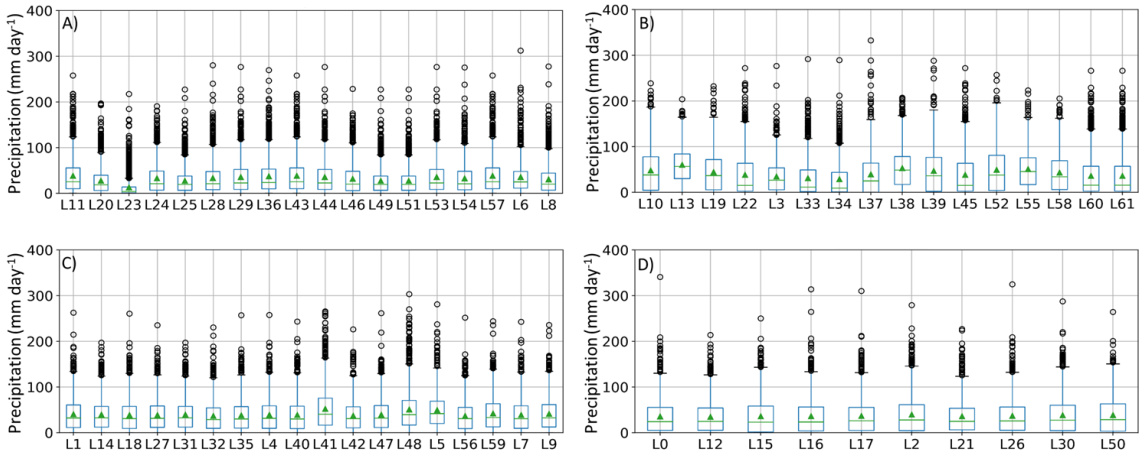

Figura 1. Precipitação acumulada das localidades produtoras de cana-de-açúcar dos (A) grupo 1, (B) grupo 2, C) grupo 3 e (D) grupo 4. Legenda: L0: Senador Guiomard-AC; L1: Plácido de Castro-AC; L2: Atalaia-AL; L3: Coruripe-AL; L4: Marechal Deodoro-AL; L5: Matriz de Camaragibe-AL; L6: Passo de Camaragibe-AL; L7: Rio Largo-AL; L8: São Luís do Quitunde-AL; L9: Campo Alegre-AL; L10: São Miguel dos Campos-AL; L11: Ipixuna -AM; L12: Caravelas- BA; L13: Juazeiro- BA; L14: Paracuru-CE; L15: Presidente Kennedy- ES; L16: Itumbiara-GO; L17: Santa Helena de Goiás-GO; L18: Timon-MA; L19:Iturama- MG; L20: Maracaju-MS; L21: Aparecida do Taboado-MS; L22:

Naviraí- MS; L23: Sidrolândia-MS; L24: Nova Andradina-MS; L25: Itaquiraí-MS; L26: Rio Brilhante- MS; L27: Diamantino-MT; L28: Cáceres- MT; L29: São José do Rio Claro-MT; L30: Tangará da Serra- MT; L31: Jaciara- MT; L32: Juripiranga-PB; L33: Rio Tinto-PB; L34: Gameleira- PE; L35: Rio Formoso-PE; L36: Sirinhaém-PE; L37: José de Freitas-PI; L38: Teresina-PI; L39: União- PI; L40: Cambará-PR; L41: Colorado-PR; L42: Cruzeiro do Oeste-PR; L43: Jacarezinho-PR; L44: Paranacity-PR; L45: Porecatu-PR; L46: Rondon-PR; L47: Tapejara-PR; L48: Porto Xavier-RS; L49: Roque Gonzales-RS; L50: Japoatã-SE; L51: Laranjeiras-SE; L52: Pacatuba-SE; L53: Rosário do Catete-SE; L54: Santo Amaro das Brota-SE; L55: Luís Antônio-SP; L56: Guararapes-SP; L57: Batatais-SP; L58: Jaboticabal-SP; L59: Jaú-SP; L60: Morro Agudo-SP; L61: Peixe-TO.

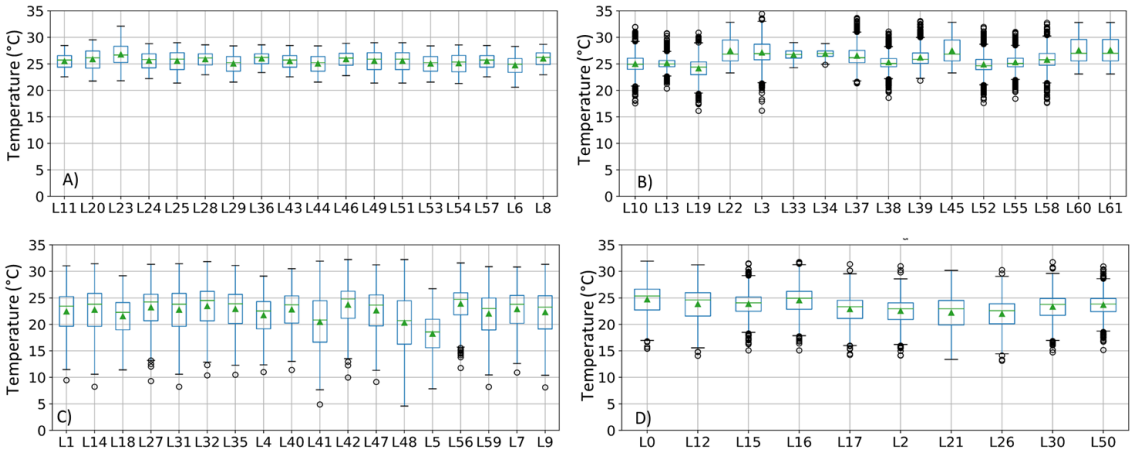

Figura 2. Temperatura média do ar das localidades produtoras de cana-de-açúcar dos (A) grupo 1, (B) grupo 2, C) grupo 3 e (D) grupo 4. Legenda: L0: Senador Guiomard-AC; L1: Plácido de Castro-AC; L2: Atalaia-AL; L3: Coruripe-AL; L4: Marechal Deodoro-AL; L5: Matriz de Camaragibe-AL; L6: Passo de Camaragibe-AL; L7: Rio Largo-AL; L8: São Luís do Quitunde-AL; L9: Campo Alegre-AL; L10: São Miguel dos Campos-AL; L11: Ipixuna -AM; L12: Caravelas- BA; L13: Juazeiro- BA; L14: Paracuru-CE; L15: Presidente Kennedy- ES; L16: Itumbiara-GO; L17: Santa Helena de Goiás-GO; L18: Timon-MA; L19:Iturama- MG; L20: Maracaju-MS; L21: Aparecida do Taboado-MS; L22: Naviraí- MS; L23: Sidrolândia-MS; L24: Nova Andradina-MS; L25: Itaquiraí-MS; L26: Rio Brilhante- MS; L27: Diamantino-MT; L28: Cáceres- MT; L29: São José do Rio Claro-MT; L30: Tangará da Serra- MT; L31: Jaciara- MT; L32: Juripiranga-PB; L33: Rio Tinto-PB; L34: Gameleira- PE; L35: Rio Formoso-PE; L36: Sirinhaém-PE; L37: José de Freitas-PI; L38: Teresina-PI; L39: União- PI; L40: Cambará-PR; L41: Colorado-PR; L42: Cruzeiro do Oeste-PR; L43: Jacarezinho-PR; L44: Paranacity-PR; L45: Porecatu-PR; L46: Rondon-PR; L47: Tapejara-PR; L48: Porto Xavier-RS; L49: Roque Gonzales-RS; L50: Japoatã-SE; L51: Laranjeiras-SE; L52: Pacatuba-SE; L53: Rosário do Catete-SE; L54: Santo Amaro das Brota-SE; L55: Luís Antônio-SP; L56: Guararapes-SP; L57: Batatais-SP; L58: Jaboticabal-SP; L59: Jaú-SP; L60: Morro Agudo-SP; L61: Peixe-TO.

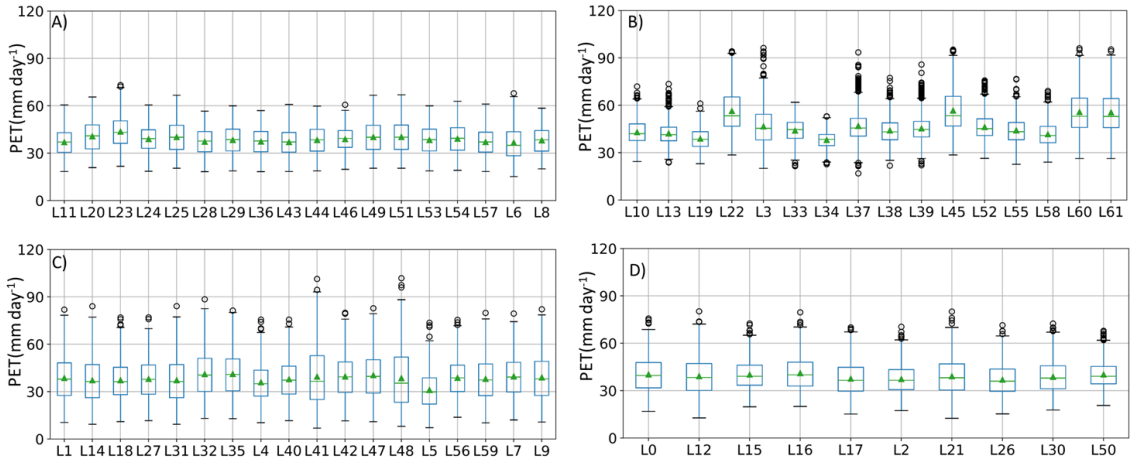

Figura 3. Valores de evapotranspiração potencial das localidades produtoras de cana-de-açúcar dos (A) grupo 1, (B) grupo 2, © grupo 3 e (D) grupo 4. Legenda: L0: Senador Guiomard-AC; L1: Plácido de Castro-AC; L2: Atalaia-AL; L3: Coruripe-AL; L4: Marechal Deodoro-AL; L5: Matriz de Camaragibe-AL; L6: Passo de Camaragibe-AL; L7: Rio Largo-AL; L8: São Luís do Quitunde-AL; L9: Campo Alegre-AL; L10: São Miguel dos Campos-AL; L11: Ipixuna -AM; L12: Caravelas- BA; L13: Juazeiro- BA; L14: Paracuru-CE; L15: Presidente Kennedy- ES; L16: Itumbiara-GO; L17: Santa Helena de Goiás-GO; L18: Timon-MA; L19:Iturama- MG; L20: Maracaju-MS; L21: Aparecida do Taboado-MS; L22: Naviraí- MS; L23: Sidrolândia-MS; L24: Nova Andradina-MS; L25: Itaquiraí-MS; L26: Rio Brilhante- MS; L27: Diamantino-MT; L28: Cáceres- MT; L29: São José do Rio Claro-MT; L30: Tangará da Serra- MT; L31: Jaciara- MT; L32: Juripiranga-PB; L33: Rio Tinto-PB; L34: Gameleira- PE; L35: Rio Formoso-PE; L36: Sirinhaém-PE; L37: José de Freitas- PI; L38: Teresina-PI; L39: União- PI; L40: Cambará-PR; L41: Colorado-PR; L42: Cruzeiro do Oeste-PR; L43: Jacarezinho-PR; L44: Paranacity-PR; L45: Porecatu-PR; L46: Rondon-PR; L47: Tapejara-PR; L48: Porto Xavier-RS; L49: Roque Gonzales-RS; L50: Japoatã-SE; L51: Laranjeiras-SE; L52: Pacatuba-SE; L53: Rosário do Catete-SE; L54: Santo Amaro das Brota-SE; L55: Luís Antônio-SP; L56: Guararapes-SP; L57: Batatais-SP; L58: Jaboticabal-SP; L59: Jaú-SP; L60: Morro Agudo-SP; L61: Peixe-TO.

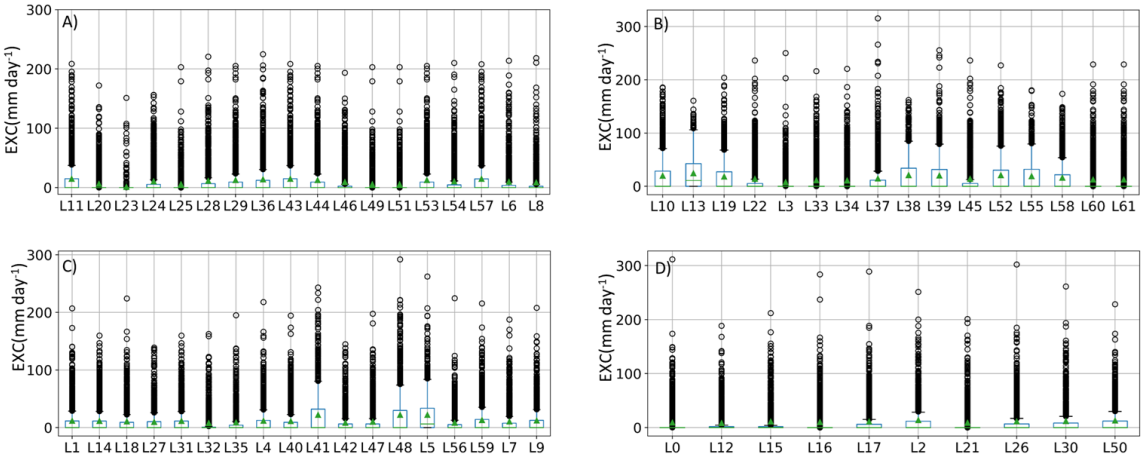

Figura 4. Excedente hídrico das localidades produtoras de cana-de-açúcar dos (A) grupo 1, (B) grupo 2, C) grupo 3 e (D) grupo 4. Legenda: L0: Senador Guiomard-AC; L1: Plácido de Castro-AC; L2: Atalaia-AL; L3: Coruripe-AL; L4: Marechal Deodoro-AL; L5:

Matriz de Camaragibe-AL; L6: Passo de Camaragibe-AL; L7: Rio Largo-AL; L8: São Luís do Quitunde-AL; L9: Campo Alegre-AL; L10: São Miguel dos Campos-AL; L11: Ipixuna -AM; L12: Caravelas- BA; L13: Juazeiro- BA; L14: Paracuru-CE; L15: Presidente Kennedy- ES; L16: Itumbiara-GO; L17: Santa Helena de Goiás-GO; L18: Timon-MA; L19:Iturama- MG; L20: Maracaju-MS; L21: Aparecida do Taboado-MS; L22: Naviraí-MS; L23: Sidrolândia-MS; L24: Nova Andradina-MS; L25: Itaquiraí-MS; L26: Rio Brilhante- MS; L27: Diamantino-MT; L28: Cáceres- MT; L29: São José do Rio Claro-MT; L30: Tangará da Serra- MT; L31: Jaciara- MT; L32: Juripiranga-PB; L33: Rio Tinto-PB; L34: Gameleira- PE; L35: Rio Formoso-PE; L36: Sirinhaém-PE; L37: José de Freitas-PI; L38: Teresina-PI; L39: União- PI; L40: Cambará-PR; L41: Colorado-PR; L42: Cruzeiro do Oeste-PR; L43: Jacarezinho-PR; L44: Paranacity-PR; L45: Porecatu-PR; L46: Rondon-PR; L47: Tapejara-PR; L48: Porto Xavier-RS; L49: Roque Gonzales-RS; L50: Japoatã-SE; L51: Laranjeiras-SE; L52: Pacatuba-SE; L53: Rosário do Catete-SE; L54: Santo Amaro das Brota-SE; L55: Luís Antônio-SP; L56: Guararapes-SP; L57: Batatais-SP; L58: Jaboticabal-SP; L59: Jaú-SP; L60: Morro Agudo-SP; L61: Peixe-TO.

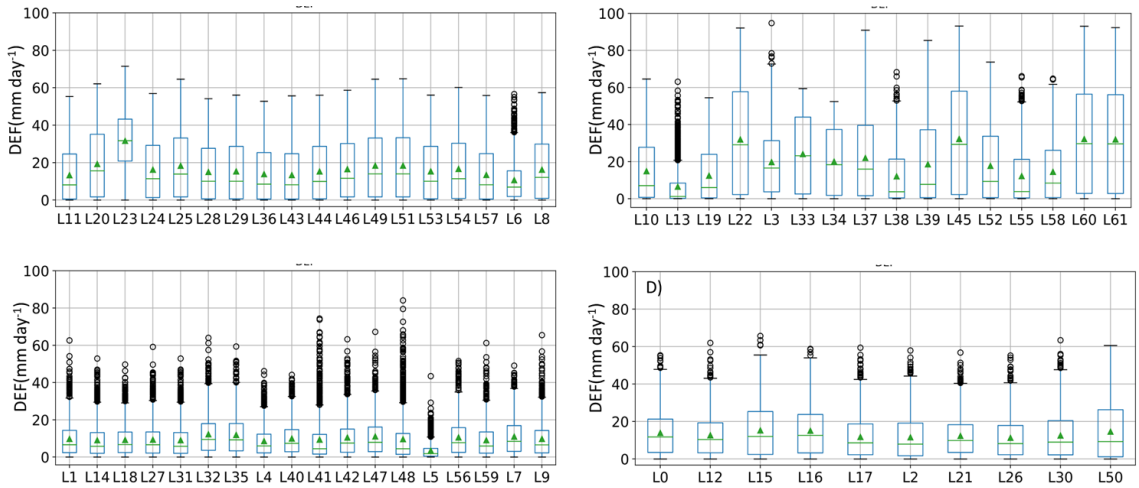

Figura 5. Deficiência hídrica das localidades produtoras de cana-de-açúcar dos (A) grupo 1, (B) grupo 2, C) grupo 3 e (D) grupo 4. Legenda: L0: Senador Guiomard-AC; L1: Plácido de Castro-AC; L2: Atalaia-AL; L3: Coruripe-AL; L4: Marechal Deodoro-AL; L5: Matriz de Camaragibe-AL; L6: Passo de Camaragibe-AL; L7: Rio Largo-AL; L8: São Luís do Quitunde-AL; L9: Campo Alegre-AL; L10: São Miguel dos Campos-AL; L11: Ipixuna -AM; L12: Caravelas- BA; L13: Juazeiro- BA; L14: Paracuru-CE; L15: Presidente Kennedy- ES; L16: Itumbiara-GO; L17: Santa Helena de Goiás-GO; L18: Timon-MA; L19:Iturama- MG; L20: Maracaju-MS; L21: Aparecida do Taboado-MS; L22: Naviraí-MS; L23: Sidrolândia-MS; L24: Nova Andradina-MS; L25: Itaquiraí-MS; L26: Rio Brilhante- MS; L27: Diamantino-MT; L28: Cáceres- MT; L29: São José do Rio Claro-MT; L30: Tangará da Serra- MT; L31: Jaciara- MT; L32: Juripiranga-PB; L33: Rio Tinto-PB; L34: Gameleira- PE; L35: Rio Formoso-PE; L36: Sirinhaém-PE; L37: José de Freitas-PI; L38: Teresina-PI; L39: União- PI; L40: Cambará-PR; L41: Colorado-PR; L42: Cruzeiro do Oeste-PR; L43: Jacarezinho-PR; L44: Paranacity-PR; L45: Porecatu-PR; L46: Rondon-PR; L47: Tapejara-PR; L48: Porto Xavier-RS; L49: Roque Gonzales-RS; L50: Japoatã-SE; L51: Laranjeiras-SE; L52: Pacatuba-SE; L53: Rosário do Catete-SE; L54: Santo Amaro das Brota-SE; L55: Luís Antônio-SP; L56: Guararapes-SP; L57: Batatais-SP; L58: Jaboticabal-SP; L59: Jaú-SP; L60: Morro Agudo-SP; L61: Peixe-TO.

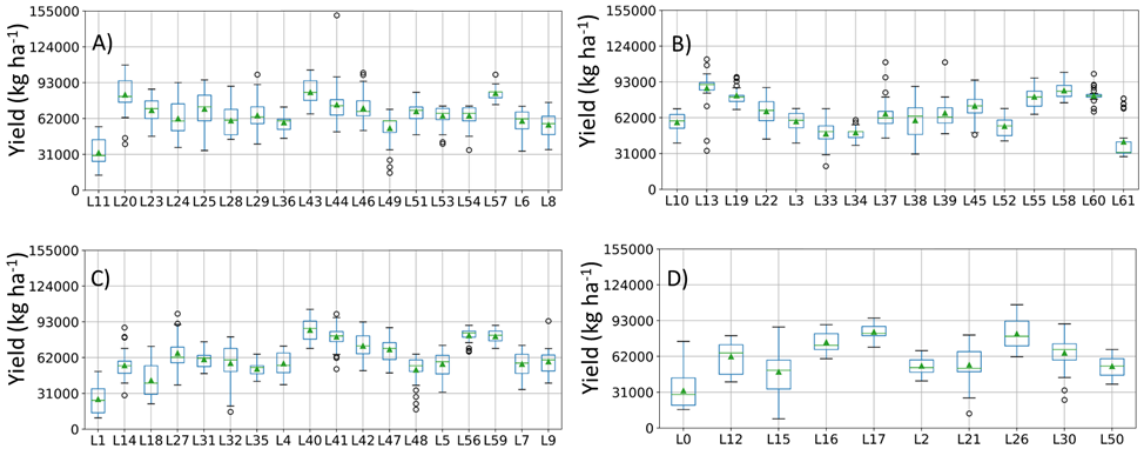

Figura 6. Produtividade de cana-de-açúcar das localidades produtoras de cana-de-açúcar dos (A) grupo 1, (B) grupo 2, C) grupo 3 e (D) grupo 4. Legenda: L0: Senador Guiomard-AC; L1: Plácido de Castro-AC; L2: Atalaia-AL; L3: Coruripe-AL; L4: Marechal Deodoro-AL; L5: Matriz de Camaragibe-AL; L6: Passo de Camaragibe-AL; L7: Rio Largo-AL; L8: São Luís do Quitunde-AL; L9: Campo Alegre-AL; L10: São Miguel dos Campos-AL; L11: Ipixuna -AM; L12: Caravelas- BA; L13: Juazeiro- BA; L14: Paracuru-CE; L15: Presidente Kennedy- ES; L16: Itumbiara-GO; L17: Santa Helena de Goiás-GO; L18: Timon-MA; L19:Iturama- MG; L20: Maracaju-MS; L21: Aparecida do Taboado-MS; L22: Naviraí- MS; L23: Sidrolândia-MS; L24: Nova Andradina-MS; L25: Itaquiraí-MS; L26: Rio Brilhante- MS; L27: Diamantino-MT; L28: Cáceres- MT; L29: São José do Rio Claro-MT; L30: Tangará da Serra- MT; L31: Jaciara- MT; L32: Juripiranga-PB; L33: Rio Tinto-PB; L34: Gameleira- PE; L35: Rio Formoso-PE; L36: Sirinhaém-PE; L37: José de Freitas- PI; L38: Teresina-PI; L39: União- PI; L40: Cambará-PR; L41: Colorado-PR; L42: Cruzeiro do Oeste-PR; L43: Jacarezinho-PR; L44: Paranacity-PR; L45: Porecatu-PR; L46: Rondon-PR; L47: Tapejara-PR; L48: Porto Xavier-RS; L49: Roque Gonzales-RS; L50: Japoatã-SE; L51: Laranjeiras-SE; L52: Pacatuba-SE; L53: Rosário do Catete-SE; L54: Santo Amaro das Brota-SE; L55: Luís Antônio-SP; L56: Guararapes-SP; L57: Batatais-SP; L58: Jaboticabal-SP; L59: Jaú-SP; L60: Morro Agudo-SP; L61: Peixe-TO.

CAPÍTULO 3 – Algoritmos para previsão da produtividade do algodão utilizando parâmetros climáticos no Brasil

RESUMO - As previsões acuradas da produtividade do algodão são de grande interesse para o desenvolvimento do mercado, aumentando a sustentabilidade do setor mundialmente. A maioria dos modelos de previsões de produtividade de cultivos não realizam previsões, mas sim estimações. Esses modelos apresentam grandes desvios quando não incluem a variabilidade climática, sugerindo que há espaço para uma nova metodologia de previsão. Os modelos de Machine Learning (ML) têm se mostrado muito robustos em previsões de variáveis de sistemas complexos, como exemplo, nos sistemas agrícolas. Portanto, os objetivos desse estudo foram: 1) Avaliar a influência dos elementos climáticos decendiais na produtividade do algodão nas principais regiões produtoras do Brasil, 2) Prever a produtividade do algodão usando Regressão linear múltipla (RLM), KNeighborsRegressor (KNN), Random Forest Regressor (RFT), Redes Neurais Artificiais - Multi-layer Perceptron (MLP), Gradient Boosting for regression (BOO) and, Extra-trees regressor (TREE) em função da variação dos elementos climáticos, 3) Calibrar modelos de machine learning para prever a produtividade do algodão baseados em dados de clima, e 4) Testar os modelos de previsão com dados independentes e mapear a produtividade do algodão prevista do modelo mais acurado. O trabalho foi realizado utilizando séries de dados de produtividade de algodão e climáticos das 18 localidades com maior produção do centro-oeste do Brasil correspondendo a uma área total de 1924000 ha. Os dados climáticos foram temperatura média do ar (T, º C), precipitação pluvial (P, mm), evapotranspiração potencial (PET, mm), armazenamento de água no solo (STO, mm), deficiência hídrica (DEF, mm) e excedente hídrico (EXC, mm). Em todas as análises a variável dependente foi à produtividade anual do algodão e as variáveis independentes foram os elementos climáticos do plantio ao florescimento dos cultivos de algodão. A influência dos elementos climáticos na produtividade do algodão foi verificada pela análise de correlação de Pearson. A previsão de produtividade em função dos elementos climáticos foi realizada usando modelos não lineares logísticos com quatro parâmetros ajustados por ordinary least squares. Os melhores algoritmos foram selecionados pelos índices de acurácia, precisão e tendência. Como resultado, os elementos climáticos que mais influenciaram a produtividade do algodão nas principais regiões produtoras do Brasil foram PET e o STO. Os modelos não lineares evidenciam que a produtividade do algodão tem tendência sigmoide em função do acúmulo de P, PET, STO e EXC durante o ciclo. É possível prever a produtividade do algodão com antecipação para as principais regiões produtoras do Brasil usando algoritmos de Machine learning. Os modelos TREE tiveram maior desempenho na previsão da produção de algodão utilizando dados climáticos do plantio até o florescimento. Com isso é possível ter uma antecipação média em torno de 80 dias, possibilitando o produtor um tempo hábil para planejar suas atividades como colheita e estratégias de venda.

PALAVRAS-CHAVE: Artificial intelligence, Random Forest; Crop modelling; Water balance; Deep learning; Bigdata

ALGORITHMS FOR FORECASTING COTTON YIELD BASED ON CLIMATIC PARAMETERS IN BRAZIL

ABSTRACT - Accurate forecasts of cotton yield is of great interest for the development of the market, increasing the sustainability of the sector worldwide. Thus, the objectives of this study were: 1) to evaluate the influence of climatic elements on cotton yield in Brazil, 2) to predict cotton yield using machine learning algorithms based on climatic elements, 3) to calibrate and test machine learning models to forecast cotton yield based on climate data, and 4) to interpolate the estimated cotton yield of the most accurate model. The cotton yield forecast as a function of climatic elements was performed using machine learning algorithms with four parameters adjusted by ordinary least squares. The models show that cotton yield has a sigmoid trend due to the accumulation of P, PET, STO, and EXC during the cycle. It is possible to forecast cotton yield for the main producing regions of Brazil using Machine learning algorithms. Extra-trees regressor models performed better in forecasting cotton yield using climatic data from planting to flowering. Therefore, it is possible to have average anticipation of around 80 days, allowing the producer time to plan his activities such as harvest and sales strategies.

Keywords: Artificial intelligence; Random Forest; Crop modelling; Water balance; Deep learning; Bigdata

3.1 Introdução

O algodão (*Gossypium hirsutum* L.) é a cultura de fibra mais importante do mundo devido ao seu grande potencial econômico e social (Feng et al., 2020). O algodão é cultivado em mais de 100 países em uma área de 33,2 milhões de hectares (Hussain et al., 2020). A produção mundial na safra de 2017/2018 foi de 26,931 milhões de toneladas (Conab, 2019). O Brasil é o quarto maior produtor de algodão depois da Índia, China e dos Estados Unidos (Barros et al., 2020), o país possui 3.590 municípios com condições adequadas para o plantio do algodoeiro (Assad et al., 2013), no entanto, a variabilidade climática tem reduzido esse número.

As atividades agrícolas são sensíveis aos elementos climáticos (Silva et al., 2020; Chou et al., 2019), com o algodão não é diferente (Li et al., 2020). O algodoeiro é uma planta de clima tropical (Hussain et al., 2020; Iqbal et al., 2017) e necessita precipitações anuais entre 500 mm e 1500 mm e bem distribuídas ao longo do ciclo. A temperatura do ar média adequada deve ser entre 20 ºC e 30 ºC, umidade relativa de 70% e insolação em 2.500 horas luz ano^{-1} (Andrade Júnior et al., 2009).

O algodão é uma cultura resistente à seca, mas a sua produtividade é afetada negativamente pelo estresse hídrico (Chen et al., 2019). Gridd-Papp (1965), Marur (1991) e Wrege et al. (2000) relataram que os períodos secos são mais prejudiciais ao algodão durante o estabelecimento da planta e o primeiro mês de floração. As alterações nesses padrões climáticas podem impactar negativamente na produção de algodão no mundo e no Brasil.

A disponibilidade de água no Brasil depende em grande parte da variação climática (Marengo et al., 2017). As alterações climáticas caracterizadas pelo aumento da temperatura do ar e mudanças nos padrões de precipitação afetam os cultivos agrícolas, além de atrapalhar os sistemas de previsão de safra atuais do Brasil.

No Brasil, há a Companhia Nacional de Abastecimento (Conab) e Instituto Brasileiro de Geografia e Estatística (IBGE) que fornecem dados sobre status das lavouras e previsões de safras baseadas em pesquisas de campo (Schwalbert et al., 2020), entretanto, essas previsões têm baixa acurácia devido a metodologia não conseguir acompanhar as alterações climáticas. Uma

alternativa, é a utilização de modelagem como machine learning para realizar essas previsões de produtividade.

Todas essas relações existentes entre os elementos climáticas e a variabilidade da produtividade dos cultivos podem ser simuladas com acurácia por meio de modelos agrometeorológicos (Aparecido et al., 2020) usando algoritmos de machine learning (Sahoo et al., 2017). Machine learning é um método que trabalha com análise de dados e busca automatizar a construção de modelos analíticos (Shekoofa et al., 2014; LI et al., 2016). As técnicas de machine learning são muitas promissoras para análises mais rápidas, eficientes e acuradas de bigdata (Rehman et al., 2019). As Técnicas de computação, como exemplo o aprendizado de máquina, é o novo paradigma em utilização na área agrícola para previsão de produtividade de cultivos (Elavarasan et al., 2018).

Há trabalhos que utilizam machine learning para fins de previsão de produtividade na literatura, por exemplo, redes neurais artificiais para a previsão da produtividade de milho (Singh, 2008), random forest para a previsão acurada da produtividade de cana-de-açúcar (Everingham et al., 2016) e previsão de produção de soja por satélite com integração de aprendizado de máquina e dados climáticos no sul do Brasil (Schwalbert et al., 2020). No entanto, existe carência de trabalhos sobre previsão da produtividade do algodão utilizando técnicas de machine learning.

Portanto, os nossos objetivos com este estudo foram: 1) Avaliar a influência dos elementos climáticos decendiais na produtividade do algodão nas principais regiões produtoras do Brasil, 2) Prever a produtividade do algodão usando Regressão linear múltipla (RLM), KNeighborsRegressor (KNN), Random Forest Regressor (RFT), Redes Neurais Artificiais - Multi-layer Perceptron (MLP), Gradient Boosting for regression (BOO) and, Extra-trees regressor (TREE) em função da variação dos elementos climáticos, 3) Calibrar e testar modelos de machine learning para prever a produtividade do algodão baseados em dados de clima, e 4) Interpolar da produtividade do algodão prevista do modelo mais acurado.

3.2 Material e Métodos

3.2.1 Região estudada

O estudo foi realizado no centro-oeste do Brasil, a maior região produtora de algodão do país. Nós utilizamos séries históricas de produtividade e dados climáticos das principais localidades produtoras de algodão da região estudada (Tabela 1 e Figura 1).

Tabela 1. Características Geográficas das principais localidades produtoras de algodão do Brasil.

LOCALIDADES	ESTADO	Longitude	Latitude	Altitude	Clima*
Acreúna	GO	-50.37	-17.39	531	B2rB'4a'
Cachoeira-Dourada	GO	-49.5	-18.51	479	B2rB'4a'
Goiatuba	GO	-49.35	-18.01	620	B2rB'4a'
Itumbiara	GO	-49.21	-18.42	553	B2rB'4a'
Montividiu	GO	-51.17	-17.44	835	B3rB'4a'
Morrinhos	GO	-49.1	-17.73	716	B2rB'4a'
Paraúna	GO	-50.44	-16.94	670	B3rB'4a'
Rio-Verde	GO	-50.92	-17.79	765	B3rB'4a'
Maracaju	MS	-55.16	-21.61	469	B2rB'4a'
Naviraí	MS	-54.19	-23.06	308	B2rB'4a'
São-Gabriel-Do-Oeste	MS	-54.56	-19.39	504	B1rA'a'
Sidrolândia	MS	-54.96	-20.93	439	B2rA'a'
Campo-Novo-Do-Parecis	MT	-57.89	-13.67	529	B3rA'a'
Diamantino	MT	-56.44	-14.4	476	B2rA'a'
Itiquira	MT	-54.15	-17.2	450	B2rA'a'
Novo-São-Joaquim	MT	-53.01	-14.9	467	B2wB'4a'
Pedra Preta	MT	-54.47	-16.62	248	B1rA'a'
Primavera-Do-Leste	MT	-54.34	-15.52	647	B3rB'4a'

* Classificação Climática pelo método de Thornthwaite (1948).

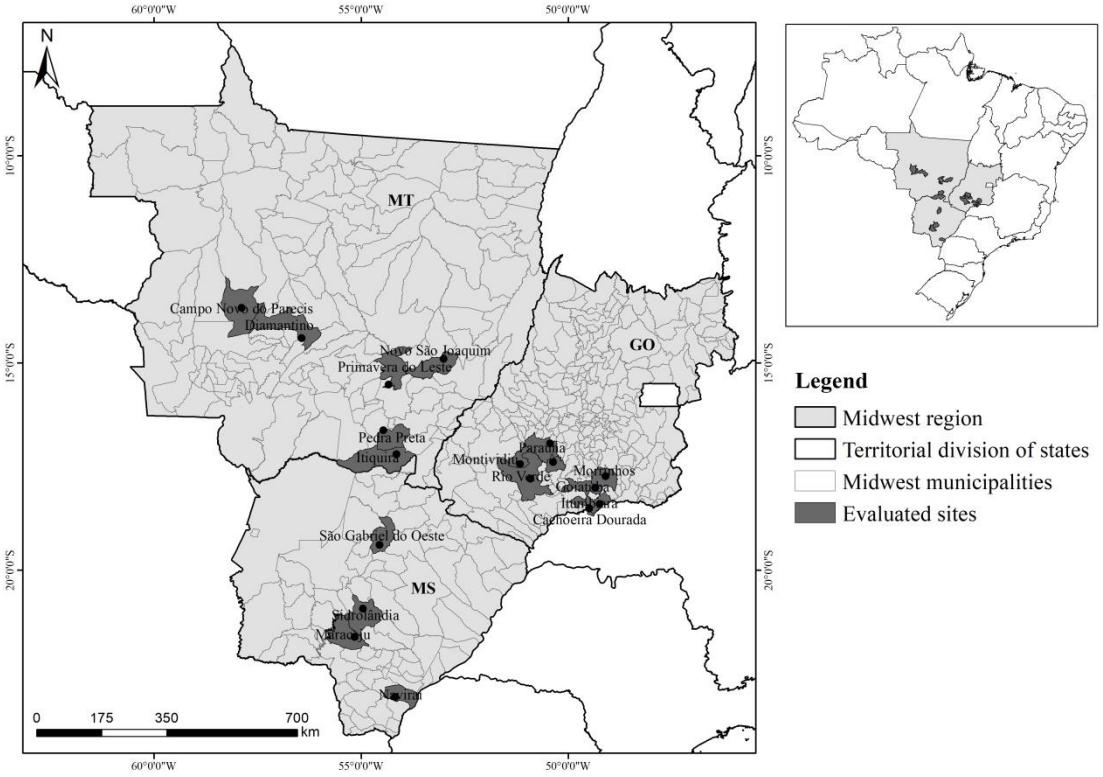

FIGURA 1. Localidades produtoras de algodão utilizadas neste estudo. Símbolos: estados do Mato Grosso (MT), Goiás (GO), Mato Grosso do Sul (MS).

3.2.2 Banco de dados

A série histórica da produtividade do algodoeiro foi do período de 1989 a 2017. Os dados foram obtidos na plataforma SIDRA do Instituto Brasileiro de Geografia e Estatística (IBGE, 2020).

Os dados de temperatura média do ar (Tm, º C), precipitação pluvial (P, mm) e evapotranspiração potencial (ETP, mm) do período de 1983-2018 foram obtidos em escala diária na plataforma National Aeronautics and Space Administration/Prediction of World Wide Energy Resources - NASA/POWER (Stackhouse et al., 2016). A plataforma fornece informações meteorológicas em grids com resolução espacial de 1°, correspondendo a aproximadamente 110,57 km.

3.2.3 Evapotranspiração potencial

A Evapotranspiração Potencial (ETP) foi calculada utilizando o método Penman e Monteith (Allen et al., 1998) na escala diária e depois estratificada na escala decendial, descrito nas Equações 1 a 4.

$$ETP = \frac{0.408 \times s \times (Rn - G) + \frac{\gamma \times 900 \times U_2 \times (es - ea)}{T + 273}}{s + \gamma \times (1 + 0.34 \times U_2)} \quad (1)$$

$$s = \frac{4098 \times es}{(T + 273)^2} \quad (2)$$

$$ea = \frac{UR \times es}{100} \quad (3)$$

$$es = 0.6108 \times e^{\frac{17.27 \times T}{237.3 + T}} \quad (4)$$

em que Rn é a radiação líquida (MJ m^{-2} d^{-1}); UR é umidade relativa; G é o fluxo de calor no solo igual a 0 MJ m^{-2} d^{-1}; T é a temperatura do ar (°C); γ é a constante psicométrica igual a 0,063 kPa $°C^{-1}$; s é a declividade da curva da pressão de vapor *versus* temperatura (kPa $°C^{-1}$) ; U_2 é a velocidade do vento (m s^{-1}) na altura de 2 m; es é a pressão da saturação de vapor (kPa); ea é a pressão de vapor atual (kPa).

3.2.4 Balanço hídrico climatológico

O armazenamento de água no solo (STO, mm), deficiência hídrica (DEF, mm) e excedente hídrico (EXC, mm) foram calculados na escala decendial para todas as localidades estudadas, segundo a metodologia de Thornthwaite e Mather (1955) (Equações 5-10). Uma capacidade média de retenção de água no solo de 60 mm foi assumida para todos os locais, já que existe uma grande variabilidade de solos nesses locais (BRASIL, 1981).

$$if\ (P - PET_i < 0 = \begin{cases} NAC_i = NAC_{i-1} + (P - PET)_i \\ STO_i = WCe^{\frac{(\mathrm{NAC}_i)}{WC}} \end{cases} \quad (5)$$

$$if\ (P - PET)_i \geq 0 = \begin{cases} STO_i = (P - PET)_i + STO_{i-1} \\ NAC_i = WCln^{\frac{(\mathrm{STO}_i)}{WC}} \end{cases} \quad (6)$$

$$ALT_i = STO_i - STO_{i-1} \quad (7)$$

$$AET_i = \begin{cases} P + |ALT_i| & ,if\ ALT < 0 \\ PET_i & ,if\ ALT \geq 0 \end{cases} \quad (8)$$

$$DEF = PET - AET \quad (9)$$

$$EXC_i = \begin{cases} 0 & ,if\ WC < 0 \\ (P - PET)_i - ALT_i & ,if\ WC = 0 \end{cases} \quad (10)$$

em que, AET is actual evapotranspiration (mm); ALT is soil water storage of the current month – soil water storage of the preceding month (mm), *i* is the monthly period, NAC is accumulated negative; P is rainfall (mm); PET is potential evapotranspiration (mm); WC is available water capacity (mm); STO is soil water storage (mm); EXC is water surplus at the soil-plant-atmosphere system (mm) and, DEF is water deficiency at the soil-plant-atmosphere system (mm).

3.2.5 Produtividade de algodão x Modelos não lineares

A análise de regressão não linear permite verificar a tendência da variação da produtividade do algodão em função dos elementos climáticos. Utilizamos análise de regressão com sigmoidal linear e não linear (logística) com quatro parâmetros (Equação 11). A produtividade foi a variável dependente e as variáveis climáticas as variáveis independentes (Tabela 2). O método de estimação dos modelos não lineares utilizado foi o dos mínimos quadrados ordinários (OLS) (Draper e Smith, 1980), através do sistema de otimização "gradiente reduzido generalizado" (GRG2) (Lasdon; Waren, 1982).

$$y = y_{max} + \frac{Ymax - Ymin}{1 + (\frac{x}{x_0})^p} \quad (11)$$

em que y é a variável cumulativa estimada; ymax é o ponto máximo da curva (amplitude); Ymin é o ponto mínimo da curva; x0 é o ponto de inflexão (valor X de crescimento máximo) e, pe taxa de crescimento máximo.

TABELA 2. Variáveis climáticas utilizadas na previsão da produtividade do algodão no Brasil.

Variables	Symbol	Unit	Concept
Air temperature	T	º C	index expressing the amount of sensitive heat in the air
Precipitation	P	mm	water that returns from atmosphere to the Earth's surface
Evapotranspiration	ETP	mm	evaporation + transpiration
Water Storage	ARM	mm	water stored in soils

Water deficit	DEF	mm	potential evapotranspiration - real evapotranspiration
Water surplus	EXC	mm	water left over in the rainy season

O clima local determina a produtividade média dos cultivos. Assim, nós aprofundamos essa análise separando as localidades por tipo climático, segundo a classificação climática de Thornthwaite (1948), pois esta classificação resume as condições hídricas normais de uma localidade separando com eficiência regiões de mesmo potencial agrícola.

As correlações de Pearson foram realizadas buscando quantificar com maiores detalhes a relação entre a produtividade do algodão e os elementos meteorológicos (Tabela 2), considerando cada estádio da fenologia do cultivo. A correlação permite a quantificação da relação entre duas variáveis permitindo conhecer a importância de cada elemento meteorológico em relação à produtividade. As correlações foram feitas considerando sete decêndios entre a plantio e o florescimento do algodoeiro. A data do plantio foi 10 de janeiro, pois representa em média grande parte das localidades estudadas.

3.2.6 Algoritmos de Machine learning

Nós utilizamos diferentes metodologias para se realizar a previsão da produtividade do algodão. Em todos os casos a produtividade foi a variável dependente e os elementos meteorológicas (ME) decendiais do plantio ao florescimento foram as variáveis independentes (Tabela 2), totalizando 35 variáveis (5 ME x 7 decêndios). Para todas as metodologias foram separados de maneira aleatória 60% dos dados para o treinamento e os demais 40% utilizados para calibração utilizando a biblioteca scikit-learn (Pedregosa et al., 2011) do Python.

Os algoritmos utilizados para a previsão foram: 1) Regressão linear múltipla (RLM); 2) K Neighbors Regressor (KNN); 3) Random Forest Regressor (RFT); 4) Redes Neurais Artificiais - Multi-layer Perceptron (MLP); 5) Gradient Boosting for regression (BOO) e 6) Extra-trees regressor (TREE).

A RLM é um método comumente utilizado para a previsão de produtividade de cultivos (Torkashvand et al., 2017; Mercante et al., 2010; Biswas et al., 2019; Bhojani et al., 2020). Nós usamos a RLM de “Ridge”, pois evita o problema de multicolinearidade sem ter que excluir variáveis regressoras, assim não existem perdas de informações.

O algoritmo KNN é um método não paramétrico usado para prever o rendimento das culturas (Gonzalez-Sanchez, Frausto-Solis e Ojeda-Bustamante 2014; Hansen e Indeje 2004; Shakil Ahamed et al. 2015). É uma técnica simples e facilmente implementada e bastante flexível, encontra um grupo de k amostras (dados de treinamento) mais próximas de amostras desconhecidas (dados de teste). Destas k amostras, as amostras desconhecidas são determinadas calculando a média da variável de resposta. O parâmetro de k é determinado usando o método de elbow criterion. No KNN foi identificado os três vizinhos mais próximos e a métrica utilizado para o cálculo das distâncias foi a distância euclidiana.

No RFT é uma técnica utilizada para previsão de cultivos (Gyamerah et al., 2020; Feng et al., 2020; Schwalbert et al 2020; Aparecido et al., 2020, Everingham et al., 2016), foi criado uma floresta de modo aleatório na qual foi utilizado uma combinação (ensemble) de 1000 árvores de decisão, para que ocorresse uma previsão da produtividade em função do clima.

A Rede Neural Artificial empregada foi a Multi-layer Perceptron (MLP), é comumente aplicada para prever produtividade das culturas (Akbar et al. 2018; Kaul et al., 2005; Torkashvand, Ahmadi e Nikravesh, 2017), com três camadas de neurônios, sendo que em cada uma dessas camadas foi empregado 10 neurônios. A função de ativação foi a sigmoide. O treinamento da MLP foi usando backpropagation com taxa de aprendizagem igual a 60% buscando a minimização de MAPE e aumento do R2

Gradient Boosting for regression (BOO) é um método que ainda não foi utilizado para prever a produtividade das culturas. É uma técnica de aprendizado de máquina para problemas de regressão e/ou classificação, que produz um modelo de previsão na forma de um conjunto de modelos de previsão, geralmente árvores de decisão. Essa técnica combina um conjunto de base-learners para estimar dependências estatísticas complexas (Thomas et al., 2017). Um base-learner, por si só, normalmente não será suficiente para ajustar

um modelo estatístico de bom desempenho aos dados, mas uma combinação aprimorada de um grande número pode competir com outros algoritmos de ponta em muitas tarefas, por exemplo, classificação (Li, 2012) ou reconhecimento de imagem (Opelt et al., 2004). Um problema remanescente do aumento do Gradient Boosting é a tendência dos algoritmos de aumento para selecionar um número relativamente alto de variáveis falso-positivas e incluir muitas covariáveis não-informativas em um modelo de regressão estatística (Thomas et al., 2017).

TREE são métodos de aprendizado de máquinas supervisionado não paramétricos, muito utilizados em tarefas de classificação e regressão, mas esse método ainda não é utilizado para previsão de produtividade de cultivos. O TREE permite classificar/prever variáveis em um número finito de classes, através de regras hierárquicas e da sua divisão em grupos e obtendo uma visão real da natureza do processo (Quinlan, 1983). Para prever a produtividade do algodão foi utilizado 1000 estimadores neste método.

3.2.7 Avaliação dos algoritmos

A seleção do melhor algoritmo calibrado foi realizada utilizando os seguintes índices estatísticos: 1) correlação de Pearson (r); 2) Coeficiente de determinação ajustado (R^2 adj); 3) erro quadrático médio (MSE); 4) Raiz quadrada do erro-médio (RMSE); 5) Média Percentual Absoluta do Erro (MAPE) (Equações 12 a 16). E para obter a maior confiabilidade nas regressões foram selecionadas apenas as regressões significativas pelo teste F a 5% de probabilidade.

$$r = \frac{\sum_{i=1}^{n}(Yobs_i - \overline{Yobs}) \times (Yest_i - \overline{Yest})}{\sqrt{\sum_{i=1}^{n}(Yobs_i - \overline{Yobs})^2} \times \sqrt{\sum_{i=1}^{n}(Yest_i - \overline{Yest})^2}} \quad (12)$$

$$R^2\ adj = \left[1 - \frac{(1-R^2)\times(n-1)}{N-k-1}\right] \quad (13)$$

$$MSE = \frac{\sum_{i=1}^{N}\left(Y_{obs_i} - Y_{est_i}\right)^2}{N} \quad (14)$$

$$RMSE = \sqrt{\frac{\sum_{i=1}^{N}\left(Y_{obs_i} - Y_{est_i}\right)^2}{N}} \quad (15)$$

$$MAPE(\%) = \frac{\sum_{i=1}^{n}\left(\left|\frac{Yest_i - Yobs_i}{Yobs_i}\right| \times 100\right)}{N} \quad (16)$$

em que, $Yest_i$: Produtividade interpolada; $Yobs_i$: Produtividade observada; N: número de dados e k: número de variáveis independentes na regressão.

Nós fizemos a espacialização da produtividade real, estimada e o desvio (real - estimada) para todas as regiões produtoras do Brasil usando o sistema de informação geográfica. O método de interpolação utilizado foi a krigagem (Krige, 1951), com o modelo esférico, um vizinho e resolução de 1° (111 km).

3.3 Resultados e Discussão

A região centro-oeste demonstrou grande variabilidade espacial nas condições térmicas e hídricas (Figura 2). O Mato Grosso foi o estado mais quente da região e de maior variabilidade (9.8 °C), no entanto, as temperaturas médias do ar da região estão dentro da faixa adequada do algodoeiro para expressar seu potencial genético, ou seja, temperaturas médias de 20 °C a 30 °C (Figura 2a).

A maior parte da região centro-oeste está entre 1501 mm e 2000 mm de precipitação anual. Estes valores estão de acordo com diversos autores como Álvares et al. (2013) e Aparecido et al. (2019). A faixa de distribuição de chuva mais homogênea foi no estado de Goiás. Já no estado do Mato Grosso, a distribuição das chuvas vai diminuindo da parte norte para o sul do estado. A precipitação anual necessária para o algodoeiro é entre 500 mm e 1500 mm (Andrade Júnior et al., 2009), sendo a maior porção dessa faixa no estado do MS (Figure 2).

A evapotranspiração potencial mais predominante no centro-oeste foi entre 1110 mm e 1200 mm. O estado de GO apresentou as menores ETP (<1100 mm) na grande parte de seu território. No oeste do estado do MT apresentou os maiores valores de ETP. Os maiores valores de armazenamento hídrico calculados na região centro-oeste foram de 110 mm, sendo que estes valores estão localizados nas partes leste e sul da região estudada. A faixa de ARM

predominante na região centro-oeste foi 61-80 mm, principalmente nos estados do MT e GO (Figure 2).

O norte da região centro-oeste do Brasil é mais seca em comparação ao sul da região. O estado do MS apresentou os menores valores de DEF (<100 mm). Enquanto o estado do MT apresentou maior variabilidade nos valores de deficiência hídrica, sendo que a maior parte desse estado apresentou DEF entre 110-200 mm. A maior parte do território da região centro-oeste apresentou valores de excedente hídrico de 360 mm a 1400 mm. O oeste do estado do Mato Grosso foi mais úmido da região, com ETP entre 1500 mm a 2500 mm (Figure 2).

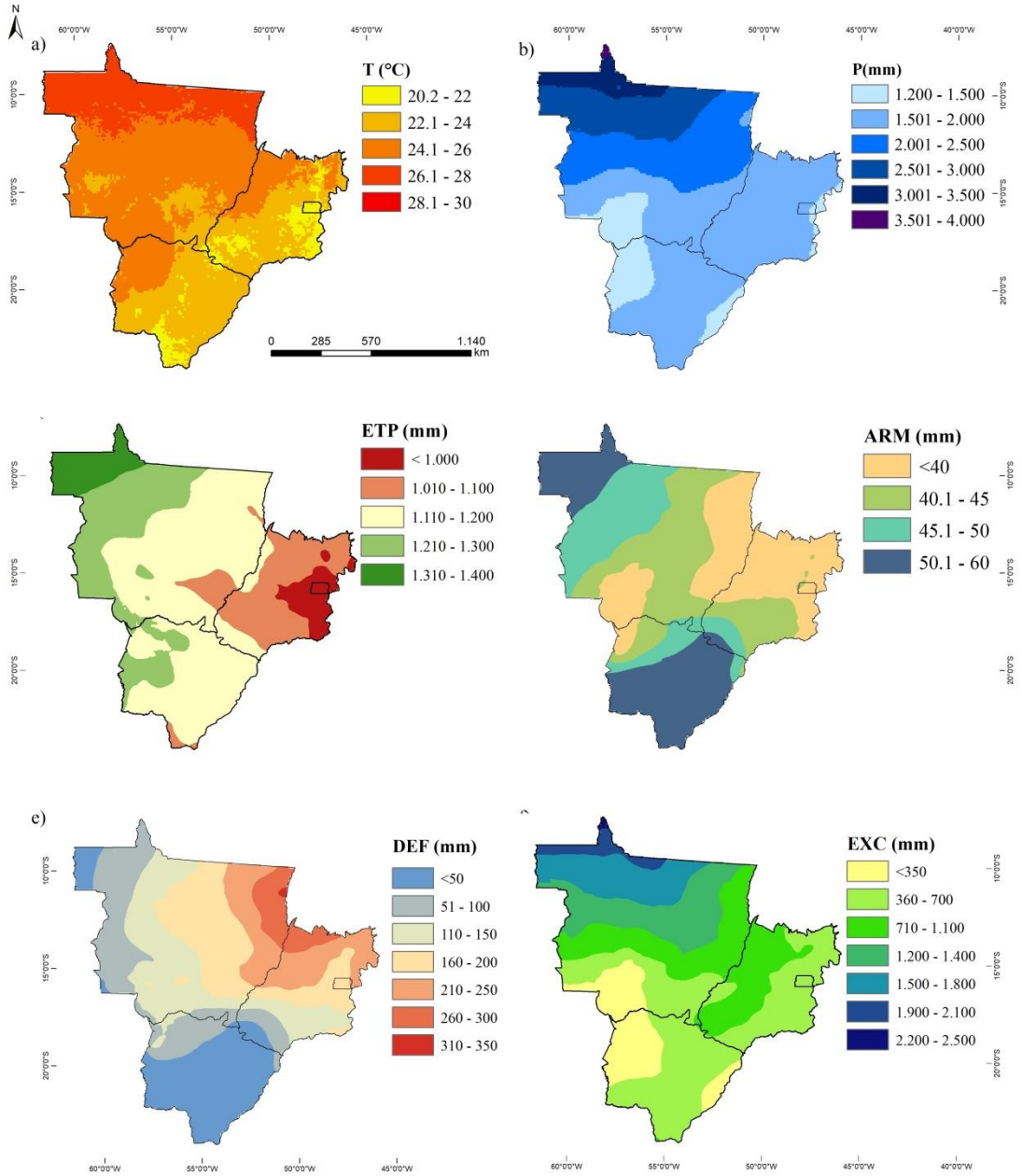

FIGURA 2. Variação espacial dos elementos climáticos nas regiões produtoras de algodão do Brasil.

3.3.1 Produtividade do algodão

A produtividade do algodão na região centro-oeste demonstrou uma grande variabilidade temporal em todo período de 1989 a 2018, chegando a um coeficiente de variação de 33.34% (Figura 3). Toda essa variação é considerada dentro da normalidade, e ocorre devido a diversos fatores, sendo que as condições climáticas é um dos principais deles. É notável que o grande aumento

na média produtiva ocorreu a partir do ano 2000. No período de 1989 a 2000 a produtividade média do algodão era de apenas 1960.41 kg $ha^{-1.}$

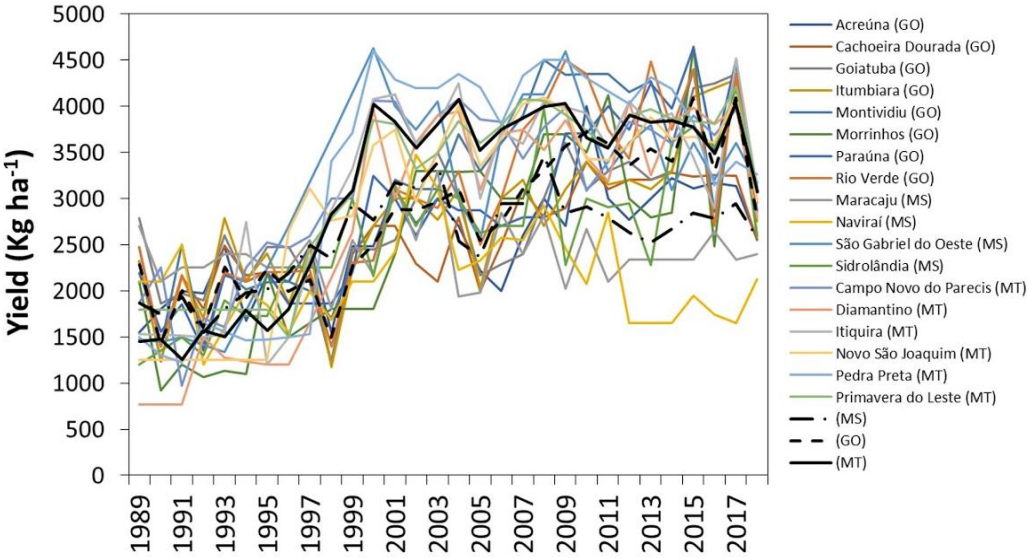

FIGURA 3. Produtividade de algodão de 1989 a 2018 nas principais localidades produtoras do Brasil.

O centro-oeste apresentou uma produtividade média de 2804.91 kg ha^{-1}. O estado do MS teve a menor e o MT a maior variabilidade na produtividade média do algodão no período de 1989-2018 (Figura 4). Esse resultado do MT pode estar relacionado a esse estado apresentar várias faixas de condições térmicas e hídricas no seu território (Figura 2). O MT apresentou a maior produtividade média do algodão em seguida dos estados de GO e MS, sendo 3074.76 kg ha^{-1}, 2769.70 kg ha^{-1} e 2570.26 kg ha^{-1}, respectivamente. Pedra Preta (MT) foi a localidade produtora de algodão na região com maior produtividade média no período estudado (3265.45 kg $ha^{-1)}$, além de apresentar alta variabilidade na produtividade média de algodão. A região centro oeste é a maior produtora de grãos do Brasil e sua economia é totalmente voltada à produção agrícola (Silva e Marujo, 2012).

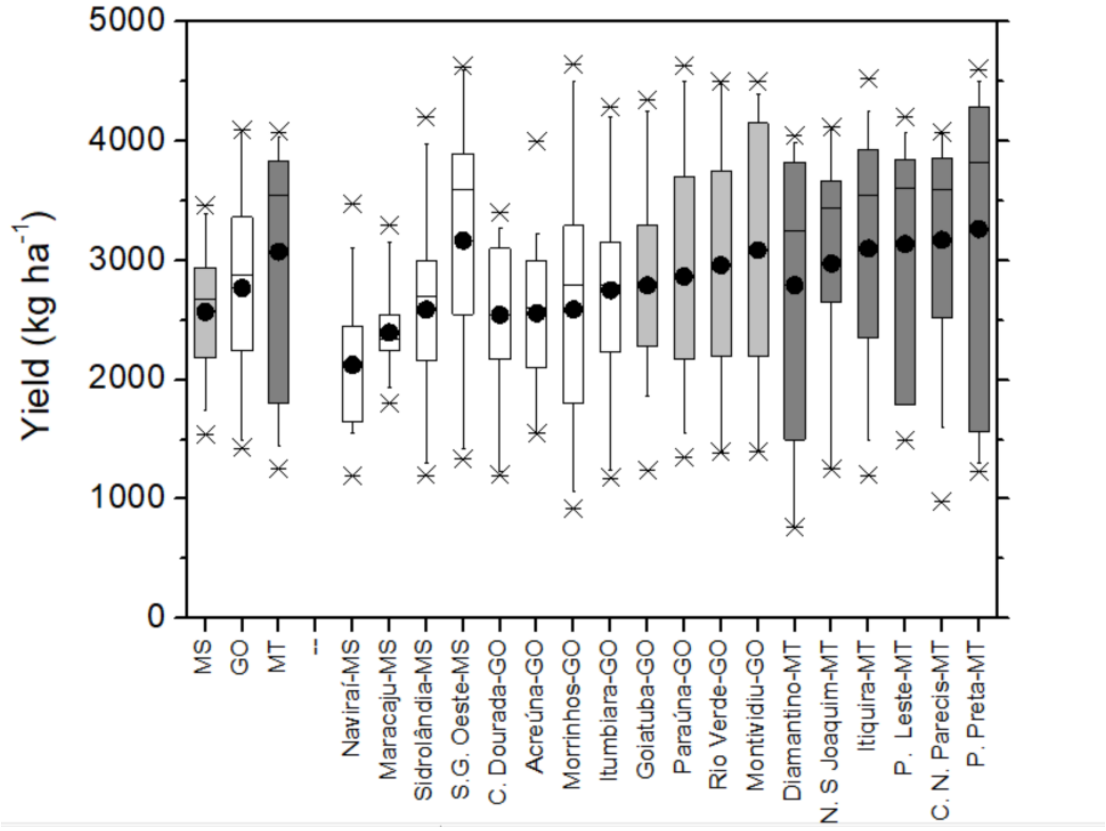

FIGURA 4. Variabilidade da produtividade de algodão de 1989 a 2018 nas principais localidades produtoras do Brasil.

Como uma tentativa de entendimento geral da influência do clima na produtividade do algodão aplicou-se a classificação climática de Thornthwaite (1948) em todos os municípios e nós analisamos a influência das classes na variabilidade da produtividade. Na região centro-oeste foi predominantes seis tipos de clima pelo sistema de classificação climática de Thornthwaite (1948): B1rA'a', B2rA'a', B2rB'4a', B2wB'4a', B3rA'a' e B3rB'4a' (Figure 5). O tipo de clima da região influência no potencial genético das culturas (Aparecido et al., 2016). A produtividade de 3900 kg ha^{-1} ocorreu em 36% dos anos na região centro-oeste e a classe B2rA´a´ foi a mais predominante nos locais com essa produtividade média do algodão (Figura 5B). Essa classe climática é caracterizada por ser úmida, megatérmica e com baixo déficit hídrico (Rahimi et al., 2019). Diamantino (MT), Itiquira (MT) e Sidrolândia (MS) são as localidades em que predominam essa classe climática (Tabela 1). Entre os climas secos a classe B1rA`a` foi a que obteve a maior frequência na produtividade média de 4600 kg ha^{-1}. Outra classe considerada seca B2rB'4a' também evidenciou

elevadas produtividades ao redor de 3160 kg ha^{-1} enquanto as demais classes foram levianas (Figure 5).

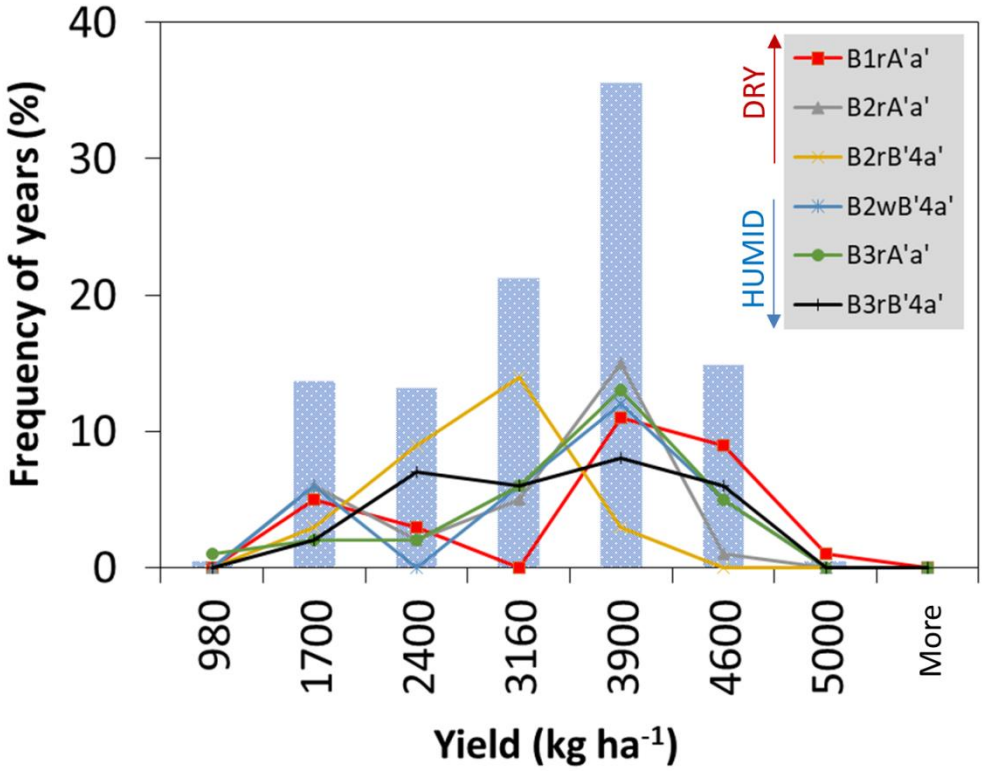

FIGURA 5. Distribuição da produtividade (B) do algodão para cada tipo climático de Thornthwaite (1948).

Os elementos meteorológicos têm influência variada na produtividade do algodão. Por exemplo, com o aumento da chuva, evapotranspiração, armazenamento de água no solo e excedente hídrico durante o ciclo ocorre uma elevação da produtividade do algodão. Já, com o aumento da temperatura do ar e da deficiência hídrica ocorre uma redução da produtividade do algodão (Figure 6). Diversos autores como Martins et al. (2015) and Moreto et al. (2017) destacam que o déficit hídrico é uma variável climática com grande influência na agricultura, e está relacionada com a produtividade e qualidade de diversas culturas agrícolas.

As maiores produtividades médias do algodão ocorrem com temperaturas em torno de 23,7 °C, acima desse valor ocorre um decréscimo da produtividade. O aumento de precipitação no ciclo promove elevação da produtividade média do algodão na região, sendo que a maior produtividade prevista (3149,2 kg ha^{-1}) foi com a precipitação de 700 mm ciclo^{-1}. A produtividade do algodão atingiu sua

produtividade máxima com evapotranspiração potencial de 280 mm acumulados após a semeadura e a partir dessa ETP a produtividade foi igual. A máxima produtividade média do algodão na região centro-oeste foi de 4.010 kg ha^{-1} e ocorre com baixo nível de deficiência hídrica. Diversos autores como Doorenbos e Kassam (1979), Passos et al., (1987) e Batista (2010) destacam que o DEF é prejudicial ao cultivo do algodão.

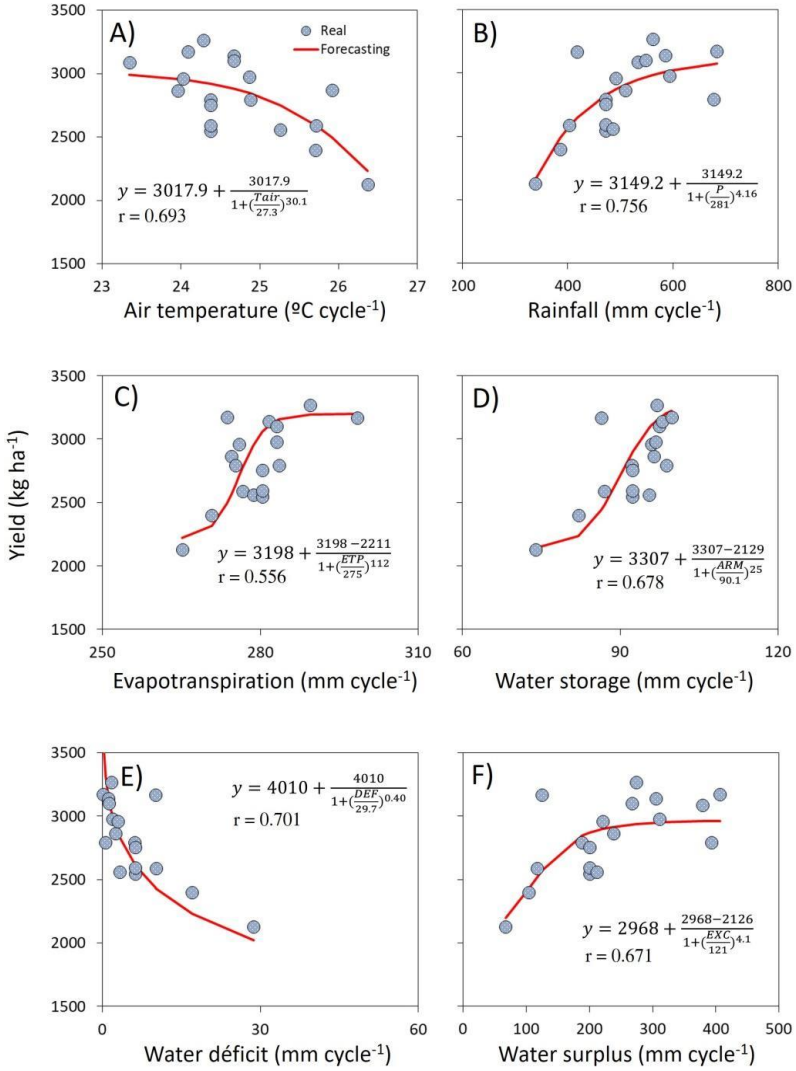

FIGURA 6. Previsão da produtividade do algodão (kg ha^{-1}) por modelo não linear em função de A) temperatura (ºC ciclo^{-1}), B) precipitação (mm ciclo^{-1}), C) evapotranspiração (mm ciclo^{-1}), D) armazenamento de água (mm ciclo^{-1}), E) déficit hídrico (mm ciclo^{-1}) e F) excedente hídrico (mm ciclo^{-1}) acumulado após a semeadura.

3.3.2 Produtividade x Correlação de Pearson

O potencial produtivo do algodão é definido entre o plantio ao florescimento. É neste momento que condições climáticas têm grande influência na produtividade da cultura. Os elementos climáticos demonstraram relações e intensidades distintas com a produtividade do algodão (Figura 7). Por exemplo, a ETP e o ARM tiveram relação direta com a produtividade desde o plantio até o florescimento, o que comprova que manter um ARM elevado proporciona produtividade mais elevadas ao cultivo. No início do crescimento vegetativo (2º decêndio) a influência do ETP+ARM chegam a 66.5% da influência do clima no cultivo (Figura 8). Para as demais variáveis as correlações variaram para cada decênio analisado.

A P demonstrou correlações positivas com a produtividade do plantio até o final da formação dos botões florais (Figure 7), mas, no florescimento a correlação foi negativa (r=0.201), além de ser a maior influência do clima no florescimento (36.4%) (Figure 8). P no florescimento inibem a atividade do inseto abelha que é o polinizador do algodoeiro (Free, 1993; Sanchez Junior e Malerbo-Souza, 2004) atrapalhando a polinização da cultura, o que proporcionalmente promove queda na produtividade do algodão.

O algodoeiro é uma cultura que apresenta tolerância relativamente alta à seca quando comparado às demais culturas anuais (Bezerra et al., 2003; Rosolem, 2007). Isso se deve à sua capacidade de aprofundamento do sistema radicular, que cresce em comprimento até a época do florescimento (Nayakekorala e Taylor, 1990; Rosolem, 2007). O período em que o DEF demonstrou correlação negativa com a produtividade do algodão foi no 5º decêndio, período que corresponde do aparecimento dos botões florais. DEF representa 14% da influência do clima neste momento e tem como principal consequência a abscisão no algodoeiro e consequentemente queda dos botões florais. Doorenbos e Kassam (1979) relatam que DEF severos próximos a floração detém o desenvolvimento da planta. Passos et al., (1987) e Batista (2010) destacam que o DEF diminui as estruturas reprodutivas e capulhos por plantas, do rendimento de fibra, da produção de algodão em caroço, e, consequentemente, da produtividade.

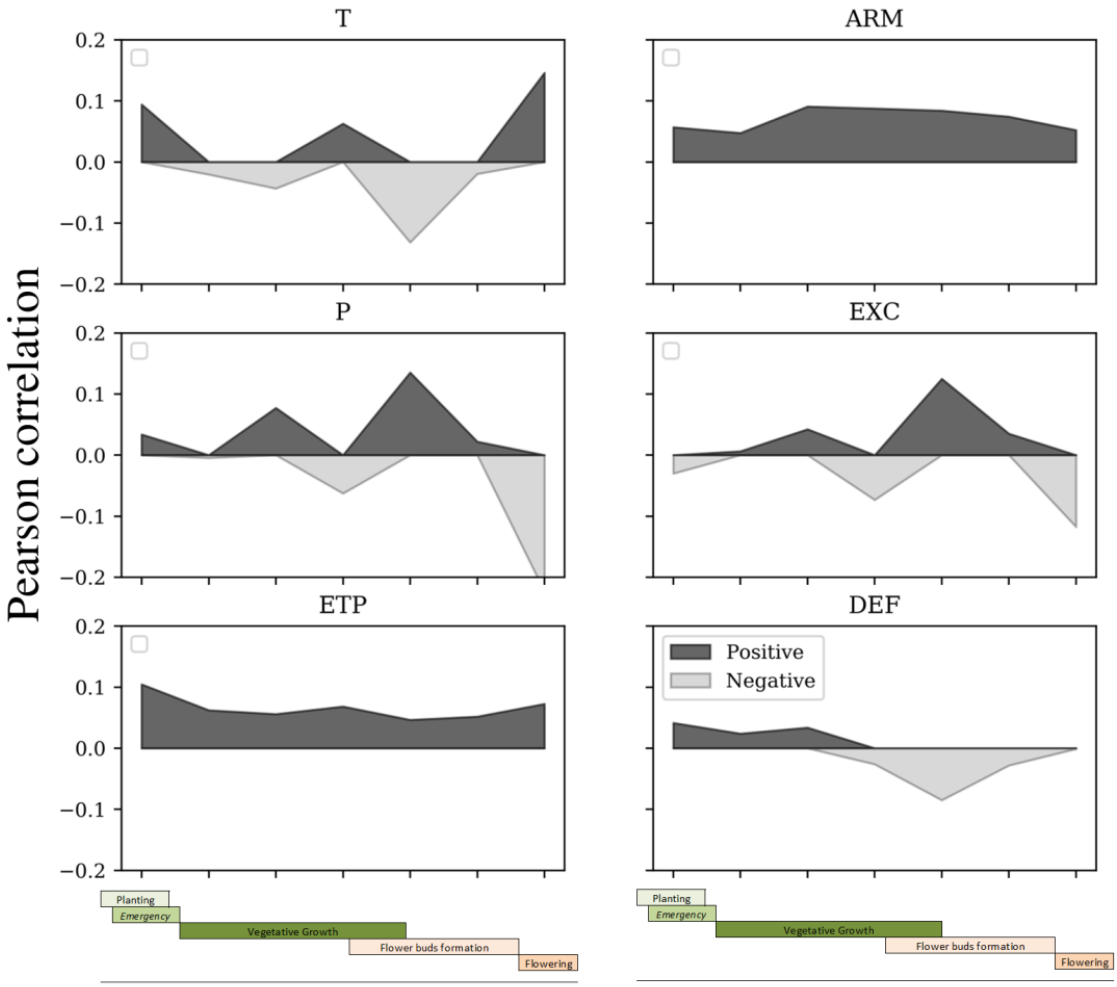

FIGURA 7. Correlação de Pearson entre a produtividade de algodão e os elementos climáticos em decêndios do plantio-florescimento.

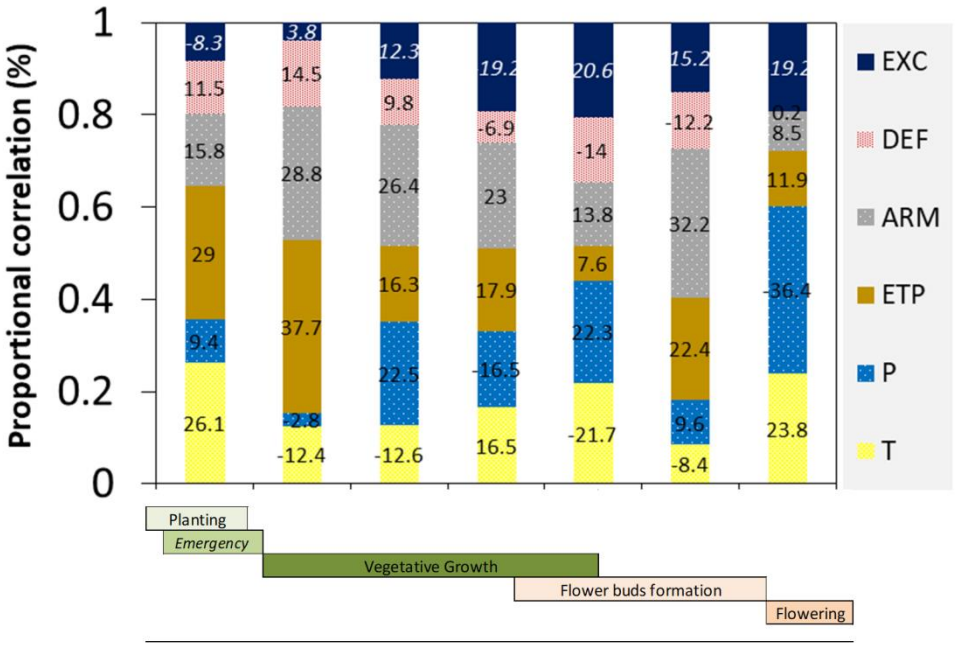

FIGURA 8. Proporção da correlação de Pearson entre a produtividade de algodão e os elementos climáticos em decêndios do plantio-florescimento.

3.3.3 Algoritmos x clima x previsão

Na previsão da produtividade do algodão usando dados climáticos os algoritmos demonstraram acurácias e precisões distintas (Table 3). Na calibração o melhor algoritmo foi o TREE que evidenciou índices de r=0.99, R^2=0.98, MSE=20294.51, RMSE=142.46, e MAPE=2.13. O algoritmo com menor desempenho foi o RLM com os seguintes índices: r=0.51, R^2=0.26, MSE=663864.87, RMSE=814.78 e MAPE=30.23. A performance de todos os algoritmos na calibração pode ser vista na Figura 9.

No teste os algoritmos TREE, RLM, RFT, KNN, BOO e MLP evidenciaram MAPEs de 18.35 %, 28.32%, 19.18%, 27.01%, 20.54% e 24.65%, respectivamente. Como a previsão está sendo realizado até florescimento, o que promove uma antecipação de ±80 dias, um MAPE de 18.35% é considerado baixo. Por exemplo, em uma produção de algodão média de 2827.06 kg ha^{-1} o algoritmo TREE com MAPE de 18.35% promove uma variação de apenas ±518.76 kg ha^{-1}. A superioridade do TREE também foi evidenciada pelos demais índices estatísticos: R=0.63, R^2=0.35, MSE=570681.50 e RMSE=755.43. A performance do TREE no teste pode ser vista na Figura 10.A. TREE é um algoritmo não paramétricos bastante utilizado em diversas áreas (Veenadhari et al., 2011; Veenadhari et al., 2014, Goyal, 2014), mas pouco utilizado na previsão de safra em agricultura. Um valor de MAPE de 18.35% é considerado adequado no teste de modelos de previsão usando como variáveis independentes os dados climáticos (Marcari et al., 2015; Moreto e Rolim, 2015).

TABELA 3. Índices estatísticos dos algoritmos na previsão da produtividade de algodão Brasil. Legenda: RLM é Regressão linear múltipla; KNN é KNeighborsRegressor; RFT é Random Forest Regressor; MLP é Redes Neurais Artificiais - Multi-layer Perceptron; BOO é Gradient Boosting for regression and TREE é Extra-trees regressor.

INDEXES	TREE	RLM	RFT	KNN	BOO	MLP

			Calibration			
r	0.99	0.51	0.97	0.76	0.99	0.94
R^2	0.98	0.26	0.94	0.58	0.98	0.88
MSE	20294.51	663864.87	98648.44	379493.60	20297.61	112904.95
RMSE	142.46	814.78	314.08	616.03	142.47	336.01
MAPE*	2.13	30.23	11.33	19.74	2.18	10.85
			Test			
r	0.59	0.42	0.62	0.45	0.57	0.60
R^2	0.35	0.17	0.38	0.20	0.32	0.36
EAmax	2162.34	2045.54	2219.36	2873.71	1851.85	2460.77
MSE	570681.50	763954.55	557810.26	784916.65	620128.57	653815.89
RMSE	755.43	874.04	746.87	885.96	787.48	808.59
MAPE*	18.35	28.32	19.18	27.01	20.54	24.65

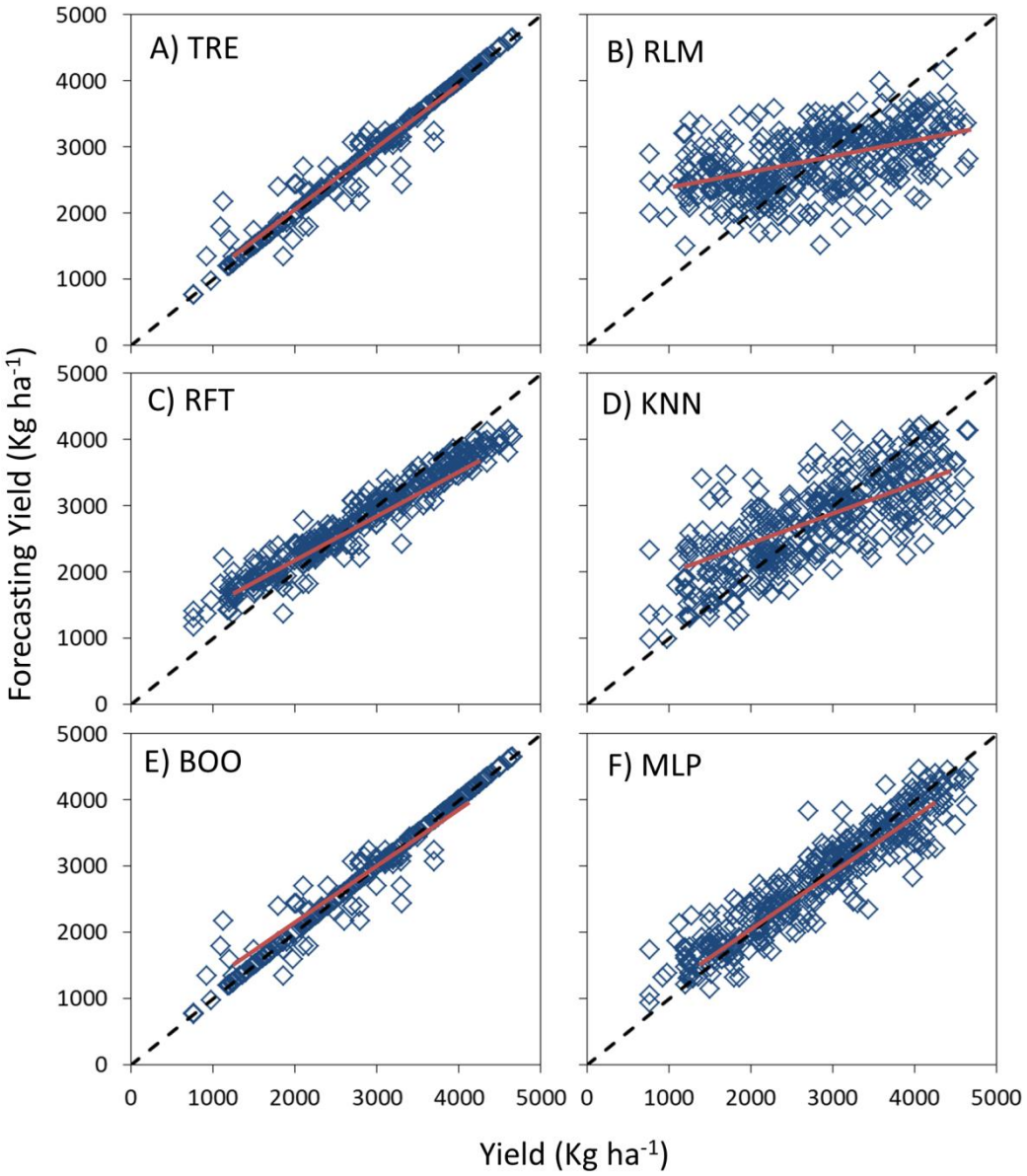

FIGURA 9. Relação entre os valores observados e previstos da produtividade do algodão na calibração por algoritmos: RLM é Regressão linear múltipla; KNN é KNeighborsRegressor; RFT é Random Forest Regressor; MLP é Redes Neurais Artificiais - Multi-layer Perceptron; BOO é Gradient Boosting para regressão e TREE é o regressor de Extra-árvores.

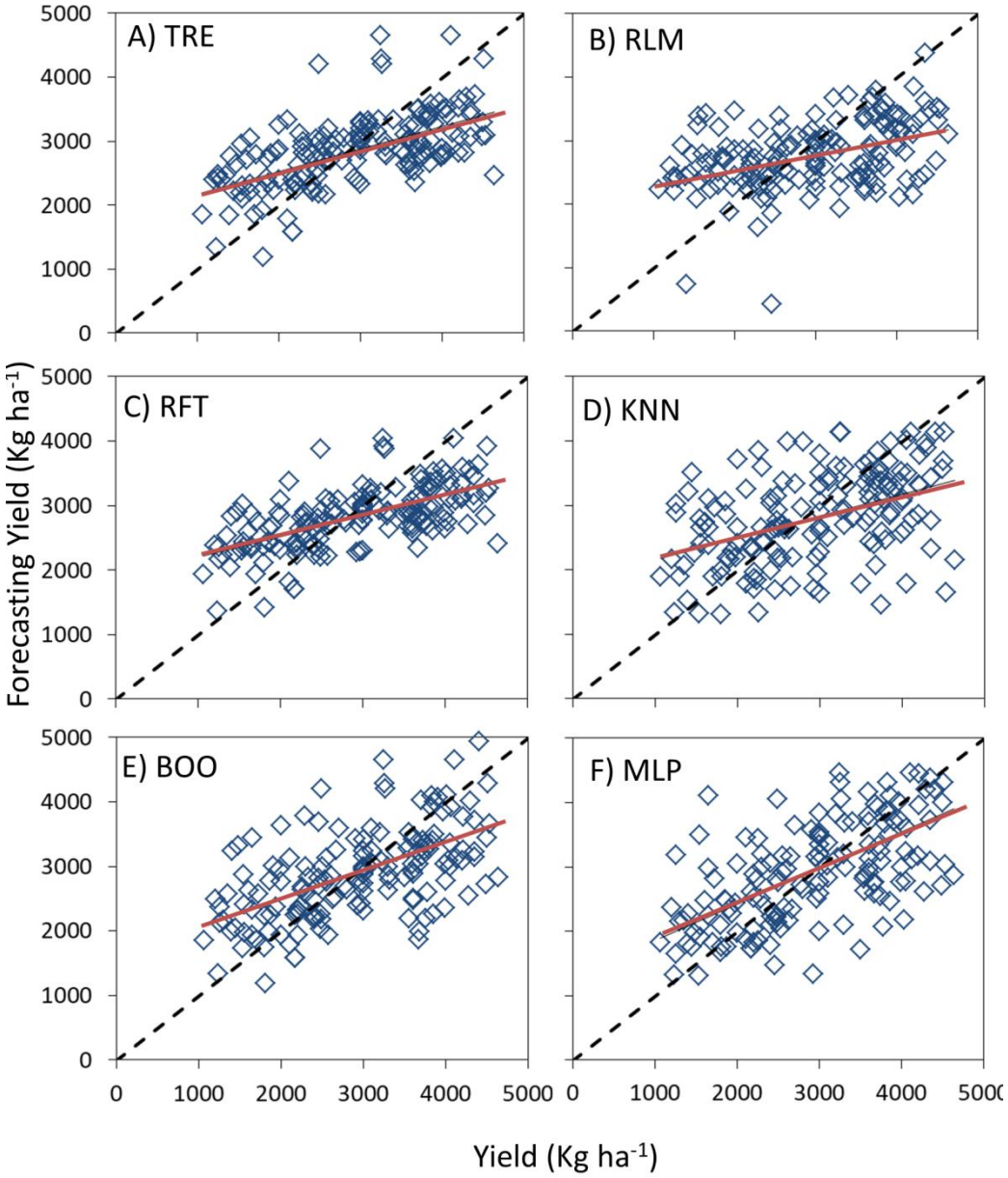

FIGURA 10. Relação entre os valores observados e previstos da produtividade do algodão em teste por algoritmos: RLM é Regressão linear múltipla; KNN é KNeighborsRegressor; RFT é Random Forest Regressor; MLP é Redes Neurais Artificiais - Multi-layer Perceptron; BOO é Gradient Boosting para regressão e TREE é o regressor de Extra-árvores.

3.3.4 Mapas de produtividade x Algoritmo TREE

A produtividade do algodão variou de 2065.05 kg ha^{-1} (Naviraí-MS) a 3046.64 kg ha^{-1} (Montividiu-GO) na região centro-oeste do Brasil. E o algoritmo TREE foi o que demonstrou o melhor desempenho para acompanhar toda a variabilidade espacial da produtividade do algodão tanto na calibração (Figura 11.AB) como também no teste (Figura 11.CD), utilizando como variáveis independentes os elementos climáticos.

Na calibração do TREE os desvios médios entre os dados reais e os dados previstos foram de apenas 57.2 kg ha^{-1} (Figura 11.F). Desvios nessa magnitude são baixos considerando uma previsão antecipada de 80 dias. No teste dos algoritmos os desvios dos dados para o TREE foram mais elevados em relação à calibração, por exemplo, para os estados de GO, MS e MT os desvios foram de 432.06 (±57.2) kg ha^{-1} (Paraúna), 755.59 (±57.2) kg ha^{-1} (São-Gabriel-do-Oeste) e 796.48 (±57.2) kg ha^{-1} (Pedra-Preta), respectivamente (Figure 11.E).

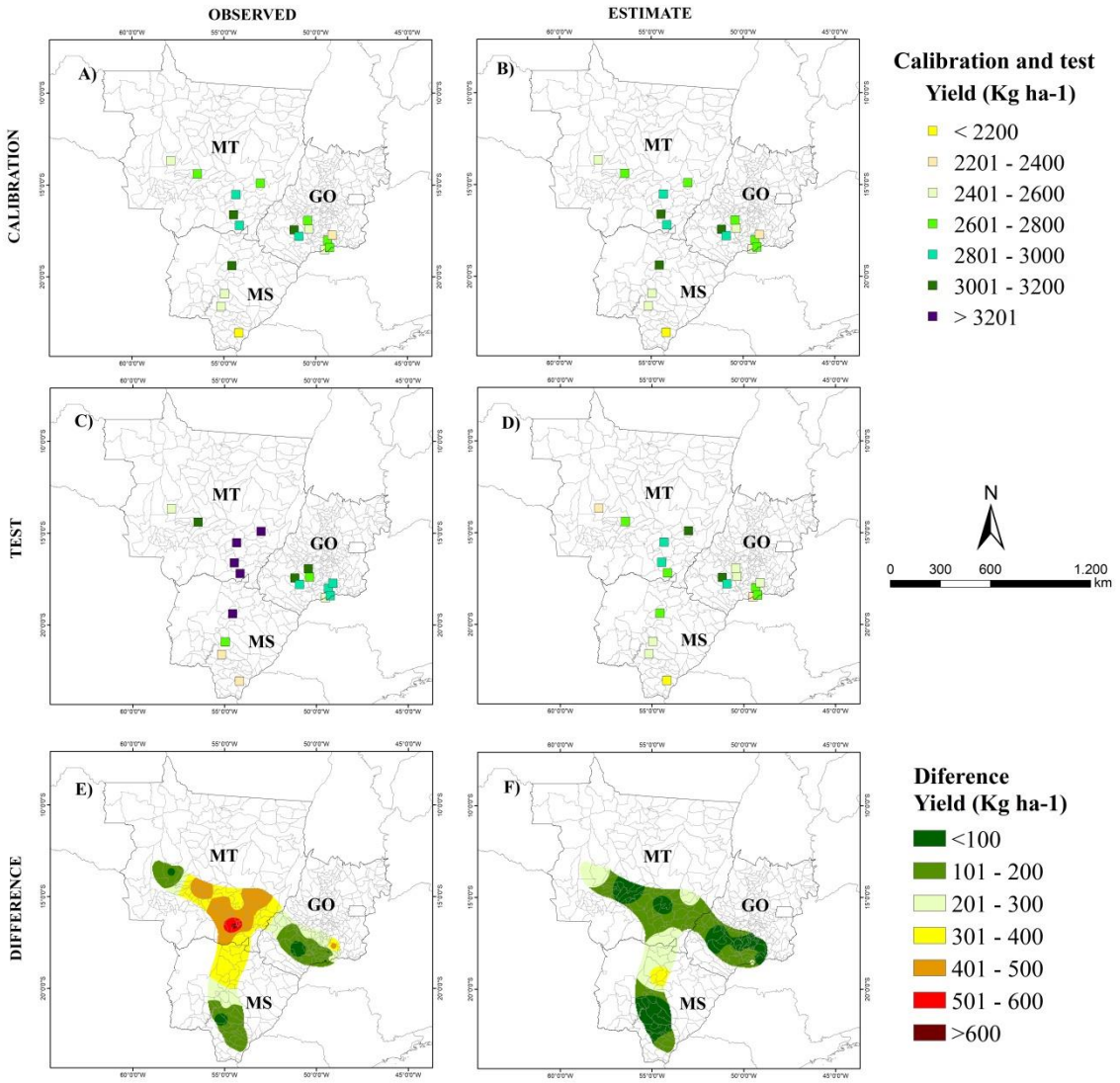

Figura 11. Produtividade real e prevista do algodão na calibração e teste pelo algoritmo TREE. A) Real calibrado; B) Previsto Calibrado; C) Real teste, D) Previsto teste, E) Desvios do teste e F) Desvios da calibração.

3.4 Conclusões

Os elementos climáticos que mais influenciam a produtividade do algodão nas principais regiões produtoras do Brasil são ETP e o ARM. Essas duas variáveis demonstraram correlações positivas e alta significância no período de plantio ao florescimento. A ocorrência de elevadas P e EXC no período do florescimento promove redução da produtividade do algodão.

Os modelos não lineares evidenciam que a produtividade do algodão tem tendência sigmoide em função do acúmulo de P, ETP, ARM e EXC durante o ciclo da cultura. Com acúmulo de 30 mm de DEF e T médias acima de 26.4º C ocorre redução drástica da produtividade do algodão.

É possível prever antecipadamente a produtividade do algodão com acurácia para as principais regiões produtoras do Brasil usando algorithms of Machine learning. O melhor algoritmo foi o TREE que evidenciou índices de r=0.99, R^2=0.98, MSE=20294.51, RMSE=142.46 e MAPE=2.13. O algoritmo com menor desempenho foi o RLM.

O algoritmo TREE teve bastante sucesso na previsão da produção de algodão com dados climáticos do plantio até o florescimento. Com isso é possível prever ter uma antecipação em torno de ±80 dias, o que possibilita o produtor um tempo hábil para planejar sua colheita.

Referências

Ahamed ATMS, Mahmood NT, Hossain N, Kabir MT, Das K, Rahman F, Rahman RM. 2015. Applying data mining techniques to predict annual yield of major crops and recommend planting different crops in different districts in Bangladesh. In: 2015 IEEE/ACIS 16th Int Conf Softw Eng Artif Intell Netw Parallel/Distributed Comput. [place unknown]; p. 1–6.

Akbar A, Kuanar A, Patnaik J, Mishra A, Nayak S. 2018. Application of Artificial Neural Network modeling for optimization and prediction of essential oil yield in turmeric (Curcuma longa L.). Comput Electron Agric [Internet]. 148:160–178. http://www.sciencedirect.com/science/article/pii/S0168169916311371

Allen RG, S PL, Raes D, Martin S. 1998. Crop evapotranspiration : Guidelines for computing crop water requirements / by Richard G. Allen ... [et al.]. FAO Irrig Drain Pap 56.:1–15.

Alvares CA, Stape JL, Sentelhas PC, de Moraes G, Leonardo J, Sparovek G. 2013. Köppen's climate classification map for Brazil. Meteorol Zeitschrift. 22(6):711–728.

Andrade Junior AS de, Silva FAM, de Lima MG, Amaral AB. 2009. Climatic aptitude zoning for cotton in Piaui State, Brazil. Rev Ciência Agronômica. 40(2):175.

Aparecido LE de O, Rolim G de S, de Moraes JR da SC, Rocha HG, Lense GHE, Souza PS. 2018. Agroclimatic zoning for urucum crops in the state of Minas Gerais, Brazil. Bragantia. 77(1):193–200.

Aparecido LE de O, de Souza Rolim G, De JR da SC, Costa CTS, de Souza PS, others. 2020. Machine learning algorithms for forecasting the incidence of Coffea arabica pests and diseases. Int J Biometeorol.:1–18.

Aparecido LE de O, Torsoni GB, Mesquita DZ, Meneses KC de, Moraes JR da SC de. 2020. Modeling safrinha corn productivity according to climatic conditions in Mato Grosso do Sul. Rev Bras Climatol. 26.

Assad ED, Martins SC, Beltrão NE de M, Pinto HS. 2013. Impacts of climate change on the agricultural zoning of climate risk for cotton cultivation in Brazil. Pesqui Agropecuária Bras. 48(1):1–8.

Barros MAL, Silva CRC Da, Lima LM De, Farias FJC, Ramos GA, Santos RC Dos. 2020. A Review on Evolution of Cotton in Brazil: GM, White, and Colored Cultivars. J Nat Fibers.:1–13.

Batista CH, de Aquino LA, Silva TR, Silva HRF. 2010. Growth and productivity of cotton culture in response to phosphorus application and irrigation methods. Rev Bras Agric Irrig. 4(4).

Bezerra, J. R. C.; Silva E Luz, M. J.; Pereira, J. R.; Santana, J. C. F.; Dias, J. M.; Santos, J. W.; Santos, T. Silva. 2003. Effect of soil water deficit on yield and fiber of herbaceous cotton, cultivar BRS 201. Revista Brasileira de Oleaginosas e Fibrosas, Campina Grande, v. 7, n. 2/3, p.727-734.

Bhojani SH, Bhatt N. 2020. Wheat crop yield prediction using new activation functions in neural network. Neural Comput Appl.:1–11.

Biswas R, Bhattacharyya B. 2019. Rice yield prediction in lower Gangetic Plain of India through multivariate approach and multiple regression analysis. J Agrometeorol. 21(1):101–103.

Brasil.1981. Ministry of Mines and Energy. General secretary. RADAMBRASIL Project. Rio de Janeiro: Survey of Natural Resources, 25, 29, 31.

CONAB (Companhia Nacional Do Abastecimento). 2019. www.conab.gov.br

Chen X, Qi Z, Gui D, Sima MW, Zeng F, Li L, Li X, Gu Z. 2020. Evaluation of a new irrigation decision support system in improving cotton yield and water productivity in an arid climate. Agric Water Manag. 234:106139.

Chou J, Xu Y, Dong W, Xian T, Wang Z. 2019. Research on the variation characteristics of climatic elements from April to September in China's main grain-producing areas. Theor Appl Climatol. 137(3–4):3197–3207.

Doorenbos J, Kassam AH. 1979. Yield response to water. Irrig Drain Pap.(33):257.

Draper NR, Smith H. 1980. Applied Regression Analysis, 2nd Edn. Chap. 1.
Elavarasan D, Vincent DR, Sharma V, Zomaya AY, Srinivasan K. 2018. Forecasting yield by integrating agrarian factors and machine learning models: A survey. Comput Electron Agric [Internet]. 155(October):257–282. https://doi.org/10.1016/j.compag.2018.10.024

Everingham Y, Sexton J, Skocaj D, Inman-Bamber G. 2016. Accurate prediction of sugarcane yield using a random forest algorithm. Agron Sustain Dev. 36(2):27.
Feng A, Zhou J, Vories ED, Sudduth KA, Zhang M. 2020. Yield estimation in cotton using UAV-based multi-sensor imagery. Biosyst Eng [Internet]. 193:101–114. http://www.sciencedirect.com/science/article/pii/S1537511020300544

Gonzalez-Sanchez A, Frausto-Solis J, Ojeda-Bustamante W. 2014. Attribute selection impact on linear and nonlinear regression models for crop yield prediction. Sci World J. 2014.

Goyal MK. 2014. Modeling of sediment yield prediction using M5 model tree algorithm and wavelet regression. Water Resour Manag. 28(7):1991–2003.
Griddi-Papp, I.L. 1965, Botany and Genetics. In - Cotton Culture and Fertilization, ed. Brazilian Institute of Potash, São Paulo, pp.117-157.

Gyamerah SA, Ngare P, Ikpe D. 2020. Probabilistic forecasting of crop yields via

quantile random forest and Epanechnikov Kernel function. Agric For Meteorol. 280:107808.

Hansen JW, Indeje M. 2004. Linking dynamic seasonal climate forecasts with crop simulation for maize yield prediction in semi-arid Kenya. Agric For Meteorol. 125(1–2):143–157.

Hussain M, Tariq AF, Nawaz A, Nawaz M, Sattar A, Ul-Allah S, Wakeel A. 2020. Efficacy of fertilizing method for different potash sources in cotton (Gossypium hirsutum L.) nutrition under arid climatic conditions. PLoS One. 15(1):1–9.

IBGE, G. Sistema IBGE de recuperação automática: SIDRA. 2020. Banco de dados agregados. http://www. sidra. ibge. gov. br/bda/tabela/protabl. asp.

Iqbal M, Ul-Allah S, Naeem M, Ijaz M, Sattar A, Sher A. 2017. Response of cotton genotypes to water and heat stress: from field to genes. Euphytica. 213(6):1–11. Kaul M, Hill RL, Walthall C. 2005. Artificial neural networks for corn and soybean yield prediction. Agric Syst. 85(1):1–18.

Krige DG. 1951. A statistical approach to some basic mine valuation problems on the Witwatersrand. J South African Inst Min Metall. 52(6):119–139.

Lasdon LS, Waren AD. 1982. General reduced gradient software for linearly and non-linearly contained problems. GRG2 users Guid Univ Texas, Austin.

Li L, Baker TE, White SR, Burke K, others. 2016. Pure density functional for strong correlation and the thermodynamic limit from machine learning. Phys Rev B. 94(24):245129.

Li N, Lin H, Wang T, Li Y, Liu Y, Chen X, Hu X. 2020. Impact of climate change on cotton growth and yields in Xinjiang, China. F Crop Res. 247:107590.
Li P. 2012. Robust logitboost and adaptive base class (abc) logitboost. arXiv Prepr arXiv12033491.

Marcari MA, Rolim G de S, Aparecido LE de O. 2015. Agrometeorological models for forecasting yield and quality of sugarcane. Aust J Crop Sci. 9(11):1049–1056. Marengo JA, Tomasella J, Nobre CA. 2017. Climate change and water resources. In: Waters of Brazil. [place unknown]: Springer; p. 171–186.

Martins E, de Oliveira Aparecido LE, Santos LPS, de Mendonça JMA, de Souza PS. 2015. Weather influence in yild and quality coffee produced in South Minas Gerais region. Coffee Sci. 10(4):499–506.

Marur, C.J. 1991. Comparison of rates of liquid photosynthesis, stomatal resistance and yield of two cotton cultivars submitted to water stress. Pesquisa Agropecuaria Brasileira, 26, 153- 161.

Mercante E, de Lima LEP, Justina DDD, Uribe-Opazo MA, Lamparelli RAC. 2012. Detection of soybean planted areas through orbital images based on culture spectral dynamics. Eng Agr{\'\i}cola. 32(5):920–931.

Moreto VB, de Souza Rolim G. 2015. Estimation of annual yield and quality of Valncia orange related to monthly water deficiencies. African J Agric Res. 10(6):543–553.

Moreto VB, de Souza Rolim G, Zacarin BG, Vanin AP, de Souza LM, Latado RR. 2017. Agrometeorological models for forecasting the qualitative attributes of “Valência” oranges. Theor Appl Climatol. 130(3–4):847–864.

Nayakekorala, H. Taylor, H. M. 1990. Phosphorus uptake rates of cotton roots at different growth stages from different soil layers. Pant and soil. Dordrecht, 122:105-110.

Opelt A, Fussenegger M, Pinz A, Auer P. 2004. Weak hypotheses and boosting for generic object detection and recognition. In: Eur Conf Comput Vis. [place unknown]; p. 71–84.

Pedregosa F, Varoquaux G, Gramfort A, Michel V, Thirion B, Grisel O, Blondel M, Prettenhofer P, Weiss R, Dubourg V, others. 2011. Scikit-learn: Machine learning in Python. J Mach Learn Res. 12:2825–2830.

Quinlan JR. 1983. Learning Efficient Classification Procedures and Their Application to Chess End Games. In: Michalski RS, Carbonell JG, Mitchell TM, editors. Mach Learn An Artif Intell Approach [Internet]. Berlin, Heidelberg: Springer Berlin Heidelberg; p. 463–482. https://doi.org/10.1007/978-3-662-12405-5_15

Rahimi J, Khalili A, Butterbach-Bahl K. 2019. Projected changes in modified Thornthwaite climate zones over Southwest Asia using a CMIP5 multi-model ensemble. Int J Climatol. 39(12):4575–4594.

Rehman TU, Mahmud MS, Chang YK, Jin J, Shin J. 2019. Current and future applications of statistical machine learning algorithms for agricultural machine vision systems. Comput Electron Agric. 156:585–605.

Rosolem, C A. 2007. Maximum soybean productivity. Rondonópolis: Foundation MT. p.237-244.

Sahoo S, Russo TA, Elliott J, Foster I. 2017. Machine learning algorithms for modeling groundwater level changes in agricultural regions of the US. Water Resour Res. 53(5):3878–3895.

Schwalbert RA, Amado T, Corassa G, Pott LP, Prasad PVV, Ciampitti IA. 2020. Satellite-based soybean yield forecast: Integrating machine learning and weather data for improving crop yield prediction in southern Brazil. Agric For Meteorol. 284:107886.

Shekoofa A, Emam Y, Shekoufa N, Ebrahimi M, Ebrahimie E. 2014. Determining the most important physiological and agronomic traits contributing to maize grain yield through machine learning algorithms: a new avenue in intelligent agriculture.

PLoS One. 9(5):e97288.

Silva KA, de Souza Rolim G, Valeriano TTB, others. 2020. Influence of El Niño and La Niña on coffee yield in the main coffee-producing regions of Brazil. Theor Appl Climatol. 139(3):1019–1029.

Silva MP da, Marujo LG. 2012. Analysis of an intermodal model for the outlet of the soy production at the brazilian midwest. J Transp Lit. 6(3):90–106.

Singh RK. 2008. Artificial neural network methodology for modelling and forecasting maize crop yield. Agric Econ Res Rev. 21(347-2016–16813):5–10.

Stackhouse PW, Westberg D, Hoell JM, Chandler WS, Zhang T. 2017. Prediction Of Worldwide Energy Resource (POWER)---Sustainable Buildings Methodology---(1.0 o Latitude by 1.0 o Longitude Spatial Resolution).

Thomas J, Mayr A, Bischl B, Schmid M, Smith A, Hofner B. 2018. Gradient boosting for distributional regression: faster tuning and improved variable selection via noncyclical updates. Stat Comput. 28(3):673–687.

Thornthwaite CW. 1948. An approach toward a rational classification of climate. [place unknown]: LWW.

Thornthwaite CW, Mather JR. 1955. Publications in climatology. water Balanc. 8:1–104.

Torkashvand AM, Ahmadi A, Nikravesh NL. 2017. Prediction of kiwifruit firmness using fruit mineral nutrient concentration by artificial neural network (ANN) and multiple linear regressions (MLR). J Integr Agric. 16(7):1634–1644.

Veenadhari S, Mishra B, Singh CD. 2011. Soybean productivity modelling using decision tree algorithms. Int J Comput Appl. 27(7):11–15.

Veenadhari S, Misra B, Singh CD. 2014. Machine learning approach for forecasting crop yield based on climatic parameters. In: 2014 Int Conf Comput Commun Informatics. [place unknown]; p. 1–5.

Wrege MS, Caramori PH, Gonçalves SL, Almeida WP de, Marur CJ, Pires JR, Yamaoka RS. 2000. Cotton zoning based on sowing periods of lower risk in Paraná State, Brazil. Brazilian Arch Biol Technol. 43(1):0.