Universidade de São Paulo
Escola Superior de Agricultura “Luiz de Queiroz”

Qualidade da fibra de diferentes cultivares brasileiras de algodão e sua relação
com as condições meteorológicas

Isabella Theresa de Almeida Martins

Dissertação apresentada para obtenção do título de
Mestra em Ciências. Área de concentração: Engenharia
de Sistemas Agrícolas

Piracicaba
2020

Isabella Theresa de Almeida Martins
Engenheira Agrícola e Ambiental

Qualidade da fibra de diferentes cultivares brasileiras de algodão e sua relação com as
condições meteorológicas

versão revisada de acordo com a resolução CoPGr 6018 de 2011

Orientador:
Prof. Dr. **PAULO CESAR SENTELHAS**

Dissertação apresentada para obtenção do título de
Mestra em Ciências. Área de concentração: Engenharia
de Sistemas Agrícolas

Piracicaba
2020

Dados Internacionais de Catalogação na Publicação
DIVISÃO DE BIBLIOTECA – DIBD/ESALQ/USP

Martins, Isabella Theresa de Almeida

Qualidade da fibra de diferentes cultivares brasileiras de algodão e sua relação com as condições meteorológicas/ Isabella Theresa de Almeida Martins. - - versão revisada de acordo com a resolução CoPGr 6018 de 2011. - - Piracicaba, 2020.

160 p.

Dissertação (Mestrado) - - USP / Escola Superior de Agricultura “Luiz de Queiroz”.

1. *Gossypium hirsutum* L. *raça latifolium Hutch* 2. Micronaire 3. Random forest 4. XGBoosting 5. Support vector machine I. Título

AGRADECIMENTOS

Primeiramente, à Deus, por sempre me iluminar, proteger e guiar os meus caminhos.

Aos meus pais, José Helvecio Martins e Leila de A.O. Martins, e aos meus irmãos, Sarah F. A. Martins e Thiago E. A. Martins, pelo amor, pelo incentivo, pelo apoio incondicional, pela união e por todos os momentos felizes.

À minha sobrinha Júlia M. Lima, por ser meu exemplo de força e superação e por sempre alegrar os meus dias.

Ao meu namorado Luis Felipe S. Santos e à sua família, pelo amor, amizade, carinho, paciência e suporte durante toda a minha caminhada profissional.

Ao Prof. Dr. Paulo Cesar Sentelhas, pela amizade, pela confiança depositada, pelos ensinamentos, pelos conselhos, pela paciência e pela oportunidade.

À Escola Superior de Agricultura “Luiz de Queiroz” (ESALQ/USP), pela formação profissional de qualidade e pelas oportunidades.

Ao Programa de Engenharia de Sistemas Agrícolas, pela oportunidade de realização do Mestrado.

À empresa Tropical Melhoramento & Genética (TMG), pela disponibilidade em fornecer os dados experimentais essenciais para a realização do presente projeto, em especial ao Eduardo Masakazu Kawakami, Paulo Hugo Aguiar, Vinicius Eduardo Nunes Ferreira e Rafael José Zeni.

Aos colegas da Cúpula, pela amizade, pelo auxílio e pelo apoio. Em especial, ao Cleverson H. Freitas, à Juliana C. Rodrigues, à Silvana I. Brand, à Paola F. Bongiovani e à Fabiani Bender que tanto me ensinaram e me ajudaram ao longo da realização do projeto.

Ao financiamento concedido pela Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (CAPES) para realização deste projeto.

Às minhas queridas amigas, Helizani C. Bazame, Jéssica G. Nascimento e Larissa B. Moura, parceiras da UFV, da ESALQ e da vida.

A todos que de alguma maneira contribuíram com a minha formação pessoal e profissional ao longo da condução desse trabalho.

Muito obrigada!

*“Sem a curiosidade que me move,
que me inquieta,
que me insere na busca,
não aprendo nem ensino.”*

Paulo Freire

SUMÁRIO

RESUMO.....	8
ABSTRACT.....	9
1. INTRODUÇÃO.....	11
2. REVISÃO BIBLIOGRÁFICA.....	15
2.1. A cultura do algodão.....	15
2.1.1. Importância socioeconômica.....	15
2.1.2. Aspectos agronômicos.....	16
2.1.3. A fibra de algodão: origem, estrutura, composição e caracterização.....	19
2.1.3.1. Origem de uma fibra de algodão.....	19
2.1.3.2. Estrutura de uma fibra de algodão.....	21
2.1.3.3. Composição de uma fibra de algodão.....	22
2.1.3.4. Caracterização de uma fibra de algodão.....	22
2.1.4. Condicionantes agrometeorológicos da qualidade da fibra de algodão.....	26
2.1.4.1. Temperatura.....	26
2.1.4.2. Radiação solar.....	29
2.1.4.3. Disponibilidade hídrica do solo.....	33
2.1.5. Modelagem agrícola e mineração de dados.....	36
2.1.5.1. Modelagem agrícola.....	36
2.1.5.2. Mineração de dados.....	36
2.1.5.3. Tarefas e técnicas de mineração de dados.....	37
2.1.5.4. Análise de agrupamentos (Clusters).....	38
2.1.5.5. Grupo de árvores de decisão.....	40
2.1.5.6. Random Forest.....	42
2.1.5.7. Boosting.....	43
2.1.5.8. Support Vector Machine (SVM).....	45
3. MATERIAL E MÉTODOS.....	51
3.1. Dados experimentais.....	51
3.1.1. Locais.....	51
3.1.2. Descrição edafoclimática das áreas de estudo.....	52
3.1.3. Cultivares.....	52
3.1.4. Variáveis analisadas.....	53
3.2. Fases fenológicas avaliadas.....	54

3.3. Dados meteorológicos.....	54
3.4. Balanço hídrico da cultura	55
3.5. Relação das variáveis meteorológicas e índices de qualidade da fibra do algodão em função da localidade e datas de semeadura.....	58
3.6. Análise de correlação e Análise de Agrupamento	58
3.7. Modelagem dos índices de qualidade da fibra de algodão por meio da mineração de dados	61
3.8. Avaliação dos resultados.....	64
4. RESULTADOS E DISCUSSÃO	65
4.1. Variação dos índices de qualidade da fibra do algodão em função do local de produção e da data de semeadura.....	65
4.2. Análise de correlação e Análise de Agrupamento	78
4.2.1. Análise de correlação (Abordagem geral)	78
4.2.1.1. Micronaire	80
4.2.1.2. Comprimento	80
4.2.1.3. Resistência	81
4.2.1.4. Uniformidade	82
4.2.1.5. Elongação.....	82
4.2.1.6. Índice de fibras curtas (SFI).....	83
4.2.2. Análise de Agrupamento.....	84
4.3. Modelagem dos índices de qualidade da fibra de algodão por meio da mineração de dados	95
4.3.1. Análise exploratória e entendimento dos dados.....	95
4.3.1.1. Análise do número de dados por cultivar.....	95
4.3.1.2. Valores médios dos índices de qualidade por cultivar	96
4.3.1.3. Variabilidade dos índices de qualidade da fibra de diferentes cultivares de algodoeiro.....	98
4.3.1.4. Variação dos índices de qualidade da fibra de algodão para as safras analisadas	102
4.3.1.5. Distribuição dos índices de qualidade da fibra de algodão	106
4.3.2. Fase de modelagem.....	109
4.3.2.1. Micronaire	110
4.3.2.2. Comprimento	116
4.3.2.3. Resistência	123

4.3.2.4. Uniformidade.....	130
4.3.2.5. Elongação	137
4.3.2.6. Índice de fibras curtas (SFI)	144
5. CONCLUSÕES.....	153
REFERÊNCIAS.....	155

RESUMO

Qualidade da fibra de diferentes cultivares brasileiras de algodão e sua relação com as condições meteorológicas

O algodoeiro (*Gossypium hirsutum* L. raça *latifolium* Hutch) é produzido por mais de 60 países e, apesar da qualidade e multiplicidade de uso de suas sementes, é cultivado essencialmente para a produção de fibras. A qualidade da fibra pode diferir entre os diferentes ambientes de produção, sendo fator chave para determinação do preço e qualidade do algodão destinado aos produtos têxteis. Essas diferenças de qualidade estão associadas principalmente às cultivares e às condições meteorológicas, as quais influenciam os parâmetros indicativos de qualidade de fibra. O conhecimento dos fatores que condicionam a qualidade da fibra de algodão é importante para a definição das regiões com potencial para produção de fibras de qualidade superior. Dessa forma, como maneira de subsidiar a produção de fibras de algodão de melhor qualidade, os objetivos do presente estudo foram identificar e ordenar os fatores que afetam a qualidade da fibra de algodão, de acordo com sua importância, bem como o desenvolvimento de modelos matemáticos de estimação para os índices de qualidade da fibra de algodão por meio de técnicas de mineração de dados (MD). Para tanto, dados das variáveis meteorológicas e de índices de qualidade da fibra de algodão de 32 cultivares brasileiras foram submetidos à análises de correlação de Pearson (para uma abordagem geral); à análise de agrupamento (“clusterização”); e à três técnicas de MD (*Random Forest*, *XGBoosting*, e *Support Vector Machine*). Para a MD, os dados foram separados em duas partes, sendo uma para o treinamento dos algoritmos e geração dos modelos (70%) e outra para a etapa de teste dos modelos gerados (30% dos dados restantes). A análise de desempenho dos modelos foi efetuada pela comparação entre os índices de qualidade da fibra estimados pelos modelos e os índices de qualidade observados por meio da análise de regressão. O desempenho dos modelos foi avaliado considerando os seguintes índices estatísticos: coeficiente de correlação (R); coeficiente de determinação (R^2); índice de Willmott (D); índice de confiança de Camargo (C); erro absoluto médio (EAM); e raiz quadrada do erro médio (RMSE). Por fim, comparou-se os resultados obtidos pelos modelos gerados com as técnicas de mineração de dados com os obtidos pelas análises de correlação e agrupamento. De modo geral, os modelos gerados por meio do emprego de técnicas de MD tiveram desempenho melhor do que as análises de correlação feitas a nível de abordagem geral e a nível de agrupamento e o melhor desempenho dos modelos foi para a elongação e o pior para o comprimento da fibra de algodão, sendo que o modelo XGB se apresentou ligeiramente melhor que o RF e o SVM foi o pior para todas as variáveis. As variáveis preditoras que foram comuns às três técnicas para os índices de qualidade da fibra avaliados variaram da seguinte forma: para o micronaire as variáveis de maior importância foram a radiação solar (Qg), o número de dias com temperatura maior ou igual a 32°C (Tmax32) e o número de dias com temperatura mínima menores ou iguais a 20°C (Tmin20); para o comprimento foram Qg, Tmin20 e velocidade do vento (U2); para a resistência foram Qg, Tmax32, Tmin 20 e temperatura mínima (Tmin); para a uniformidade foram U2, excedente hídrico ($\sum EXC$), déficit hídrico ($\sum DEF$)(mm), Tmin20 e precipitação pluvial (P); para a elongação foram Qg, número de dias com excedente hídrico maior do que zero (EXC0); temperatura máxima (Tmax), número de dias com temperatura menor ou igual a 18°C (Tmin18) e P; e para o índice de fibras curtas (SFI) foram umidade relativa do ar (UR); Tmax; e P.

Palavras-chave: *Gossypium hirsutum* L. raça *latifolium* Hutch, Micronaire, Random forest, XGBoosting, Support vector machine

ABSTRACT

Fiber quality of different Brazilian cotton cultivars and their relationship with weather conditions

Cotton (*Gossypium hirsutum* L. race *latifolium* Hutch) is produced by more than 60 countries and despite the quality and multiplicity of use of its seeds, it is grown mainly to produce fibers. The quality of the fiber can differ between different production environments, being a key factor in determining the price and quality of cotton destined for textile products. These differences in quality are mainly associated with cultivars and weather conditions, which influence the indicating parameters of fiber quality. The knowledge of the factors that affect the quality of the cotton fiber is important for the definition of the regions with potential to produce that with superior quality. Thus, as a way to subsidize the production of better quality cotton fibers, the objectives of the present study were to identify and classify the factors that affect the quality of the cotton fiber, according to its importance, as well as the development of mathematical models for estimating the indices of cotton fiber quality using data mining (DM) techniques. To accomplish this, data of meteorological variables, and indices of cotton fiber quality from 32 Brazilian cultivars, were submitted to Pearson's correlation analysis (for a general approach); cluster analysis ("clustering"); and three DM techniques (Random Forest; XGBoosting; and Support Vector Machine). For DM, the data were separated in two parts, one for training the algorithms and generating the models (70%), and the other for testing the generated models (30% of the remaining data). The performance analysis of the models was achieved by comparing the fiber quality indices estimated using the models with the quality indices observed through the regression analysis. The performance of the models was evaluated considering the following statistical indices: correlation coefficient (R); determination coefficient (R^2); Willmott index (D); Camargo confidence index (C); mean absolute error (MAE); and root mean square error (RMSE). Finally, the results obtained by the models generated with data mining techniques were compared with those obtained by correlation and cluster analysis. In general, the models generated using DM performed better than the correlation analyzes made using both the general approach and the grouping. The best performance of the models was for elongation and the worst for the length of the cotton fiber. In addition, the XGB model showed slightly better results than the RF model, and the SVM model presented the worst results for all indices. The predictor variables that were common to the three DM techniques for the evaluated indices of fiber quality varied as follows: for micronaire, the most important variables were solar radiation (Qg), the number of days with temperature greater than or equal to 32°C (Tmax32), and the number of days with minimum temperature less than or equal to 20°C (Tmin20); for the length of the fibers, the variables with the greatest influence were Qg, Tmin20 and wind speed (U2); for fibers resistance the most important variables were Qg, Tmax32, Tmin 20, and minimum temperature (Tmin); for uniformity of the fibers these variables were U2, water surplus ($\sum EXC$), water deficit ($\sum DEF$), Tmin20, and rainfall (P); for fibers elongation these variables were Qg, number of days with water surplus greater than zero (EXC0), maximum temperature (Tmax), number of days with a temperature less than or equal to 18°C (Tmin18), and P; and for short fiber index (SFI) the most influencing variables were relative humidity (UR), Tmax, and P.

Keywords: *Gossypium hirsutum* L. race *latifolium* Hutch, Micronaire, Random forest, XGBoosting, Support vector machine

1. INTRODUÇÃO

O algodoeiro (*Gossypium hirsutum* L. raça *latifolium* Hutch) é uma das plantas mais cultivadas no mundo, devido à grande utilização e aplicabilidade da sua fibra na indústria têxtil. Além de ser a mais importante fonte de fibra natural como matéria-prima da indústria têxtil, o algodão é um produto de extrema importância socioeconômica para o Brasil, já que toda a sua cadeia produtiva (têxtil, alimentação animal e humana, moda, entre outras) é uma importante fonte de renda e empregos.

Segundo a Associação Brasileira dos Produtores de Algodão – ABRAPA (2020), a produção mundial para a safra 2019/2020, está projetada em 26,4 milhões de toneladas e o consumo em 26,2 milhões de toneladas. Além disso, o ranking de maiores produtores de algodão do mundo traz o Brasil na quarta colocação mundial, ficando atrás apenas da Índia, da China e dos Estados Unidos.

O Brasil produziu 2,726 milhões de toneladas de pluma na safra 2018/19 em uma área plantada de 1,618 milhões de hectares, resultando em uma produtividade média de 1.685 kg ha⁻¹. Para esta mesma safra, o Mato Grosso foi o estado brasileiro que mais produziu (1,815 milhões de toneladas), seguido da Bahia, Goiás, Mato Grosso do Sul e Minas Gerais. Já para a safra de 2019/2020, a previsão de produção de algodão em pluma no país está estimada em 2,880 milhões de toneladas, em uma área plantada de 1,677 milhões de hectares, com uma produtividade de 1.717 kg ha⁻¹ (ABRAPA, 2020).

Até o ano de 2050, estima-se que a população mundial crescerá de dois a três bilhões de pessoas. Este incremento populacional, acompanhado pela expansão da classe média, resultará em um expressivo aumento da demanda por alimentos, fibras e energia (FOLEY et al., 2011). Além disso, de acordo com o Painel Intergovernamental para a Mudança de Clima – IPCC (2019), as influências das mudanças climáticas já são evidentes em sistemas naturais e humanos, sugerindo que a sociedade enfrentará condições climáticas alteradas no futuro. Posto isso, diante da importância econômica e social da produção de fibras no Brasil e no mundo, torna-se relevante estudar os impactos das condições climáticas na qualidade da fibra de algodão, com o intuito de entender as possíveis consequências nos aspectos socioeconômicos de importantes regiões produtoras.

A cultura do algodão é sensível às variações das condições ambientais, como temperatura do ar, radiação solar e disponibilidade de água no solo, sendo que cada fase de seu desenvolvimento apresenta uma determinada exigência climática. Portanto, condições adversas em qualquer fase do ciclo da cultura podem levar as plantas de algodão a sofrerem estresses (momentâneos ou prolongados), afetando o desenvolvimento, o crescimento, a produtividade e a qualidade das fibras. Por intermédio do melhoramento genético, as plantas de algodão tornaram-se mais tolerantes aos estresses ambientais, o que lhes conferiu uma maior capacidade de produzir plumas (SOARES, 2015).

A qualidade da fibra do algodão é uma característica que depende de uma série de fatores, incluindo aqueles relativos à cultivar, ao solo, ao clima, à época de semeadura, ao controle de pragas, de doenças e plantas infestantes, à nutrição das plantas, ao tipo de colheitadeira e ao processo de beneficiamento e armazenamento. Assim, cada cultivar possui uma qualidade intrínseca de fibra,

controlada geneticamente, mas que pode ser alterada pelas condições ambientais e de manejo (FREIRE, 2015). A qualidade da fibra é uma característica que determina o valor da matéria prima na indústria têxtil, influenciando nos processos de comercialização e industrialização.

As características tecnológicas da fibra de algodão, tais como percentagem de fibra, índice de fibras, comprimento, uniformidade do comprimento, resistência e finura (índice micronaire), são importantes para o melhoramento genético do algodoeiro, pois são caracteres determinantes da qualidade desse produto, o que influencia em sua remuneração (RESENDE et al., 2014).

No cultivo comercial do algodoeiro procura-se alcançar a maior quantidade de fibra possível por unidade de área, em um determinado intervalo de tempo (safra). Para potencializar a produção de fibra, é necessário otimizar a produção por planta, que por sua vez depende da quantidade de frutos e qualidade da fibra, os quais dependem do acúmulo de estresses ocorridos ao longo das diferentes fases de desenvolvimento da cultura. Cada uma dessas fases tem necessidades climáticas específicas, e suas variações podem afetar tanto a produtividade como a qualidade da fibra de algodão (SOARES, 2015).

Diante desse contexto, estudos que visem auxiliar em questões relacionadas ao manejo da cultura em função das condições climáticas, de modo a aumentar a eficiência produtiva, são fundamentais (BATTISTI, 2013). Além disso, o entendimento da hierarquia das variáveis climáticas e de manejo, de acordo com sua importância nos caracteres tecnológicos de qualidade da fibra de algodão, pode auxiliar tanto na otimização do planejamento agrícola, quanto no aprimoramento das previsões de safra.

Uma das formas de se abordar a influência das condições meteorológicas nas características intrínsecas da fibra de algodão é por meio do uso de modelos empíricos ou fisiológicos-matemáticos. Esses modelos são baseados em dados observados e são capazes de representar, de forma simplificada, as relações entre a resposta fisiológica das plantas e as variáveis ambientais, durante os diferentes estádios fenológicos de uma cultura (DA FONSECA et al., 2007; TONATO et al., 2010). A decisão sobre as variáveis a serem utilizadas baseia-se em relações empíricas, obtidas por meio do conhecimento experimental, porém, qualquer modelo de simulação pode ser empregado, desde que represente a realidade (DIAS, 2016).

O conhecimento dos fatores que condicionam a qualidade da fibra de algodão é importante para a definição das regiões com potencial para produção de fibras de qualidade superior. Dessa forma, visando a maximização da qualidade da fibra de algodão produzida, as hipóteses deste estudo são que as condições meteorológicas afetam a qualidade da fibra de algodão e que modelos de estimação com base em variáveis meteorológicas são capazes de prever os indicadores de qualidade para diferentes cultivares brasileiras de algodão.

Diante dessas hipóteses, os objetivos do presente estudo foram:

- a) Identificar os fatores que afetam a qualidade da fibra de algodão por meio do uso da técnica de Mineração de dados (MD), a partir de dados de índices de qualidade observados a campo, em ensaios de competição de cultivares brasileiras de algodão;

- b) Desenvolver modelos empíricos de estimação da qualidade da fibra de algodão em função das condições meteorológicas para diferentes cultivares brasileiras, empregando-se as seguintes técnicas de mineração de dados: *Random Forest* (RF), *XGBoosting* (XGB) e *Support Vector Machine* (SVM); e
- c) Testar os modelos propostos com dados independentes, avaliando-se sua viabilidade em prever os índices de qualidade da fibra de algodão de diferentes cultivares brasileiras.

2. REVISÃO BIBLIOGRÁFICA

2.1. A cultura do algodão

2.1.1. Importância socioeconômica

O algodoeiro herbáceo ou anual (*Gossypium hirsutum* L. raça *latifolium* Hutch.) é uma espécie fibrosa e oleaginosa das mais antigas e de grande importância para a humanidade, conhecida mundialmente por produzir a mais importante fibra têxtil, além da sua capacidade de produção de óleo vegetal e torta para ração animal (ZONTA et al., 2016).

O algodão é uma cultura nativa dos trópicos e explorada economicamente, principalmente, em países com clima subtropical. Além disso, o algodoeiro está entre as culturas mais cultivadas no Brasil e é responsável pelo sustento de milhões de pessoas em todo o mundo gerando empregos e renda tanto para pequenos quanto para grandes produtores (VIDAL; FREIRE, 2013).

O algodão tornou-se a mais importante fibra cultivada do mundo, plantado em 100 países e em cinco continentes, envolvendo mais de 350 milhões de pessoas em sua produção, desde as fazendas até a logística, descaroçamento, processamento e embalagem. A média de plantio, nos últimos anos, foi ao redor de 35 milhões de hectares. Desde a década de 1950, a demanda mundial tem aumentado gradativamente. Segundo a Empresa Brasileira de Pesquisa Agropecuária – EMBRAPA (2020), o comércio mundial do produto movimentou, anualmente, cerca de US\$ 12 bilhões.

A cotonicultura em larga escala, em função dos avanços tecnológicos, a exemplo do melhoramento genético, permitiu que o Brasil modificasse sua condição de maior importador mundial de algodão para quarto maior produtor, ficando atrás apenas da Índia, China e Estados Unidos, conforme informações da Companhia Nacional de Abastecimento - CONAB (2019).

Em 2019, o Brasil produziu 2,726 milhões de toneladas de pluma, atingindo uma produtividade média de 1.685 kg ha⁻¹ em uma área de 1,618 milhões de hectares. Do total de pluma produzido, cerca de 1,815 milhões de toneladas foi no estado do Mato Grosso, principal produtor brasileiro, seguido dos estados da Bahia (597,60 mil toneladas), Goiás (68,50 mil toneladas), Mato Grosso do Sul (67,70 mil toneladas) e Minas Gerais (mil toneladas) (ABRAPA, 2020).

Em termos de cenários futuros, projetar a produção em períodos de até 10 anos não é uma tarefa trivial, principalmente no caso de produtos da agricultura brasileira que ainda não se encontram em processo de estabilização e são fortemente dependentes da demanda internacional. Do lado da produção, fatores climáticos como secas ou excesso de chuvas afetam a produtividade e a produção total. No caso do algodão, há ainda a concorrência de fibras sintéticas que depende fortemente do preço do petróleo; quando este se torna barato, compete com fibras naturais. Além disso, políticas nacionais de principais consumidores, como a China, afetam a produção local e consequentemente as exportações de países produtores, como o Brasil. A China, por exemplo, ao priorizar a produção de alimentos para o consumo de sua crescente população irá depender de importações para sua indústria de tecidos (EMBRAPA, 2020).

Diante do exposto, é importante e útil conhecer projeções nacionais e internacionais sobre a produção, consumo e exportações de algodão do Brasil. Segundo a Organização para Cooperação e Desenvolvimento Econômico (OECD) e a Organização das Nações Unidas para a Alimentação e a Agricultura (FAO) – OECD-FAO (2017), o consumo de algodão no Brasil deve apresentar ligeira redução nos próximos dez anos situando-se em 620 mil toneladas. As exportações mundiais de algodão em pluma, de acordo com as previsões do Departamento de Agricultura dos Estados Unidos - USDA (2017), irão se expandir em 54,3% entre 2016/17 e 2027/2028 e, nos próximos dez anos, o algodão do Brasil deverá representar cerca de 13,7% do comércio mundial, além de que os principais exportadores ao final destas projeções deverão ser os Estados Unidos, Índia e Brasil.

Diante disso, a cadeia produtiva da cotonicultura, tem apresentado expressivos avanços tecnológicos, a fim de fornecer subsídios aos produtores brasileiros, destacando-se a implantação de novas tecnologias e cultivares geneticamente superiores e com maior produtividade e resistência a pragas e doenças e mais adaptáveis às condições ambientais dos locais de produção.

2.1.2. Aspectos agronômicos

O algodoeiro herbáceo ou anual (*Gossypium hirsutum* L.) é uma espécie dicotiledônea da família *Malvaceae*, gênero *Gossypium*, de origem tropical entre o México e a América Central, e cultivado em diversos países do mundo, mas principalmente em regiões de clima quente (FAO, 2016). Dentre as 52 espécies catalogadas no gênero *Gossypium*, apenas quatro são cultivadas comercialmente, sendo elas: *G. hirsutum*; *G. barbadense*; *G. herbaceum*; e *G. arboreum*. Destas, *G. hirsutum* é a principal espécie, ocupando quase toda a totalidade das áreas produtoras do mundo. Esta espécie se diferenciou em raças e, no Brasil, assim como no mundo, o algodoeiro herbáceo ou anual (*Gossypium hirsutum* L. raça *latifolium* Hutch.) é o mais importante. O algodoeiro arbóreo (mocó) ou perene (*Gossypium hirsutum* L. raça *Marie galante* Hutch.), de grande importância até a década de 80, encontra-se, atualmente, restrito a pequenas áreas de clima semi-árido do Nordeste brasileiro (CHIAVEGATO et al., 2009).

As espécies *G. hirsutum* L. representam plantas pequenas, arbustivas, com altura entre 60 e 100 cm, pouco ramificadas e caules variando entre as cores verde e marrom. Além disso, possuem estrutura organográfica peculiar, apresentando ramos frutíferos e vegetativos e dois tipos de folhas, associadas ao ramo reprodutivo ou ao vegetativo (BORÉM; FREIRE, 2014). A diferença básica entre um ramo vegetativo (monopodial) e um frutífero (simpodial) é que, neste último, o meristema apical, após originar o prófio (folha sem bainha com duas estípulas) e uma folha verdadeira, termina com uma flor, enquanto o ramo vegetativo continua emitindo folhas (BELTRÃO; AZEVEDO, 2008).

A duração do ciclo do algodoeiro é uma característica genética, fortemente influenciada pelo ambiente. Manejando-o como cultura anual e tomando-se por referência o tempo necessário para que 90% dos seus frutos estejam abertos, classifica-se o ciclo do algodoeiro em três grupos de maturação no

Brasil: precoce (120 a 130 dias); médio (140 a 160 dias); e tardio (acima de 170 dias) (CHIAVEGATO et al., 2009).

Embora haja diferenças genéticas entre os grupos de maturação e as condições ambientais dos locais de cultivo, o algodoeiro anual inicia o florescimento com cerca de 50 dias, estendendo-se até 120 dias ou mais, com pico da curva de florescimento ao redor de 70 a 80 dias. O pegamento das flores é maior até 30 a 40 dias de florescimento, sendo favorecido pelas condições de vigor da planta e do ambiente (CHIAVEGATO et al., 2009). Diante disso, e considerando que o algodoeiro é estruturalmente complexo, devido ao seu hábito de crescimento indeterminado e aos seus padrões de desenvolvimento dos ramos simpodiais e de florescimento, o conhecimento detalhado do ciclo de produção do algodão é fundamental para seu manejo, já que durante parte do ciclo da cultura as plantas apresentam sobreposição de eventos fenológicos, como crescimento vegetativo, aparecimento de gemas reprodutivas, florescimento, crescimento e maturação dos frutos (Figura 1) (ROSOLEM, 2001).

Figura 1. Variação dos órgãos reprodutivos do algodoeiro. Adaptado de Rosolem (2001).

Conforme apresentado na Figura 1, a transição entre os processos de desenvolvimento (estabelecimento da planta e raízes, área foliar e estabelecimento do dossel, florescimento e maturação) não é claramente distinguível (CHIAVEGATO et al., 2009). Entretanto, como cada um desses eventos fenológicos é importante para a produção final, é necessário que eles ocorram de modo balanceado. Durante boa parte do ciclo da planta do algodoeiro ocorre uma forte competição interna pelos fotoassimilados. Assim, se houver, por exemplo, uma queda excessiva de estruturas reprodutivas, haverá crescimento vegetativo exagerado, aumentando o auto-sombreamento que, por sua vez, causará maior queda de estruturas reprodutivas (ROSOLEM, 2001).

Para melhor entendimento do crescimento e desenvolvimento das plantas do algodoeiro durante o ciclo, as escalas fenológicas são ferramentas úteis e indispensáveis. Câmara e Chiavegato (2001) subdividem o ciclo do algodoeiro em duas grandes fases: fase vegetativa (V), que se inicia com a emergência da plântula e termina com a formação do primeiro ramo frutífero; e fase reprodutiva (R), que

tem início com o aparecimento do primeiro botão floral e termina quando as fibras nos capulhos atingem o ponto de maturidade fisiológica, possibilitando a colheita (Tabela 1).

Tabela 1. Escala fenológica do algodoeiro herbáceo (*Gossypium hirsutum* L. raça *latifolium* Hutch).

Símbolo	DAS*	Denominação	Descrição
VE	6	Emergência	Saída do gancho da plântula (alça do hipocótilo), elevando os cotilédones acima do solo.
VC	7	Afastamento dos cotilédones	Exposição da gema apical vegetativa à radiação solar. Primeira folha enrolada entre os cotilédones.
V1	8	Primeiro nó vegetativo	Primeira folha cordiforme com 30% a 50% de expansão e a segunda folha também em forma de coração aberta.
V2	8	Segundo nó vegetativo	Segunda folha cordiforme com 30% a 50% de expansão e a primeira folha com lóbulos (lobada) aberta.
V3	14	Terceiro nó vegetativo	Primeira folha lobada com 30% a 50% de expansão e segunda aberta. Devido ao hábito de crescimento indeterminado, o algodoeiro tende a vegetar indefinidamente, emitindo sucessivos nós vegetativos com folhas lobadas. As cultivares anuais associadas ao ambiente e manejo completam o ciclo cultural entre V20 e V25.
VR		Primeiro ramo frutífero	A partir de V5 ou V6, surge o primeiro ramo frutífero (simpodial) com botão floral e folha correspondente fechados. Desta fase em diante, o algodoeiro acelera o crescimento e desenvolvimento vegetativo, emitindo novos ramos frutíferos (RF) nos nós vegetativos subsequentes.
R1		Primeiro botão floral	Primeiro botão floral (BF), com 5 mm de comprimento, encobertos pelas brácteas, na primeira posição do primeiro RF. Intensifica-se o acúmulo de matéria seca na planta. Na sequência, surgem BF's na seguinte ordem (padrão espiral): 2º BF na primeira posição do 2º RF; 3º BF na segunda posição do 1º RF, e assim sucessivamente.
R2	40	Primeira flor	Início do florescimento com abertura da primeira flor, na primeira posição do 1º RF. Plantas entre V14 e V16. O florescimento prossegue seguindo o padrão espiral.
R3		Crescimento da primeira maçã	Início da frutificação. Primeira maçã com 1 cm de diâmetro, na primeira posição do 1º RF.
R4		Primeira maçã visível	Plantas em florescimento pleno fechando o dossel. Maçã visível na primeira posição do 1º RF, sobressaindo-se às brácteas, rica em água, macia ao tato e com sementes e fibras em desenvolvimento.
R5		Primeira maçã cheia	Primeira maçã na primeira posição do 1º RF, iniciando a pigmentação (antocianina), consistente ao tato, aquosa, com sementes e fibras imaturas (alongamento das fibras).
R6		Final do florescimento	Fertilização da última flor viável, isto é, capaz de originar um capulho possível de ser colhido. Flor localizada a partir do 5º nó vegetativo (referência básica para manejo) do ponteiro para baixo. Planta com altura final definida (V20 a V25), predominando as maçãs.
R7		Primeiro capulho	Final da frutificação e maturidade fisiológica. Primeiro capulho na primeira posição do 1º RF. Translocação intensa devido à carga pendente. Acentuada queda de folhas a partir do baixeiro da planta. Final da deposição de celulose nas fibras, maturação das sementes e desidratação das maçãs cheias, consistentes e pigmentadas.
R8		Maturidade plena	Planta com 2/3 de desfolha, contendo 60% a 70% de capulhos. Colheita viável desde que a umidade nas fibras esteja entre 12 e 15%.

*Dias após semeadura (valor médio de referência). Fonte: Câmara e Chiavegato (2001).

Embora apresente um complexo hábito de crescimento, o algodoeiro segue um padrão de desenvolvimento, que, por razões práticas de manejo e por demandas distintas do ambiente e do ciclo, pode ser dividido em cinco estádios: germinação e emergência da plântula; emergência ao primeiro botão

floral; primeiro botão floral à abertura da primeira flor; primeira flor ao primeiro capulho; e primeiro capulho à colheita (maturação e deiscência). A duração média entre esses estádios são apresentados na Tabela 2 (CHIAVEGATO et al., 2009).

Tabela 2. Duração média de cada fase de desenvolvimento do algodoeiro herbáceo (*Gossypium hirsutum* L. raça *latifolium* Hutch).

Fase de desenvolvimento	Duração (dias)
Germinação - Emergência	5 a 10
Emergência - Primeiro botão floral	30 a 35
Primeiro botão floral - Primeira flor	20 a 25
Primeira flor - Primeiro capulho	50 a 55
Primeiro capulho - Colheita	45 a 50

Fonte: Chiavegato et al., (2009).

2.1.3. A fibra de algodão: origem, estrutura, composição e caracterização

2.1.3.1. Origem de uma fibra de algodão

A semente do algodão é piriforme, geralmente de cor pardo-escuro, com massa variando de 0,10 a 0,13 gramas (PASSOS, 1977). A epiderme é formada por uma camada de células esclerenquimáticas (células originadas a partir de células parenquimáticas) algumas das quais se diferenciam em fibras, que, quando menores são chamadas de línter (KOHEL; LEWIS, 1993).

A fibra é o principal produto do algodão, sendo esta constituída de uma única célula que contém mais de 95% de celulose, apresentando, quando madura, 25 camadas de celulose, cristalina, fibrilar e amorfa, situadas na sua parede secundária. Há ainda em torno de 1,3% de proteínas de elevada importância, pois são elas os mordentes da fibra e que recebem no processo de tingimento, os corantes que darão cor ao tecido. Além disso, há as ceras que ocorrem em pequenas quantidades, sendo importantes no processo de fiação, pois atuam como lubrificante (BELTRÃO, 2006a).

O algodoeiro possui um padrão particular de florescimento, tendo início com a antese da primeira flor, na primeira posição do primeiro ramo frutífero localizado mais abaixo na planta. A partir deste ramo, as flores vão surgindo de maneira espiral ascendente. Quando as flores atingem o ápice da planta, grande parte dos nutrientes e fotoassimilados são direcionados prioritariamente aos frutos em desenvolvimento, consequentemente reduzindo ou cessando a produção de flores. Este evento é denominado de ponto de maturidade fisiológica da planta e acontece quando o suprimento de carboidratos é igual à demanda, cessando o desenvolvimento vegetativo e reprodutivo (CHIAVEGATO et al., 2009).

O período compreendido entre a deiscência do primeiro ao último fruto denomina-se fase de maturação e deiscência. Esta fase tem duração de cerca de 50 dias e é nela que ocorre o desenvolvimento

das sementes e fibras e abertura dos frutos. Consequentemente, é a fase responsável pela qualidade da fibra de algodão produzida (CHIAVEGATO et al., 2009).

O fruto apresenta de três a cinco lóculos (cavidades que contêm as sementes) e é formado a partir do ovário, após o processo de fecundação. Quando jovem é chamado de maçã e, depois que se abre, de capulho. Cada lóculo geralmente contém de seis a oito sementes, cuja massa, associada ao das fibras, é chamada de peso de capulho, sendo este um dos principais componentes da produção. O fruto também pode ser chamado de carimã (ou fruto mumificado) em situações que ocorreram ataque de pragas e, ou, condições climáticas adversas (BORÉM; FREIRE, 2014).

Os frutos são compostos por uma casca (semente), a qual protege uma amêndoa e carrega a fibra nas espécies cultivadas. Dessa forma, as fibras são pelos unicelulares que se desenvolvem a partir da epiderme externa dessa semente. Cerca de 30% das células da epiderme externa diferenciam-se em fibra, sendo que as fibras mais longas (ou lint), em *G. hirsutum*, medem entre 25 a 32 mm e as fibras menores (linter) estão presentes apenas em determinadas espécies ou variedades e formam uma penugem que recobre total ou parcialmente os caroços (BACHELIER; GOURLOT, 2018).

As células que originam as fibras, embora estejam presentes no óvulo, iniciam seu crescimento imediatamente antes da antese. Esse fenômeno tem origem na chalaza, que é o ponto de entrada no óvulo dos vasos que o ligam ao ovário e representa a extremidade arredondada do caroço. Em seguida, as células continuam se propagando até a micrópila, orifício do óvulo pelo qual penetra o tubo polínico e que representa a extremidade pontiaguda do caroço. Todo esse processo gera primeiro o lint e, posteriormente, o linter, a partir do quinto dia após a antese. Um caroço/semente carrega cerca de 10 mil fibras de lint e entre 5 mil a 10 mil fibras de linter (BACHELIER; GOURLOT, 2018). A estrutura de um caroço/semente é apresentada na Figura 2.

Figura 2. Corte longitudinal de um caroço de algodão (*Gossypium hirsutum* L.). Adaptado de Bachelier (1998).

Na primeira fase de desenvolvimento do fruto do algodoeiro, ocorre a alongação das fibras, as quais são formadas a partir de células individuais localizadas na epiderme das sementes (CHIAVEGATO et al., 2009). Após a alongação das fibras, ocorre seu espessamento, a partir da deposição diária de fios de celulose na parede celular secundária das células. Ao contrário do crescimento das árvores, onde a deposição de celulose ocorre anualmente e sem sobrepor o crescimento do ano anterior, os anéis de crescimento da fibra de algodão são diários e ocorrem dentro do crescimento do dia anterior. Dessa forma, esses anéis (fios de celulose quase puros) diários da parede secundária são depositados em diferentes ângulos, construindo uma resistência de fibra semelhante à construção de um compensado de madeira (HAKE, 1990). A resistência construída da fibra será responsável pela capacidade de torção dos fios e deve ser alta o suficiente para que, ao se aplicar determinada carga nos fios, durante as etapas de beneficiamento, a fibra não se rompa.

A deposição de celulose em sucessivas camadas no interior das fibras prolonga-se até a completa maturação do fruto (maturidade fisiológica), que ocorre 40 a 45 dias após a fertilização dos óvulos (CHIAVEGATO et al., 2009). A partir daí, os frutos passam então a ser deiscentes, abrem-se e deixam aparecer caroços e suas fibras, que constituem o algodão em caroço, objeto da colheita (BACHELIER; GOURLOT, 2018).

Por fim, na fase de alongação define-se o comprimento da fibra e, na fase de deposição de celulose, define-se o diâmetro da fibra. A quantidade de celulose depositada determina as características da fibra, como a maturidade, a resistência e índice micronaire (índice que relaciona a finura e a maturidade). Essas características definem as qualidades tecnológicas da fibra, e são fortemente dependentes da interação do genótipo com o ambiente (CHIAVEGATO et al., 2009).

2.1.3.2. Estrutura de uma fibra de algodão

As fibras de algodão são estruturalmente diferenciadas, apresentam-se dentro de zonas concêntricas e possuem um miolo central profundo, chamado lúmen. Por se tratar de um pelo unicelular, a fibra de algodão tem a estrutura de uma célula não clorofiliana, ou seja, contém um núcleo, um citoplasma e organelas celulares associados. Estes componentes são circundados por uma membrana plasmática e duas paredes, uma externa ou primária e outra interna ou secundária. A parede primária é coberta por uma camada externa (cutícula) e a parede secundária regula três camadas, dispostas em espiral - uma externa, uma média e uma interna, sendo a camada média a mais grossa e essencialmente composta por celulose (Figura 3) (BACHELIER; GOURLOT, 2018).

Durante a fase de maturação no âmbito do fruto, a fibra de algodão é turgescente e cilíndrica. Porém, ao alcançar a maturidade, ela torna-se deiscente e desidrata-se e, conseqüentemente, o citoplasma retrai-se, deixando um canal central vazio na fibra, o lúmen. Assim, submetida às limitações decorrentes das diferenças de composição e orientação de suas diversas camadas, bem como à perda de pressão osmótica interna, a fibra se contorce e o lúmen confere-lhe propriedades mecânicas de fiabilidade,

permitindo que um feixe de fibra de algodão consiga conservar sua coesão após a torção para formar um fio resistente (BACHELIER; GOURLOT, 2018).

Figura 3. Representação esquemática da estrutura de uma fibra madura de algodão. Adaptado de Parry (1981).

2.1.3.3. Composição de uma fibra de algodão

A composição da fibra de algodão varia conforme seu estágio de desenvolvimento e a estrutura que está sendo formada. A primeira estrutura formada é a cutícula, constituída de ceras e graxas. Durante o crescimento da fibra são gerados os componentes da parede primária, pectinas, hemicelulose, lignina e celulose. Na fase de maturação é produzida apenas celulose, único composto que preenche a parede secundária da fibra. Em termos de proporção, a fibra é composta por: 94 a 96% de celulose; 1 a 2 % de proteínas; 0,7 a 1,2% de pectinas; 0,4 a 1% de ceras; 0,3% de açúcares fisiológicos; e entre 0,5 e 3% de outros compostos (MARTINS et al., 2011; BACHELIER; GOURLOT, 2018).

2.1.3.4. Caracterização de uma fibra de algodão

As fibras são obtidas a partir do beneficiamento do algodão em caroço. A pluma é separada do caroço por um processo mecânico denominado descaroçamento, por meio do uso de descaroçadores, do tipo rolo ou de serra (ALMEIDA et al., 2011). A qualidade da fibra tem grande influência na comercialização e industrialização do algodão, garantindo valorização e rentabilidade para o agricultor. Para a indústria, a qualidade da fibra é influenciada por diversos fatores e sua determinação depende do completo conhecimento de suas diversas características (FREDDI et al., 2014).

Até 1995, as análises de qualidade de fibra do algodão eram feitas manualmente, o que dificultava a determinação de certas características, e atrasava a obtenção dos dados. Posteriormente, a qualidade da fibra começou a ser determinada por aparelhos como o *High Volume Instrument* (HVI), o qual fornece rapidamente informações relacionadas à qualidade de fibra a partir de uma pequena quantidade de amostras (USTER TECHNOLOGIES, 2008). Atualmente, os maiores países produtores e consumidores de algodão adotam o padrão universal proposto pelo USDA para a classificação da fibra do algodoeiro. Esses padrões são atualizados anualmente e os aparelhos do tipo HVI são regulados de acordo com essas mudanças (SESTREN; LIMA, 2015).

As principais características de interesse para a qualidade da fibra de algodão são:

a) Micronaire

O índice micronaire é adimensional e definido pela relação entre a finura e a maturidade da fibra. A sua determinação se dá por meio de um ensaio no qual a massa de fibras do corpo de prova é submetida a um fluxo de ar, à pressão constante, em câmara de volume definido. Para realizar esse ensaio, a massa de fibras do corpo de prova é inserida na câmara do aparelho micronaire, onde é submetida a uma pressão de ar comprimido. Os valores da massa do corpo de prova, do volume da câmara e da pressão do ar comprimido injetado são constantes. Em seguida, o aparelho determina a vazão do ar que passou pela massa de fibras, que é variável e que está em função do diâmetro externo e da espessura da parede secundária das fibras, formada por camadas de celulose, e que também estão relacionadas diretamente com a maturidade das fibras (DE LIMA, 2018).

A finura da fibra pode ser representada pelo seu perímetro, diâmetro, ou massa linear, expressa em militex (mtex), que é a milésima parte do tex, equivalente à massa em gramas de mil metros de material têxtil na forma de fibra, manta fita, pávio e fio (BACHELIER; GOURLOT, 2018).

A maturidade, por sua vez, indica a espessura das camadas de celulose da parede secundária em relação ao diâmetro externo da fibra. Em outras palavras, é a relação da área em corte da parede secundária de uma fibra sobre a área de um círculo de mesmo perímetro (valores compreendidos entre zero e um). Fibras com parede secundária espessa e lúmen reduzido são classificadas como maduras, enquanto fibras que não apresentam espessamento na parede secundária são classificadas como imaturas (BACHELIER; GOURLOT, 2018; FONSECA; FARIAS, 2002).

Desta forma, o micronaire, apesar de indireto, é o método mais prático de se medir a relação entre a finura e a maturidade da fibra. A finura e maturidade também podem ser medidas por métodos diretos. No caso da finura, por meio de observações microscópicas das fibras dispostas longitudinalmente, sob iluminação intensa. Já no caso da maturidade, por meio de observações microscópicas de cortes em seções retas de fibra para avaliar Theta (grau de maturidade das fibras) (BACHELIER; GOURLOT, 2018). Portanto, deve-se ressaltar que a escala do índice micronaire foi estabelecida empiricamente com um conjunto de algodão padrão e não é linear, já que depende da finura e da maturidade (DE LIMA, 2018).

Quando as fibras são pouco maduras, significa que tiveram o processo de maturação prejudicado por alguma razão e, apresentarão uma forma achatada quando observadas de modo transversal, por conta da ausência das paredes internas de celulose. Por outro lado, se estiverem bem desenvolvidas, terão formas arredondadas ou elípticas, indicando o preenchimento bem sucedido das camadas internas (CORREA, 1965). Para entender melhor essa configuração, a Figura 4 mostra a representação de uma fibra imatura (A), de uma fibra de maturação média (B) e de uma fibra madura (C). De acordo com a Figura 4, é possível entender melhor a relação entre a maturidade e a finura e o índice micronaire das fibras de algodão. Fibras imaturas representam fibras finas, de diâmetro externo pequeno, parede secundária estreita e um lúmen acentuado (micronaire muito baixo); fibras de maturação média representam fibras finas e pouco maduras, de diâmetro externo pequeno, parede secundária espessa e lúmen pequeno (micronaire baixo); e fibras maduras representam fibras grossas, de diâmetro externo grande, parede secundária espessa e lúmen pequeno (micronaire alto) (BACHELIER; GOURLOT, 2018).

Figura 4. Representação esquemática de fibra imatura (A); fibra de maturação média (B); e fibra madura de aspecto cilíndrico e com lúmen reduzido. Fonte: Correa (1965).

Os valores de micronaire geralmente variam de 2,0 a 6,0. As fibras com valores abaixo de 2,9 são consideradas muito finas; as com valores entre 3,0 e 3,6 finas; entre 3,7 a 4,7 médias; entre 4,8 e 5,4 grossas; e valores acima de 5,5 representam fibras muito grossas (BACHELIER; GOURLOT, 2018).

b) Comprimento

O comprimento da fibra é a extensão média ou a mais constante ao longo do eixo de um corpo de prova de fibras paralelas previamente preparado. As fibras são paralelizadas por escova rotatória e então mensuradas por aparelhos como os do tipo HVI. O comprimento médio da metade superior (*Upper Half Mean Length* – UHML) é o mais utilizado no processo de classificação e comercialização do algodão e é definido como o comprimento médio dos 50% das fibras mais longas, expresso em polegadas ou milímetros. Já o comprimento médio (*Mean Length* - ML) é o comprimento médio de 100% das fibras (BACHELIER; GOURLOT, 2018).

Em termos de qualidade, o comprimento de fibra influencia no número de torções por unidade de comprimento dos fios. Quanto menor este número, maior a produtividade das máquinas e melhor a obtenção de tecidos macios e com melhor caimento (DE LIMA, 2018).

c) Resistência

A resistência específica (ou tenacidade à rotura da fibra) é a força requerida, em gramas, para romper um feixe de fibras de um Tex. Define-se como Tex a massa em gramas de mil metros de material têxtil em forma de fibras, mantas, fitas, pavios e fios. A força máxima ou força de rotura (*Max Force*), mensurada pelos aparelhos do tipo HVI, é a força necessária para romper o feixe de fibras amostrado e é expressa em gramas-força por tex (gf tex^{-1}) (BACHELIER; GOURLOT, 2018). A resistência pode variar de fraca (menor ou igual a 23 gf tex^{-1}) até muito resistente (acima de 31 gf tex^{-1}) (FONSECA; FARIAS, 2002).

d) Uniformidade

A uniformidade do comprimento da fibra (%UI) é a relação entre o comprimento médio de 100% das fibras (ML) e o comprimento médio dos 50% das fibras mais longas (UHML) e é expresso em porcentagem (%) (BACHELIER; GOURLOT, 2018). Essa característica representa a homogeneidade do comprimento das fibras do fardo e, portanto, é uma inferência da qualidade da fiação. O índice de uniformidade pode variar de muito baixo (menor que 77%) a muito alto (maior que 85%) (SESTREN; LIMA, 2015).

e) Elongação

O alongamento ou elongação à rotura da fibra (% Elg) é o quanto o feixe de fibras cede no sentido longitudinal até o momento da rotura e é expresso em percentual, em função do comprimento inicial do corpo de prova. A elongação pode variar de muito baixa (valores abaixo de 5%) até muito alta (valores acima de 7,7%) (BACHELIER; GOURLOT, 2018).

Por representar o quanto o material de fibra cede no sentido longitudinal até o momento da rotura, a elongação é uma característica associada à resistência. Fibras resistentes e alongadas geram fios e tecidos mais resistentes, capazes de suportar a máxima velocidade de trabalho do maquinário industrial, o que pode aumentar a produtividade têxtil, bem como diminuir os custos de produção (VIOLATTI, 2016).

f) Índice de fibras curtas

O índice de fibras curtas (%SFI ou %SF) é o percentual de fibras menores que 12,7 mm presentes nos corpos de prova. Seus valores podem variar de muito baixos ($\leq 6\%$) até muito altos ($\geq 18\%$) (BACHELIER; GOURLOT, 2018).

O índice de fibras age em conjunto com o índice de uniformidade, no sentido de que um alto conteúdo de fibras curtas e um baixo índice de uniformidade podem provocar a flutuação de fibrilas e de impurezas nas zonas de estiação de diversas máquinas no processo de fiação, gerando acúmulo de pó e micropó, que contribuem de forma negativa para a produção, qualidade e custo do processo e dos produtos têxteis (FONSECA; FARIAS, 2002).

Os valores de referência para os indicadores de qualidade da fibra do algodão, tendo como referência as exigências da indústria têxtil, são apresentados na Tabela 3.

Tabela 3. Valores de referência dos indicadores de qualidade da fibra de algodão com base nas exigências da indústria têxtil.

Caracteres	Valores de referência da indústria têxtil
Micronaire (adimensional)	3,8 - 4,5
Comprimento (mm)	> 30
Resistência (gf tex ⁻¹)	> 31
Uniformidade (%)	85 - 87
Alongamento ou Elongação (%)	6,8 - 7,7
Índice de fibras curtas (SFI) (%)	< 6

Fonte: De Lima (2018).

2.1.4. Condicionantes agrometeorológicos da qualidade da fibra de algodão

A fibra do algodão é uma célula simples, sem vida, resultante de um processo biológico que se realiza por um período que varia entre cinquenta e setenta dias após a fertilização da flor. No entanto, durante seu desenvolvimento, ela é um organismo vivo, decisivamente afetado por fatores ambientais, de modo que suas propriedades podem apresentar grande variação dependendo do ano, da época de semeadura, do local, das propriedades do solo etc. Assim, é fundamental entender a influência desses fatores sobre as propriedades intrínsecas finais da fibra de algodão, que serão determinantes para definição de seu valor comercial para matéria-prima de fiação (ESCHER et al., 2018).

2.1.4.1. Temperatura

A temperatura é o principal fator ambiental que interfere no crescimento e no desenvolvimento do algodoeiro e o mais importante para a formação da qualidade da fibra (ESCHER et al., 2018). O algodoeiro possui mecanismos de ajustes que possibilitam o seu cultivo em regiões com temperaturas

entre 18°C e 40°C, porém a faixa ideal é de 20°C a 30°C (REDDY et al., 1992). A temperatura interfere no desenvolvimento da cultura do algodão, afetando expressivamente a fenologia, a expansão foliar, a alongação dos internódios, a produção de biomassa e a partição dos assimilados pelas diferentes partes da planta, entre outros aspectos (CHIAVEGATO et al., 2009).

A maioria das cultivares de algodoeiro requer mais de 150 dias, com temperaturas acima de 15°C (temperatura base), para produzir satisfatoriamente, com paralisação do metabolismo em temperaturas inferiores à basal. Além disso, é bastante sensível às condições mais frias e, principalmente às geadas (FREELAND, 2006).

Temperaturas inferiores a 20°C durante o desenvolvimento do fruto podem reduzir o comprimento das fibras, além de afetar outras características tecnológicas dependentes da deposição de celulose, já que esta variável induz à redução do metabolismo celular que envolve organelas comprometidas na síntese dos componentes da fibra, dos quais a celulose é o mais importante, representando mais de 94% da fibra madura (BELTRÃO; AZEVEDO, 1993). A Tabela 4 mostra os limites de temperaturas para as diferentes fases do algodoeiro herbáceo.

Tabela 4. Limites de temperatura para diferentes etapas de algodoeiro.

Etapas de desenvolvimento	Limite mínimo	Limite ideal	Limite máximo
Germinação	14°C	18 a 30 °C	40°C
Formação de gemas e floração	Diurna - 20°C	30°C	Diurna - 40°C
	Noturna - 12°C		Noturna- 27°C
Desenvolvimento e maturação dos frutos	20°C	27 a 32 °C	38°C

Fonte: Embrapa (2006).

Bange et al. (2010) desenvolveram um modelo para estimar o efeito das temperaturas mínimas e médias nos valores de micronaire, o qual apresentou uma acurácia de 68%. Dessa forma, para obtenção de fibras com micronaire na faixa premium (3,7 a 4,2), os valores de temperatura mínima e média deveriam ser da ordem de 15 a 18°C e de 22 a 25°C, respectivamente (Figura 5).

Figura 5. Relação entre as temperaturas mínima e média e o micronaire das fibras do algodoeiro. Fonte: Bange et al. (2010).

Temperaturas elevadas durante o período de desenvolvimento das fibras tem grande influência na expressão do genótipo quanto às propriedades da fibra produzida. Características como o comprimento, a uniformidade e a finura da fibra são favorecidas quando produzidas sem estresse térmico (temperaturas elevadas) (CHIAVEGATO et al., 2009). Segundo Lokhande e Reddy (2014), a resistência da fibra aumenta linearmente com o aumento da temperatura e isso ocorre devido à menor retenção de capulhos sob condições de temperaturas maiores do que 30°C, devido à modificação da relação fonte-dreno.

Além disso, também sob temperaturas elevadas, ocorre o aumento da taxa respiratória e, consequentemente, da fotorrespiração, características das plantas C3, como é o caso do algodoeiro. A fotorrespiração ocorre com menor intensidade quando a temperatura ambiente se encontra abaixo de 25°C, porém pode representar aproximadamente 50% da fotossíntese aparente quando a temperatura alcança 35°C, tornando a assimilação líquida ineficiente. Em termos práticos, temperaturas elevadas, ocorrendo em qualquer fase de desenvolvimento do algodoeiro, comprometem a demanda de fotoassimilados e, consequentemente, a produção final (CHIAVEGATO et al., 2009).

Yeates et al. (2014) mensuraram o efeito da época de semeadura e a variação da temperatura sobre os seguintes parâmetros de qualidade da fibra: comprimento; resistência; e índice micronaire. Para atingir os objetivos do trabalho, foram analisados resultados de experimentos conduzidos durante três anos na Austrália com duas cultivares de *G. hirsutum* L., em quatro épocas de semeadura. Os autores concluíram que o comprimento de fibra foi influenciado de maneira quadrática ($R^2 = 0,82$) pela temperatura mínima durante o período de formação da fibra, sendo os melhores resultados obtidos entre as temperaturas 17 e 19°C. Além disso, houve correlação positiva e significativa entre a temperatura média e a resistência da fibra e não houve efeito significativo da época de semeadura e da temperatura no índice micronaire.

2.1.4.2. Radiação solar

O algodoeiro é uma planta de metabolismo fotossintético tipo C3 e apresenta alta taxa de fotorrespiração e baixa eficiência na fixação de CO₂ (BELTRÃO; AZEVEDO, 1993). Para que a planta mantenha elevadas taxas fotossintéticas durante seu ciclo, é recomendável que a nebulosidade seja inferior a 50%, sem ocorrência de inversão térmica (dias muito quentes e noites muito frias) e inexistência de alta umidade relativa associada a altas temperaturas (GONIAS et al., 2012).

Sob baixas condições de luminosidade, as plantas produzem menor quantidade de frutos e fibras de qualidade inferior, principalmente com baixo índice micronaire. A baixa qualidade da fibra está associada à diminuição dos níveis de carboidratos disponíveis, indicativo da redução no nível de fotoassimilados produzidos sob baixa condição de radiação solar (FREELAND, 2006).

Durante a fase reprodutiva do algodoeiro, a redução da produção de carboidratos nas folhas pela fotossíntese devido ao sombreamento limita o fluxo de sacarose para frutos jovens (YANG et al., 2016). Essa limitação nos frutos pode desencadear uma série de respostas fisiológicas que levam à menor produtividade e qualidade das fibras. Esses efeitos, de modo geral, podem ser divididos em direto e indireto.

O efeito direto é o suprimento insuficiente de carboidratos aos frutos, reduzindo a produção de polímeros de celulose, levando à formação de fibras imaturas. Já o efeito indireto é decorrente da influência de outros fatores ambientais associados ou subsequentes ao período de restrição de luz (ESCHER et al., 2018).

No efeito indireto, a redução do fluxo de carboidratos para botões florais e frutos recém-formados resulta em sua abscisão, podendo levar à redução do número de capulhos por planta e por área (ECHER; ROSOLEM, 2014; CHEN et al., 2017). A abscisão de tais estruturas aumenta a proporção de frutos tardios, sujeitos nas condições tropicais a ambientes mais limitantes em temperatura e disponibilidade hídrica do solo durante sua formação. Esse efeito é intensificado quando a produção perdida é parcial ou totalmente compensada em posições de frutificação mais tardias (PETTIGREW et al., 1992).

Se as temperaturas e a disponibilidade hídrica do solo forem baixas, o peso dos capulhos poderá ser reduzido, as fibras poderão ser imaturas e apresentar menor resistência. Chen et al. (2017) demonstraram, em um experimento com duração de dois anos, no qual a redução da radiação solar no dossel do algodoeiro foi proporcionado pelo uso de telas, com 62 e 79% de sombreamento, redução do peso do capulho e da qualidade intrínseca das fibras, quando o sombreamento foi imposto durante o período de formação da maçã (48 dias). A Figura 6 ilustra os resultados desse experimento, em que se observa a redução de todas as características do algodoeiro avaliadas com a redução da disponibilidade de radiação solar.

De acordo com Zhao et al. (2012), a redução do índice micronaire associado ao aumento da abscisão de frutos e da proporção de frutos tardios do algodoeiro podem ocorrer se as plantas forem sombreadas em diferentes épocas de florescimento. Segundo os mesmos autores, a redução de 63% da

radiação fotossintética ativa com o uso de tela de cor preta sobre o dossel durante oito dias, em diferentes fases do desenvolvimento reprodutivo do algodoeiro, resultou em aumento do número de abortamento de frutos e redução da produtividade e do índice micronaire, como ilustra a Figura 7.

Figura 6. Relação entre a disponibilidade relativa de luz e o peso de capulho e a qualidade intrínseca das fibras do algodoeiro durante todo o período de formação da maçã (48 dias). Adaptado de Chen et al. (2017).

Figura 7. Efeito de oito dias de sombreamento em diferentes fases do desenvolvimento reprodutivo do algodoeiro (primeira flor, pico do florescimento e frutificação) e seus efeitos na abscisão relativa dos frutos, produtividade relativa e índice micronaire (adimensional), em comparação à condição sem sombreamento. Adaptado de Zhao et al. (2012).

Assim, as épocas mais críticas de sombreamento incluem o início e o pico de florescimento e o período inicial de frutificação (ESCHER et al., 2018). Chen et al. (2014) observaram, durante o período de formação das fibras, que, quando as plantas ficam sujeitas aos efeitos de sombreamento, as fibras apresentaram menor acúmulo de sacarose e menor atividade das enzimas sacarose sintase e invertase, que apresentam importante função na elongação e na formação dos polímeros de celuloses. Nesse experimento, os referidos autores avaliaram o efeito do sombreamento sobre o comprimento das fibras de algodão, visando a obtenção de respostas fisiológicas durante o processo de sua formação em duas cultivares e três épocas de semeadura. Para isso, a disponibilidade relativa de luz foi estabelecida pela ausência (100%) ou pela presença de telas de sombreamento durante toda a formação do fruto para obtenção de 60 a 80% da radiação fotossinteticamente ativa original (Figura 8).

Figura 8. Efeito relativo da disponibilidade relativa de radiação fotossinteticamente ativa no comprimento e respostas fisiológicas durante o processo de formação das fibras de algodão. Fonte: Chen et al. (2014).

Por fim, em relação aos efeitos da restrição de luz sobre a qualidade intrínseca da fibra, resultados experimentais indicam que o sombreamento prolongado nas fases críticas de florescimento e frutificação pode reduzir o peso dos capulhos, a resistência e o índice micronaire, sendo este último o atributo mais frequentemente afetado (ZHAO et al., 2012; LV et al., 2013; LIU et al., 2015; CHEN et al., 2017). Assim, os maiores riscos estão relacionados à formação de fibras imaturas. No entanto, embora menos provável, também é possível que o sombreamento possibilite um aumento no índice micronaire, mesmo com a compensação da produção em épocas mais tardias. Esse fenômeno é possível, por exemplo, se os frutos mais tardios se desenvolverem em condições menos limitantes de temperatura ou umidade do solo do que os que foram abortados (ECHER; ROSOLEM, 2015).

Os efeitos sobre o comprimento da fibra não são consistentes e apresentam elevada variabilidade. Em alguns experimentos, o comprimento não foi afetado pelo sombreamento (ZHAO et al., 2012; ECHER; ROSOLEM, 2015). Todavia, aumentos de comprimento foram observados na China, em ambientes de maiores temperaturas (LV et al., 2013; CHEN et al., 2014), estando associados à maior duração do período de fechamento dos plasmodesmos das células da fibra, gerando maior turgor osmótico interno para a elongação. Por outro lado, em temperaturas baixas, a redução na atividade de enzimas e no fluxo de carboidratos levou à diminuição da taxa de elongação, reduzindo o comprimento final (CHEN et al., 2014). Nesse sentido, é possível que semeaduras muito tardias resultem em fibras mais curtas se houver sombreamento prolongado durante a formação das maçãs, uma vez que a redução da temperatura pode tornar-se restritiva para a elongação (ESCHER et al., 2018). No entanto, ressalta-se que temperaturas baixas são mais restritivas para a maturação do que para a elongação e, além disso, é possível

também que, dependendo do ambiente, o sombreamento não afete a qualidade da fibra. Se o período de restrição luminosa for muito curto, ou se as condições após esse período não forem limitantes, os frutos remanescentes na lavoura não terão sua formação comprometida (ECHER; ROSOLEM, 2015).

2.1.4.3. Disponibilidade hídrica do solo

As necessidades hídricas do algodoeiro variam com as fases fenológicas, em função do crescimento da cultura, assim como das condições meteorológicas durante seu ciclo. Nas fases iniciais da cultura, logo após a emergência, a necessidade hídrica é menor. Ao passo que as plantas crescem, a demanda também se eleva até a fase reprodutiva, a qual coincide com a máxima área foliar, quando a cultura atinge o pico da demanda por água (CHIAVEGATO et al., 2009).

Durante o ciclo da cultura, que varia de 130 a 170 dias, a necessidade hídrica geralmente é de 700 a 1000 mm, valor que pode variar conforme as condições meteorológicas locais (FUZZATO et al., 2014). Ressalta-se que, para o algodoeiro, não é suficiente apenas a quantidade, mas também, a distribuição de água ao longo do ciclo, no sentido de que a oferta hídrica seja adequada para atender às demandas específicas de cada estágio ou fase fenológica (CHIAVEGATO et al., 2009). Na Tabela 5 são apresentados os valores médios de evapotranspiração máxima diária para cada fase do desenvolvimento da cultura do algodoeiro.

Tabela 5. Valores médios de evapotranspiração diária para cada fase de desenvolvimento da planta de algodoeiro.

Estádio de desenvolvimento	Necessidade hídrica (mm)
Semeadura	< 1
Desenvolvimento vegetativo	1 a 2
Primeiro botão floral	2 a 4
Fase reprodutiva	3 a 8
Pico de florescimento	8
Primeiro capulho	4 a 8
Fase de maturação	4

Fonte: Freeland (2006).

O estresse hídrico ocorre quando a normalidade de um processo fisiológico da planta é afetada pelo déficit ou pelo excesso de água. No contexto agrônomo, o estresse hídrico se refere a um conjunto de respostas das plantas que levam a uma redução do crescimento e produtividade da cultura (YEATES, 2014). Em outras palavras, caracteriza-se por uma oferta em maior ou menor quantidade de água demandada pelo algodoeiro em determinada fase fenológica, ou seja, eventos de seca ou de encharcamento. Cabe ressaltar que durante o ciclo da cultura do algodão, em uma mesma safra, existe a possibilidade da ocorrência dessas duas situações (ESCHER et al., 2018).

A fibra de algodão, possui duas fases principais de desenvolvimento: o crescimento e o enchimento. Por ser uma célula, como qualquer outra da planta, depende de água para sua expansão e,

portanto, para seu crescimento. Dessa forma, o que se espera é que, em condições de seca, haja uma redução do comprimento da fibra. Além disso, o secamento do solo induz à produção de ácido abscísico, fito-hormônio que controla a abertura e o fechamento dos estômatos, reduzindo o potencial hídrico da folha, interferindo negativamente na interceptação luminosa, na assimilação de carbono e, conseqüentemente, na disponibilidade de carboidratos para os órgãos reprodutivos (ALLEN; ALEMAN, 2011).

Com o objetivo de verificar o impacto do déficit hídrico sobre as características qualitativas da fibra de algodão, Davis et al. (2014) observaram que o micronaire aumentou em função do déficit hídrico e Lokhande e Reddy (2014) observaram redução da biomassa total, número de capulhos, peso de sementes e produtividade do algodão em resposta ao déficit hídrico, o qual provocou declínio da atividade fotossintética. Em adição, houve redução linear do comprimento da fibra, da resistência e da uniformidade com o decréscimo do potencial hídrico das folhas (quanto mais negativo, mais seco) e aumento do micronaire em resposta ao aumento do déficit hídrico (Figura 9). Os autores pontuaram que a resistência e o micronaire foram as características mais responsivas ao estresse por seca, seguido pelo comprimento e pela uniformidade. Também o conteúdo de fibras imaturas aumentou e a maturidade diminuiu como resposta à redução do potencial hídrico das folhas.

Figura 9. Relação entre o potencial hídrico da folha e as características da fibra de algodão. Fonte: Lokhande e Reddy (2014).

Eventos de encharcamento do solo são frequentes em algumas regiões produtoras de algodão, principalmente no estado do Mato Grosso, em microrregiões que sofrem grande influência da região Amazônica, o que leva à redução da disponibilidade de oxigênio às raízes do algodoeiro, dificultando a absorção de nutrientes, especialmente o nitrogênio e o ferro, diminuindo, assim, a área foliar, a interceptação luminosa, a fixação de carbono e, por fim, o crescimento e o desenvolvimento (ESCHER et al., 2018). Entretanto, resultados da literatura são controversos com relação à resposta da

qualidade da fibra ao encharcamento do solo. Por exemplo, Bange et al. (2004) não observaram efeito do encharcamento sobre o comprimento da fibra e sobre o micronaire em um estudo de campo durante três safras na Austrália. Por outro lado, Wang et al. (2017), em ambiente controlado, observaram que o encharcamento do solo no florescimento reduziu o comprimento da fibra, aumentou o micronaire (exceto no estágio de botões florais) e reduziu a resistência quando imposto nas fases de florescimento e de formação do capulho. Particularmente, a resistência da fibra foi inversamente proporcional à quantidade de capulhos retidos na planta, assim, as plantas com menos capulhos apresentaram maior resistência, o que se justifica pela alteração da relação fonte dreno (Figura 10) (ESCHER et al., 2018).

Figura 10. Efeito do período de encharcamento do solo (2, 4, 6, 8 e 10 dias) em diferentes fases fenológicas do algodoeiro sobre as características tecnológicas da fibra e número de capulhos por planta. Adaptado de Wang et al. (2017).

2.1.5. Modelagem agrícola e mineração de dados

2.1.5.1. Modelagem agrícola

Os modelos agrícolas integram conhecimentos de várias disciplinas, incluindo agrometeorologia, física e química dos solos, fisiologia das culturas, melhoramento vegetal e agronomia em um conjunto de equações matemáticas com o objetivo de estimar o crescimento e o desenvolvimento vegetal bem como as produtividades das culturas (HOOGENBOOM, 2000). De acordo com Thornley e France (2017), os modelos podem ser categorizados como sendo empíricos, fisiológicos-matemáticos, estocásticos ou mecanísticos.

Os modelos empíricos são baseados na relação entre dados observados e variáveis do ambiente. Já os fisiológicos-matemáticos explicam por meio de equações matemáticas alguns dos processos fisiológicos associados à produção de biomassa, em função de variáveis meteorológicas, enquanto os modelos estocásticos são baseados em processos descritos por leis de probabilidade. Finalmente, os modelos mecanísticos são baseados em processos que envolvem leis físicas, químicas e biológicas, sendo, portanto, mais próximos do que ocorre na realidade. Desde que produzam resultados compatíveis com a realidade do local, qualquer modelo de simulação pode ser utilizado (DIAS, 2016).

Embora a análise estatística de informações oriundas de dados de campo para o entendimento dos padrões e mecanismos que influenciam a variabilidade da produtividade ou qualidade de uma dada cultura traga resultados promissores, tal abordagem com o uso de ferramentas estatísticas tradicionais muitas vezes é inviável, em função de problemas na estrutura dos bancos de dados, interações de ordem mais elevada e até mesmo da não linearidade dos dados. Para contornar esses problemas, métodos não paramétricos de análise estatística estão sendo utilizados. Dentre esses métodos, destacam-se os baseados em mineração de dados e inteligência artificial, os quais viabilizam as análises de relações em grandes bancos de dados (Big Data) e que envolvem uma grande gama de variáveis (FERRARO et al., 2009).

2.1.5.2. Mineração de dados

Nos últimos anos, importantes ferramentas de coleta e armazenamento de dados foram desenvolvidas e aplicadas de forma acelerada em diversos campos do conhecimento, produzindo um enorme volume de dados. Entretanto, ao mesmo tempo que esse grande volume de dados é importante para a ciência, ele também pode ser um problema, visto que a capacidade de interpretação humana nem sempre acompanha tal demanda. Nesse sentido, a Mineração de Dados (MD) surge como uma ferramenta capaz de elucidar os padrões implícitos nesses grandes volumes de dados (WITTEN et al., 2011).

A MD, também conhecida como descoberta de conhecimento a partir de dados (*Knowledge Discovery from Data* - KDD), pode ser definida como a extração de padrões que representam o conhecimento armazenado em grandes bancos de dados. O processo de MD integra técnicas oriundas de

diferentes áreas de conhecimento, como tecnologia de bancos de dados, estatísticas, computação, processamento e visualização de dados (HAN et al., 2011). As fases do processo de KDD são apresentados na Figura 11.

Figura 11. Fases do processo de descoberta de conhecimento a partir de dados (KDD). Adaptado de Han et al. (2011).

Em resumo, as etapas do processo de KDD são:

- Identificação do problema - compreender o domínio da aplicação e o tipo de conhecimento que se busca no banco de dados, de modo a traçar um objetivo;
- Realização da seleção do conjunto de dados - o analista define o conjunto de dados no qual a descoberta deve ser realizada;
- Pré-processamento dos dados - etapa na qual são realizadas operações básicas de limpeza de dados, remoção dos ruídos e tratamento dos valores faltantes;
- Transformação dos dados - é a etapa importante para promover melhor representação dos dados, visando à redução de dimensionamento ou balanceamento entre as classes existentes, para então serem aplicados algoritmos específicos na etapa de MD (modelagem) em busca dos padrões interessantes e
- Interpretação dos padrões encontrados - tem como finalidade descobrir se os resultados apresentam avanço do conhecimento e avaliar o modelo, a partir de métricas que garantam sua validação.

2.1.5.3. Tarefas e técnicas de mineração de dados

A tarefa de MD consiste na especificação do que se pretende buscar nos dados, quais os tipos de regularidades ou categorias de padrões que se deseja encontrar. A partir da escolha da tarefa, pode-se

definir qual o melhor algoritmo (método) que se aplica ao conjunto de dados para descobrir os padrões surpreendentes. De maneira geral, as possíveis tarefas em que o algoritmo pode atuar estão divididas em dois grupos: tarefas preditivas e tarefas descritivas (REZENDE et al., 2002).

As tarefas preditivas consistem em se fazer previsões a partir de inferências sobre os dados em questão e se dividem em dois grupos: classificação e regressão. A tarefa preditiva de classificação pode ser definida como um processo para encontrar modelos que descrevem e distinguem classes. Esses modelos são gerados com base na análise de um banco de dados de treinamento (em que os dados são conhecidos) e depois são utilizados para prever a categoria dos objetos, cujas classes não são conhecidas, podendo ser representadas de diferentes formas, como por regras, árvores de decisão, fórmulas matemáticas ou redes neurais. Por outro lado, a tarefa preditiva de regressão trata-se da análise de regressão, metodologia estatística usada principalmente para previsões numéricas. Enquanto a tarefa preditiva de classificação estima fatores categóricos (discretos), os modelos de regressão estimam funções de valores contínuos (HAN et al., 2011).

As tarefas descritivas têm o foco na caracterização das propriedades de forma geral dos dados e podem ser divididas em: associação ou “clusterização” (agrupamento) (HAN et al., 2011). A tarefa descritiva de associação consiste em identificar o relacionamento entre atributos (CAMILO; SILVA, 2009; BARROS, 2013). Ao contrário das técnicas de classificação e regressão, que analisam conjunto de dados categorizados, as análises de agrupamentos (*cluster*) são destinadas aos dados cuja distribuição em classes é desconhecida. A “clusterização” pode ser utilizada para gerar um agrupamento de dados com o intuito de formar diferentes classes. Os objetos são agrupados baseados no princípio de maximização das similaridades dentro de uma classe (*intracluster similarity*) ou da minimização das similaridades entre as classes (*intercluster similarity*). Os agrupamentos são gerados de forma que os objetos dentro de um cluster tenham alta similaridade em comparação aos demais, ao mesmo tempo em que são diferentes dos objetos de outros clusters. Assim, cada cluster pode ser visto como uma classe de objetos, da qual podem ser derivadas regras (HAN et al., 2011).

É importante ressaltar que, para a realização de cada tarefa, diferentes técnicas podem ser utilizadas. Não há uma técnica ideal, cada uma delas apresenta suas particularidades, vantagens e desvantagens. Desse modo, será a análise do problema em questão que definirá a escolha da técnica ou das técnicas a serem utilizadas, uma vez que se pode aplicar mais de um método para solucionar um mesmo problema (VIEIRA, 2014). Dentre as diversas técnicas disponíveis para realizar a tarefa de previsão, destacam-se as seguintes: *Boosting* e *Random forest* (pertencentes ao grupo de árvores de decisão) e *Support Vector Machine*.

2.1.5.4. Análise de agrupamentos (Clusters)

A análise de agrupamento, ou análise de clusters, é uma técnica usada para classificar objetos ou casos em grupos relativamente homogêneos chamados de agrupamentos ou conglomerados. Assim, os

objetos em cada agrupamento tendem a ser semelhante entre si, mas diferentes de objetos em outros agrupamentos. Dessa forma, três questões fundamentais devem ser consideradas na aplicação da análise de agrupamento: primeira - como será medida a similaridade de dados; segunda - como formar os agrupamentos; e terceira - como decidir quantos grupos formar (MALHOTRA, 2006). O procedimento para se efetuar a análise de agrupamentos é mostrado na Figura 12.

Figura 12. Etapas para efetuar a análise de agrupamento. Adaptado de Malhotra (2006).

Segundo Hair et al. (2005), as características de cada objeto são combinadas em uma medida de semelhança, que pode ser de similaridade ou dissimilaridade, calculada para todos os pares de objetos, possibilitando a comparação de qualquer objeto com outro pela medida de similaridade e a associação dos objetos semelhantes por meio da análise de agrupamento. As medidas de distância representam a similaridade, que é representada pela proximidade entre as observações ao longo das variáveis.

A distância euclidiana é a medida de distância mais frequentemente empregada quando todas as variáveis são quantitativas. A distância euclidiana é utilizada para calcular medidas específicas, assim como a distância euclidiana simples e a distância euclidiana quadrática ou absoluta, que consiste na soma dos quadrados das diferenças, sem calcular a raiz quadrada. A distância euclidiana quadrática (DE) é dada por:

$$DE = \sum_{j=1}^p (x_{ij} - x_{ij})^2 \quad (1)$$

em que: x_{ij} é a j-ésima característica do i-ésimo indivíduo; e $x_{i'j}$ é a j-ésima característica do i'-ésimo indivíduo. Quanto mais próximo de zero for a distância euclidiana, mais similares são os objetos comparados (HAIR et al., 2005).

Existem dois tipos de aglomeração: a hierárquica e a não-hierárquica. A aglomeração hierárquica se caracteriza pelo estabelecimento de uma hierarquia ou estrutura em forma de árvore e interliga os objetos por suas associações, produzindo uma representação gráfica denominada de dendrograma, onde os objetos semelhantes, segundo as variáveis estudadas, são agrupados entre si. Já na aglomeração não hierárquica, assume-se um centro de agrupamento e, em seguida, agrupam-se todos os objetos que estão a menos de um valor pré-estabelecido do centro (MANLY, 2008).

Dentre os métodos de agrupamento hierárquico, o mais conhecido é o método de Ward. O método de Ward consiste em um procedimento de agrupamento hierárquico no qual a medida de similaridade usada para juntar agrupamentos é calculada como a soma dos quadrados entre os dois agrupamentos feita sobre todas as variáveis. Esse método tende a resultar em agrupamentos de tamanhos aproximadamente iguais devido à sua minimização de variação interna. Assim, em cada estágio, combinam-se os dois agrupamentos que apresentarem menor aumento na soma global de quadrados dentro dos agrupamentos (HAIR et al., 2005).

Os dados para uma análise de agrupamentos consistem dos valores de p variáveis $X_1, X_2, \dots, X_p$ para n objetos. Para algoritmos hierárquicos, estes valores são então usados para produzir um arranjo de distâncias entre os objetos. Entretanto, usualmente as variáveis devem ser padronizadas de alguma maneira antes das distâncias serem calculadas, de modo que todas as p variáveis sejam igualmente importantes na determinação dessas distâncias. Isto pode ser feito codificando as variáveis de modo que as médias sejam todas zero e as variâncias sejam todas um (MANLY, 2008).

2.1.5.5. Grupo de árvores de decisão

O grupo de árvores de decisão engloba métodos usualmente simples, de fácil interpretação e podem ser aplicados tanto para tarefas preditivas, quanto para descritivas. As árvores de decisão são estruturas muito utilizadas em MD e podem ser chamadas de árvores de classificação (quando a variável dependente é categórica) ou árvores de regressão (quando a variável dependente é contínua) (SPINDEL, 2008).

As árvores de decisão captam relações não lineares do sistema estudado, em situações em que o uso de abordagens lineares, tais como a regressão, fica restrito (JAMES et al., 2013). Para tanto, utiliza-se o método denominado de particionamento recursivo, no qual ocorre a divisão do conjunto de dados em subgrupos (ou regiões) que são tratados posteriormente por modelos específicos (SPINDEL, 2008). A representação gráfica é efetuada por nós e ramos, como uma árvore, porém em sentido inverso (Figura 13) (WITTEN et al., 2011).

O ciclo de construção da árvore de decisão começa com um único nó de decisão, situado mais acima na árvore e chamado de nó raiz (ou nó inicial). Nos nós internos (ou intermediários), são realizados testes sobre os valores das variáveis independentes, gerando ramos, de maneira que cada ramo representa uma saída do teste e particiona os sub-grupos (nós intermediários) em mais sub-grupos. Essa ramificação continua a ser executada até o momento em que a divisão de nós gere um ganho considerado suficiente. Assim, de forma resumida, o dado entrante é submetido a sucessivas avaliações e, por fim, um ou mais modelos localizados nos nós terminais (finais), ou folhas, são combinados para gerar um resultado final (SPINDEL, 2008; TACONELI, 2008).

Figura 13. Representação esquemática de uma árvore de decisão. Adaptado de Taconeli (2008).

Em árvores de regressão e classificação, deve-se buscar por uma árvore de tamanho (número de nós finais) reduzido, com baixa heterogeneidade em seus nós finais, a fim de se evitar uma alta complexidade da árvore e garantir sua capacidade preditiva (TACONELI, 2008). Um dos meios para reduzir o número excessivo de nós é um procedimento denominado de poda, que deve ser aplicado após a construção da árvore (JAMES et al., 2013). A poda consiste em desfazer, uma a uma, as partições que menos contribuem para a explicação da variável resposta (TACONELI, 2008). Apesar de apresentarem grandes vantagens, como a facilidade de explicação e representação gráfica de simples interpretação, as árvores de regressão e classificação não são competitivas com outras técnicas, como o aprendizado de máquina, por exemplo. Isso acontece porque o poder preditivo das árvores de decisão pode ser melhorado consideravelmente utilizando-se técnicas como *Random Forest* e *Boosting*, nas quais são geradas grandes quantidades de árvores múltiplas que, posteriormente, são re combinadas para se obter uma predição de consenso (JAMES et al., 2013).

2.1.5.6. Random Forest

A técnica *Random Forest* se baseia na aplicação de *bagging* em árvores de decisão, com uma importante extensão - além de amostrar os registros, o algoritmo também amostra as árvores. Nas árvores de decisão tradicionais, para determinar como criar uma sub-repartição de uma repartição A, o algoritmo escolhe uma variável e um ponto de divisão por meio da minimização de um critério a ser escolhido. Entretanto, no caso das florestas aleatórias (ou *Random Forest*), em cada estágio do algoritmo, a escolha de uma variável é limitada a um subconjunto aleatório de variáveis. Dessa forma, comparado com o algoritmo de árvore básico, o algoritmo de *Random Forest* adiciona mais dois passos: o *bagging* e a amostragem *bootstrap* das variáveis em cada divisão (BRUCE; BRUCE, 2019).

O primeiro passo é o *bagging* (ou *bootstrap* agregador), utilizado quando buscamos uma resposta Y e temos P variáveis preditoras, ou seja, Y é função de X1, X2, ..., Xp com n registros. *Bagging* é como o algoritmo básico de agrupamentos, exceto pelo fato de que, em vez ajustar diversos modelos aos mesmos dados, cada modelo é ajustado a uma reamostra *bootstrap*, seguindo os seguintes passos:

1. Inicializar M (o número de modelos a serem ajustados) e n (o número de registros a escolher), com a condição $n < M$. Ajustar a iteração $m=1$;
2. Tirar uma reamostra *bootstrap* (ou seja, com reposição) de n registros dos dados de treinamento para formar uma subamostra Y_m a X_m (a *bag*);
3. Treinar um modelo usando Y_m e X_m para criar um conjunto de regras de decisão $\hat{f}_m(X)$;
4. Incrementar o contador do modelo $m = m + 1$. Se $m \leq M$, voltar ao Passo 1.

No caso em que $\hat{f}_m$ prevê a probabilidade $Y = 1$, a estimativa *bagged* é dada por:

$$\hat{f}_m = \frac{1}{M} \left(\hat{f}_1(X) + \hat{f}_2(X) + \dots + \hat{f}_m(X) \right) \quad (2)$$

A segunda etapa é a amostragem *bootstrap* (com reposição) dos registros;

1. Tirar uma subamostra *bootstrap* (com reposição) dos registros;
2. Para a primeira divisão, amostrar $p < P$ variáveis aleatoriamente sem reposição;
3. Para cada uma das variáveis amostradas $X_{j(1)}, X_{j(2)}, \dots, X_{j(p)}$, aplicar o algoritmo da divisão:
 - a) Para cada valor $s_{j(k)}$ de $X_{j(k)}$:
 - Dividir os registros na repartição A com $X_{j(k)} < s_{j(k)}$ como uma repartição, e os registros restantes em que $X_{j(k)} \geq s_{j(k)}$ como outra repartição
 - Meça a homogeneidade de classes dentro de cada sub-repartição de A

- b) Selecionar o valor de $s_{j(k)}$ que produzam máxima homogeneidade de classes dentro da repartição.
4. Selecionar a variável $X_{j(k)}$ e o valor da divisão $s_{j(k)}$ que produzam máxima homogeneidade de classes dentro da repartição;
5. Prossiga à próxima divisão e repita os passos anteriores, começando com o passo 2;
6. Continuar com divisões adicionais seguindo o mesmo procedimento até que a árvore tenha crescido;
7. Voltar ao passo 1, retirar outra amostra *bootstrap*, e recomeçar todo o processo.

Para se escolher o número de variáveis que se deve amostrar em cada passo utiliza-se a regra $\sqrt{P}$, em que P é o número de variáveis preditoras.

A potência do algoritmo de *Random Forest* se mostra quando se constrói modelos preditivos para dados com muitas características e registros, uma vez que ele tem a habilidade de determinar automaticamente quais variáveis preditoras são importantes e descobrir relacionamentos complexos entre elas, correspondentes a termos de interação. Entretanto, da mesma forma que muitos algoritmos estatísticos de aprendizado de máquina (*Machine Learning*), *Random Forest* pode ser considerado um algoritmo de “caixa-preta”, ou seja, com manivelas para ajustar o funcionamento dessa caixa. Essas “manivelas” são chamadas de hiperparâmetros, os quais dizem respeito aos parâmetros que devem ser ajustados antes de se desenvolver um modelo. Esses hiperparâmetros não são otimizados como parte do processo de treinamento e, diferentemente dos modelos estatísticos tradicionais, os hiperparâmetros para *Random Forest* não exigem escolhas (por exemplo, a escolha dos preditores a serem usados em um modelo de regressão) e são mais críticos, especialmente para evitar sobreajuste (BRUCE; BRUCE, 2019). Os dois hiperparâmetros mais importantes para *Random Forest* são:

- **Nodesize** - tamanho mínimo para os nós terminais (folhas na árvore). O padrão é 1 para árvores de classificação e 5 para árvores de regressão.
- **Maxnodes** - número máximo de nós em cada árvore de decisão. Não existe um limite padrão e árvores maiores deverão ser ajustadas conforme as restrições de *nodesize*.

2.1.5.7. Boosting

O *Boosting* é uma técnica geral usada para criar um agrupamento de modelos e foi desenvolvido na mesma época que o *Bagging*. Assim como o *Bagging*, o *Boosting* também é usado em árvores de decisão, entretanto, apesar de suas similaridades, o *Boosting* assume uma abordagem com muito mais detalhes. Como resultado, enquanto o *Bagging* pode ser feito relativamente com poucos ajustes, o *Boosting* exige muito mais cuidados em sua aplicação (BRUCE; BRUCE, 2019).

Em modelos de regressão linear, os resíduos costumam ser examinados para ver se o ajuste pode ser melhorado. O *Boosting* leva esse conceito muito além e ajusta uma série de modelos com cada modelo sucessivo ajustado para minimizar o erro dos anteriores. Algumas variantes do algoritmo costumam ser usadas: *Adaboost*, *Boosting* gradiente e *Boosting* gradiente estocástico. O último, *Boosting* gradiente estocástico, é o mais geral e amplamente usado, uma vez que incorpora a reamostragem de registros e colunas em cada rodada. A ideia básica por trás dos diversos algoritmos de *Boosting* é essencialmente a mesma. Porém, o mais fácil de entender é o *Adaboost*, descrito a seguir (BRUCE; BRUCE, 2019):

1. Inicializar M , o número máximo de modelos a serem ajustados, e ajustar o contador de iterações $m = 1$. Inicializar os pesos de observação por meio da Eq.3:

$$w_i = \frac{1}{N} \quad (3)$$

para $i = 1, 2, \dots, N$. Inicializar o modelo de agrupamento $\hat{F}_0 = 0$;

2. Treinar um modelo usando $\hat{f}_m$ utilizando os pesos de observações $w_1, w_2, \dots, w_N$, que minimize o erro ponderado e_m definido pela soma dos pesos para as observações classificadas erroneamente;
3. Adicionar o modelo ao grupo $\hat{F}_m$, dado pela equação a seguir:

$$\hat{F}_m = \hat{F}_{m-1} + \alpha_m \times \hat{f}_m \quad (4)$$

em que α_m é dado por:

$$\alpha_m = \frac{\log(1 - e_m)}{e_m} \quad (5)$$

4. Atualizar os pesos $w_1, w_2, \dots, w_N$ de forma que eles sejam aumentados para as observações que eram mal classificadas. O tamanho do aumento depende de α_m , com valores maiores de α_m levando a maiores pesos;
5. Incrementar o contador de modelo $m = m + 1$. Se $m \leq M$, voltar ao passo 1.

Em geral, a estimativa *Boosted* é dada pela seguinte equação:

$$\hat{F} = \alpha_1 \times \hat{f}_1 + \alpha_2 \times \hat{f}_2 + \dots + \alpha_M \times \hat{f}_M \quad (6)$$

Ao aumentar os pesos para as observações mal classificadas, o algoritmo força os modelos a treinarem mais pesadamente nos dados que tiveram um mal desempenho. Dessa forma, o fator α_m garante que os modelos com menor erro tenham maior peso. No âmbito do *Random Forest*, o *Boosting*

gradiente estocástico inclui a aleatoriedade ao algoritmo visto acima por meio da amostragem de observações e variáveis preditoras em cada estágio (BRUCE; BRUCE, 2019).

O software de domínio público mais usado para *Boosting* é o *XGBoost*, uma implementação do *Boosting* gradiente estocástico. Trata-se de uma implementação computacional eficiente com muitas opções e está disponível como um pacote para as principais linguagens de software de ciência de dados. Na linguagem R, por exemplo, o *XGBoost* está disponível como o pacote *XGboost* (BRUCE; BRUCE, 2019).

A função *XGboost* tem muitos parâmetros que devem ser ajustados, sendo os dois mais importantes o *subsample* (controla a fração de observações que devem ser amostradas em iteração) e o *eta* (fator de encolhimento aplicado em α_m no algoritmo de *Boosting*). Usar o *subsample* faz o *Boosting* agir como *Random Forest*, exceto pelo fato de a amostragem ser feita sem reposição. Já o parâmetro de encolhimento *eta* é útil para prevenir o sobreajuste por meio da redução da mudança de pesos, já que uma mudança menor nos pesos significa que o algoritmo é menos propenso a sobreajustar o conjunto de treinamento (BRUCE; BRUCE, 2019).

2.1.5.8. Support Vector Machine (SVM)

A abordagem SVM representa um procedimento de aprendizado não paramétrico para classificação e regressão de alto desempenho (LIMA, 2004). Esse grupo engloba funções ou classificadores que implementam modificações espaciais nos dados, levando-os a um plano onde a classificação se torna mais fácil (DI GIROLAMO NETO, 2013).

Para entender melhor o conceito de SVM, deve-se supor que seja necessário, em um plano bidimensional, escolher um classificador que separe os círculos dos triângulos na ilustração da Figura 14. Pode-se observar que na Figura 14a tem-se um classificador com um erro considerável, em função de sua baixa capacidade em distinguir pontos próximos pertencentes às diferentes classes. Entretanto, sua taxa de acerto é bem inferior à função da Figura 14c, que classifica corretamente a maioria dos exemplos do conjunto de treinamento, mas que pode ser muito específica a um determinado conjunto de dados, podendo causar uma situação de *overfitting*, que ocorre quando o modelo se ajusta muito bem ao conjunto em que foi treinado, porém, quando confrontado com novos dados, fica muito suscetível a erros. Por outro lado, a função do classificador na Figura 14a é muito mais simples do que o da Figura 14c, enquanto um classificador mais complexo do que o da Figura 14a e não tão sobre-ajustado quanto o da Figura 14c seria o classificador da Figura 14b, o qual tem uma complexidade intermediária e, por isso, classifica bem grande parte dos exemplos (LORENA; CARVALHO, 2003).

Figura 14. Exemplo de conjunto de dados de treinamento pertencentes a duas classes (círculos e triângulos) e classificados segundo três diferentes hipóteses pela técnica SVM: (a) com erro considerável; (b) com complexidade intermediária; e (c) com elevada complexidade. Adaptado de Lorena e Carvalho (2003).

As SVM foram desenvolvidas com base na Teoria do Aprendizado Estatísticos, que estabelece condições matemáticas que auxiliam na escolha de um classificador específico, a partir de um conjunto de dados de treinamento que tenha bom desempenho, tanto para os conjuntos de treinamento, quanto para os de teste ou validação. Originalmente, foram desenvolvidas para a solução de problemas que envolviam a classificação de dados em duas categorias, tanto para conjuntos linearmente separáveis quanto para os casos não lineares, mas foram expandidas também para aplicações que envolviam mais do que duas classes, por meio de técnicas específicas para gerar classificadores multiclasse a partir de preditores binários (LORENA; CARVALHO, 2003).

Para encontrar um hiperplano separador que possa ser considerado ótimo, entre os diferentes hiperplanos capazes de classificar um conjunto de dados, é importante que ele tenha não somente a habilidade de discriminação dos dados, mas também sua margem maximizada (MUCHERINO et al., 2009). A margem de um classificador é definida como a menor distância entre as observações de um conjunto de treinamento e o hiperplano utilizado na categorização desses dados (LORENA; CARVALHO, 2003). Na figura 15, a margem aparece como a distância entre as linhas contínuas e o hiperplano separados (linha pontilhada). Quanto maior a margem do classificador, maior será sua capacidade de generalização, ou seja, de classificar corretamente as observações do conjunto de teste (MUCHERINO et al., 2009).

Figura 15. Exemplo de observações separadas por hiperplano com suas respectivas margens para classificação por meio da técnica SVM. Adaptado de Sayad (2010).

Em casos de classificação linear, a separação dos dados se dá por meio de hiperplanos definidos com base no conjunto de dados de treinamento (MUCHERINO et al., 2009). Exemplificando-se o problema, com foco na utilização de uma SVM para a classificação de duas classes $y_i = +1, -1$, se o conjunto de treinamento é linearmente separável, então existe um par $(\vec{w}, b)$, tal que:

$$(\vec{w} \cdot \vec{x}_i + b) \geq 1, \forall \vec{x}_i, \text{ para } y_i = 1.$$

$$(\vec{w} \cdot \vec{x}_i + b) \leq -1, \forall \vec{x}_i, \text{ para } y_i = -1.$$

As duas inequações acima podem ser combinadas para se obter a seguinte função objetivo:

$$y_i(\vec{w} \cdot \vec{x}_i + b) \geq 1, \forall \vec{x}_i.$$

A aprendizagem das SVMs consiste, então, em encontrar $\vec{w}$ e b que satisfaçam o seguinte problema de otimização (HASTIE et al., 2009; SAYAD, 2010):

$$\min \frac{1}{2} \|\vec{w}\|^2.$$

Como a solução que minimiza a função objetivo consiste em um problema convexo, então existe um valor mínimo global (HASTIE et al., 2009; SAYAD, 2010). A figura 16 apresenta o problema de otimização de uma SVM para classificação linear.

Figura 16. Exemplo de otimização de uma SVM para classificação linear. Adaptado de Sayad (2010).

Assim, a análise de SVM ideal é aquela capaz de produzir um hiperplano que separe perfeitamente os vetores (observações) em duas classes não sobrepostas. No entanto, uma separação perfeita não é sempre possível e existe a possibilidade de obtenção de modelos que não classifiquem corretamente grandes quantidades de observações. Nesses casos, as SVMs procuram encontrar um hiperplano que maximize as margens ao mesmo tempo que minimize os erros de classificação (HAMMER, 2016).

Quando o conjunto de dados não é linearmente separável, a utilização de uma SVM para a classificação de duas classes $y_i = +1, -1$, é possível por meio da adição de uma variável de penalização ϵ_i . A figura 17 apresenta um exemplo de aplicação de uma SVM em que os dados não são linearmente separáveis (SAYAD, 2010).

Figura 17. Exemplo de um problema de otimização de uma SVM para o caso em que o conjunto de dados não é linearmente separável. Adaptado de Sayad (2010).

No caso em que os dados não são linearmente separáveis, a variável ε_i será responsável por penalizar as observações que venham a ficar fora das margens. Neste caso, a função objetivo pode ser definida como:

$$y_i(\vec{w} \cdot \vec{x}_i + b) \geq 1 - \varepsilon_i, \forall \vec{x}_i \text{ e } \varepsilon_i \geq 0$$

sendo que neste caso a aprendizagem das SVMs consiste em solucionar o seguinte problema de otimização (HASTIE et al., 2009; SAYAD, 2010):

$$\min \frac{1}{2} \|\vec{w}\|^2 + c \sum_i \varepsilon_i$$

em que c é um *tuning parameter* e quanto maior o seu valor, maiores são as chances de as observações caírem do lado “errado” das margens.

Para tarefas de classificação não linear, pode-se usar o artifício de elevação dos dados de treinamento para um espaço hiperdimensional, em que eles se tornam linearmente separáveis, por meio de um hiperplano de margem maximizada. Esse processo de generalização de SVMs para tarefas não lineares é possível graças à utilização de funções *Kernel*, que permitem o acesso a espaços complexos de forma simplificada (HAMMER, 2016).

A elevação para um espaço hiperdimensional ocorre por meio de um produto interno chamado de *Kernel*, que pode ser do tipo linear, não linear (Polinomiais), RBF (*Radial-Basis Function*) e sigmóides (LORENA; CARVALHO, 2003). Os resultados da aplicação dessa técnica são comparáveis e, muitas vezes, superiores aos obtidos por outros algoritmos de aprendizado, como as redes neurais artificiais, por exemplo (HAYKIN, 1999; BRAGA et al., 2007). De forma geral, o desempenho da abordagem SVM depende da seleção dos parâmetros adequados dentro da função *Kernel* do classificador. Embora essa

escolha de parâmetros seja importante para garantir a acurácia do modelo, ainda não há um método universal para a determinação dos mesmos (BONESSO, 2013).

Apesar de as SVMs terem sido desenvolvidas, à princípio, para a solução de problemas envolvendo a classificação binária, outros tipos de SVMs foram posteriormente formulados, como SVM para Regressão ou “Clusterização”. No entanto, tais modelos de SVM ainda mantiveram todas as principais características que caracterizam o algoritmo, sendo a maximização das margens a principal delas (HAMMER, 2016).

3. MATERIAL E MÉTODOS

3.1. Dados experimentais

3.1.1. Locais

Os dados experimentais utilizados neste estudo foram provenientes de ensaios de competição de cultivares de algodoeiro conduzidos pela empresa Tropical Melhoramento & Genética (TMG). Os experimentos foram conduzidos em cinco localidades do estado do Mato Grosso: Campo Novo do Parecis; Campo Verde; Pedra Preta; Sapezal e Sorriso, entre os anos de 2011 a 2019, totalizando oito safras (Tabela 6 e Figura 18).

Tabela 6. Localidades do estado do Mato Grosso em que foram realizados os ensaios de competição de cultivares do algodoeiro, com suas respectivas coordenadas geográficas e altitude.

Localidade	Latitude (graus e décimos)	Longitude (graus e décimos)	Altitude (m)
Campo Novo do Parecis	-13,7	-57,9	664
Campo Verde	-15,5	-50,2	625
Pedra Preta	-16,6	-54,5	740
Sapezal	-13,6	-58,8	570
Sorriso	-12,6	-55,9	414

Figura 18. Localização geográfica dos ensaios de competição de cultivares de algodoeiro conduzidos pela empresa TMG no estado do Mato Grosso.

3.1.2. Descrição edafoclimática das áreas de estudo

De acordo com o Sistema Brasileiro de Classificação de Solos – SIBCS (2018), os solos da área de estudo foram classificados como Latossolo Vermelho Amarelo. Já os climas dessas áreas são classificados, segundo Köppen, como Aw para as localidades de Campo Verde e Pedra Preta e como uma transição entre Aw e Am para as demais (ALVARES et al., 2013). As condições meteorológicas médias de cada uma das localidades, onde os ensaios foram conduzidos, e safras são apresentados na Tabela 7.

Tabela 7. Condições meteorológicas no período dos ensaios conduzidos em cinco localidades do Mato Grosso, nas safras 2011-2012 a 2018-2019.

Localidade	Safrá							
	2011-12	2012-13	2013-14	2014-15	2015-16	2016-17	2017-18	2018-19
Temperatura média (°C)								
Campo Novo do Parecis	24,80	24,46	24,54	25,02	24,80	24,97	24,68	25,23
Campo Verde	24,80	24,02	24,17	24,42	24,70	24,63	25,74	26,17
Pedra Preta	24,38	24,16	24,23	24,51	24,85	24,52	25,67	26,26
Sapezal	24,52	24,63	24,49	25,00	25,08	24,68	24,49	25,27
Sorriso	25,86	25,88	26,05	26,06	26,60	26,15	26,33	26,62
Radiação Solar (MJ m ⁻² dia ⁻¹)								
Campo Novo do Parecis	17,89	16,76	15,62	16,31	15,89	15,66	17,56	17,58
Campo Verde	18,77	16,86	14,57	16,10	14,39	14,90	18,52	19,57
Pedra Preta	18,01	16,59	16,63	16,49	15,97	15,27	18,21	18,57
Sapezal	17,49	16,14	15,24	15,78	15,41	15,30	15,41	17,63
Sorriso	18,00	18,16	17,48	17,63	17,17	16,77	18,02	18,55
Chuva (mm)								
Campo Novo do Parecis	1493,10	1101,80	1417,23	1193,83	1219,23	1340,18	1441,29	1343,02
Campo Verde	1172,57	1187,03	1243,21	1594,23	1189,61	814,41	1083,65	1026,26
Pedra Preta	1313,62	1403,41	590,80	1099,85	862,40	876,91	1056,18	982,05
Sapezal	1558,25	1061,21	1611,40	1086,46	1009,15	115,55	1476,56	1606,73
Sorriso	1466,86	1624,00	1552,02	1548,23	998,42	1378,15	1546,86	1444,20

3.1.3. Cultivares

Foram utilizados dados de caracteres tecnológicos de qualidade da fibra de algodão de 32 cultivares brasileiras (Tabela 8), referentes à oito safras: 2011-12; 2012-13; 2013-14; 2014-15; 2015-16; 2016-17; 2017-18; e 2018-19. Além disso, para todos os locais e safras, foram consideradas três datas de semeadura e quatro repetições. O número total de dados avaliados foi 1085 (considerando os dados de todas as cultivares). A relação do número de dados por cultivar é apresentada na Tabela 9.

Em todos os locais e anos, as datas de semeadura não foram idênticas para o dia do mês. No entanto, a primeira semeadura ocorreu aproximadamente em dezembro, a segunda em janeiro e a terceira entre o final de janeiro e fevereiro, conforme a Tabela 10.

Tabela 8. Cultivares brasileiras de algodão empregadas no presente estudo.

Cultivares de algodão			
DP1240B2RF	FM951LL	FMT705	TMG44B2RF
DP1243B2RF	FM954GLT	FMT707	TMG45B2RF
DP1536B2RF	FM975WS	FMT709	TMG46B2RF
DP1648B2RF	FM980GLT	IMA8405GLT	TMG47B2RF
DP1746B2RF	FM983GLT	TMG11WS	TMG61RF
FM910	FM993	TMG41WS	TMG62RF
FM940GLT	FMT523	TMG42WS	TMG81WS
FM944GL	FMT701	TMG43WS	TMG82WS

Tabela 9. Número de dados por cultivar de algodão (n) e total utilizado no presente estudo.

Cultivar	n	Cultivar	n	Cultivar	n	Cultivar	n
DP1240B2RF	15	FM951LL	24	FMT705	39	TMG44B2RF	62
DP1243B2RF	35	FM954GLT	20	FMT707	24	TMG45B2RF	47
DP1536B2RF	35	FM975WS	83	FMT709	39	TMG46B2RF	15
DP1648B2RF	35	FM980GLT	15	IMA8405GLT	32	TMG47B2RF	47
DP1746B2RF	28	FM983GLT	20	TMG11WS	21	TMG61RF	26
FM910	11	FM993	11	TMG41WS	15	TMG62RF	29
FM940GLT	30	FMT523	13	TMG42WS	65	TMG81WS	74
FM944GL	47	FMT701	39	TMG43WS	59	TMG82WS	30
SOMA (n)	236	225		294		330	
TOTAL (n)				1085			

Tabela 10. Período em que as épocas de semeadura (1, 2 e 3) dos ensaios de competição de cultivares foram realizados nas cinco localidades do estado do Mato Grosso.

Semeadura	Período
1	3 a 28 de dezembro
2	29 de dezembro a 22 de janeiro
3	23 de janeiro a 13 de fevereiro

3.1.4. Variáveis analisadas

Os dados experimentais utilizados neste trabalho referem-se aos índices de qualidade da fibra de algodão medidos pelo equipamento *High Volume Instrument* (HVI). Os caracteres tecnológicos utilizados foram:

- 1) Micronaire (adimensional);
- 2) Comprimento (mm);
- 3) Resistência (gf tex⁻¹);
- 4) Uniformidade (%);

- 5) Elongação (%); e
- 6) Índice de fibras curtas (SFI) (%).

3.2. Fases fenológicas avaliadas

Câmara e Chiavegato (2001) subdividiram o ciclo do algodoeiro em duas grandes fases: vegetativa, que se inicia com a emergência das plântulas e termina com a formação do primeiro ramo frutífero; e reprodutiva, que tem início com o aparecimento do primeiro botão floral e termina quando as fibras nos capulhos atingem o ponto de maturidade fisiológica para a colheita. Pela tabela 1 é possível observar que as fibras iniciam seu desenvolvimento quando a primeira maçã está visível (estádio R4), ou seja, após a frutificação. Dessa maneira, acredita-se que as características intrínsecas das fibras de algodão sofram maiores influências das variáveis meteorológicas a partir dessa fase, seguindo até o momento em que o algodão completa o seu ciclo e está pronto para a colheita.

Para o presente estudo, conforme os dados fornecidos pela empresa TMG, considerou-se o ciclo médio do algodoeiro como sendo de 170 dias. Dessa forma, a partir da Tabela 2, pode-se observar que o momento em que a primeira maçã está visível acontece aproximadamente nos últimos 100 dias do ciclo. Assim, levando-se em conta os aspectos citados, foram considerados para as análises de correlação entre as variáveis meteorológicas e os caracteres de qualidade das fibras de algodão os seguintes períodos:

- Ciclo total (170 dias);
- Últimos 100 dias do ciclo; e
- Últimos 50 dias do ciclo.

3.3. Dados meteorológicos

Os dados meteorológicos empregados neste estudo foram: precipitação pluvial (P), temperatura máxima do ar ($T_{\max}$), temperatura mínima do ar ($T_{\min}$), radiação solar global ou insolação (Q_g), umidade relativa do ar (UR) e velocidade do vento a 2 m de altura (U2), todos na escala diária para os locais em que foram realizados os experimentos. Os dados meteorológicos foram obtidos junto às estações meteorológicas automáticas (Modelo HOBO U-30-NRC) instaladas nas áreas experimentais da empresa TMG. Além dessas variáveis meteorológicas, foram também determinadas outras, conforme apresentado a seguir:

- Temperatura média do ar (T_{med} , °C);
- Amplitude térmica (AT, °C);
- Número de dias com $T_{\max} \geq 30^\circ\text{C}$ ($T_{\max 30}$);

- Número de dias com $T_{\max} \geq 32^{\circ}\text{C}$ ($T_{\max}32$);
- Número de dias com $T_{\min} \leq 18^{\circ}\text{C}$ ($T_{\min}18$);
- Número de dias com $T_{\min} \leq 20^{\circ}\text{C}$ ($T_{\min}20$);
- Número de dias com $AT \geq 10^{\circ}\text{C}$ ($AT10$);
- Número de dias com $AT \geq 12^{\circ}\text{C}$ ($AT12$);
- Número de dias com $AT \geq 13^{\circ}\text{C}$ ($AT13$); e
- Número de dias com chuva (ou número de dias com $P > 0$ mm) (NDC).

Esses dados foram utilizados para compor a matriz de correlação e para determinar a relação das variáveis meteorológicas com os índices de qualidade de fibra dos algodões, avaliados em todos os períodos do ciclo do algodoeiro mencionados anteriormente. Neste caso, para a geração dos modelos agrometeorológicos, as variáveis meteorológicas foram tratadas como as variáveis independentes (preditoras) e os índices de qualidade da fibra como as variáveis dependentes. Os patamares de temperatura foram definidos com base na literatura, considerando-se a faixa ótima para cultivo do algodoeiro entre 20 e 30°C (CHIAVEGATO et al., 2009).

3.4. Balanço hídrico da cultura

As variáveis meteorológicas foram também empregadas para a elaboração do balanço hídrico sequencial da cultura, considerando-se todas as datas de semeadura, por meio do método proposto por Thornthwaite e Mather (1955) e considerando as condições da cultura do algodão, a fim de se determinar as seguintes variáveis:

- Déficit hídrico (DEF) total do ciclo ($\sum DEF, mm$)
- Número de dias com $DEF > 0$ ($DEF0$)
- Excedente hídrico (EXC) total do ciclo ($\sum EXC, mm$)
- Número de dias com $EXC > 0$ ($EXC0$)
- Número de dias com $\frac{ARM}{CAD} \geq 0,5$ (ARM/CAD)

As etapas do balanço hídrico climatológico sequencial proposto por Thornthwaite e Mather (1955) são apresentados a seguir:

1º) Precipitação pluvial (P) (mm) menos a evapotranspiração da cultura (ET_C) (mm), ou seja: (P- ET_C). A determinação da ET_C é feita conforme a Eq.7:

$$ET_C = ET_0 \times k_C \quad (7)$$

em que ET_0 é a evapotranspiração de referência e k_C é o coeficiente da cultura. A estimação da ET_0 foi feita com base no método de Penman-Monteith, conforme recomendado por Allen et al. (1998). Já os valores de k_C foram estimados para cada dia do ciclo do algodoeiro com base nos índices de área foliar diários, conforme feito por Bezerra et al. (2012). Os valores diários de k_C e de IAF são apresentados na Figura 19.

Figura 19. Valores diários do coeficiente da cultura (k_C) e do índice foliar (IAF) da cultura do algodoeiro, para o ciclo de 170 dias. *DAS: dias após a semeadura.

2º) Cálculo do negativo acumulado (N_{AC}) e do armazenamento (ARM) simultaneamente. Para isso, considera-se duas situações:

- Se $(P - ET_C) < 0$, calcula-se o negativo acumulado primeiro e em seguida o armazenamento de água no solo, por meio das seguintes equações:

$$N_{AC(i)} = (P - ET_C) + N_{AC(i-1)} \quad (8)$$

em que $N_{AC(i)}$ é o negativo acumulado do dia atual; $(P - ET_C)$ é o termo já determinado no 1º passo; e $N_{AC(i-1)}$ é o negativo acumulado do dia anterior.

$$ARM_i = CAD \times e^{-|N_{AC}|/CAD} \quad (9)$$

em que ARM_i é o armazenamento do dia atual; CAD é a capacidade de água disponível; e N_{AC} é o negativo acumulado do dia atual.

- Se $(P - ET_C) > 0$, calcula-se o armazenamento de água no solo primeiro e em seguida o negativo acumulado, através das Eq.10 e Eq.11, respectivamente:

$$ARM_i = ARM_{i-1} + (P - ET_C) \quad (10)$$

em que ARM_i é o armazenamento do dia atual; ARM_{i-1} é o armazenamento dia anterior; e $(P - ET_C)$ é o termo já determinado no 1º passo.

$$N_{ACi} = CAD \times \ln \left(\frac{ARM}{CAD} \right) \quad (11)$$

em que N_{ACi} é o negativo acumulado do dia atual; CAD é a capacidade de água disponível; e ARM é o armazenamento de água no solo.

3º) Alteração do armazenamento (ALT), determinado pela Eq. 12:

$$ALT = ARM_i - ARM_{i-1} \quad (12)$$

em que: ARM_i é o armazenamento do dia atual; e ARM_{i-1} é o armazenamento dia anterior.

4º) Cálculo da evapotranspiração real (ET_r), considerando as seguintes condições:

- Se $(P - ET_C) > 0$, $ET_r = ET_C$; ou
- Se $(P - ET_C) < 0$, ET_r é determinada pela Eq.13:

$$ET_r = P + |ALT| \quad (13)$$

5º) Cálculo da deficiência hídrica (DEF) é feito por meio da Eq.14:

$$DEF = ET_C - ET_r \quad (14)$$

6º) Cálculo do excedente hídrico (EXC), considerando as seguintes condições:

- Se $ARM < CAD$, o excedente é igual a zero; ou
- Se $ARM = CAD$, o excedente é determinado pela Eq. 15:

$$EXC = (P - ET_C) - ALT \quad (15)$$

Assim, o balanço hídrico da cultura foi determinado para cada uma das datas de semeadura avaliadas e o extrato do balanço foi determinado para cada uma das fases do ciclo do algodoeiro avaliadas: ciclo total (170 dias); últimos 100 dias do ciclo; e últimos 50 dias do ciclo. Para a determinação do balanço

hídrico da cultura do algodoeiro foi considerada uma capacidade de água disponível (CAD) representativa dos solos em que foram realizados os experimentos. De posse dos valores de ARM e CAD determinou-se o armazenamento relativo de água no solo (ARM/CAD).

Conforme descrito pela Portaria N°131, de 13 de agosto de 2019, relacionada ao Zoneamento Agrícola de Risco Climático para a cultura do algodão herbáceo no Estado do Mato Grosso (ano safra 2019/2020), a capacidade de água disponível (CAD) foi estimada em função da profundidade efetiva das raízes e da reserva útil dos solos, para três tipos de solos. Foram considerados os solos Tipo 1 (textura arenosa), Tipo 2 (textura média) e Tipo 3 (textura argilosa), respectivamente com CADs de 42, 66 e 90 mm (MARQUES, 2019). Diante dessa informação, neste estudo optou-se por adotar a CAD de 90 mm, uma vez que os solos das áreas experimentais da TMG foram classificados como Latossolos Vermelho Amarelo, com textura argilosa.

3.5. Relação das variáveis meteorológicas e índices de qualidade da fibra do algodão em função da localidade e datas de semeadura

Foram analisadas as variáveis meteorológicas e dos índices de qualidade da fibra do algodão em função das localidades e das datas de semeadura avaliadas. Essa análise teve o intuito de se observar a existência de algum tipo de relação e/ou padrão de associação entre as variáveis mencionadas.

Após essa análise, os índices de qualidade foram separados em determinado padrão (em relação às localidades e épocas de semeadura) e as variáveis meteorológicas que se associaram a esses índices. Assim, para os índices de qualidade de fibra e as variáveis meteorológicas que apresentaram um mesmo padrão, foram feitas análises de correlação e foram determinados coeficientes de correlação de Pearson (r).

Sejam X_i e Y_i os valores de duas variáveis X e Y . $\bar{X}$ e $\bar{Y}$ são respectivamente as médias dos valores X_i e Y_i . Assim, a fórmula para determinação do coeficiente de correlação de Pearson (r) é dado por:

$$r = \frac{\sum (x - \bar{x}) \times (y - \bar{y})}{\sqrt{\sum (x - \bar{x})^2 \times \sum (y - \bar{y})^2}} \quad (16)$$

3.6 Análise de correlação e Análise de Agrupamento

Foi elaborada uma matriz de correlação de Pearson entre os caracteres de qualidade da fibra de algodão e todas as variáveis meteorológicas e aquelas geradas a partir dessas, considerando os dados de todas as cultivares e épocas de semeadura (abordagem geral), para cada uma das três fases do ciclo do algodoeiro avaliadas (ciclo total, últimos 100 dias e últimos 50 dias).

As análises de agrupamento objetivaram usar os valores de variáveis para planejar um esquema para agrupar os objetos em classes, de modo que aqueles similares estejam na mesma classe. Como neste

estudo existe uma grande número de objetos (neste caso, as cultivares) e também de variáveis (os índices de qualidade da fibra), realizou-se uma análise de agrupamento de modo a tentar reunir, em um mesmo grupo, os objetos (cultivares) mais próximos no que se refere aos parâmetros de qualidade avaliados. Assim, foi feita uma análise de agrupamento considerando dados de todos os índices de qualidade de todas as cultivares avaliadas e após definir os grupos e quais cultivares estavam inclusas em cada um deles, foi feita uma matriz de correlação por grupo (abordagem por grupo) entre os índices de qualidade avaliados e as variáveis meteorológicas correspondentes aos dados das cultivares de cada grupo. Na abordagem por grupo, também foram consideradas as três fases do ciclo do algodoeiro supracitadas.

Como relatado anteriormente, existem vários métodos de análise de agrupamento, no entanto, o mais utilizando é o método de Ward, uma vez que o mesmo, diferentemente dos outros, fornece um dendrograma demonstrativo dos agrupamentos formados. Por esse motivo, optou-se por utilizar o método de Ward.

O método de Ward, embora forneça o dendrograma, apresenta a desvantagem de que o número de grupos deve ser indicado pelo usuário, o que é muitas vezes inviável e demanda muito tempo, já que a escolha do número de grupos teria que ser feita por meio de tentativas, até que o melhor resultado fosse encontrado. Como forma de resolver esse problema, utilizou-se, em conjunto com o método de Ward, o algoritmo de K-médias (*K-means*) para a escolha do número de grupos.

As K-médias dividem os dados em K grupos através da minimização da soma das distâncias quadráticas de cada registro à média de seu grupo atribuído. Isso é chamado de soma dos quadrados dentro do grupo ou SS dentro do grupo. As K-médias não garantem que os grupos tenham o mesmo tamanho, mas encontra grupos que sejam melhores quando separados (BRUCE; BRUCE, 2019).

Para exemplificar, segundo Bruce e Bruce (2019), considere um conjunto de dados de n registros e apenas duas variáveis, por exemplo, x e y. Suponha que queira dividir os dados em K=4 grupos. Isso significa atribuir cada registro das variáveis (x_i, y_i) a um grupo K. Assim, dada uma atribuição de n_k registros a um grupo K, o centro do grupo $(\bar{x}_k, \bar{y}_k)$ é a média dos pontos no grupo:

$$\bar{x}_k = \frac{1}{n_{k_i}} \in \sum_{Cluster\ k} x_i \quad (17)$$

$$\bar{y}_k = \frac{1}{n_{k_i}} \in \sum_{Cluster\ k} y_i \quad (18)$$

No agrupamento de registro com múltiplas variáveis, o termo média do grupo não se refere a um único número, mas ao vetor das médias das variáveis. Dessa forma, a soma dos quadrados dentro de um

grupo é dada por:

$$SS_k = i \in \sum_{Cluster k} (x_i + \bar{x}_k)^2 + (y_i + \bar{y}_k)^2 \quad (19)$$

As K-médias encontram, assim, a atribuição dos registros que minimizam a soma dos quadrados dentro do agrupamento ao longo de todos os quatro grupos $SS_1 + SS_2 + SS_3 + SS_4$, ou seja:

$$\sum_{k=1}^4 SS_i$$

No programa R, o agrupamento por K-médias pode ser realizado usando a função *kmeans* e pode ser aplicado em um conjunto de dados com p variáveis $X_1, X_2, \dots, X_p$. O algoritmo começa com um K especificado pelo usuário e um conjunto inicial de médias de grupo e itera os seguintes passos:

- Atribui cada registro à média de grupo mais próxima conforme a medida da distância quadrada.
- Calcula a nova média do grupo com base na atribuição de registros.

O algoritmo converge quando a atribuição de registros a grupos não muda. Para a primeira iteração, é necessário especificar um conjunto inicial de médias de grupo, o que geralmente é feito pela atribuição aleatória de cada registro a um dos K grupos e, então, encontrando a média de tais grupos. Quando se usa mais de um conjunto de iterações, os resultados das K-médias é dado pela iteração que tenha a menor soma dos quadrados dentro do grupo (BRUCE; BRUCE, 2019).

Neste estudo, o método K-médias foi utilizado apenas para a escolha do número de grupos a ser utilizada no método de Ward, uma vez que esse procedimento permite a visualização do número de grupos que vai gerar a menor soma dos quadrados dentro do grupo, o que pode contribuir para um melhor resultado da análise de agrupamento. Para isso, o número de iterações foi definido por meio da criação de um vetor para que o K variasse de 1 a 10 grupos, as variáveis utilizadas foram os índices de qualidade da fibra de algodão (micronaire, comprimento, resistência, uniformidade, alongação e índice de fibras curtas) e os registros foram os dados referentes às 32 cultivares de algodão. Assim, o algoritmo gerou um gráfico com a relação da soma dos quadrados dentro dos grupos e o número de grupos.

Tanto a matriz de correlação quanto a análise de agrupamentos foram feitas a partir de algoritmos executados no software R (<http://www.r-project.org/>). Para a matriz de correlação, empregou-se o pacote *corrplot*, para o K-médias a função *kmeans* para o método Ward, foi gerado um script seguindo as seguintes etapas:

- a) Leitura dos dados;
- b) Padronização dos dados, de modo que p variáveis sejam igualmente importantes na

determinação das distâncias entre os objetos. Isto foi feito codificando as variáveis de modo que as médias fossem todas zero e as variâncias fossem iguais a um;

- c) Aplicação da função *k-means* como forma de se obter o número de grupos;
- d) Geração do gráfico que mostra a relação da soma dos quadrados dentro dos grupos e o número de grupos. A partir disso, foi definido o número de agrupamentos a ser utilizado;
- e) Análise de agrupamento pelo método Ward (agrupamento não-hierárquico), utilizando-se a distância euclidiana para determinar as distâncias entre os objetos e
- f) Geração do dendrograma.

3.7. Modelagem dos índices de qualidade da fibra de algodão por meio da mineração de dados

A mineração de dados (MD) é a exploração e análise de dados por meios automáticos ou semi-automáticos, em grandes quantidades de dados, com o objetivo de descobrir regras ou padrões interessantes (BERRY; LINOFF, 1997). A MD é parte de um processo maior de pesquisa denominado “Busca de conhecimento em banco de dados” (*Knowledge Discovery in Database - KDD*), o qual possui uma metodologia própria para preparação e exploração dos dados, interpretação de seus resultados e assimilação dos conhecimentos minerados. No entanto, a MD se tornou mais conhecida que o próprio processo de KDD em função de ser a etapa onde são aplicadas as técnicas de busca de conhecimentos (CÔRTEZ et al., 2002).

Os procedimentos realizados para modelagem dos índices de qualidade da fibra de algodão por meio da mineração de dados foram realizados considerando-se as etapas que constituem o processo CRISP-DM (*Cross Industry Standard Process for Data Mining*). Segundo Camilo e Silva (2009), essa metodologia encontra ampla aceitação em projetos de mineração de dados e é descrita a seguir.

- a) Fase de compreensão do contexto (*business understanding*): compreensão e entendimento dos objetivos e necessidades do projeto que deverão ser atendidos por meio da mineração de dados. É o planejamento inicial do projeto;
- b) Fase de entendimento dos dados (*data understanding*): uma vez que os dados podem vir de diversas fontes e estarem disponíveis em diferentes formatos, essa fase possibilita a familiarização com os dados e suas estruturas, bem como o entendimento de suas especificidades, com o objetivo de identificar possíveis problemas de qualidade dos dados;
- c) Fase de preparação dos dados (*data preparing*): esta fase compreende as atividades relacionadas à construção do conjunto final de dados, que será utilizado para análise por meio de técnicas de mineração de dados, como por exemplo, alterações de estruturas para adequação aos softwares de mineração de dados, transformações de variáveis, inclusão de novas variáveis, mudanças de escala, eliminação de inconsistências e de dados faltantes;

- d) Fase de modelagem (*modeling*): aplicação de diferentes algoritmos de mineração de dados. Tipicamente, para cada tipo de problema há diferentes soluções possíveis, sendo que algumas técnicas têm exigências específicas com relação ao formato dos dados;
- e) Fase de avaliação (*evaluation*): uma vez desenvolvidos os modelos, consiste em revisar o processo e verificar se os objetivos foram alcançados, de forma que, ao final desta etapa, sejam tomadas decisões sobre o uso dos resultados;
- f) Fase de implantação (*deployment*): um projeto de mineração de dados não termina com a geração de modelos, mas deve também incorporar a aplicação e distribuição do conhecimento gerado durante o processo. Portanto, nessa fase os resultados devem ser organizados para sua apresentação ao usuário final, de forma que este possa fazer uso prático do conhecimento adquirido.

Além disso, todos os modelos foram gerados a partir da execução dos algoritmos no software R (<http://www.r-project.org/>). Três diferentes técnicas de mineração de dados foram empregadas neste estudo:

- *Random Forest* (RF);
- *Boosting -XGBoosting* (XGB); e
- *Support Vector Machine* (SVM).

Como interface para a aplicação dessas técnicas, foi utilizado o pacote *Caret* (*Classification and Regression Training*). O *Caret* é um conjunto de funções desenvolvidas para simplificar o processo de criação de modelos preditivos, e contém ferramentas para diversas funcionalidades, como a seleção automática de parâmetros para as diferentes técnicas de mineração de dados, como por exemplo, o número de árvores ideal no *Random Forest*, bem como a caracterização das variáveis preditoras mais importantes dos modelos. Além disso, o pacote proporciona uma visualização mais amigável dos resultados (KUHN, 2008).

Em uma primeira etapa, realizou-se uma abordagem geral, em que os algoritmos acima foram aplicados na base integral dos dados, que contém os dados de todas as regiões, safras e datas de semeadura analisadas. Em uma segunda etapa, as mesmas técnicas foram utilizadas para análises mais específicas (cultivar, localidade e safras), com o intuito de se capturar possíveis particularidades que poderiam estar ocultas pela análise geral dos dados.

A definição dos conceitos da funcionalidade da mineração de dados é fundamental para se obter os resultados esperados em um trabalho, bem como para entender o processo como um todo. A classificação da funcionalidade em mineração de dados é dividida em Análise Descritiva e Análise de Prognóstico. A análise descritiva é composta pelas etapas de análise prévia e descobrimento dos dados e a análise de prognóstico pelas etapas de estimação, predição e classificação. Uma das vantagens de se focar nas funcionalidades da mineração de dados dessa forma, diz respeito às facilidades que podem ser obtidas quando surge uma nova necessidade de análise de dados. Neste caso, por exemplo, basta identificar a que resultado se deseja chegar e imediatamente partir para a identificação de que técnica aplicar (HAN et al.,

2011; WEIS; INDURKHYA, 1999).

Dentre as etapas de análise prévia da análise descritiva está inclusa a análise de *outliers* (ou detecção de desvios). Esta funcionalidade objetiva encontrar conjuntos de dados que não obedecem ao comportamento ou modelo dos dados. Uma vez encontrados, podem ser tratados ou descartados para a utilização na mineração de dados. Assim, trata-se de uma importante avaliação nos dados, no sentido de descobrir probabilidades crescentes de desvios ou riscos associados aos vários objetivos traçados inicialmente na mineração de dados. A detecção desses desvios é análoga às técnicas utilizadas em análises estatísticas convencionais, onde são aplicados testes de significância que assumem uma distribuição, utilizando medidas estatísticas do tipo média aritmética e desvio padrão para aferir essas diferenças (CÔRTEZ et al., 2002).

Salvador et al. (2009) desenvolveram um método de verificação de dados climatológicos apoiado em modelos minerados (VEDALOGIC) para o Instituto de Controle do Espaço Aéreo Brasileiro (ICEA). O VEDALOGIC consiste em uma verificação de dados, utilizando-se modelos criados com algoritmos de MD, e o método utiliza modelos de *clustering*, gerados a partir de uma série histórica, que propiciam a identificação de grupos homogêneos em uma Base de Dados Climatológicos (BDC). Segundo os autores, a partir desses modelos, torna-se possível a detecção de inconformidades nos dados, denominados pontos estranhos (*outliers*). Após a detecção de um *outlier*, este é classificado/predito, de acordo com o modelo de árvore de decisão, gerado também a partir de uma série histórica. Assim, o valor encontrado com base na árvore de decisão é adotado como sugestão para a correção do *outlier*, de modo a contribuir com a consistências dos dados no BDC.

Barbosa et al. (2018) propuseram uma técnica de detecção de *outliers* via Distância de Mahalanobis. Para a geração dos dados foram utilizados simulações via método de Monte Carlo e a técnica de mistura de distribuições normais multivariadas. Os resultados apresentados nas simulações mostraram que o método proposto foi superior ao método de Mahalanobis (medida de distância baseada nas correlações entre variáveis com os quais distintos padrões podem ser identificados e analisados) tanto para a sensibilidade quanto para especificidade, ou seja, ele apresentou maior capacidade de diagnosticar corretamente os *outliers* e os não *outliers*. Além disso, a metodologia foi aplicada em dados reais provenientes da área de saúde.

Dessa forma, é evidente que, embora seja um procedimento simples, a detecção de *outliers* é uma etapa importante não só para a MD, já que vem sendo extensivamente utilizada em diversas aplicações. A identificação de *outliers* desempenha um papel importante na análise estatística, pois tais observações podem conter informações importantes em relação a dados contendo valores atípicos e os resultados podem ser enganosos e decisões equivocadas podem ser tomadas. Além disso, em situações práticas, os próprios *outliers* são muitas vezes os pontos especiais de interesse e sua identificação pode ser o principal objetivo da investigação (BARBOSA et al., 2018).

Diante desse contexto, como uma das etapas da MD, foram elaborados gráficos do tipo boxplot para detecção de *outliers* e para avaliar as variabilidades dos índices de qualidade da fibra de algodão em

função das cultivares estudadas em cinco localidades do estado do Mato Grosso, durante oito safras (2011/12 a 2018/19) e três épocas de semeadura por safra. Também foram feitos gráficos que mostram a variabilidade dos índices de qualidade da fibra de algodão ao longo dos anos considerados.

Com base nisso, os dados de índices de qualidade da fibra de algodão de diferentes cultivares brasileiras, bem como os dados meteorológicos, foram organizados para que fossem submetidos ao processo de treinamento dos algoritmos de mineração de dados (*Random Forest* – RF; *XGBoosting* – XGB; e *Support Vector Machine* – SVM).

3.8. Avaliação dos resultados

Para se elaborar a modelagem por meio de MD, é necessário primeiramente separar a base de dados em duas etapas, sendo uma utilizada para o treinamento dos algoritmos, em que são determinadas automaticamente as características necessárias para se representar a informação desejada, e outra para o teste dos modelos gerados (REFAEILZADEH et al., 2009). A primeira etapa seleciona aleatoriamente 70% dos dados e é denominada de “Treinamento”. A segunda etapa considera os 30% restantes dos dados para fazer o teste dos modelos e é denominada de “Teste”. Neste estudo, foram utilizados um total de 1085 dados para cada índice de qualidade de fibra do algodão avaliado, sendo que para a etapa de “Treinamento” foram utilizados 760 dados e para a etapa de “Teste” 325 dados.

A análise do desempenho dos modelos foi efetuada pela comparação entre os índices de qualidade da fibra estimados pelos modelos e os índices de qualidade observados por meio da análise de regressão. O desempenho dos modelos foi avaliado considerando-se os seguintes índices estatísticos: coeficiente de correlação (r); coeficiente de determinação (r^2); coeficiente de concordância de Willmott (d) (WILLMOTT, 1981), que representa a distância relativa dos valores estimados em relação à reta 1:1; raiz quadrada do erro médio (*Root Mean Square Error*, RMSE), que indica a acurácia do modelo; erro absoluto médio (EAM), que indica a magnitude dos erros associados à estimação dos índices de qualidade da fibra; e coeficiente de confiança (C) de Camargo e Sentelhas (1997), que representa o grau de concordância do modelo, englobando em um único índice a precisão (r) e a acurácia (d).

4. RESULTADOS E DISCUSSÃO

4.1. Variação dos índices de qualidade da fibra do algodão em função do local de produção e da data de semeadura

A fim de se obter uma visão geral da variação dos dados dos índices de qualidade da fibra do algodoeiro obtidos nas áreas experimentais, em função das diferentes localidades e épocas de semeadura, foi feita uma análise descritiva, apresentadas na Tabela 11 e nas Figuras 20 a 25.

Tabela 11. Índices de qualidade da fibra do algodoeiro (média e desvio padrão) para cada uma das localidades avaliadas e nas três diferentes épocas de semeadura.

Índice de qualidade	Localidade				
	Campo Novo do Parecis	Campo Verde	Pedra Preta	Sapezal	Sorriso
Semeadura 1 (3 a 28/12)					
Micronaire	4,34 ± 0,36	4,28 ± 0,30	4,20 ± 0,25	4,23 ± 0,42	4,47 ± 0,35
Comprimento	29,52 ± 0,94	29,57 ± 0,97	29,41 ± 0,82	29,62 ± 0,89	29,61 ± 0,98
Resistência	30,38 ± 1,79	29,85 ± 1,52	30,34 ± 1,65	31,07 ± 1,33	30,63 ± 1,31
Uniformidade	83,74 ± 1,38	83,30 ± 1,81	83,73 ± 1,37	83,58 ± 1,27	83,96 ± 1,24
Elongação	7,05 ± 0,83	6,84 ± 0,78	6,95 ± 0,93	7,06 ± 0,65	6,57 ± 0,73
Índice de fibras curtas (SFI)	7,12 ± 1,03	7,95 ± 1,56	7,53 ± 1,06	7,53 ± 0,93	7,22 ± 1,14
Semeadura 2 (29/12 a 22/01)					
Micronaire	4,17 ± 0,40	4,18 ± 0,34	4,07 ± 0,37	4,18 ± 0,44	4,31 ± 0,33
Comprimento	29,65 ± 0,91	29,57 ± 0,87	29,48 ± 0,95	29,55 ± 0,79	29,63 ± 0,90
Resistência	31,09 ± 2,32	29,77 ± 2,10	30,59 ± 1,77	30,51 ± 1,69	30,97 ± 2,13
Uniformidade	83,71 ± 1,12	83,36 ± 1,41	83,31 ± 1,30	83,13 ± 1,43	83,69 ± 1,24
Elongação	6,85 ± 0,88	7,05 ± 0,83	7,26 ± 0,91	7,00 ± 0,66	6,71 ± 0,84
Índice de fibras curtas (SFI)	7,43 ± 0,83	7,71 ± 1,10	7,93 ± 1,27	7,81 ± 1,05	7,17 ± 1,00
Semeadura 3 (23/01 a 13/02)					
Micronaire	3,99 ± 0,33	4,11 ± 0,31	3,96 ± 0,39	4,12 ± 0,46	4,07 ± 0,31
Comprimento	29,62 ± 0,86	29,46 ± 0,95	29,41 ± 1,01	29,57 ± 0,85	29,55 ± 0,92
Resistência	30,23 ± 1,28	30,56 ± 2,11	30,60 ± 1,98	30,39 ± 1,41	31,21 ± 2,42
Uniformidade	83,76 ± 1,07	83,11 ± 1,68	83,35 ± 1,12	83,16 ± 1,43	84,33 ± 1,08
Elongação	7,02 ± 1,20	6,92 ± 0,90	7,40 ± 0,92	7,19 ± 0,75	6,79 ± 0,81
Índice de fibras curtas (SFI)	7,20 ± 1,00	7,76 ± 1,28	7,85 ± 1,11	7,75 ± 0,87	7,02 ± 0,79

Figura 20. Variação do micronaire da fibra de algodão em função da localidade produtora no estado do Mato Grosso e da época de semeadura. Semeadura 1 corresponde ao período entre 3 e 28/12, Semeadura 2 entre 29/12 e 22/01, e Semeadura 3 entre 23/01 e 13/02.

Figura 21. Variação do comprimento da fibra de algodão em função da localidade produtora no estado do Mato Grosso e da época de semeadura. Semeadura 1 corresponde ao período entre 3 e 28/12, Semeadura 2 entre 29/12 e 22/01, e Semeadura 3 entre 23/01 e 13/02.

Figura 22. Variação da resistência da fibra de algodão em função da localidade produtora no estado do Mato Grosso e da época de semeadura. Semeadura 1 corresponde ao período entre 3 e 28/12, Semeadura 2 entre 29/12 e 22/01, e Semeadura 3 entre 23/01 e 13/02.

Figura 23. Variação da uniformidade da fibra de algodão em função da localidade produtora no estado do Mato Grosso e da época de semeadura. Semeadura 1 corresponde ao período entre 3 e 28/12, Semeadura 2 entre 29/12 e 22/01, e Semeadura 3 entre 23/01 e 13/02.

Figura 24. Variação da elongação da fibra de algodão em função da localidade produtora no estado do Mato Grosso e da época de semeadura. Semeadura 1 corresponde ao período entre 3 e 28/12, Semeadura 2 entre 29/12 e 22/01, e Semeadura 3 entre 23/01 e 13/02.

Figura 25. Variação do índice de fibras curtas (SFI) de algodão em função da localidade produtora no estado do Mato Grosso e da época de semeadura. Semeadura 1 corresponde ao período entre 3 e 28/12, Semeadura 2 entre 29/12 e 22/01, e Semeadura 3 entre 23/01 e 13/02.

Por meio das Figuras 20 a 25, é possível se observar que dentre os índices de qualidade da fibra de algodão avaliados o único que mostrou padrão bem definido de variação com a época de semeadura, para todas as localidades, foi o índice micronaire, havendo um decréscimo de seus valores médios com o atraso da data de semeadura (Figura 20). Para os demais índices de qualidade, a variação com a época de semeadura foi muito variada (Figuras 21 a 25).

Em relação aos outros índices de qualidade da fibra de algodão, nenhum deles apresentou padrão definido de variação de seus valores médios em relação às diferentes datas de semeadura e localidades. Dessa forma, a fim de se investigar se esse decréscimo do índice micronaire em função do atraso da data de semeadura teve relação com as condições meteorológicas ou de balanço hídrico, também foram avaliadas, para cada época de semeadura, 21 variáveis, sendo que oito delas (T_{min} ; T_{med} ; UR; P; NDC; $\sum EXC$; $EXC0$; e ARM/CAD) apresentaram relação positiva com índice micronaire (Figuras 26 a 33), enquanto que outras oito (T_{min18} ; T_{min20} ; AT; AT10; AT12; AT13; $\sum DEF$; e $DEF0$) apresentaram relação inversa com esse índice de qualidade da fibra do algodão (Figuras 34 a 41).

Por se tratar de um grande número de variáveis meteorológicas e do balanço hídrico que afetam os valores médios do índice micronaire em função da localidade e da data de semeadura, foram determinados os coeficientes de correlação de Pearson entre essas variáveis (Tabela 12).

Figura 26. Variação da média da temperatura mínima do ar durante o ciclo do algodão em diferentes localidades do estado do Mato Grosso em função da época de semeadura. Semeadura 1 corresponde ao período entre 3 e 28/12, Semeadura 2 entre 29/12 e 22/01, e Semeadura 3 entre 23/01 e 13/02.

Figura 27. Variação da temperatura média do ar durante o ciclo do algodão em diferentes localidades do estado do Mato Grosso em função da época de semeadura. Semeadura 1 corresponde ao período entre 3 e 28/12, Semeadura 2 entre 29/12 e 22/01, e Semeadura 3 entre 23/01 e 13/02.

Figura 28. Variação da média da umidade relativa do ar durante o ciclo do algodão em diferentes localidades do estado do Mato Grosso em função da época de semeadura. Semeadura 1 corresponde ao período entre 3 e 28/12, Semeadura 2 entre 29/12 e 22/01, e Semeadura 3 entre 23/01 e 13/02.

Figura 29. Variação da chuva total durante o ciclo do algodão em diferentes localidades do estado do Mato Grosso em função da época de semeadura. Semeadura 1 corresponde ao período entre 3 e 28/12, Semeadura 2 entre 29/12 e 22/01, e Semeadura 3 entre 23/01 e 13/02.

Figura 30. Variação do número de dias com chuva durante o ciclo do algodão em diferentes localidades do estado do Mato Grosso em função da época de semeadura. Semeadura 1 corresponde ao período entre 3 e 28/12, Semeadura 2 entre 29/12 e 22/01, e Semeadura 3 entre 23/01 e 13/02.

Figura 31. Variação do total de excedente hídrico (CAD = 90 mm) durante o ciclo do algodão em diferentes localidades do estado do Mato Grosso em função da época de semeadura. Semeadura 1 corresponde ao período entre 3 e 28/12, Semeadura 2 entre 29/12 e 22/01, e Semeadura 3 entre 23/01 e 13/02.

Figura 32. Variação do número de dias com excedente hídrico (CAD = 90 mm) maior do que zero durante o ciclo do algodão em diferentes localidades do estado do Mato Grosso em função da época de semeadura. Semeadura 1 corresponde ao período entre 3 e 28/12, Semeadura 2 entre 29/12 e 22/01, e Semeadura 3 entre 23/01 e 13/02.

Figura 33. Variação do número de dias com ARM/CAD maior ou igual a 0,5 (CAD = 90 mm) durante o ciclo do algodão em diferentes localidades do estado do Mato Grosso em função da época de semeadura. Semeadura 1 corresponde ao período entre 3 e 28/12, Semeadura 2 entre 29/12 e 22/01, e Semeadura 3 entre 23/01 e 13/02.

Figura 34. Variação do número de dias com Tmin menor ou igual a 18°C durante o ciclo do algodão em diferentes localidades do estado do Mato Grosso em função da época de semeadura. Semeadura 1 corresponde ao período entre 3 e 28/12, Semeadura 2 entre 29/12 e 22/01, e Semeadura 3 entre 23/01 e 13/02.

Figura 35. Variação do número de dias com T_{min} menor ou igual a 20°C durante o ciclo do algodão em diferentes localidades do estado do Mato Grosso em função da época de semeadura. Semeadura 1 corresponde ao período entre 3 e 28/12, Semeadura 2 entre 29/12 e 22/01, e Semeadura 3 entre 23/01 e 13/02.

Figura 36. Variação da amplitude térmica durante o ciclo do algodão em diferentes localidades do estado do Mato Grosso em função da época de semeadura. Semeadura 1 corresponde ao período entre 3 e 28/12, Semeadura 2 entre 29/12 e 22/01, e Semeadura 3 entre 23/01 e 13/02.

Figura 37. Variação do número de dias com amplitude térmica maior ou igual a 10°C durante o ciclo do algodão em diferentes localidades do estado do Mato Grosso em função da época de semeadura. Semeadura 1 corresponde ao período entre 3 e 28/12, Semeadura 2 entre 29/12 e 22/01, e Semeadura 3 entre 23/01 e 13/02.

Figura 38. Variação do número de dias com amplitude térmica maior ou igual a 12°C durante o ciclo do algodão em diferentes localidades do estado do Mato Grosso em função da época de semeadura. Semeadura 1 corresponde ao período entre 3 e 28/12, Semeadura 2 entre 29/12 e 22/01, e Semeadura 3 entre 23/01 e 13/02.

Figura 39. Variação do número de dias com amplitude térmica maior ou igual a 13°C durante o ciclo do algodão em diferentes localidades do estado do Mato Grosso em função da época de semeadura. Semeadura 1 corresponde ao período entre 3 e 28/12, Semeadura 2 entre 29/12 e 22/01, e Semeadura 3 entre 23/01 e 13/02.

Figura 40. Variação do déficit hídrico (CAD = 90 mm) durante o ciclo do algodão em diferentes localidades do estado do Mato Grosso em função da época de semeadura. Semeadura 1 corresponde ao período entre 3 e 28/12, Semeadura 2 entre 29/12 e 22/01, e Semeadura 3 entre 23/01 e 13/02.

Figura 41. Variação do número de dias com déficit hídrico maior do que zero (CAD = 90 mm) durante o ciclo do algodão em função da época de semeadura. Semeadura 1 corresponde ao período entre 3 e 28/12, Semeadura 2 entre 29/12 e 22/01, e Semeadura 3 entre 23/01 e 13/02.

Tabela 12. Coeficientes de correlação de Pearson para a relação entre os valores médios do micronaire da fibra do algodão e das variáveis meteorológicas e de balanço hídrico durante o ciclo da cultura. Para as variáveis do balanço hídrico foi considerada uma capacidade de água disponível (CAD) de 90 mm.

Variável meteorológica ou do balanço hídrico	Coeficiente de correlação de Pearson (r)
Temperatura mínima (Tmin) (°C)	0,78**
Umidade relativa (UR) (%)	0,47
Precipitação (P) (mm)	0,83**
Amplitude térmica (AT) (°C)	-0,37
Temperatura média (Tmédia) (°C)	0,62**
Déficit hídrico (Σ DEF) (mm)	-0,79**
Excedente hídrico (Σ EXC) (mm)	0,75**
Nº de dias com Tmin $\leq$ 18 °C (Tmin18)	-0,83**
Nº de dias com Tmin $\leq$ 20 °C (Tmin20)	-0,75**
Nº de dias com AT $\leq$ 10 °C (AT10)	-0,24
Nº de dias com AT $\leq$ 12 °C (AT12)	-0,21
Nº de dias com AT $\leq$ 13 °C (AT13)	-0,28
Nº de dias com chuva (NDC)	0,65**
Nº de dias com EXC > 0 (EXC0)	0,77**
Nº de dias com DEF > 0 (DEF0)	-0,78**
Nº de dias com ARM/CAD $\geq$ 0,5 (ARM/CAD)	0,84**

**representa correlação significativa a 1% de probabilidade pelo teste t.

De acordo com a Tabela 12 todas as variáveis meteorológicas e de balanço hídrico que apresentaram correlação positiva com o índice micronaire, com exceção apenas para a UR, todas as outras sete variáveis (Tmin; Tmed; P; NDC; Σ EXC; EXC0; e ARMCAD) apresentaram forte correlação ($P < 0,01$), com valores de r variando entre 0,62 e 0,84.

Já em relação às variáveis meteorológicas que apresentaram correlação negativa com o índice micronaire, as variáveis AT, AT10, AT12 e AT13 não foram significativas, enquanto que as variáveis Tmin18, Tmin20, Σ DEF e DEF0 apresentaram alta significância ($P < 0,01$), com valores de r variando entre -0,75 e -0,83 (Tabela 12).

Em função dessa análise, observa-se que o atraso da data de semeadura influencia as condições meteorológicas e de balanço hídrico ao longo do ciclo do algodoeiro, o que por sua vez afeta o índice micronaire. Além disso, pode-se inferir que para os demais índices de qualidade da fibra do algodão, devem haver outras variáveis que as afetam, já que não houve correlação entre o atraso da semeadura e a variação desses índices (Figuras 20 a 25). Essa abordagem foi uma maneira alternativa de identificar quais são as variáveis meteorológicas que mais interferem nos valores do micronaire.

4.2. Análise de correlação e de agrupamento

4.2.1. Análise de correlação (Abordagem geral)

O resultado da matriz de correlação de Pearson elaborada para cada caractere tecnológico de qualidade da fibra de algodão, considerando os dados de todas as cultivares (abordagem geral), para cada uma das três fases do ciclo do algodoeiro avaliadas (ciclo total, últimos 100 dias e últimos 50 dias) estão dispostos na Tabela 13.

Tabela 13. Coeficientes de correlação de Pearson (r) para a relação entre os caracteres de qualidade da fibra de algodão e as variáveis meteorológicas e do balanço hídrico (CAD = 90 mm), para três períodos do ciclo do algodoeiro: ciclo total; últimos 100 dias do ciclo; e últimos 50 dias do ciclo).

Variável *	Coeficiente de correlação de Pearson (r)		
	Ciclo total	Últimos 100 dias	Últimos 50 dias
Micronaire			
Tmin20	- 0,34**	- 0,27**	- 0,28**
Tmin	0,24**	0,18**	0,15**
Tmed	0,23**	0,21**	0,15**
Tmax32	0,22**	0,21**	0,13**
Comprimento			
UR	0,54**	0,47**	0,40**
Qg	- 0,29**	- 0,18**	- 0,29**
U2	0,25**	0,28**	0,25**
Tmed	- 0,25**	- 0,18**	- 0,16**
ΣDEF	- 0,24**	- 0,27**	- 0,42**
Tmax30	- 0,29**	- 0,25**	- 0,28**
AT13	- 0,22**	- 0,24**	- 0,29**
ARM/CAD	0,21**	0,22**	0,29**
Resistência			
Qg	- 0,25**	- 0,19**	- 0,15**
NDC	- 0,23**	- 0,20**	- 0,24**
EXC0	- 0,26**	- 0,13**	- 0,15**
DEF0	0,27**	0,16**	0,19**
ARM/CAD	- 0,20**	- 0,19**	- 0,15**
Uniformidade			
P	0,21**	0,17**	0,26**
U2	0,29**	0,28**	0,22**
ΣEXC	0,22**	0,15**	0,20**
Elongação			
Qg	-0,40**	- 0,41**	- 0,41**
Tmax	- 0,26**	- 0,29**	-0,29**
Tmin	- 0,22**	- 0,14**	-0,13**
P	-0,27**	- 0,12**	-0,16**
Tmed	- 0,28**	- 0,27**	-0,27**
Tmax30	- 0,29**	- 0,29**	-0,27**
NDC	- 0,37**	- 0,29**	-0,28**
EXC0	- 0,40**	- 0,17**	-0,19**
DEF0	0,35**	0,16**	0,15**
Índice de fibras curtas (SFI)			
UR	-0,22**	-0,20**	-0,17**
P	-0,23**	-0,19**	-0,15**
ΣEXC	-0,23**	-0,18**	-0,13**

* os significados das variáveis meteorológicas e do balanço hídrico se encontram nos itens 3.3 e 3.4.

**representa correlação significativa a 1% de probabilidade pelo teste t.

4.2.1.1. Micronaire

Na Tabela 13, pode-se observar que o micronaire, para os três períodos do ciclo do algodoeiro avaliados (ciclo total, últimos 100 dias e últimos 50 dias), apresentou correlação negativa com a variável meteorológica T_{min20} e essa correlação foi maior quando se considerou o ciclo total, em comparação com os outros períodos do ciclo do algodoeiro. Além disso, pode-se observar que esse resultado foi coerente com aqueles apresentados pela Tabela 12, na qual os valores dos dados médios de micronaire (ciclo total) apresentaram forte correlação negativa ($P < 0,01$) com os valores médios de T_{min20} .

Soares (2015) determinou a correlação entre a frequência de temperaturas mínimas inferiores a 20°C e o micronaire para três fases de desenvolvimento do algodoeiro (emergência à primeira flor; primeira flor ao primeiro capulho, mais 21 dias; e primeiro capulho mais 21 dias) e obteve expressiva correlação negativa para duas das cinco cultivares avaliadas. Dessa forma, embora neste estudo as correlações entre micronaire e T_{min20} não tenham sido tão expressivas, por conta do tamanho do banco de dados avaliados (1085 dados de 32 cultivares diferentes), elas se mantiveram negativas ($P < 0,01$) para os três períodos do ciclo considerados.

A correlação entre as variáveis meteorológicas T_{min} , T_{med} e T_{max32} e o micronaire foram positivas e, da mesma forma que para a T_{min20} , mais expressivas quando se considerou o ciclo total do algodoeiro. Soares (2015) também obteve correlações positivas significativas entre T_{min} e micronaire para as cinco cultivares consideradas em seu estudo, mas somente em duas das fases de desenvolvimento avaliadas (primeira flor ao primeiro capulho, mais 21 dias; e primeiro capulho mais 21 dias).

A influência da temperatura mínima do ar no micronaire foi mais expressiva do que a causada pela temperatura máxima. Isso foi mais evidente quando se considerou a temperatura mínima abaixo de 20°C . Isso corrobora resultados da literatura que também verificaram que a elevação da temperatura mínima promove aumento do índice micronaire (GIPSON; JOHAM, 1968; LIAKATAS et al., 1998; BANGE et al., 2010). Uma explicação para isso, segundo Hake et al. (1990), é que a elevação de temperatura mínima pode melhorar a respiração dos fotoassimilados e promover melhor deposição de celulose para formar a fibra, uma vez que noites mais quentes aumentam a respiração e a partir de certo valor diminuem a fotossíntese líquida.

4.2.1.2. Comprimento

O comprimento da fibra do algodão apresentou correlação positiva com a UR, ARM/CAD e U2 ($P < 0,01$) e correlação negativa com Q_g , T_{med} , ΣDEF , T_{max30} e AT13 ($P < 0,01$) (Tabela 13). Dentre essas, a variável que apresentou melhor correlação com o comprimento foi a UR, sendo mais representativa quando se considerou o ciclo total do algodoeiro. Depois da UR, as variáveis de maior efeito no comprimento das fibras do algodoeiro foram ΣDEF (nos últimos 50 dias do ciclo do

algodoeiro) e Q_g ($P < 0,01$). No caso da Q_g , para o ciclo total e nos últimos 50 dias do ciclo, as correlações com o comprimento foram idênticas, enquanto para os últimos 100 dias do ciclo essa foi menor. Em relação às outras variáveis meteorológicas e do balanço hídrico, embora tenham apresentado correlações significativas ($P < 0,01$) com o comprimento, os resultados para cada fase de desenvolvimento foram muito similares.

De acordo com os resultados apresentado acima, os efeitos das condições ambientais no comprimento da fibra do algodoeiro não são consistentes e variam. Estudos tem mostrado que o comprimento da fibra não foi afetado pelo sombreamento (ZHAO et al 2012; ECHER; ROSOLEM, 2015). Todavia, aumentos de comprimento foram observados na China, em ambientes de maiores temperaturas (CHEN et al., 2014; LV et al., 2013), o que está associado à maior duração do período de fechamento dos plasmodesmos das células da fibra, gerando maior turgor osmótico interno para a alongação. Por outro lado, sob temperaturas mais baixas, a redução da atividade das enzimas e do fluxo de carboidratos levou à diminuição da taxa de alongação, reduzindo o comprimento final (CHEN et al., 2014). Embora alguns trabalhos tenham evidenciado correlação entre aumento de temperatura mínima e comprimento de fibra, neste estudo, as variações de temperatura não influenciaram de maneira consistente a resposta ao comprimento da fibra (Tabela 13).

Soares (2015) concluiu que o comprimento da fibra não oscilou em função das variações de chuva, o que não aconteceu neste estudo, uma vez que o $\sum DEF$, dependente da chuva, foi a segunda variável mais correlacionada com o comprimento, especialmente nos últimos 50 dias do ciclo (Tabela 13). Esse resultado concorda com aquele obtidos por Lokhande e Reddy (2014), que também observaram redução linear do comprimento da fibra com o aumento do déficit hídrico. Uma das explicações para a elevada correlação entre o $\sum DEF$ nos últimos 50 dias do ciclo do algodoeiro e o comprimento da fibra é que as necessidades hídricas do algodoeiro variam de acordo com as fases fenológicas, sendo a de maior demanda hídrica a fase reprodutiva, a qual engloba o florescimento e formação do fruto (CHIAVEGATO et al., 2009). Dessa forma, um estresse hídrico durante essa fase, que corresponde aproximadamente aos últimos 50 dias do ciclo do algodoeiro, poderia afetar o desenvolvimento da cultura e, conseqüentemente, o comprimento da fibra de algodão.

4.2.1.3. Resistência

A resistência apresentou correlação positiva ($P < 0,01$) com DEF_0 e correlação significativa negativa ($P < 0,01$) com as variáveis Q_g , NDC , EXC_0 e ARM/CAD (Tabela 13).

Diferentemente do trabalho de Soares (2015), neste estudo, a resistência da fibra não foi tão influenciada por elevação da temperatura, mas sim por variações de chuva. Lokhande e Reddy (2014) encontram resultados similares, em que houve influência da chuva na resistência da fibra de algodão. Esses autores observaram uma redução linear da resistência da fibra em resposta ao déficit hídrico, sendo esse caractere de qualidade da fibra de algodão o que foi mais responsivo ao estresse hídrico.

Em relação ao excedente hídrico, Wang et al. (2017), em estudo em ambiente controlado, observaram que o encharcamento do solo nas fases de florescimento e de formação de capulhos reduziu a resistência da fibra do algodão. A resistência da fibra foi inversamente proporcional à quantidade de capulhos retidos na planta, o que se justifica pela alteração da relação fonte - dreno (ESCHER et al., 2018) e pode explicar a correlação negativa encontrada entre EXC0 e a resistência neste estudo.

4.2.1.4. Uniformidade

A uniformidade apresentou correlação positiva ($P < 0,01$) com as variáveis P; U2; e EXC (Tabela 13). A uniformidade, como dito anteriormente, é a relação entre o comprimento médio de 100% das fibras (*Mean Length* – ML) e o comprimento médio dos 50% das fibras mais longas (*Upper Half Mean Length* – UHML), expresso em porcentagem (BACHELIER; GOURLOT, 2018). Dessa forma, por se tratar de uma característica de homogeneidade do comprimento das fibras do fardo e por ser uma medida indireta (dependente do comprimento), há poucos estudos sobre o efeito das variáveis meteorológicas e de balanço hídrico na uniformidade. Embora Lokhande e Reddy (2014) tenham observado uma redução linear da uniformidade em resposta ao aumento do déficit hídrico, o que em parte explica a correlação encontrada com P, não há nada no estudo dos referidos autores que mencione a influência do excedente hídrico e da velocidade do vento na uniformidade das fibras do algodoeiro.

4.2.1.5. Elongação

A elongação (ou alongamento) à rotura da fibra é quanto um feixe de fibras cede no sentido longitudinal até o momento da rotura, em relação ao comprimento inicial do corpo de prova. Da mesma forma que a uniformidade, é uma característica intrínseca da fibra associada a outra característica, neste caso, a resistência. Assim, a melhor maneira de se analisar os resultados das correlações entre a elongação e as variáveis meteorológicas e do balanço hídrico é comparando-os com os resultados encontrados para a resistência. Neste estudo, a elongação apresentou correlação negativa ($P < 0,01$) com oito variáveis, sendo elas: Qg; Tmax; Tmin; P; Tmed; Tmax30; NDC; e EXC0. Por outro lado, apresentou correlação positiva ($P < 0,01$) apenas com a variável DEF0. Para as variáveis Qg, Tmax, Tmed, e Tmax30, os valores dos coeficientes de correlação foram aproximadamente iguais em todos os períodos da cultura analisados. Por outro lado, para a Tmin, P, NDC, EXC0 e DEF0, os coeficientes de correlações diferiram entre os períodos da cultura do algodoeiro, sendo o período total sempre o que apresentou as maiores correlações (Tabela 13).

Em relação às fases de desenvolvimento do algodoeiro avaliadas, houve diferenças consideráveis entre as correlações com as variáveis P, NDC, EXC0 e DEF0, sendo que para todas elas os coeficientes de correlação foram maiores quando se considerou o ciclo total do algodoeiro. Para as demais variáveis, os resultados para todas os períodos considerados foram muito próximos (Tabela 13).

Comparando-se a resistência com a elongação, das oito variáveis meteorológicas com as quais a elongação apresentou correlações significativas, quatro delas foram comuns com as correlações encontradas para a resistência. Tanto a elongação, quanto a resistência, apresentaram correlações negativas ($P < 0,01$) com Qg, NDC e EXC0 e correlação positiva ($P < 0,01$) com DEF0. Assim, pode-se afirmar que existe relação entre a elongação e a resistência e que ambas são mais sensíveis às variações de radiação solar e do estresse hídrico.

4.2.1.6. Índice de fibras curtas (SFI)

O índice de fibras curtas (SFI) é o percentual de fibras menores que 12,7 mm presentes nos corpos de prova. Além disso, o SFI age em conjunto com o índice de uniformidade, no sentido de que um alto conteúdo de fibras curtas aliado a um baixo índice de uniformidade, podem provocar a flutuação de fibrilas e de impurezas nas zonas de estiragem de diversas máquinas no processo de fiação, gerando acúmulo de pó e micropó (FONSECA; FARIAS, 2002). Dessa forma, é possível que haja correlações em comum e contrárias entre esses dois índices de qualidade da fibra de algodão e as variáveis meteorológicas. Além do mais, como o SFI diz respeito ao percentual de fibras curtas, também há chances desse índice apresentar correlações com variáveis ambientais que tenham afetado o comprimento da fibra de algodão.

O SFI apresentou correlações ($P < 0,01$) somente com as variáveis Qg, P e $\sum EXC$, todas elas negativas (Tabela 13). Ademais, para todas essas variáveis, os valores de correlações foram maiores quando se considerou o ciclo total do algodoeiro.

Comparando-se os resultados das correlações de SFI com os de índice de uniformidade, cada uma delas apresentou três correlações significativas. Dentre essas, duas foram comuns: P e $\sum EXC$. Entretanto, como já dito anteriormente, SFI e uniformidade agem em conjunto, mas de maneira contrária, uma vez que o baixo valor de uma, aliado ao alto valor de outra, gera alterações na qualidade da fibra. Assim, é esperado que haja correlações comuns, porém contrárias, entre elas. Nesse sentido, como era de se esperar, para o SFI, as correlações entre P e $\sum EXC$ foram negativas, enquanto para o índice de uniformidade, foram positivas (Tabela 13).

Em relação ao comprimento, a única correlação comum foi com a UR, sendo que para o SFI a correlação com UR foi negativa, enquanto para o comprimento foi positiva. Dessa forma, fica evidente que o SFI age em conjunto com o índice de uniformidade, porém de maneira contrária, como observado nos resultados da Tabela 13 e citado por Fonseca e Farias (2002). Além disso, foi possível observar que o SFI é mais sensível à precipitação e suas variações e que essa sensibilidade é mais evidente durante o ciclo total do algodoeiro (Tabela 13).

4.2.2. Análise de Agrupamento

Os índices de qualidades da fibra do algodão (micronaire, comprimento, resistência, uniformidade, alongação e SFI) foram submetidos a uma análise de agrupamento, considerando-se os dados de todas as cultivares avaliadas. Após a definição dos grupos e de quais cultivares estavam incluídas em cada um deles, foi feita uma matriz de correlação por grupo entre os dados de índices de qualidade da fibra e as variáveis meteorológicas e do balanço hídrico. Além disso, também foram consideradas os três períodos do ciclo do algodoeiro, ou seja, ciclo total; últimos 100 dias do ciclo; e últimos 50 dias do ciclo.

Primeiramente, definiu-se o número de grupos a serem considerados por meio da utilização do método *k-means*, que fornece um gráfico com a relação da soma dos quadrados dentro dos grupos (soma das distâncias quadráticas de cada registro à média de seu grupo atribuído) e o número de grupos (Figura 42).

Figura 42. Relação entre a soma dos quadrados dentro dos grupos (soma das distâncias quadráticas de cada registro em relação à média de seu grupo atribuído) e o número de grupos considerados na análise de agrupamento.

Conforme mencionado, quanto menor a soma dos quadrados dentro dos grupos, melhor é o agrupamento. Na Figura 42 é possível observar que de 1 a 10, o melhor número de grupos seria o de 10. Entretanto, nota-se que a partir de seis grupos, a soma dos quadrados dentro dos grupos ficou praticamente estável, em torno de 200. Assim, por questões de praticidade e viabilidade, optou-se por empregar seis grupos (valor intermediário) para compor a análise de agrupamento.

Definido o número de grupos, efetuou-se a análise de agrupamento em si, pelo método Ward, um método não-hierárquico no qual o critério utilizado para determinar as distâncias entre os objetos (cultivares) foi a distância euclidiana, apresentada na Eq. 1. O agrupamento hierárquico se presta a uma exibição gráfica natural, o dendrograma. O nome vem das palavras gregas *dendro* (árvores) e *gramma* (desenho). O dendrograma está na Figura 43 e as folhas das árvores correspondem aos registros (cultivares) e o comprimento dos galhos das árvore indica o grau de similaridade entre os grupos correspondentes (BRUCE; BRUCE, 2019).

De acordo com o dendrograma da Figura 43, é possível notar que as cultivares estão representadas dentro de cada grupo por números e não por nomes. Assim, a identificação das cultivares, bem como a composição de cada grupo estão apresentadas na Tabela 14.

Figura 43. Dendrograma do agrupamento das cultivares da cultura do algodoeiro em relação aos índices de qualidade da fibra de algodão. Os números correspondem às cultivares que se encontram na Tabela 14.

Tabela 14. Grupos de cultivares de algodoeiro classificados de acordo com a análise de agrupamento para os índices de qualidade da fibra do algodão. Os números das cultivares correspondem aos que são apresentados na Figura 43 e o número de dados se refere à quantidade de dados disponíveis por grupo considerando-se o número de cultivares, de caracteres de qualidade, épocas de semeadura e safras.

Grupos	Grupo 1	Grupo 2	Grupo 3
Cultivares	5- DP1746B2RF	3-DP1536B2RF	2- DP1243B2RF
	8-FM944GL	4- DP1648B2RF	7- FM940GLT
	25- TMG44B2RF	10-FM954GLT	11- FM975WS
	28- TMG47B2RF	13-FM983GLT	21- TMG11WS
	29-TMG61RF	20- IMA8405GLT	23- TMG42WS
	30- TMG62RF		26- TMG45B2RF
			31- TMG81WS
Número de dados	239	142	355
Grupos	Grupo 4	Grupo 5	Grupo 6
Cultivares	6-FM910	1- DP1240B2RF	15-FMT523
	9-FM951LL	12-FM980GLT	16- FMT701
	14-FM993	27-TMG46B2RF	17-FMT705
	18-FMT707	32-TMG82WS	19- FMT709
			22- TMG41WS
			24-TMG43WS
Número de dados	70	75	204

Os resultados das matrizes de correlação feitas após a análise de agrupamento são apresentados nas Tabelas 15 a 20, organizados de acordo com o índice de qualidade da fibra do algodão e pelos períodos do ciclo da cultura.

Tabela 15. Coeficiente de correlação de Pearson para a relação entre o micronaire e as variáveis meteorológicas e do balanço hídrico (CAD = 90 mm).

Coeficientes de correlação de Pearson (r)			
Variável*	Ciclo todo	Últimos 100 dias	Últimos 50 dias
Grupo 1			
Tmin	0,25**	0,28**	0,29**
UR	0,26**	0,29**	0,34**
P	0,24**	0,39**	0,39**
AT	- 0,19**	- 0,24**	- 0,26**
Σ DEF	- 0,24**	- 0,27**	- 0,33**
Σ EXC	0,22**	0,39**	0,35**
ARMCAD	0,25**	0,26**	0,34**
Tmin18	- 0,29**	- 0,31**	- 0,32**
Tmin20	- 0,31**	- 0,35**	- 0,42*
AT13	- 0,18**	- 0,22**	- 0,22**
Grupo 2			
Σ DEF	- 0,22**	- 0,23**	- 0,38**
Tmin20	- 0,28**	- 0,24**	- 0,24**
Grupo 3			
U2	- 0,18**	- 0,20**	- 0,23**
Tmax30	0,32**	0,28**	0,18**
Tmax32	0,30**	0,28**	0,17**
Tmin18	- 0,17**	- 0,14**	- 0,16**
Tmin20	- 0,43**	- 0,30**	- 0,25**
Grupo 5			
Tmin	0,3**	0,36**	0,36**
Tmin18	- 0,43**	- 0,42**	- 0,39**
Grupo 6			
Qg	0,27**	0,36**	0,32**
Tmax	0,37**	0,37**	0,32**
Tmin	0,37**	0,31**	0,29**
Tmed	0,40**	0,40**	0,37**
Tmax30	0,36**	0,36**	0,31**
Tmax32	0,39**	0,38**	0,34**
Tmin18	- 0,28**	- 0,26**	- 0,29**
Tmin20	- 0,40**	- 0,31**	- 0,30**

* os significados das variáveis meteorológicas e do balanço hídrico se encontram nos itens 3.3 e 3.4

**representa correlação significativa a 1% de probabilidade pelo teste t.

Tabela 16. Coeficiente de correlação de Pearson para a relação entre o comprimento e as variáveis meteorológicas e do balanço hídrico (CAD = 90 mm).

Coeficientes de correlação de Pearson (r)			
Variável*	Ciclo todo	Últimos 100 dias	Últimos 50 dias
Grupo 1			
Qg	-0,61**	-0,51**	-0,63**
Tmax	-0,37**	-0,38**	-0,49**
UR	0,74**	0,67**	0,58**
U2	0,28**	0,33**	0,28**
AT	-0,21**	-0,24**	-0,36**
Tmed	-0,37**	-0,31**	-0,32**
Σ DEF	-0,31**	-0,36**	-0,60**
Tmax30	-0,45**	-0,45**	-0,52**
Tmax32	-0,35**	-0,38**	-0,48**
AT12	-0,18**	-0,24**	-0,42**
AT13	-0,31**	-0,33**	-0,43**
NDC	-0,52**	-0,42**	-0,22**
ARM/CAD	0,26**	0,27**	0,39**
Grupo 3			
Qg	-0,23**	-0,16**	-0,22**
Tmax	-0,18**	-0,17**	-0,19**
UR	0,38**	0,33**	0,27**
U2	0,18**	0,20**	0,17**
Σ DEF	-0,15**	-0,18**	-0,29**
Tmax30	-0,21**	-0,18**	-0,19**
Tmax32	-0,18**	-0,18**	-0,20**
AT13	-0,15**	-0,16**	-0,19**
Grupo 4			
AT	-0,53**	-0,47**	-0,43**
Tmax30	-0,42**	-0,38**	-0,44**
AT10	-0,55**	-0,47**	-0,38**
AT12	-0,54**	-0,51**	-0,46**
AT13	-0,50**	-0,49**	-0,49**
Grupo 6			
P	0,20**	0,21**	0,19**
AT	-0,23**	-0,21**	-0,19**
AT10	-0,24**	-0,23**	-0,19**
AT12	-0,21**	-0,22**	-0,22**
AT13	-0,19**	-0,19**	-0,20**

* os significados das variáveis meteorológicas e do balanço hídrico se encontram nos itens 3.3 e 3.4

**representa correlação significativa a 1% de probabilidade pelo teste t.

Tabela 17. Coeficiente de correlação de Pearson para a relação entre a resistência e as variáveis meteorológicas e do balanço hídrico (CAD = 90 mm).

Coeficientes de correlação de Pearson (r)			
Variável*	Ciclo todo	Últimos 100 dias	Últimos 50 dias
Grupo 1			
UR	0,37**	0,31**	0,27**
NDC	-0,20**	-0,19**	-0,17**
Grupo 2			
U2	0,32**	0,32**	0,27**
Grupo 6			
Qg	-0,39**	-0,30**	-0,20**
Σ DEF	0,25**	0,22**	0,23**
NDC	-0,39**	-0,31**	-0,41**
DEF0	0,40**	0,23**	0,33**
ARM/CAD	-0,24**	-0,22**	-0,19**

* os significados das variáveis meteorológicas e do balanço hídrico se encontram nos itens 3.3 e 3.4

**representa correlação significativa a 1% de probabilidade pelo teste t.

Tabela 18. Coeficiente de correlação de Pearson para a relação entre a uniformidade e as variáveis meteorológicas e do balanço hídrico (CAD = 90 mm).

Coeficientes de correlação de Pearson (r)			
Variável*	Ciclo todo	Últimos 100 dias	Últimos 50 dias
Grupo 1			
U2	0,19**	0,23**	0,19**
Grupo 4			
Tmax	-0,37**	-0,41**	-0,46**
P	0,30**	0,40**	0,40**
AT	-0,53**	-0,57**	-0,52**
Σ DEF	-0,34**	-0,36**	-0,30**
Tmax30	-0,39**	-0,46**	-0,50**
Tmax32	-0,32**	-0,36**	-0,37**
AT10	-0,48**	-0,56**	-0,45**
AT12	-0,56**	-0,61**	-0,55**
AT13	-0,53**	-0,58**	-0,55**
Grupo 5			
Tmin	0,36**	0,37**	0,34**
UR	0,30**	0,34**	0,43**
P	0,33**	0,37**	0,45**
Σ EXC	0,30**	0,31**	0,40**
Tmin18	-0,34**	-0,38**	-0,42**
Tmin20	-0,36**	-0,46**	-0,41**
Grupo 6			
Tmax	-0,21**	-0,19**	-0,19**
AT	-0,24**	-0,22**	-0,22**
Tmax30	-0,23**	-0,23**	-0,23**
AT12	-0,22**	-0,22**	-0,23**
AT13	-0,23**	-0,23**	-0,24**

* os significados das variáveis meteorológicas e do balanço hídrico se encontram nos itens 3.3 e 3.4

**representa correlação significativa a 1% de probabilidade pelo teste t.

Tabela 19. Coeficiente de correlação de Pearson para a relação entre a elongação e as variáveis meteorológicas e do balanço hídrico (CAD = 90 mm).

Coeficientes de correlação de Pearson (r)			
Variável*	Ciclo todo	Últimos 100 dias	Últimos 50 dias
Grupo 1			
Qg	-0,57**	-0,54**	-0,60**
Tmax	-0,46**	-0,50**	-0,56**
Tmin	-0,36**	-0,24**	-0,18**
UR	0,49**	0,42**	0,30**
U2	0,25**	0,28**	0,26**
AT	-0,21**	-0,21**	-0,29**
Tmed	-0,49**	-0,47**	-0,48**
Σ DEF	- 0,22**	- 0,25**	- 0,41**
Tmax30	- 0,51**	- 0,52**	- 0,56**
Tmax32	- 0,43**	- 0,45**	- 0,53**
AT12	-0,24**	-0,24**	-0,36**
AT13	-0,30**	-0,30**	-0,36**
NDC	-0,40**	-0,34**	-0,22**
Grupo 2			
Qg	- 0,41**	- 0,41**	- 0,41**
Tmax	- 0,47**	- 0,51**	- 0,51**
U2	0,34**	0,32**	0,30**
AT	-0,28**	-0,23**	-0,26**
Tmed	-0,44**	-0,44**	-0,42**
Tmax30	-0,47**	-0,52**	-0,53**
Tmax32	-0,38**	-0,38**	-0,44**
AT12	-0,30**	-0,23**	-0,26**
AT13	-0,31**	-0,27**	-0,29**
Grupo 3			
Qg	- 0,30**	- 0,30**	- 0,34**
Tmax	-0,23**	-0,26**	-0,26**
U2	0,16**	0,16**	0,16**
Tmed	- 0,24**	- 0,25**	- 0,23**
Tmax30	- 0,25**	- 0,29**	- 0,28**
Tmax32	-0,22**	-0,23**	-0,24**
AT12	-0,17**	-0,14**	-0,15**
AT13	-0,19**	-0,17**	-0,17**
NDC	-0,29**	-0,23**	-0,17**
Grupo 4			
Qg	-0,42**	-0,42**	-0,32**
ARM/CAD	-0,41**	-0,41**	-0,33**
Grupo 6			
Qg	-0,24**	-0,24**	-0,23**
P	-0,27**	-0,21**	-0,20**
Σ DEF	0,27**	0,25**	0,20**
NDC	-0,40**	-0,30**	-0,30**
EXC0	-0,44**	-0,25**	-0,21**
DEF0	0,41**	0,26**	0,23**
ARM/CAD	- 0,30**	- 0,27**	- 0,20**

* os significados das variáveis meteorológicas e do balanço hídrico se encontram nos itens 3.3 e 3.4.

**representa correlação significativa a 1% de probabilidade pelo teste t.

Tabela 20. Coeficiente de correlação de Pearson para a relação entre o índice de fibras curtas e as variáveis meteorológicas e do balanço hídrico (CAD = 90 mm).

Coeficientes de correlação de Pearson (r)			
Variável*	Ciclo todo	Últimos 100 dias	Últimos 50 dias
Grupo 1			
UR	-0,36**	-0,34**	-0,30**
Tmin20	0,21**	0,20**	0,23**
Grupo 4			
Tmax	0,33**	0,34**	0,40**
P	-0,30**	-0,40**	-0,36**
AT	0,51**	0,56**	0,50**
Σ DEF	0,38**	0,39**	0,31**
Tmax30	0,37**	0,41**	0,46**
AT10	0,48**	0,55**	0,43**
AT12	0,56**	0,59**	0,52**
AT13	0,51**	0,57**	0,54**
Grupo 5			
Tmax	-0,45**	-0,40**	-0,37**
Tmin	-0,40**	-0,37**	-0,34**
UR	-0,50**	-0,44**	-0,38**
U2	0,43**	0,43**	0,46**
Tmed	-0,46**	-0,44**	-0,44**
Tmax30	-0,45**	-0,39**	-0,33**
Tmax32	-0,43**	-0,40**	-0,39**
Tmin18	0,33**	0,32**	0,35**
Tmin20	0,48**	0,49**	0,49**

* os significados das variáveis meteorológicas e do balanço hídrico se encontram nos itens 3.3 e 3.4

**representa correlação significativa a 1% de probabilidade pelo teste t.

Em relação ao índice micronaire, pode-se observar que, tanto para a análise de agrupamento (Tabela 15), quanto para a análise de correlação com abordagem geral (Tabela 13), houve concordância, em alguns grupos para as correlações com as variáveis meteorológicas Tmin20, Tmin, Tmed e Tmax32. O Grupo 1 apresentou, da mesma forma que a abordagem geral, correlação negativa com Tmin20 ($P < 0,01$) e correlação positiva com Tmin ($P < 0,01$). O Grupo 2 apresentou concordância de correlação em relação à Tmin20, sendo que nas duas abordagens a correlação com essa variável foi negativa ($P < 0,01$) e melhor quando considerado o ciclo total do algodoeiro. O Grupo 3, por sua vez, apresentou concordância entre as duas abordagens para a relação entre o micronaire e as variáveis Tmin20, com correlação negativa ($P < 0,01$), e Tmax32, com correlação positiva ($P < 0,01$). Além disso, para os dois casos as correlações para essas duas variáveis meteorológicas foram maiores considerando-se o ciclo total do algodoeiro. O Grupo 4 não apresentou nenhuma correlação significativa, enquanto o Grupo 5 apresentou correlação positiva somente com a Tmin ($P < 0,01$), da mesma forma que a abordagem geral (Tabela 13). Por fim, o Grupo 6

apresentou concordância com a abordagem geral em termos das correlações entre o micronaire e as variáveis Tmin20, com correlação negativa ($P < 0,01$) e mais intensa para o ciclo todo, Tmed, com correlação positiva ($P < 0,01$), Tmin, com correlação positiva ($P < 0,01$) e melhor para o ciclo todo, e Tmax32, com correlação positiva ($P < 0,01$) (Tabela 15).

Dessa forma, pode-se observar que, para as duas abordagens (geral e agrupamento), houve concordância entre as correlações do micronaire e as variáveis ambientais, sendo as principais Tmin, Tmin20, Tmed e Tmax32 (Tabela 15). Além disso, para todas as correlações entre o micronaire e essas variáveis, o ciclo total foi o período que demonstrou maior influência na qualidade da fibra de algodão, especialmente nos Grupos 2, 3 e 6. Esses resultados evidenciam a importância da temperatura do ar sobre o micronaire da fibra de algodão, especialmente quando se trata da temperatura mínima.

Para o comprimento da fibra (Tabela 16), houve muita variação dos coeficientes de correlação entre os diferentes grupos. Entretanto, independentemente do grupo, houve concordância com a abordagem geral para as correlações entre o comprimento e as variáveis UR, Qg, U2, Tmed, Σ DEF, Tmax30, AT13 e ARM/CAD. O Grupo 1 apresentou, para as duas abordagens, correlações significativas ($P < 0,01$) entre comprimento e as seguintes variáveis: UR (positiva e maior para o ciclo total); Qg (negativa); U2 (positiva); Tmed (negativa e maior para o ciclo total); Σ DEF (negativa e maior para os últimos 50 dias do ciclo do algodoeiro); Tmax30 (negativa); AT13 (negativa e maior para os últimos 50 dias do ciclo); e ARM/CAD (positiva e maior nos últimos 50 dias do ciclo). O Grupo 2 não apresentou nenhuma correlação significativa na abordagem de agrupamento e o Grupo 3 apresentou correlações significativas entre comprimento e as variáveis Qg (negativa); UR (positiva); U2 (positiva); Σ DEF (negativa e maior nos últimos 50 dias do ciclo); Tmax30 (negativa); e AT13 (negativa). Já o Grupo 4 apresentou correlações significativas do comprimento com as variáveis Tmax30 (negativa) e AT 13 (negativa), enquanto o Grupo 5 não apresentou nenhuma correlação significativa e o Grupo 6 apresentou correlação negativa ($P < 0,01$) apenas com AT13. Embora tenha havido bastante variação de correlações entre os grupos, pode-se observar que todas as correlações que ocorreram do comprimento com as variáveis meteorológicas foram concordantes entre as duas abordagens estudadas (Tabelas 13 e 16).

Para a resistência, não ocorreram muitas correlações significativas na abordagem por agrupamento (Tabela 17), do mesmo modo que também não ocorreram na abordagem geral (Tabela 13). Entretanto, quando houveram correlações, elas foram correspondentes entre as duas abordagens. Em termos comparativos, na abordagem em geral, a resistência apresentou correlação significativa ($P < 0,01$) com as seguintes variáveis: Qg (negativa e maior para o ciclo todo); NDC (negativa); EXC0 (negativa e melhor para o ciclo total); DEF0 (positiva e maior para o ciclo total); e ARM/CAD (negativa). Já na abordagem de agrupamento, o Grupo 1 apresentou correlação ($P < 0,01$) apenas com NDC (negativa), enquanto os Grupos 2, 3, 4 e 5 não apresentaram nenhuma correlação significativa. Finalmente, o Grupo 6 apresentou correlações significativas ($P < 0,01$) entre a resistência e as seguintes variáveis meteorológicas: Qg (negativa e melhor para o ciclo total); NDC (negativa); DEF0 (positiva e maior para o ciclo todo); e ARM/CAD (negativa).

No caso do índice de uniformidade, considerando-se a concordância entre as abordagens geral (Tabela 13) e por agrupamento (Tabela 18), foi possível observar que o Grupo 1 apresentou correlação ($P < 0,01$) apenas com U2 (positiva) nas duas situações; o Grupo 4 correlação positiva com P ($P < 0,01$); e o Grupo 5 correlação positiva ($P < 0,01$) com P e $\sum EXC$.

Na abordagem geral (Tabela 13), a elongação apresentou correlações com as seguintes variáveis meteorológicas ($P < 0,01$): Qg (negativa); Tmax (negativa); Tmin (negativa e melhor no ciclo total); P (negativa e melhor no ciclo total); Tmed (negativa); Tmax30 (negativa); NDC (negativa e melhor no ciclo total); EXC0 (negativa e melhor no ciclo total); e DEF0 (positiva e maior no ciclo total). Já na abordagem de agrupamento (Tabela 19), os resultados foram variáveis entre os grupos. O Grupo 1 apresentou correlações da elongação com as seguintes variáveis meteorológicas ($P < 0,01$): Qg (negativa); Tmax (negativa e melhor nos últimos 50 dias); Tmin (negativa e melhor no ciclo todo); Tmed (negativa); Tmax30 (negativa e melhor nos últimos 50 dias); e NDC (negativa e melhor no ciclo todo). O Grupo 2 apresentou correlações com a Qg (negativa), a Tmax (negativa), a Tmed (negativa) e Tmax30 (negativa), todos significativas ($P < 0,01$). O Grupo 3, por sua vez, apresentou correlações ($P < 0,01$) com: Qg (negativa); Tmax (negativa); Tmed (negativa); Tmax30 (negativa); e NDC (negativa e melhor para o ciclo todo). O Grupo 4 apresentou correlação ($P < 0,01$) apenas com Qg (negativa) e o Grupo 5 não apresentou nenhuma correlação significativa. Por fim, o Grupo 6 apresentou correlações com as seguintes variáveis ($P < 0,01$): Qg (negativa); P (negativa e melhor para o ciclo total); NDC (negativa e melhor para o ciclo total); EXC0 (negativa e melhor para o ciclo total); e DEF0 (positiva e maior para o ciclo total). Assim, pode-se observar que houve muita variação entre os grupos na análise de agrupamento para a elongação (Tabela 19), entretanto, os valores encontrados foram coerentes com os encontrados para a abordagem em geral (Tabela 13).

O índice de fibras curtas (SFI) apresentou, na abordagem geral (Tabela 13), correlações ($P < 0,01$) com três variáveis meteorológicas: UR (negativa e melhor para o ciclo todo); P (negativa e melhor para o ciclo total); e $\sum EXC$ (negativa e melhor para o ciclo total). Em relação à análise de agrupamento para o SFI (Tabela 20), não ocorreram muitas correlações significativas, mas, da mesma forma que aconteceu com os outros índices de qualidade, as que ocorreram concordaram com a abordagem geral. O Grupo 1 apresentou correlação ($P < 0,01$) apenas com UR (negativa e melhor para o ciclo total); o Grupo 4 apenas com P (negativa) ($P < 0,01$); o Grupo 5 apenas com UR (negativa e melhor para o ciclo total) ($P < 0,01$), enquanto que os Grupos 2, 3 e 6 não apresentaram nenhuma correlação significativa.

Dessa forma, constata-se que, embora tenha havido variações nos coeficientes de correlação entre os grupos na análise de agrupamento, para todos os índices de qualidade da fibra avaliados, as correlações que ocorreram foram similares às da abordagem geral. Além disso, apesar dos valores dos coeficientes de correlação encontrados nem sempre terem sido elevados, todos apresentaram-se significativos ($P < 0,01$), considerando-se o número de dados utilizados para este estudo, o que é um indício de que os índices de qualidade da fibras sofrem influência das condições ambientais, sendo que uns mais e outros menos, e que modelos empíricos de estimação desses índices de qualidade em função dessas condições podem ser

desenvolvidos, sendo que uma das formas de se fazer isso é por meio de técnicas de Mineração de dados.

4.3. Modelagem dos índices de qualidade da fibra de algodão por meio da mineração de dados

4.3.1. Análise exploratória e entendimento dos dados

4.3.1.1. Análise do número de dados por cultivar

A Figura 44 mostra a variação do número de dados por cultivar, em porcentagem. Neste estudo foram utilizados 1085 dados de 32 cultivares e foi possível notar que a cultivar mais representativa em relação ao número de dados foi a FM975WS (83 dados), seguida das cultivares TMG81WS (74 dados) e TMG42WS (65 dados). Em termos percentuais, elas representam, respectivamente: 7,7%, 6,8%, e 6% do total de dados. Já as cultivares que apresentaram menor quantidade de dados foram a FM910 e a FM993, ambas com apenas 11 dados (1% do total de dados).

Figura 44. Porcentagem de dados de índices de qualidade da fibra por cultivar do algodoeiro analisada em relação ao total de dados.

4.3.1.2. Valores médios dos índices de qualidade por cultivar

A Tabela 21 apresenta os valores médios dos índices de qualidade avaliados para cada uma das cultivares de algodoeiro empregadas na mineração de dados.

Tabela 21. Valores médios dos índices de qualidade de fibra do algodão para todas as cultivares analisadas, considerando-se todos os locais e datas de semeadura.

Cultivar	Micronaire (adimensional)	Comprimento (mm)	Resistência (gf tex⁻¹)	Uniformidade (%)	Elongação (%)	SFI (%)
DP1240B2RF	4,37	30,87	32,29	83,34	7,66	7,16
DP1243B2RF	4,06	29,32	29,89	81,67	7,22	9,23
DP1536B2RF	4,60	30,14	32,48	84,25	7,08	6,78
DP1648B2RF	4,57	30,09	30,93	83,50	7,15	7,44
DP1746B2RF	4,02	29,94	30,64	83,17	6,75	7,69
FM910	4,10	30,21	28,43	84,00	5,91	7,03
FM940GLT	4,00	29,27	30,73	83,22	8,13	7,66
FM944GL	4,20	29,96	31,33	83,48	6,89	7,39
FM951LL	4,30	29,90	28,62	83,71	6,48	7,38
FM954GLT	4,10	30,87	30,54	82,55	6,56	7,56
FM975WS	4,04	29,79	30,89	83,18	7,07	7,79
FM980GLT	3,69	28,85	31,01	82,79	7,99	7,93
FM983GLT	4,12	29,46	29,11	81,03	6,55	9,51
FM993	4,20	30,16	28,67	84,08	5,56	7,10
FMT523	4,07	31,15	31,11	85,16	6,20	6,32
FMT701	4,17	29,31	30,44	84,31	6,79	7,13
FMT705	4,53	29,25	29,53	85,02	6,71	6,66
FMT707	3,98	30,30	28,98	84,32	5,86	6,99
FMT709	4,22	29,19	28,80	84,20	6,39	7,09
IMA8405GLT	4,23	30,12	30,75	82,47	7,15	8,39
TMG11WS	3,71	30,84	31,45	84,86	6,93	7,04
TMG41WS	4,16	29,50	32,87	86,28	6,61	5,94
TMG42WS	4,03	29,42	29,95	83,38	7,21	7,97
TMG43WS	4,13	29,18	31,64	83,99	7,00	7,37
TMG44B2RF	4,08	30,20	30,96	83,26	6,74	7,58
TMG45B2RF	4,47	28,92	29,49	83,35	8,07	7,66
TMG46B2RF	4,26	28,73	29,40	82,75	7,93	7,94
TMG47B2RF	3,99	30,23	30,81	83,10	6,57	7,88
TMG61RF	4,33	29,71	31,28	83,11	7,24	7,51
TMG62RF	4,31	29,62	30,02	83,56	6,93	7,29
TMG81WS	4,37	28,77	30,73	83,64	7,21	7,65
TMG82WS	3,77	29,41	30,08	83,79	6,92	7,45
Média	4,17	29,66	30,51	83,53	6,99	7,56
Desvio padrão	0,38	0,91	1,85	1,38	0,88	1,13
CV (%)	9,11	3,07	6,06	1,65	12,59	14,95

O maior CV encontrado para os índices de qualidade da fibra de algodão foi de 14,95% para o SFI, enquanto o menor (1,65%) foi para a uniformidade (Tabela 21). Apesar da pouca variação encontrada para a maioria dos índices de qualidade, esses devem ser confrontados em relação aos valores de referência, como apresentado na Tabela 3.

Para os valores de micronaire, o ideal é que os valores estejam entre 3,8 a 4,5. Dentre as 32 cultivares consideradas, cinco delas não apresentaram valores dentro dessa faixa (DP1536B2RF; DP1648B2RF; FM980GLT; TMG11WS; e TMG82WS). A média de todas as cultivares, 4,17 (Tabela 21), está dentro do padrão de qualidade apresentado na Tabela 3.

Para a indústria têxtil, é ideal que o comprimento esteja acima de 30 mm (Tabela 3). Das 32 cultivares, apenas 12 apresentaram valores acima de 30 mm (Tabela 21), entretanto, o menor valor encontrado foi o de 28,73 mm (TMG46B2RF), que é bem próximo mínimo requerido. Além disso, a média de todas as cultivares foi de 29,66 mm, praticamente igual ao valor mínimo.

Para a resistência, os valores ideais para a indústria têxtil devem estar acima de 31 gf tex⁻¹ (Tabela 3), no entanto, os valores da Tabela 21 mostram que apenas nove entre as 32 cultivares apresentaram valores acima dessa referência. No entanto, o menor valor encontrado foi de 28,43 gf tex⁻¹ (FM910), enquanto a média de todas as cultivares foi 30,51 gf tex⁻¹, bem próximo do mínimo exigido.

A uniformidade deve apresentar valores entre 85 e 87% para representar uma fibra de qualidade (Tabela 3). Dentre as cultivares analisadas, apenas três apresentaram valores dentro dessa faixa (FMT523; FMT705; e TMG41WS). Consequentemente, a média de todas as cultivares não apresentou o valor requerido para o padrão de qualidade da indústria têxtil, ficando por volta de 83,5% (Tabela 21).

A elongação, de acordo com os valores de referência apresentados na Tabela 3, deve estar entre 6,8 e 7,7 % para atender aos critérios de qualidade exigidos pela indústria têxtil. Dentre as 32 cultivares, 18 apresentaram valores fora dessa faixa (Tabela 21), sendo o menor valor igual a 5,56 % e o maior chegando a 8,13 %. O valor de elongação médio para todas as cultivares (6,99%) está dentro do intervalo exigido.

Em relação ao índice de fibras curtas (SFI), o ideal é que ele seja menor do que 6% (Tabela 3). Sendo assim, somente a cultivar TMG41WS se adequou à faixa padrão, apresentando um SFI igual a 5,94% (Tabela 21). Todas as demais cultivares apresentaram valores maiores do que 6%, sendo a média de todas as cultivares igual a 7,56%, muito acima do limite considerado como ideal para a qualidade da fibra de algodão.

Resumidamente, foi possível observar que para a média de todas as cultivares de algodoeiro, apenas o índice micronaire e a elongação apresentaram valores que se adequam às exigências de qualidade de fibra da indústria têxtil, enquanto o comprimento e a resistência apresentaram valores bem próximos da condição ideal. Por outro lado, a uniformidade e o SFI foram os índices de qualidade que apresentaram as maiores discrepâncias em relação aos valores de referência de qualidade para a fibra de algodão, como apresentado nas Tabelas 3 e 21.

4.3.1.3. Variabilidade dos índices de qualidade da fibra de diferentes cultivares de algodoeiro

As Figuras 45 a 50 apresentam os resultados (boxplot) para detecção de *outliers* e para avaliação das variabilidades dos índices de qualidade da fibra de algodão em função das cultivares estudadas em cinco localidades do estado do Mato Grosso, durante oito safras (2011/12 a 2018/19) e três épocas de semeadura por safra. Já as informações que permitem avaliar a variabilidade dos índices de qualidade da fibra de algodão ao longo dos anos considerados são apresentadas no item 4.3.1.4.

A Figura 45 mostra a variação do micronaire entre as cultivares e, para uma mesma cultivar, entre as localidades e safras/semaduras analisadas. O micronaire variou de 3,05 a 5,15 e para esse índice foram contabilizados 15 *outliers*, distribuídos entre nove cultivares: DP1243B2RF (dois *outliers*); FM944GL (dois *outliers*); FM951LL (dois *outliers*); FM954GLT (um *outlier*); FM983GLT (um *outlier*); FMT705 (dois *outliers*); TMG44B2RF (um *outlier*); TMG47B2RF (dois *outliers*); e TMG62RF (um *outlier*).

A Figura 46 apresenta a variação do comprimento da fibra de algodão em função das cultivares e das localidades e safras/semaduras analisadas. O comprimento variou de 26,92 mm a 32,58 mm. Foram contabilizados 21 *outliers*, distribuídos entre doze cultivares: DP1240B2RF (três *outliers*); DP1243B2RF (um *outlier*); FM944GL (dois *outliers*); FM975WS (três *outliers*); FM993 (um *outlier*); FMT523 (três *outliers*); FMT705 (um *outlier*); FMT707 (três *outliers*); TMG42WS (um *outlier*); TM44B2RF (um *outlier*); TMG45B2RF (um *outlier*); e TMG81WS (um *outlier*).

Para a resistência da fibra (Figura 47), os valores variaram de 24,6 a 37,38 gf.tex⁻¹ considerando-se todas as cultivares, locais, safras e épocas de semeadura. Para esse índice de qualidade da fibra do algodão houveram 13 *outliers*, distribuídos entre 10 cultivares: DP1746B2RF (um *outlier*); FM944GL (três *outliers*); FM983GLT (um *outlier*); FMT709 (um *outlier*); TMG41WS (um *outlier*); TMG42WS (dois *outliers*); TMG43WS (um *outlier*); TMG45B2RF (um *outlier*); TMG46B2RF (um *outlier*); e TMG62RF (um *outlier*).

A uniformidade, por sua vez, variou de 78,4% a 87,35% (Figura 48) quando se considerou todas as cultivares, locais, safras e épocas de semeadura. O número de *outliers* para a uniformidade foi igual a 16 e eles foram distribuídos entre onze cultivares: DP1243B2RF (dois *outliers*); DP1536B2RF (um *outlier*); DP1648B2RF (dois *outliers*); FM910 (um *outlier*); FM951LL (um *outlier*); FM954GLT (um *outlier*); FMT523 (dois *outliers*); FMT707 (dois *outliers*); TMG41WS (um *outlier*); TMG44B2RF (dois *outliers*); e TMG47B2RF (um *outlier*).

Para o alongamento (ou alongação) os dados mostraram uma variação de 4,9% a 9,7% (Figura 49), para a cominação de condições analisadas (cultivar, local, safra e semeadura). Foram contabilizados 14 *outliers* para o alongamento, sendo esses para 12 cultivares: DP1243B2RF (um *outlier*); DP1536B2RF (um *outlier*); FM940GLT (um *outlier*); FM944GL (dois *outliers*); FM954GLT (um *outlier*); TMG41WS (um *outlier*); TMG42WS (um *outlier*); TMG43WS (dois *outliers*); TMG44B2RF (um *outlier*); TMG46B2RF (um *outlier*); TMG47B2RF (um *outlier*) e TMG82WS (um *outlier*).

Por fim, o índice de fibras curtas (SFI) variou entre 4,2% e 13% (Figura 50), para as situações

Figura 46. Variação do comprimento da fibra de algodão entre 32 cultivares de algodoeiro, considerando-se as cinco localidades de cultivo no estado do Mato Grosso, oito safras (2011/12 a 2018/19) e três épocas de semeadura. Os asteriscos vermelhos representam os *outliers*.

Figura 47. Variação da resistência da fibra de algodão entre 32 cultivares de algodoeiro, considerando-se as cinco localidades de cultivo no estado do Mato Grosso, oito safras (2011/12 a 2018/19) e três épocas de semeadura. Os asteriscos vermelhos representam os *outliers*.

Figura 48. Variação da uniformidade da fibra de algodão entre 32 cultivares de algodoeiro, considerando-se as cinco localidades de cultivo no estado do Mato Grosso, oito safras (2011/12 a 2018/19) e três épocas de semeadura. Os asteriscos vermelhos representam os *outliers*.

Figura 49. Variação da elongação da fibra de algodão entre 32 cultivares de algodoeiro, considerando-se as cinco localidades de cultivo no estado do Mato Grosso, oito safras (2011/12 a 2018/19) e três épocas de semeadura. Os asteriscos vermelhos representam os *outliers*.

Figura 50. Variação do índice de fibras curtas (SFI) de algodão entre 32 cultivares de algodoeiro, considerando-se as cinco localidades de cultivo no estado do Mato Grosso, oito safras (2011/12 a 2018/19) e três épocas de semeadura. Os asteriscos vermelhos representam os *outliers*.

4.3.1.4. Variação dos índices de qualidade da fibra de algodão para as safras analisadas

As Figuras 51 a 56 apresentam os resultados relativos aos índices de qualidade da fibra do algodoeiro avaliada por ano. Considerando-se todos os índices de qualidade, foram contabilizados 86 *outliers* ao longo das oito safras avaliadas (2011/12 a 2018/19). Considerando-se os anos, a maior quantidade de *outliers* observada foi 22 em 2016, e a menor igual a quatro, em 2012 e em 2019. Já quando se considerou os índices de qualidade separadamente, o SFI foi o que mais teve mais *outliers*, totalizando 29, enquanto o micronaire foi índice que menos outliers teve, com apenas dois.

Figura 51. Variação dos dados de micronaire da fibra de algodão para as oito safras, com colheitas entre 2012 e 2019, considerando-se 32 cultivares de algodoeiro, cinco localidades do estado do Mato Grosso, e três épocas de semeadura. Os pontos pretos representam os *outliers*.

Figura 52. Variação dos dados de comprimento da fibra de algodão para as oito safras, com colheitas entre 2012 e 2019, considerando-se 32 cultivares de algodoeiro, cinco localidades do estado do Mato Grosso, e três épocas de semeadura. Os pontos pretos representam os *outliers*.

Figura 53. Variação dos dados de resistência da fibra de algodão para as oito safras, com colheitas entre 2012 e 2019, considerando-se 32 cultivares de algodoeiro, cinco localidades do estado do Mato Grosso, e três épocas de semeadura. Os pontos pretos representam os *outliers*.

Figura 54. Variação dos dados de uniformidade da fibra de algodão para as oito safras, com colheitas entre 2012 e 2019, considerando-se 32 cultivares de algodoeiro, cinco localidades do estado do Mato Grosso, e três épocas de semeadura. Os pontos pretos representam os *outliers*.

Figura 55. Variação dos dados de alongação da fibra de algodão para as oito safras, com colheitas entre 2012 e 2019, considerando-se 32 cultivares de algodoeiro, cinco localidades do estado do Mato Grosso, e três épocas de semeadura. Os pontos pretos representam os *outliers*.

Figura 56. Variação dos dados de índice de fibras curtas (SFI) de algodão para as oito safras, com colheitas entre 2012 e 2019, considerando-se 32 cultivares de algodoeiro, cinco localidades do estado do Mato Grosso, e três épocas de semeadura. Os pontos pretos representam os *outliers*.

4.3.1.5. Distribuição dos índices de qualidade da fibra de algodão

A distribuição Normal é uma das mais importantes características das variáveis aleatórias contínuas, cuja distribuição é também chamada de distribuição Gaussiana, a qual serve como modelo de distribuição para muitos problemas da vida real, mas também aparece em muitas investigações teóricas, pois sua importância nas análises matemáticas se deve ao fato de que muitas técnicas estatísticas, como análise de variância e regressão, além de alguns testes de hipóteses, assumem ou exigem a Normalidade dos dados (MORETTIN; BUSSAB, 2003). Diante disso, e considerando-se que a geração de modelos agrometeorológicos de estimação da qualidade da fibra do algodão por meio de técnicas estatísticas de regressão exigem que os dados empregados no desenvolvimento desses modelos tenham distribuição normal ou próximo dela, foram feitas as análises da distribuição dos dados de todos os índices de qualidade do algodoeiro. As Figuras 57 a 62 apresentam essas distribuições, por meio das quais é possível observar que todos os índices de qualidade avaliados apresentaram distribuição normal ou muito próxima dela. Assim, todas essas variáveis são passíveis de serem modeladas por meio de métodos estatísticos que envolvam regressões de diferentes naturezas (MORETTIN; BUSSAB, 2003).

Figura 57. Histograma da distribuição dos dados de micronaire da fibra de algodão em cinco localidades do estado do Mato Grosso, nas safras de 2011/12 a 2018/19 e em três épocas de semeadura.

Figura 58. Histograma da distribuição dos dados de comprimento da fibra de algodão em cinco localidades do estado do Mato Grosso, nas safras de 2011/12 a 2018/19 e em três épocas de semeadura.

Figura 59. Histograma da distribuição dos dados de resistência da fibra de algodão em cinco localidades do estado do Mato Grosso, nas safras de 2011/12 a 2018/19 e em três épocas de semeadura.

Figura 60. Histograma da distribuição dos dados de uniformidade da fibra de algodão em cinco localidades do estado do Mato Grosso, nas safras de 2011/12 a 2018/19 e em três épocas de semeadura.

Figura 61. Histograma da distribuição dos dados de elongação da fibra de algodão em cinco localidades do estado do Mato Grosso, nas safras de 2011/12 a 2018/19 e em três épocas de semeadura.

Figura 62. Histograma da distribuição dos dados de índice de fibras curtas (SFI) de algodão em cinco localidades do estado do Mato Grosso, nas safras de 2011/12 a 2018/19 e em três épocas de semeadura.

4.3.2. Fase de modelagem

Como já mencionado, o pacote *caret* possibilita a identificação das variáveis preditoras segundo sua importância nos modelos de mineração de dados, ordenando-os da variável mais relevante a menos relevante. As etapas seguintes para cada uma das técnicas utilizadas (RF; XGB; e SVM) foram: identificação das dez variáveis preditoras mais importantes para o modelo; seleção aleatória de 70% dos dados para treinar o modelo (Fase de treinamento); e utilização dos 30% restantes dos dados para testar o modelo (Fase de teste).

Dessa forma, os resultados foram apresentados separadamente para cada índice de qualidade da fibra de algodão avaliado. Os procedimentos detalhados de cada etapa estão considerando apenas os resultados referentes ao ciclo total do algodoeiro, uma vez que, pelas análises de correlação feitas anteriormente e pela execução dos algoritmos de mineração de dados, foi constatado que esse período foi o que gerou as melhores correlações.

4.3.2.1. Micronaire

- **Identificação das variáveis preditoras**

As Figuras 63, 64 e 65 apresentam os resultados da Fase de identificação das variáveis preditoras mais importantes pelos modelos *Random Forest*, *XGBoosting* e *Support Vector Machine*, respectivamente. Já a Figura 66 mostra o Diagrama de Venn, o qual apresenta as dez variáveis selecionadas pelos três modelos empregados e a concordância entre eles para selecionar as variáveis que mais afetaram o micronaire.

Pelo diagrama de Venn (Figura 66), é possível observar que houve concordância entre os modelos de mineração de dados empregados na classificação das variáveis preditoras determinantes do micronaire em três variáveis meteorológicas: Tmax32; Qg; e Tmin20.

Soares (2015), em seu trabalho, concluiu que o micronaire foi influenciado negativamente pela variação de dias com temperatura inferior a 20°C, o que foi condizente com o resultado encontrado no presente estudo. O autor também observou efeito significativo da variação da temperatura máxima para três das cinco cultivares de algodoeiro que avaliou, de maneira que o efeito da temperatura gerou regressão linear significativa positiva com o micronaire. Além disso, segundo Soares (2015), de acordo com as correlações testadas, foi possível inferir que extremos de temperatura máxima, especificamente as superiores a 35°C não representaram problemas à qualidade de fibra do algodoeiro no estado do Mato Grosso, uma vez que as correlações significativas, incluindo às do micronaire, foram muito baixas. Neste estudo, entretanto, os resultados encontrados para extremos de temperatura máxima foram diferentes dos encontrados pelo autor supracitado, uma vez que dentre as variáveis preditoras mais importantes detectadas pelas três técnicas de MD utilizadas, uma delas foi o número de dias com temperatura máxima maior ou igual a 32°C.

Raphael (2019) avaliou a qualidade da fibra de algodão de plantas sombreadas submetidas a doses de nitrogênio. Quatro experimentos de campo foram conduzidos no Brasil nos anos de 2016 e 2017: Primavera do Leste, MT (I); Itapeva, SP, em áreas não irrigada (II) e irrigada (III); e Chapadão do Sul, MS (IV). As plantas foram sombreadas no início do florescimento, reduzindo-se em 42% a radiação global por um período entre 16 e 25 dias, e realizaram-se duas adubações de N em cobertura durante a fase reprodutiva, nas doses de 0, 60, 120 e 180 kg ha⁻¹ (I, II e III); e 0, 80 e 160 kg ha⁻¹ (IV), testando-se também no último experimento os cultivares EB 1405 GLT (ciclo precoce) e FM 983 GLT (ciclo tardio). O autor concluiu que a ausência de interações na maioria das análises indicou que os efeitos do sombreamento e da adubação nitrogenada sobre a qualidade de fibra são independentes e que o sombreamento pouco influenciou os índices de qualidade avaliados, com efeitos pontuais e inconsistentes sobre o micronaire. O resultado encontrado no presente estudo foi diferente, uma vez que a radiação solar (Qg), para o micronaire, se apresentou como uma das variáveis preditoras de maior importância.

Figura 63. Classificação das dez principais variáveis determinantes do micronaire pelo método *Random Forest* considerando-se os dados do ciclo total do algodoeiro.

Figura 64. Classificação das dez principais variáveis determinantes do micronaire pelo método *XGBoosting* considerando-se os dados do ciclo total do algodoeiro.

Figura 65. Classificação das dez principais variáveis determinantes do micronaire pelo método *Support Vector Machine* considerando-se os dados do ciclo total do algodoeiro.

Figura 66. Diagrama de Venn para a indicação da concordância entre os modelos de mineração de dados empregados na classificação das variáveis preditoras determinantes do microneurite, considerando-se o ciclo total do algodoeiro.

- **Etapa de Treinamento**

As Figuras 67, 68 e 69 apresentam a relação entre os dados observados de micronaire e aqueles estimados, respectivamente, pelas técnicas do *Random Forest*, *XGBoosting* e *Support Vector Machine*, durante a fase de treinamento dos modelos.

Figura 67. Relação entre o micronaire observado e predito pelo algoritmo gerado pela técnica *Random Forest*, na fase de treinamento, considerando-se os dados do ciclo total do algodoeiro.

Figura 68. Relação entre o micronaire observado e predito pelo algoritmo gerado pela técnica *XGBoosting*, na fase de treinamento, considerando-se os dados do ciclo total do algodoeiro.

Figura 69. Relação entre o micronaire observado e predito pelo algoritmo gerado pela técnica *Support Vector Machine*, na fase de treinamento, considerando-se os dados do ciclo total do algodoeiro.

Pelos resultados da fase de treinamento, foi possível observar que as três técnicas utilizadas apresentaram índices estatísticos de avaliação praticamente iguais ou muito próximos, sendo assim, não houve diferenças expressivas de desempenho entre os três modelos. Em relação aos índices estatísticos R^2 , R e D , quanto mais próximos eles se apresentarem da unidade, melhor é o desempenho do modelo. Já para os erros (RMSE e EAM), quanto mais próximos de zero eles forem, melhor o desempenho do modelo. Segundo Camargo e Sentelhas (1997), valores de “C” maiores do que 0,85 indicam modelos de desempenho ótimo; entre 0,76 e 0,85 indicam modelos de desempenho muito bom; entre 0,66 a 0,75 são indicativos de bons modelos; valores entre 0,61 e 0,65 indicam modelos medianos; valores entre 0,51 e 0,60 indicam modelos sofríveis; entre 0,41 e 0,50 indicam modelos ruins; e valores menores ou iguais a 0,40 classificam os modelos como péssimos.

Embora os índices “C” encontrados sejam indicativos de modelos sofríveis, segundo a classificação supracitada, quando se considera todos os índices estatísticos obtidos, pode-se considerar que os três modelos gerados pelas técnicas RF, XGBoosting e *Support Vector Machine* tiveram desempenho satisfatório para determinação das variáveis preditoras mais relevantes para o micronaire.

- **Etapa de Teste**

As figuras 70, 71 e 72 apresentam a relação entre os dados observado de micronaire e os estimados na fase de teste para as técnicas *Random Forest*, *XGBoosting* e *Support Vector Machine*, respectivamente.

Figura 70. Relação entre o micronaire observado e predito pelo algoritmo gerado pela técnica *Random Forest*, na fase de teste, considerando-se os dados do ciclo total do algodoeiro.

Figura 71. Relação entre o micronaire observado e predito pelo algoritmo gerado pela técnica *XGBoosting*, na fase de teste, considerando-se os dados do ciclo total do algodoeiro.

Figura 72. Relação entre o micronaire observado e predito pelo algoritmo gerado pela técnica *Support Vector Machine*, na fase de teste, considerando-se os dados do ciclo total do algodoeiro.

Assim como para a fase de treinamento, na fase de teste houve concordância dos índices estatísticos entre as três técnicas utilizadas. Além disso, observou-se desempenho similar dos modelos entre as duas fases.

Comparando-se as etapas de treinamento e de teste foi possível observar os seguintes resultados: os valores de R^2 variaram entre 0,46 e 0,51 na etapa de treinamento e na etapa de teste esse índice praticamente não se alterou e se manteve em 0,43; os valores de R variaram de 0,68 a 0,71 na etapa de treinamento e se mantiveram estáveis e igual a 0,66 na etapa de teste; os valores de RMSE variaram de 0,26 a 0,28 na etapa de treinamento e se mantiveram em 0,30 na etapa de teste; os valores do índice de Willmott (D) variaram entre 0,76 e 0,82 na etapa de treinamento e na etapa de teste variaram entre 0,74 e 0,78; e, por fim, os valores do índice de confiança C variaram de 0,51 a 0,56 na etapa de treinamento e na etapa de teste variaram entre 0,48 e 0,51. Dessa forma, o fato de as duas etapas (treinamento e teste) terem apresentado resultados tão parecidos nos modelos gerados pelas três técnicas (RF; XGB e SVM), é um indicativo que os modelos são estáveis.

4.3.2.2. Comprimento

- **Identificação das variáveis preditoras**

As Figuras 73, 74 e 75 mostram os resultados obtidos na Fase de identificação das variáveis preditoras mais importantes para os modelos *Random Forest*, *XGBoosting* e *Support Vector Machine*, respectivamente. Já a Figura 76 apresenta o Diagrama de Venn, o qual apresenta as dez variáveis

selecionadas pelos modelos empregados e a concordância entre eles para selecionar as variáveis que mais afetaram o comprimento.

Pelo diagrama de Venn (Figura 76), é possível observar que houve concordância entre os modelos de mineração de dados empregados na classificação das variáveis preditoras determinantes do comprimento, em três variáveis meteorológicas: Qg; Tmin20; e U2.

Em relação ao comprimento da fibra, Soares (2015) não encontrou correlações significativas para nenhuma das cultivares testadas para a frequência de temperatura mínima inferior a 20°C, o que não aconteceu neste estudo, já que Tmin20 se apresentou como umas das variáveis meteorológicas mais relevantes para esse índice de qualidade da fibra de algodão. Raphael (2019) também não encontrou influência dos efeitos do sombreamento nos valores de comprimento da fibra de algodão, diferentemente dos resultados deste trabalho, no qual a radiação solar (Qg) foi uma das variáveis meteorológicas que mais afetou o comprimento. Por outro lado, a velocidade do vento, embora tenha ficado entre as variáveis preditoras mais importantes para o comprimento, é pouco considerada nos estudos envolvendo qualidade da fibra de algodão.

Figura 73. Classificação das dez principais variáveis determinantes do comprimento pelo método *Random Forest* considerando-se os dados do ciclo total do algodoeiro.

Figura 74. Classificação das dez principais variáveis determinantes do comprimento pelo método *XGBoosting* considerando-se os dados do ciclo total do algodoeiro.

Figura 75. Classificação das dez principais variáveis determinantes do comprimento pelo método *Support Vector Machine* considerando-se os dados do ciclo total do algodoeiro.

Figura 76. Diagrama de Venn para a indicação da concordância entre os modelos de mineração de dados empregados na classificação das variáveis preditoras determinantes do comprimento, considerando-se o ciclo total do algodoeiro.

- **Etapa de Treinamento**

As Figuras 77, 78 e 79 mostram os resultados obtidos para o comprimento na fase de treinamento para as técnicas *Random Forest*, *XGBoosting* e *Support Vector Machine*, respectivamente.

Figura 77. Relação entre o comprimento observado e predito pelo algoritmo gerado pela técnica *Random Forest*, na fase de treinamento, considerando-se os dados do ciclo total do algodoeiro.

Figura 78. Relação entre o comprimento observado e predito pelo algoritmo gerado pela técnica *XGBoosting*, na fase de treinamento, considerando-se os dados do ciclo total do algodoeiro.

Figura 79. Relação entre o comprimento observado e predito pelo algoritmo gerado pela técnica *Support Vector Machine*, na fase de treinamento, considerando-se os dados do ciclo total do algodoeiro.

Para a determinação do comprimento na fase de treinamento, considerando-se o ciclo total do algodoeiro, pode-se observar que as três técnicas utilizadas apresentaram índices estatísticos de avaliação praticamente iguais ou muito próximos, sendo assim, não houve diferenças significativas de desempenho

entre os três modelos. Segundo Camargo e Sentelhas (1997), valores de “C” menores ou iguais a 0,40 classificam os modelos como péssimos. Embora os índices “C” encontrados classifiquem os modelos como péssimos, pode-se deduzir que, como os três modelos apresentaram resultados bem próximos, o problema não está na técnica utilizada e sim, no índice de qualidade avaliado (comprimento), no sentido de que as variáveis meteorológicas não influenciaram fortemente seus valores.

- **Fase de Teste**

As Figuras 80, 81 e 82 mostram os resultados obtidos na Fase de teste para as técnicas *Random Forest*, *XGBoosting* e *Support Vector Machine*, respectivamente.

Figura 80. Relação entre o comprimento observado e predito pelo algoritmo gerado pela técnica *Random Forest*, na fase de teste, considerando-se os dados do ciclo total do algodoeiro.

Figura 81. Relação entre o comprimento observado e predito pelo algoritmo gerado pela técnica *XGBoosting*, na fase de teste, considerando-se os dados do ciclo total do algodoeiro.

Figura 82. Relação entre o comprimento observado e predito pelo algoritmo gerado pela técnica *Support Vector Machine*, na fase de teste, considerando-se os dados do ciclo total do algodoeiro.

Da mesma forma que os resultados obtidos na etapa de treinamento, os resultados obtidos na etapa de teste, praticamente não variaram entre uma técnica e outra. Assim, houve concordância dos índices estatísticos entre as três técnicas utilizadas, tanto na etapa de treinamento, quanto na etapa de teste.

Comparando-se as etapas de treinamento e de teste foi possível observar os seguintes resultados: os valores de R^2 variaram entre 0,23 e 0,30 na etapa de treinamento e na etapa de teste esse índice praticamente não se alterou e se manteve em 0,17; os valores de R variaram de 0,48 a 0,55 na etapa de

treinamento e se mantiveram estáveis e igual a 0,42 na etapa de teste; os valores de RMSE variaram de 0,77 a 0,81 na etapa de treinamento e se mantiveram em 0,84 na etapa de teste; os valores de EAM variaram entre 0,60 e 0,63 na etapa de treinamento e entre 0,53 e 0,61 na etapa de teste; os valores do índice de Willmott (D) variaram entre 0,56 e 0,68 na etapa de treinamento e na etapa de teste variaram entre 0,53 e 0,61; e, por fim, os valores do índice de confiança C variaram de 0,27 a 0,37 na etapa de treinamento e na etapa de teste variaram entre 0,21 e 0,26. Dessa forma, para o comprimento, a diferença entre os resultados das três técnicas utilizadas, para cada uma das etapas (treinamento e teste) foi pequena, com o modelo *XGBoosting* apresentando um desempenho ligeiramente melhor do que os outros. Entretanto, quando se analisa cada técnica separadamente, as duas etapas (treinamento e teste) não apresentaram valores dos indicadores estatísticos próximos, como o que aconteceu com os modelos determinados para o micronaire, por exemplo. Isso é um indicativo que os modelos gerados para determinação das variáveis preditoras de mais importância para o comprimento não foram satisfatórios, não por conta do desempenho dos modelos (já que eles apresentaram resultados parecidos), mas pelo fato de o comprimento não ser fortemente influenciado pelas variáveis meteorológicas avaliadas.

4.3.2.3. Resistência

- **Identificação das variáveis preditoras**

As Figuras 83, 84 e 85 mostram os resultados obtidos na Fase de identificação das variáveis preditoras mais importantes para a resistência, segundo os modelos *Random Forest*, *XGBoosting* e *Support Vector Machine*, respectivamente. Já a Figura 86 mostra o Diagrama de Venn, o qual apresenta as dez variáveis selecionadas pelos modelos empregados e a concordância entre eles para selecionar as variáveis que mais afetaram a resistência.

Pelo diagrama de Venn (Figura 86), é possível observar que houve concordância entre os modelos de mineração de dados empregados na classificação das variáveis preditoras determinantes da resistência em quatro variáveis meteorológicas: Qg; Tmax32; Tmin20; e Tmin.

Soares (2015) encontrou correlações significativas entre resistência e frequência de temperatura mínima inferior a 20°C para uma das cultivares analisadas em seu estudo, o que condiz com o resultado encontrado no presente trabalho, no qual a Tmin20 está entre as principais variáveis preditoras para a resistência. Para a temperatura mínima (Tmin), o mesmo autor também encontrou correlações significativas para duas das quatro cultivares consideradas. Conforme já citado anteriormente, para a frequência de temperatura máxima superior a 35°C, Soares (2015) não encontrou relação com os índices de qualidade da fibra de algodão. Já no presente estudo, Tmax32 mostrou ser uma das variáveis mais relevantes para a definição da qualidade da fibra do algodoeiro, tendo sido classificada como uma das preditoras mais importantes, até o momento, não somente para a resistência, mas também para o micronaire. Da mesma forma que para os outros índices de qualidade avaliados até o momento

(micronaire e comprimento), Raphael (2019) também não encontrou influência dos efeitos do sombreamento nos valores de resistências da fibra de algodão.

Figura 83. Classificação das dez principais variáveis determinantes da resistência pelo método *Random Forest* considerando-se os dados do ciclo total do algodoeiro.

Figura 84. Classificação das dez principais variáveis determinantes da resistência pelo método *XGBoosting* considerando-se os dados do ciclo total do algodoeiro.

Figura 85. Classificação das dez principais variáveis determinantes da resistência pelo método *Support Vector Machine* considerando-se os dados do ciclo total do algodoeiro.

Figura 86. Diagrama de Venn para a indicação da concordância entre os modelos de mineração de dados empregados na classificação das variáveis preditoras determinantes da resistência, considerando-se o ciclo total do algodoeiro.

- **Fase de Treinamento**

As figuras 87, 88 e 89 mostram os resultados obtidos para a resistência na fase de treinamento para as técnicas *Random Forest*, *XGBoosting* e *Support Vector Machine*, respectivamente.

Figura 87. Relação entre a resistência observada e predita pelo algoritmo gerado pela técnica *Random Forest*, na fase de treinamento, considerando-se os dados do ciclo total do algodoeiro.

Figura 88. Relação entre a resistência observada e predita pelo algoritmo gerado pela técnica *XGBoosting*, na fase de treinamento, considerando-se os dados do ciclo total do algodoeiro.

Figura 89. Relação entre a resistência observada e predita pelo algoritmo gerado pela técnica *Support Vector Machine*, na fase de treinamento, considerando-se os dados do ciclo total do algodoeiro.

Para a determinação da resistência na fase de treinamento, os índices estatísticos de avaliação, entre as técnicas utilizadas, não tiveram diferenças de valores significativas, sendo as maiores diferenças encontradas para o RMSE (variou de 1,21 a 1,35) e para o índice C (variou de 0,54 a 0,65). Entretanto, embora as diferenças tenham sido pequenas, segundo a classificação de Camargo e Sentelhas (1997) para o índice C, os modelos RF e XGB encontram-se dentro da faixa de modelos considerados medianos (0,61 a 0,65) e o SVM encontra-se dentro da faixa de modelos sofríveis (0,51 a 0,60), o que indica que o SVM teve um desempenho inferior aos outros modelos para determinação das variáveis preditoras da resistência.

- **Fase de teste**

As Figuras 90, 91 e 92 mostram os resultados obtidos na Fase de teste para as técnicas *Random Forest*, *XGBoosting* e *Support Vector Machine*, respectivamente.

Figura 90. Relação entre a resistência observada e predita pelo algoritmo gerado pela técnica *Random Forest*, na fase de teste, considerando-se os dados do ciclo total do algodoeiro.

Figura 91. Relação entre a resistência observada e predita pelo algoritmo gerado pela técnica *XGBoosting*, na fase de teste, considerando-se os dados do ciclo total do algodoeiro.

Figura 92. Relação entre a resistência observada e predita pelo algoritmo gerado pela técnica *Support Vector Machine*, na fase de teste, considerando-se os dados do ciclo total do algodoeiro.

Para a determinação da resistência na fase de teste, os índices estatísticos de avaliação, entre as técnicas utilizadas, tiveram pequenas diferenças. Entretanto, diferentemente do que aconteceu na fase de treinamento, segundo a classificação de Camargo e Sentelhas (1997) para o índice C, todos os modelos encontram-se dentro da faixa de modelos sofríveis (0,51 a 0,60).

Comparando-se as etapas de treinamento e de teste foi possível observar os seguintes resultados: os valores de R^2 variaram entre 0,49 e 0,58 na etapa de treinamento e na etapa de teste entre 0,44 e 0,52; os valores de R variaram de 0,70 e 0,76 na etapa de treinamento e entre 0,66 e 0,72 na etapa de teste; os valores de RMSE variaram de 1,21 a 1,35 na etapa de treinamento e entre 1,26 e 1,34 na etapa de teste; os valores de EAM variaram entre 0,93 e 1,02 na etapa de treinamento e entre 0,98 e 1,03 na etapa de teste; os valores do índice de Willmott (D) variaram entre 0,78 e 0,85 na etapa de treinamento e na etapa de teste variaram entre 0,78 e 0,84; e, por fim, os valores do índice de confiança C variaram de 0,54 a 0,65 na etapa de treinamento e na etapa de teste variaram entre 0,54 e 0,60.

Embora as etapas de treinamento e teste tenham apresentado pequenas diferenças em seus resultados, considerando-se os índices estatísticos como um todo, os modelos gerados para determinação das variáveis preditoras de mais importância para a resistência foram satisfatórios, sendo Qg, Tmax32, Tmin20 e Tmin, as variáveis preditoras que mais influenciaram a resistência, já que foram comuns aos três modelos.

4.3.2.4. Uniformidade

- **Identificação das variáveis preditoras**

As Figuras 93, 94 e 95 mostram os resultados obtidos na Fase de identificação das variáveis preditoras mais importantes para a uniformidade, segundo os modelos *Random Forest*, *XGBoosting* e *Support Vector Machine*, respectivamente. Já a Figura 96 mostra o Diagrama de Venn, o qual apresenta as dez variáveis selecionadas pelos modelos empregados e a concordância entre eles para selecionar as variáveis que mais afetaram a uniformidade.

Pelo diagrama de Venn (Figura 96), foi possível observar que houve concordância entre os modelos de mineração de dados empregados na classificação das variáveis preditoras determinantes da uniformidade em cinco variáveis meteorológicas: U2; $\sum$ EXC; P; T_{min20}; e $\sum$ DEF.

Conforme já citado anteriormente, a uniformidade é uma característica de homogeneidade do comprimento das fibras do fardo e, por esse motivo, é uma medida indireta e dependente do comprimento. Assim, há poucos estudos que mostrem a relação entre as variáveis meteorológicas e de balanço hídrico na uniformidade.

Comparando-se os resultados obtidos para o comprimento, em relação às variáveis preditoras de maior importância e comuns às três técnicas utilizadas (RF; XGB; e SVM), foi observado que tanto para o micronaire, quanto para a uniformidade, a T_{min20} esteve entre as variáveis meteorológicas de maior relevância.

Soares (2015) não testou os efeitos das condições meteorológicas na uniformidade, mas para o comprimento, o autor não encontrou correlações significativas para nenhuma das quatro cultivares com relação à frequência de temperatura mínima inferior a 20°C. Entretanto, no presente estudo, tanto o comprimento, quanto a uniformidade (dependente do comprimento), apresentaram relação com T_{min20}.

Lokhande e Reddy (2014) observaram uma redução linear da uniformidade em resposta ao aumento do déficit hídrico. Neste, a variável $\sum$ DEF esteve entre as preditoras de mais relevância para a uniformidade, concordando com os resultados dos autores supracitados.

Figura 93. Classificação das dez principais variáveis determinantes da uniformidade pelo método *Random Forest* considerando-se os dados do ciclo total do algodoeiro.

Figura 94. Classificação das dez principais variáveis determinantes da uniformidade pelo método *XGBoosting* considerando-se os dados do ciclo total do algodoeiro.

Figura 95. Classificação das dez principais variáveis determinantes da uniformidade pelo método *Support Vector Machine* considerando-se os dados do ciclo total do algodoeiro.

Figura 96. Diagrama de Venn para a indicação da concordância entre os modelos de mineração de dados empregados na classificação das variáveis preditoras determinantes da uniformidade, considerando-se o ciclo total do algodoeiro.

- Fase de Treinamento

As Figuras 97, 98 e 99 mostram os resultados obtidos para a resistência na fase de treinamento para as técnicas *Random Forest*, *XGBoosting* e *Support Vector Machine*, respectivamente.

Figura 97. Relação entre a uniformidade observada e predita pelo algoritmo gerado pela técnica *Random Forest*, na fase de treinamento, considerando-se os dados do ciclo total do algodoieiro.

Figura 98. Relação entre a uniformidade observada e predita pelo algoritmo gerado pela técnica *XGBoosting*, na fase de treinamento, considerando-se os dados do ciclo total do algodoieiro.

Figura 99. Relação entre a uniformidade observada e predita pelo algoritmo gerado pela técnica *Support Vector Machine*, na fase de treinamento, considerando-se os dados do ciclo total do algodoeiro.

Para a determinação da uniformidade na fase de treinamento, os índices estatísticos de avaliação, entre as técnicas utilizadas, tiveram pequenas diferenças. Segundo a classificação de Camargo e Sentelhas (1997) para o índice C, os modelos RF e XGB encontram-se dentro da faixa de modelos considerados bons (0,66 a 0,75) e o SVM encontra-se dentro da faixa de modelos medianos (0,61 a 0,65), o que indica que o SVM teve um desempenho inferior aos outros modelos para determinação das variáveis preditoras da uniformidade.

- **Fase de teste**

As Figuras 100, 101 e 102 mostram os resultados obtidos na Fase de teste para as técnicas *Random Forest*, *XGBoosting* e *Support Vector Machine*, respectivamente.

Figura 100. Relação entre a uniformidade observada e predita pelo algoritmo gerado pela técnica *Random Forest*, na fase de teste, considerando-se os dados do ciclo total do algodoeiro.

Figura 101. Relação entre a uniformidade observada e predita pelo algoritmo gerado pela técnica *XGBoosting*, na fase de teste, considerando-se os dados do ciclo total do algodoeiro.

Figura 102. Relação entre a uniformidade observada e predita pelo algoritmo gerado pela técnica *Support Vector Machine*, na fase de teste, considerando-se os dados do ciclo total do algodoeiro

Na fase de teste dos modelos gerados para a uniformidade, os índices estatísticos de avaliação, entre as técnicas utilizadas, não tiveram diferenças de valores significativas. Diferentemente da fase de treinamento, segundo a classificação de Camargo e Sentelhas (1997) para o índice C, na fase de teste, todos os modelos foram classificados como sofríveis, o que mostrou discrepância entre os resultados obtidos pelas duas etapas (Treinamento e Teste).

Comparando-se as etapas de treinamento e de teste foi possível observar os seguintes resultados: os valores de R^2 variaram entre 0,56 e 0,62 na etapa de treinamento e na etapa de teste entre 0,50 e 0,52; os valores de R variaram de 0,75 e 0,79 na etapa de treinamento e se manteve em 0,72 na etapa de teste; os valores de RMSE variaram de 0,84 a 0,92 na etapa de treinamento e se manteve na faixa de 1,0 na etapa de teste; os valores de EAM variaram entre 0,63 e 0,68 na etapa de treinamento e entre 0,73 e 0,76 na etapa de teste; os valores do índice de Willmott (D) variaram entre 0,82 e 0,86 na etapa de treinamento e na etapa de teste variaram entre 0,79 e 0,83; e, por fim, os valores do índice de confiança C variaram de 0,61 a 0,69 na etapa de treinamento e na etapa de teste variaram entre 0,55 e 0,60.

As etapas de treinamento e teste apresentaram resultados discrepantes em relação à classificação dos modelos para o índice C, segundo Camargo e Sentelhas (1997). Entretanto, tomando-se como base todos os índices estatísticos, pode-se considerar que os modelos tiveram desempenho satisfatório e que as variáveis meteorológicas que mais influenciaram a uniformidade foram: U2; $\sum$ EXC; P; T_{min20}; e $\sum$ DEF.

4.3.2.5. Elongação

- **Identificação das variáveis preditoras**

As Figuras 103, 104 e 105 mostram os resultados obtidos na Fase de identificação das variáveis preditoras mais importantes para a elongação, segundo os modelos *Random Forest*, *XGBoosting* e *Support Vector Machine*, respectivamente. Já a Figura 106 mostra o Diagrama de Venn, o qual apresenta as dez variáveis selecionadas pelos modelos empregados e a concordância entre eles para selecionar as variáveis que mais afetaram a uniformidade.

Pelo diagrama de Venn (Figura 106), foi possível observar que houve concordância entre os modelos de mineração de dados empregados na classificação das variáveis preditoras determinantes da elongação em cinco variáveis meteorológicas: Qg; EXC0; Tmax; Tmin18; e P.

Na abordagem geral (Tabela 12), a elongação apresentou correlações com as seguintes variáveis meteorológicas: Qg; Tmax; Tmin; P; Tmed; Tmax30; NDC; e EXC0. Dentre essas, Qg, Tmax, P e EXC0 se apresentaram como as variáveis preditoras de mais importância e comum entre as três técnicas de mineração de dados utilizadas, o que comprova que houve concordância entre os resultados encontrados por essas duas técnicas utilizadas neste estudos (correlação de Pearson/abordagem geral e mineração de dados).

Figura 103. Classificação das dez principais variáveis determinantes da elongação pelo método *Random Forest* considerando-se os dados do ciclo total do algodoeiro.

Figura 104. Classificação das dez principais variáveis determinantes da elongação pelo método *XGBoosting* considerando-se os dados do ciclo total do algodoeiro.

Figura 105. Classificação das dez principais variáveis determinantes da elongação pelo método *Support Vector Machine* considerando-se os dados do ciclo total do algodoeiro.

Figura 106. Diagrama de Venn para a indicação da concordância entre os modelos de mineração de dados empregados na classificação das variáveis preditoras determinantes da elongação, considerando-se o ciclo total do algodoeiro.

- **Fase de Treinamento**

As Figuras 107, 108 e 109 mostram os resultados obtidos para a resistência na fase de treinamento para as técnicas *Random Forest*, *XGBoosting* e *Support Vector Machine*, respectivamente.

Figura 107. Relação entre a elongação observada e predita pelo algoritmo gerado pela técnica *Random Forest*, na fase de treinamento, considerando-se os dados do ciclo total do algodoeiro.

Figura 108. Relação entre a elongação observada e predita pelo algoritmo gerado pela técnica *XGBoosting*, na fase de treinamento, considerando-se os dados do ciclo total do algodoeiro.

Figura 109. Relação entre a elongação observada e predita pelo algoritmo gerado pela técnica *Support Vector Machine*, na fase de treinamento, considerando-se os dados do ciclo total do algodoeiro.

Para a determinação da elongação na fase de treinamento, os índices estatísticos de avaliação, entre as técnicas utilizadas, tiveram pequenas diferenças. Segundo a classificação de Camargo e Sentelhas (1997) para o índice C, os modelos RF e XGB encontram-se dentro da faixa de modelos considerados bons (0,66 a 0,75) e o SVM encontra-se dentro da faixa de modelos medianos (0,61 a 0,65), o que indica que o SVM teve um desempenho inferior aos outros modelos para determinação das variáveis preditoras da elongação.

- **Fase de teste**

As Figuras 110, 111 e 112 mostram os resultados obtidos na Fase de teste para as técnicas *Random Forest*, *XGBoosting* e *Support Vector Machine*, respectivamente.

Figura 110. Relação entre a elongação observada e predita pelo algoritmo gerado pela técnica *Random Forest*, na fase de teste, considerando-se os dados do ciclo total do algodoeiro.

Figura 111. Relação entre a elongação observada e predita pelo algoritmo gerado pela técnica *XGBoosting*, na fase de teste, considerando-se os dados do ciclo total do algodoeiro.

Figura 112. Relação entre a elongação observada e predita pelo algoritmo gerado pela técnica *Support Vector Machine*, na fase de teste, considerando-se os dados do ciclo total do algodoeiro.

Para a elongação, na fase de teste, os índices estatísticos de avaliação, entre as técnicas utilizadas, não tiveram diferenças de valores significativas. Segundo a classificação de Camargo e Sentelhas (1997) para o índice C, da mesma forma que ocorreu na etapa de treinamento, na etapa de teste os modelos RF e XGB encontram-se dentro da faixa de modelos considerados bons (0,66 a 0,75) e o SVM encontra-se dentro da faixa de modelos medianos (0,61 a 0,65), o que indica que o SVM teve um desempenho inferior aos outros modelos para determinação das variáveis preditoras da elongação. Dessa forma, houve conformidade entre os resultados encontrados para as duas etapas (treinamento e teste).

Comparando-se as etapas de treinamento e de teste foi possível observar os seguintes resultados: os valores de R^2 variaram entre 0,55 e 0,65 na etapa de treinamento e na etapa de teste entre 0,55 e 0,60; os valores de R variaram de 0,74 e 0,80 na etapa de treinamento e entre 0,78 na etapa de teste; os valores de RMSE variaram de 0,52 a 0,60 na etapa de treinamento e entre 0,57 e 0,61 na etapa de teste; os valores de EAM variaram entre 0,39 e 0,43 na etapa de treinamento e entre 0,44 e 0,46 na etapa de teste; os valores do índice de Willmott (D) variaram entre 0,82 e 0,88 na etapa de treinamento e na etapa de teste variaram entre 0,82 e 0,87; e, por fim, os valores do índice de confiança C variaram de 0,61 a 0,71 na etapa de treinamento e na etapa de teste variaram entre 0,61 e 0,67.

4.3.2.6. Índice de fibras curtas (SFI)

- **Identificação das variáveis preditoras**

As figuras 113, 114 e 115 mostram os resultados obtidos na Fase de identificação das variáveis preditoras mais importantes para o SFI, segundo os modelos *Random Forest*, *XGBoosting* e *Support Vector Machine*, respectivamente. Já a Figura 116 mostra o Diagrama de Venn, o qual apresenta as dez variáveis selecionadas pelos modelos empregados e a concordância entre eles para selecionar as variáveis que mais afetaram o SFI.

Pelo diagrama de Venn (Figura 116), foi possível observar que houve concordância entre os modelos de mineração de dados empregados na classificação das variáveis preditoras determinantes da elongação em três variáveis meteorológicas: UR; Tmax; e P.

O índice de fibras curtas (SFI) apresentou, na abordagem geral (Tabela 12), correlações ($P < 0,01$) com três variáveis meteorológicas: UR; P; e $\sum EXC$. Já nas técnicas utilizadas na abordagem de mineração de dados, as variáveis preditoras de maior importância para o SFI foram UR, Tmax e P. Assim, pode-se concluir que UR e Tmax foram concordantes entre essas duas abordagens consideradas neste estudo e, além do mais, a P é uma variável meteorológica que tem relação direta com $\sum EXC$, o que também explica o aparecimento desta última como umas das variáveis preditoras de mais relevância para a elongação.

Figura 113. Classificação das dez principais variáveis determinantes do SFI pelo método *Random Forest* considerando-se os dados do ciclo total do algodoeiro.

Figura 114. Classificação das dez principais variáveis determinantes do SFI pelo método *XGBoosting* considerando-se os dados do ciclo total do algodoeiro.

Figura 115. Classificação das dez principais variáveis determinantes do SFI pelo método *Support Vector Machine* considerando-se os dados do ciclo total do algodoeiro.

Figura 116. Diagrama de Venn para a indicação da concordância entre os modelos de mineração de dados empregados na classificação das variáveis preditoras determinantes do SFI, considerando-se o ciclo total do algodoeiro.

- Fase de Treinamento

As Figuras 117, 118 e 119 mostram os resultados obtidos para o SFI na fase de treinamento para as técnicas *Random Forest*, *XGBoosting* e *Support Vector Machine*, respectivamente.

Figura 117. Relação entre o índice de fibras curtas observado e predito pelo algoritmo gerado pela técnica *Random Forest*, na fase de treinamento, considerando-se os dados do ciclo total do algodoeiro.

Figura 118. Relação entre o índice de fibras curtas observado e predito pelo algoritmo gerado pela técnica *XGBoosting*, na fase de treinamento, considerando-se os dados do ciclo total do algodoeiro.

Figura 119. Relação entre o índice de fibras curtas observado e predito pelo algoritmo gerado pela técnica *Support Vector Machine*, na fase de treinamento, considerando-se os dados do ciclo total do algodoeiro.

Para a determinação do SFI na fase de treinamento, os índices estatísticos de avaliação, entre as técnicas utilizadas, tiveram pequenas diferenças. Segundo a classificação de Camargo e Sentelhas (1997) para o índice C, os modelos RF e XGB encontram-se dentro da faixa de modelos considerados medianos (0,61 a 0,65) e o SVM encontra-se dentro da faixa de modelos ruins (0,41 a 0,50), o que indica que o SVM teve um desempenho inferior aos outros modelos para determinação das variáveis preditoras do SFI.

- **Fase de Teste**

As Figuras 120, 121 e 122 mostram os resultados obtidos na Fase de teste para as técnicas *Random Forest*, *XGBoosting* e *Support Vector Machine*, respectivamente.

Figura 120. Relação entre o índice de fibras curtas observado e predito pelo algoritmo gerado pela técnica *Random Forest*, na fase de teste, considerando-se os dados do ciclo total do algodoeiro.

Figura 121. Relação entre o índice de fibras curtas observado e predito pelo algoritmo gerado pela técnica *XGBoosting*, na fase de teste, considerando-se os dados do ciclo total do algodoeiro.

Figura 122. Relação entre o índice de fibras curtas observado e predito pelo algoritmo gerado pela técnica *Support Vector Machine*, na fase de teste, considerando-se os dados do ciclo total do algodoeiro.

Para a fase de teste, os índices estatísticos de avaliação, entre as técnicas utilizadas, não tiveram pequenas diferenças. Segundo a classificação de Camargo e Sentelhas (1997) para o índice C, diferentemente do que aconteceu na etapa de treinamento, todos os modelos na etapa de teste foram classificados como ruins (0,41 a 0,50). Na etapa de treinamento, o SVM teve desempenho inferior em relação aos modelos RF e XGB e na etapa de testes não houve diferença entre os três modelos, o que mostra que houve discrepância dos resultados entre o desempenho dos modelos nas duas etapas (treinamento e teste).

Comparando-se as etapas de treinamento e de teste foi possível observar os seguintes resultados: os valores de R^2 variaram entre 0,44 e 0,51 na etapa de treinamento e na etapa de teste se manteve em 0,41; os valores de R variaram de 0,66 a 0,71 na etapa de treinamento e se manteve em 0,64 na etapa de teste; os valores de RMSE variaram de 0,77 a 0,84 na etapa de treinamento e entre 0,95 e 0,98 na etapa de teste; os valores de EAM variaram entre 0,56 e 0,60 na etapa de treinamento e entre 0,70 e 0,78 na etapa de teste; os valores do índice de Willmott (D) variaram entre 0,72 e 0,82 na etapa de treinamento e na etapa de teste variaram entre 0,67 e 0,76; e, por fim, os valores do índice de confiança C variaram de 0,47 a 0,58 na etapa de treinamento e na etapa de teste variaram entre 0,43 e 0,49.

- **Análise geral**

A Tabela 22 apresenta a análise geral dos modelos, considerando-se os valores do índice estatístico C. Pela Tabela 22 é possível observar que o melhor desempenho dos modelos foi para a

elongação e o pior para o Comprimento. Além, disso, em geral, o modelo XGB foi ligeiramente melhor do que o RF, enquanto que o modelo SVM foi o pior para todas as variáveis.

Tabela 22. Valores do índice estatístico C para cada modelo utilizado (RF; XGB; e SVM) e referentes aos índices de qualidade da fibra avaliados (micronaire; comprimento; resistência; uniformidade; elongação; e SFI) nas etapas de treinamento e teste.

Modelo	Valores do índice estatístico C					
	Etapa de treinamento					
	Micronaire	Comprimento	Resistência	Uniformidade	Elongação	SFI
RF	0,56	0,33	0,64	0,66	0,70	0,55
XGB	0,58	0,37	0,65	0,69	0,71	0,58
SVM	0,51	0,27	0,54	0,61	0,61	0,48
Etapa de teste						
RF	0,51	0,25	0,54	0,60	0,67	0,49
XGB	0,51	0,26	0,52	0,59	0,67	0,49
SVM	0,48	0,21	0,49	0,55	0,61	0,43

Hammer et al. (2020) utilizou os modelos RF, XGB, e SVM para modelagem da produtividade da cultura da cana de açúcar. O autor concluiu que os três modelos gerados por meio do emprego de técnicas de MD tiveram desempenho similar entre si, o que não aconteceu neste estudo, no qual o modelo XGB foi ligeiramente melhor do que o RF e o SVM foi o pior entre os três para todos os índices de qualidade da fibra de algodão. Além disso, os melhores valores de índice C encontrados por Hammer et al. (2020) em cada uma das técnicas empregadas foram: 0,56 para *Random Forest*, 0,56 para *XGBoosting*; e 0,57 para *Support Vector Machine*. Já no presente estudo, os maiores valores de C encontrados para cada uma das técnicas foram: 0,70 para *Random Forest*, 0,71 para *XGBoosting*; e 0,61 para *Support Vector Machine*. Dessa forma, por mais que se tratem de culturas e variáveis diferentes, os valores de C encontrados neste trabalho foram superiores aos encontrados por Hammer et al. (2020) em cada uma das técnicas utilizadas de mineração de dados.

A Tabela 23 apresenta as variáveis preditoras que foram comuns aos três modelos (RF, XGB e RF) para cada índice de qualidade da fibra de algodão. É possível observar que as variáveis meteorológicas que mais estiveram presentes como preditoras entre os índices de qualidade da fibra de algodão foram Qg, Tmin20 e P. A Qg foi preditora para quatro índices de qualidade (micronaire, comprimento, resistência e elongação). A Tmin20 também em quatro (micronaire, comprimento, resistência e uniformidade) e P em três (uniformidade, elongação e SFI).

Tabela 23. Variáveis preditoras comuns aos três modelos de mineração de dados (RF, XGB e SVM) para cada um dos índices de qualidade da fibra de algodão avaliados.

Micronaire	Comprimento	Resistência	Uniformidade	Elongação	SFI
Tmax32	Qg	Qg	U2	Qg	UR
Tmin20	Tmin20	Tmax32	$\sum$ EXC	EXC0	Tmax
Qg	U2	Tmin20	$\sum$ DEF	Tmax	P
		Tmin	Tmin20	Tmin18	
			P	P	

5. CONCLUSÕES

Os resultados obtidos no presente estudo, por meio de técnicas estatísticas clássicas e de mineração de dados, permitiram concluir que:

- a) O atraso na semeadura da cultura do algodão nas regiões analisadas promove decréscimo do índice micronaire, diminuindo a qualidade da fibra do algodoeiro. Algumas das variáveis meteorológicas e do balanço hídrico apresentam relação diretamente proporcional com esse decréscimo do índice micronaire (T_{min} , P , $T_{média}$, $\sum EXC$, NDC , $EXC0$ e ARM/CAD), enquanto outras apresentam relação inversamente proporcional (AT , $\sum DEF$, T_{min18} , T_{min20} e $DEF0$);
- b) Considerar o ciclo total da cultura do algodoeiro permite obter melhores correlações estatísticas entre as variáveis meteorológicas e de qualidade da fibra do algodão;
- c) Pela técnica de Mineração de Dados constata-se que Radiação Solar, temperatura noturna abaixo de $20^{\circ}C$ e precipitação são as variáveis mais importantes para a definição da qualidade tecnológica da fibra de algodão;
- d) Os modelos gerados por meio do emprego das técnicas de mineração de dados têm melhor desempenho na estimação das variáveis de qualidade da fibra do algodoeiro do que os modelos baseados em análises de correlação, a nível de abordagem geral (incluindo dados de todas as cultivares) e a nível de agrupamento;
- e) O modelo de melhor desempenho é aquele empregado na estimação da elongação, enquanto o pior é para o comprimento da fibra. O modelo XGB tem desempenho ligeiramente superior do que os modelos RF e SVM, sendo este último o de pior performance;
- f) Para a fase de treinamento, o modelo XGB é o que apresenta o melhor desempenho para a estimação de todos os índices indicadores de qualidade da fibra de algodão.

REFERÊNCIAS

- ABRAPA (ASSOCIAÇÃO BRASILEIRA DOS PRODUTORES DE ALGODÃO). **Paper Presidência Abrapa – Principais indicadores da cotonicultura brasileira e mundial**. Disponível em: <<http://abapa.com.br/wp-content/uploads/2020/01/Paper-Presidência-Principais-Indicadores-Cotonicultura-15jan2020vsfi....pdf>>. Acesso em: 20 abr. 2020.
- ALLEN, R. D.; ALEMAN, L. Abiotic stress and cotton fiber development. In: OOSTERHUIS, D. M. (Ed.). **Stress physiology in cotton**. Cordova, Tennessee: Number Seven: The Cotton Foundation Reference Book Series, 2011. p. 149–160.
- ALLEN, R. G. et al. Crop evapotranspiration: Guidelines for computing crop requirements. **Irrigation and Drainage Paper No. 56, FAO**, 1998.
- ALMEIDA, F. de A. C. et al. Desenvolvimento e avaliação de descaroçador para o beneficiamento do algodão. **Revista Brasileira de Engenharia Agrícola e Ambiental**, Campina Grande, PB, v.15, n.6, p.607-614, 2011.
- ALVARES, C. A. et al. Köppen's climate classification map for Brazil. **Meteorologische Zeitschrift**, v.22, n.6, p.711-728, 2013.
- BACHELIER, B. **Contribution à l'étude de la variabilité et du déterminisme génétique de la teneur en fragments de coque de la fibre de coton. Premières applications pratiques en sélection chez Gossypium hirsutum L.** École Nationale Supérieure Agronomique de Rennes (ENSAR), 1998.
- BACHELIER, B.; GOURLOT, J. P. A fibra de algodão. In: **Safra 2018- Manual de Qualidade da Fibra da AMPA**. Cuiabá, MT: IMAmt- Instituto Matogrossense do Algodão, p. 28–153, 2018.
- BANGE, M. P. et al. A method to estimate the effects of temperature on cotton micronaire. **Journal of Cotton Science**, v.14, n.14, p. 164-172, 2010.
- BANGE, M. P.; MILROY, S. P.; THONGBAI, P. Growth and yield of cotton in response to waterlogging. **Field Crops Research**, v.88, p.129-142, 2004.
- BARBOSA, J. J.; PEREIRA, T. M.; OLIVEIRA, F. L. P. de. Uma proposta para identificação de outliers multivariados. **Ciência e Natura**, v.40, e40, 2018.
- BARROS, F. M. M. **Um sistema de recomendação de páginas web sobre a cultura da cana de açúcar**. 2013. (Dissertação de Mestrado). Universidade Estadual de Campinas, Campinas, 2013.
- BATTISTI, R. **Épocas de semeadura da cultura da soja com base no risco climático e na rentabilidade líquida para as principais regiões produtoras do Brasil**. 2013. (Dissertação de Mestrado). Escola Superior de Agricultura “Luiz de Queiroz” - Universidade de São Paulo, Piracicaba, 2013.
- BELTRÃO, N. E. de M. Fisiologia da produção do algodoeiro. In: **Circular Técnica 94**. Campina Grande, PB: Embrapa- CNPA, 2006. p. 8.
- BELTRÃO, N. E. de M.; AZEVEDO, D. M. P. Fitologia do algodoeiro herbáceo (Sistemática, organografia e anatomia). In: **O agronegócio do algodão no Brasil**. 2. ed. Brasília, DF: Embrapa Informação Tecnológica, 2008. p. 181–218.
- BELTRÃO, N. E. M.; AZEVEDO, D. M. **Defasagem entre as produtividades real e potencial em algodoeiro herbáceo: limitações morfológicas, fisiológicas e ambientais**. Campina Grande, PB: Embrapa- CNPA, 1993.
- BERRY, M. J.; LINOFF, G. Data Mining Techniques: For Marketing, Sales, and Customer Support. **John Wiley & Sons, Inc.**, 1997.

- BEZERRA, B. G. et al. Evapotranspiration and crop coefficient for sprinkler-irrigated cotton crop in Apodi Plateau semiarid lands of Brazil. **Agricultural Water Management**, v.107, p.86-93, 2012.
- BONESSO, D. **Estimação dos parâmetros do Kernel em um classificador SVM na classificação de imagens hiperespectrais em uma abordagem multiclasse**. 2013. (Dissertação de Mestrado). Universidade Federal do Rio Grande do Sul, Porto Alegre, 2013.
- BORÉM, A.; FREIRE, E. C. **Algodão: do plantio à colheita**. 1. ed. Viçosa, MG: UFV, 2014.
- BRAGA, A. P.; LUDERMIR, T. B.; CARVALHO, A. C. P. L. F. **Redes neurais artificiais: teoria e aplicações**. 2. ed. Rio de Janeiro: LTC, 2007.
- BRUCE, P.; BRUCE, A. **Estatística Prática para Cientistas de Dados: 50 conceitos essenciais**. 1. ed. Rio de Janeiro: Alta Books, 2019.
- CÂMARA, G. M. de S.; CHIAVEGATO, E. J. Algodão: Origem, plantas e perspectivas para a cultura. In: **O agronegócio das plantas oleaginosas: algodão, amendoim, girassol e mamona**. Piracicaba, SP: ESALQ, 2001. p. 51–73.
- CAMARGO, A. P.; SENTELHAS, P. C. Avaliação do desempenho de diferentes métodos de estimativas da evapotranspiração potencial no Estado de São Paulo, Brasil. **Revista Brasileira de Agrometeorologia**, v.5, n.1, p.89-97, 1997.
- CAMILO, C. O.; SILVA, J. C. **Mineração de dados: conceitos, tarefas, métodos e ferramentas: relatório técnico**. Goiânia: Universidade Federal de Goiás, Instituto de Informática, 2009.
- CHEN, B. L. et al. Effect of shading on yield, fiber quality and physiological characteristics of cotton subtending leaves on different fruiting positions. **Photosynthetica**, v.55, p. 240-250, 2017.
- CHEN, J. et al. Effects of different planting dates and low light on cotton fibre length formation. **Acta Physiologiae Plantarum**, v.36, p.2581-2595, 2014.
- CHIAVEGATO, E. J.; SALVATIERRA, D. K.; GOTTARDO, L. C. B. **Agrometeorologia dos cultivos- O fator meteorológico na produção agrícola- Algodão**. 1. ed. Brasília, DF: INMET, 2009.
- CONAB. COMPANHIA NACIONAL DE ABASTECIMENTO. **Planilhas de custo de produção. Séries históricas**. Disponível em: <<https://www.conab.gov.br/info-agro/custos-de-producao/planilhas-de-custo-de-producao/itemlist/category/414-planilhas-de-custos-de-producao-series-historicas>>. Acesso em: 20 abr. 2020.
- CORREA, F. A. A fibra e os subprodutos. In: **Cultura e adubação do algodoeiro**. São Paulo: Instituto Brasileiro De Potassa, 1965. p. 509–540.
- CÔRTEZ, S. DA C.; PORCARO, R. M.; LIFSCHITZ, S. **Mineração de dados- Funcionalidades, Técnicas e Abordagens**. Maio/2002 ed. Rio de Janeiro: PUC RIO Inf, 2002.
- DA FONSECA, E. L.; FORMAGGIO, A. R.; PONZONI, F. J. Estimativa da disponibilidade de forragem do bioma Campos Sulinos a partir de dados radiométricos orbitais: Parametrização do submodelo espectral. **Ciência Rural**, v.37, n.6, p.1668-1674, 2007.
- DAVIS, R. F.; EARL, H. J.; TIMPER, P. Effect of simultaneous water deficit stress and *Meloidogyne incognita* infection on cotton yield and fiber quality. **Journal of Nematology**, v.46, n.2, p.108-118, 2014.
- DE LIMA, J. J. Classificação do algodão em pluma. In: **Safra 2018- Manual de Qualidade da Fibra da AMPA**. Cuiabá, MT: IMAmt- Instituto Matogrossense do Algodão, p. 58–115, 2018.
- DI GIROLAMO NETO, C. **Desenvolvimento e avaliação de modelos de alerta para a ferrugem do cafeeiro**. 2013. (Dissertação de Mestrado). Universidade Estadual de Campinas, Campinas, 2013.

- DIAS, H. B. **Intercomparação de modelos de simulação da cana-de-açúcar e seu uso na avaliação da quebra de produtividade e dos impactos da irrigação em diferentes regiões do Brasil**. 2016. (Dissertação de Mestrado). Escola Superior de Agricultura “Luiz de Queiroz” - Universidade de São Paulo, Piracicaba, 2016.
- ECHER, F. R.; ROSOLEM, C. A. Efeitos do estresse luminoso na fisiologia do algodoeiro. In: ECHER, F. R. (Ed.). **O algodoeiro e os estresses abióticos: temperatura, luz, água e nutrientes**. Cuiabá, MT: IMAmt- Instituto Matogrossense do Algodão, p. 31–41, 2014.
- ECHER, F. R.; ROSOLEM, C. A. Cotton yield and fiber quality affected by row spacing and shading at different growth stages. **European Journal of Agronomy**, v.65, p.18-26, 2015.
- EMBRAPA. **Cultivo do algodoeiro irrigado**. Brasília, DF: Embrapa Algodão, Sistemas de Produção, v.3, 2006.
- EMBRAPA. **Série Desafios do Agronegócio- Algodão**. Disponível em: <<https://ainfo.cnptia.embrapa.br/digital/bitstream/item/198192/1/SerieDesafiosAgronegocioBrasileiroNT3Algodao.pdf>>. Acesso em: 20 abr. 2020.
- ESCHER, F.; ROSOLEM, C. A.; RAPHAEL, J. P. A. Desenvolvimento da planta e qualidade da fibra. In: **Safra 2018- Manual de Qualidade da Fibra da AMPA**. Cuiabá, MT: IMAmt- Instituto Matogrossense do Algodão, p. 206–237, 2018.
- FAO (FOOD AND AGRICULTURE ORGANIZATION). **Statistics Division (2016)**. Disponível em: <<http://www.fao.org/economic/ess>>. Acesso em: 20 abr. 2020.
- FERRARO, D. O.; RIVERO, D. E.; GHERSA, C. M. An analysis of the factors that influence sugarcane yield in Northern Argentina using classification and regression trees. **Field Crops Research**, v.112, p.149-157, 2009.
- FOLEY, J. A. et al. Solutions for a cultivated planet. **Nature**, v.478, p.337-342, 2011.
- FONSECA, R. G.; FARIAS, J. C. S. Resultados de ensaio HVI e suas interpretações (ASTM D-4605). In: **Circular Técnica 66**. [s.l.] EMBRAPA, 2002.
- FREDDI, O. S.; STOQUERO, A.; PERES, F. S. C. Saturação do Solo por Bases na Produtividade e Qualidade de Fibra do Algodoeiro. **Nativa**, v.2, n.1, p.6-12, 2014.
- FREELAND, T. B. Agrometeorology and cotton productions. In: **Agricultural Research Service**. Washington, DC: United States Department of Agriculture, p. 17, 2006.
- FREIRE, E. C. Algodão no cerrado do Brasil. In: **ABRAPA**. 3. ed. Brasília: Gráfica e Editora Positiva, p. 301–332, 2015.
- FUZZATO, M. G. et al. **Boletim 200**. Campinas, SP: Instituto Agrônomo de Campinas, 2014.
- GIPSON, J. R.; JOHAM, H. E. Influence of Night Temperature on Growth and Development of Cotton (*Gossypium hirsutum* L.). I. Fruiting and Boll Development 1. **Agronomy Journal**, v.60, p.292-295, 1968.
- GONIAS, E. D.; OOSTERHUIS, D. M.; BIBI, A. C. Cotton radiation use efficiency response to plant growth regulators. **Journal of Agricultural Science**, v.150, p.595-602, 2012.
- HAIR, J. F. et al. **Análise multivariada de dados**. 5. ed. Porto Alegre: Bookman, 2005.
- HAKKE, K. Fiber development and HVI Quality. **Physiology Today**, v. 2, n. 2, 1990.
- HAKKE, K. et al. Causes of high and low micronaire. **Cotton Physiology Today**, v. 1, n. 12, p. 4, 1990.
- HAMMER, R. G. **Modelagem da produtividade da cultura da cana de açúcar por meio do uso de técnicas de mineração de dados**. 2016. (Dissertação de Mestrado). Escola Superior de Agricultura “Luiz de Queiroz”- Universidade de São Paulo, Piracicaba, 2016.
- HAMMER, R. G.; SENTELHAS, P. C.; MARIANO, J. C. Q. Sugarcane Yield Prediction Through Data Mining and Crop Simulation Models. **Sugar Tech**, v.22, p.216-225, 2020.

- HAN, J.; KAMBER, M.; PEI, J. **Data Mining: Concepts and Techniques**. San Francisco: Morgan Kaufmann Publishers, 2011.
- HASTIE, T.; TIBSHIRANI, R.; FRIEDMAN, J. **The elements of statistical learning: prediction, inference and data mining**. Berlin: Springer Verlag, 2009.
- HAYKIN, S. **Neural networks: a comprehensive foundation**. New Jersey: Prentice Hall, 1999.
- HOOGENBOOM, G. Contribution of agrometeorology to the simulation of crop production and its applications. **Agricultural and Forest Meteorology**, v.103, p.137-157, 2000.
- IPCC. The Ocean and Cryosphere in a Changing Climate. **In press**, 2019.
- JAMES, G. et al. **An introduction to statistical learning: with applications in R**. London: Springer, 2013.
- KOHEL, R. ; LEWIS, C. . **Cotton**. Madison: American Society of Agronomy, 1993.
- KUHN, M. Building predictive models in R using the caret package. **Journal of Statistical Software**, v.28, n.5, p.1-25, 2008.
- LIAKATAS, A.; ROUSSOPOULOS, D.; WHITTINGTON, W. J. Controlled-temperature effects on cotton yield and fibre properties. **Journal of Agricultural Science**, v.130, n.4, p.463-471, 1998.
- LIMA, C. A. M. **Comitê de máquinas: uma abordagem unificada empregando máquina de vetores-suporte**. 2004. (Tese de Doutorado). Universidade Estadual de Campinas, Campinas, 2004.
- LIU, J. et al. Effect of late planting and shading on cotton yield and fiber quality formation. **Field Crops Research**, v.183, p.1-13, 2015.
- LOKHANDE, S.; REDDY, K. R. Quantifying temperature effects on cotton reproductive efficiency and fiber quality. **Agronomy Journal**, v.106, n.4, p.1275-1282, 2014.
- LORENA, A. C.; CARVALHO, A. C. P. L. F. Introdução às máquinas de vetores suporte (support vector machines). In: **Relatórios Técnicos**, 192. São Carlos: USP, ICMC, p. 58, 2003.
- LV, F. et al. Effect of shading on cotton yield and quality on different fruiting branches. **Crop Science**, v.53, p.2670-2678, 2013.
- MALHOTRA, N. **Pesquisa de marketing: uma orientação aplicada**. 4. ed. Porto Alegre: Bookman, 2006.
- MANLY, B. J. F. **Métodos estatísticos multivariados**. 3. ed. Porto Alegre: Bookman, 2008.
- MARQUES, E. S. **PORTARIA Nº 131, DE 13 DE AGOSTO DE 2019**. Disponível em: <<https://apps.agr.br/zoneamento-agricola-2019-2020/>>. Acesso em: 30 abr. 2020.
- MARTINS, M. A. et al. Extraction and characterization of cellulose whiskers from commercial cotton fibers. **Journal of Materials Science**, v.46, p.7858-7864, 2011.
- MORETTIN, P. A.; BUSSAB, W. **Estatística básica**. 5. ed. São Paulo: Saraiva, 2003.
- MUCHERINO, A.; PAPAJOGEJI, P. J.; PARDALOS, P. M. **Data mining in agriculture**. Gainesville: Springer, 2009.
- OECD-FAO. **Medium-term prospects for major agricultural commodities 2017-2026**. México, 2017.
- PARRY, G. Le cotonnier et ses produits. In: COSTE, R. (Ed.). **Techniques Agricoles et Productions Tropicales**. Paris, France: G-P Maisonneuve et Larose Paris, p. 502, 1981.
- PASSOS, S. M. DE G. **Algodão**. Campinas, SP: Instituto Campineiro de Ensino Agrícola, 1977.
- PETTIGREW, W. T.; HEITHOLT, J. J.; MEREDITH, W. R. Early Season Floral Bud Removal and Cotton Growth, Yield, and Fiber Quality. **Agronomy Journal**, v.84, n.2, p.209-214, 1992.
- RAPHAEL, J. P. A. **Produtividade e qualidade do algodão em função da adubação nitrogenada sob restrição luminosa**. 2019. (Tese de Doutorado). UNESP (Universidade Estadual Paulista), Botucatu-SP, 2019.

- REDDY, K. R.; REDDY, V. R.; HODGES, H. F. Temperature Effects on Early Season Cotton Growth and Development. **Agronomy Journal**, v.84, n.2, p.229-237, 1992.
- REFAELZADEH, P.; TANG, L.; LIU, H. Cross-validation. In: LIU, L.; ÖZSU, M. T. (Ed.). **Encyclopedia of database systems**. New York: Springer US, p. 532–538, 2009.
- RESENDE, M. A. V. de et al. Divergência genética e índice de seleção via BLUP em acessos de algodoeiro para características tecnológicas da fibra. **Pesquisa Agropecuária Tropical**, v.44, n.3, p.334-340, 2014.
- REZENDE, S. O. et al. Mineração de dados. In: REZENDE, S. O. (Ed.). **Sistemas inteligentes: fundamentos e aplicações**. Barueri: Ed.Manole, p. 307–335, 2002.
- ROSOLEM, C. A. Ecofisiologia e manejo da cultura do algodoeiro. **Informações Agronômicas**, 2001.
- SALVADOR, H. G.; CUNHA, A. M. Da; CORRÊA, C. S. Vedalogic: um método de Verificação de Dados Climatológicos Apoiado em Modelos Minerados. **Revista Brasileira de Meteorologia**, v.24, n.4, p.448-460, 2009.
- SAYAD, S. **Support vector machines**. Disponível em: <<http://chem-eng.utoronto.ca/~datamining/Presentations/SVM.pdf>>. Acesso em: 27 abr. 2020.
- SESTREN, J. A.; LIMA, J. J. Características e classificação da fibra de algodão. In: FREIRE, E. C. (Ed.). **Algodão no cerrado do Brasil**. 3. ed. Brasília, DF: ABRAPA, p. 653–750, 2015.
- SIBCS. **Sistema Brasileiro de Classificação de Solos**. 5. ed. Brasília, DF: Embrapa Solos, 2018.
- SOARES, L. C. da S. **Efeito da temperatura e chuva sobre a qualidade da fibra e produtividade de algodão no estado de Mato Grosso**. 2015. (Dissertação de Mestrado). Escola Superior de Agricultura "Luiz de Queiroz", Universidade de São Paulo, Piracicaba-SP, 2015.
- SPINDEL, B. R. **Análise e desenvolvimento de sistema de estimação de modelos da classe STAR-Tree**. 2008. (Dissertação de Mestrado). Pontifícia Universidade Católica do Rio de Janeiro, Rio de Janeiro, 2008.
- TACONELI, C. A. **Árvores de classificação multivariadas fundamentadas em coeficientes de dissimilaridade e entropia**. 2008. (Tese de Doutorado). Escola Superior de Agricultura "Luiz de Queiroz", Universidade de São Paulo, Piracicaba-SP, 2008.
- THORNLEY, J. H. M.; FRANCE, J. Crop models. In: **Mathematical models in agriculture: quantitative methods for the plant, animal and ecological sciences**, 2017.
- THORNTHWAITE, C.W; MATHER, J. R. The water balance. **Nex Jersey: Drexel Institute of Technology**, p. 104, 1955.
- TONATO, F. et al. Desenvolvimento de modelos preditores de acúmulo de forragem em pastagens tropicais. **Pesquisa Agropecuária Brasileira**, v.45, n.5, p.522-529, 2010.
- USDA. **USDA Agricultural projections to 2026: Interagency Agricultural Projections Committee Global Agriculture: Projections to 2020 and U.S. Trade Trends**, 2017.
- USTER TECHNOLOGIES. USTER HVI 1000 – Para Seleção e Gerenciamento da Fibra de Algodão. In: **Anais...FIBER SYMPOSIUM**, 2008.
- VIDAL, C.; FREIRE, E. C. Melhoramento genético do algodoeiro. In: **Melhoramento genético de plantas no Nordeste**, 2013.
- VIEIRA, F. D. **Modelos baseados em técnicas de mineração de dados para suporte à certificação racial de ovinos**. 2014. (Dissertação de Mestrado). Universidade Estadual de Campinas, Campinas-SP, 2014.
- VIOLATTI, M. R. **Qualidade da fibra e diversidade genética em algodoeiro de fibra branca**. 2016. (Dissertação de Mestrado). Universidade Federal de Uberlândia, Uberlândia-MG, 2016.

- WANG, X. et al. Effect of waterlogging duration at different growth stages on the growth, yield and quality of cotton. **PLoS ONE**, p.1-14, 2017.
- WEIS, S. M.; INDURKHYA, N. **Predict Data Mining**. Morgan Kaufmann Publishers, 1999.
- WILLMOTT, C. J. On the validation of models. **Physical Geography**, p.184-194, 1981.
- WITTEN, I. H.; FRANK, E.; HALL, M. a. **Data Mining: Practical Machine Learning Tools and Techniques (Google eBook)**, 2011.
- YANG, J. S. et al. Soil Potassium Deficiency Reduces Cotton Fiber Strength by Accelerating and Shortening Fiber Development. **Scientific Reports**, v.6, n.28856, 2016.
- YEATES, S. J. Efeitos do estresse hídrico na fisiologia do algodoeiro. In: ECHER, F. R. (Ed.). **O algodoeiro e os estresses abióticos:temperatura, luz, água e nutrientes**. Cuiabá, MT: Instituto Mato-Grossense do Algodão, p. 63–77, 2014.
- ZHAO, W. et al. A model for cotton (*Gossypium hirsutum* L.) fiber length and strength formation considering temperature-radiation and N nutrient effects. **Ecological Modelling**, v.243, p.112-122, 2012.
- ZONTA, J. et al. Irrigation and nitrogen effects on seed cotton yield, water productivity and yield response factor in semi-arid environment. **Australian Journal of Crop Science**, v.10, n.1, p.118-126, 2016.