Revista de Informática Teórica e Aplicada - RITA - ISSN 2175-2745
Vol. 28, Num. 02 (2021) 66-89

RESEARCH ARTICLE

Classification based on rules for the study of cotton productivity in the state of Mato Grosso

Classificação baseada em regras para estudo da produtividade do algodão no estado do Mato Grosso

Alexandra Virgínia Valente da Silva[1]*, Carlos Manoel Pedro Vaz[2], Ednaldo Jose Ferreira[2], Rafael Galbieri[3]

Abstract: The advance of cotton farming in the Brazilian savannah boosted and made possible a highly technified, efficient and profitable production, elevating the country from the condition of cotton fiber importer in the 70s to one of the main exporters so far. Despite the increasing contribution of technologies such as transgenic cultivars, machines, inputs and more efficient data management, in recent years there has been a stagnation of cotton productivity in the State of Mato Grosso (MT). Data Mining (MD) techniques offer an excellent opportunity to assess this problem. Through the rules-based classification applied to a real database (BD) of cotton production in MT, factors were identified that were affecting and consequently limiting the increase in productivity. In the pre-processing of the data, we perform the attributes, selection, transformation and identification of outliers. Numerical attributes were discretized using automatic techniques: Kononenko (KO), Better Encoding (BE) and combination: KO + BE. In modeling the rule algorithms used were PART [1] and JRip, both implemented in the WEKA tool. Performance was assessed using statistical metrics: accuracy, recall, cost and their combination using the I_{FC} index (created by the authors). Results showed better performance for the PART classifier, with discretization by the KO + BE technique, followed by binary conversion. The analysis of the rules made it possible to identify the attributes that most impact productivity. This article is an excerpt from an ICMC/USP Professional Master's Dissertation in Science carried out in São Carlos-SP/BR.
Keywords: Cotton productivity — Data mining — Classification based on rules — Machine learning

Resumo: O avanço da cotonicultura no cerrado brasileiro impulsionou e viabilizou uma produção altamente tecnificada, eficiente e lucrativa, elevando o país da condição de importador de fibra de algodão na década de 70 a um dos principais exportadores até o momento. Apesar do aporte cada vez maior de tecnologias como cultivares transgênicas, máquinas, insumos e gestão de dados mais eficientes, nos últimos anos tem-se verificado a estagnação da produtividade de algodão no Estado do Mato Grosso (MT). Técnicas de Mineração de Dados (MD) oferecem excelente oportunidade para avaliar este problema. Através da classificação baseada em regras aplicada a um banco de dados (BD) real de produção de algodão no MT, foram identificados fatores que estavam afetando e consequentemente limitando o aumento da produtividade. No pré-processamento dos dados realizamos nos atributos, seleção, transformação e identificação de *outliers*. Atributos numéricos foram discretizados utilizando técnicas automáticas: Kononenko (KO), Better Encoding (BE) e combinação: KO+BE. Na modelagem os algoritmos de regras utilizados foram o PART [1] e JRip, ambos implementados na ferramenta WEKA. O desempenho foi avaliado pelas métricas estatísticas: precisão, revocação, custo e a combinação delas pelo índice I_{FC} (criado pelos autores). Resultados mostraram melhor desempenho para o classificador PART, com discretização pela técnica de KO+BE, seguida pela conversão binária. A análise das regras possibilitou a identificação dos atributos que mais impactam na produtividade. Este artigo é um recorte de uma dissertação de Mestrado Profissional em Ciências do ICMC/USP realizado em São Carlos-SP/BR.
Palavras-Chave: Produtividade do algodão — Mineração de dados — Classificação baseada em regras — Aprendizado de Máquina

[1] *Instituto de Ciências Matemáticas e Computação da Universidade de São Paulo (ICMC/USP), São Carlos - São Paulo, Brasil*
[2] *Embrapa Instrumentação, São Carlos - São Paulo, Brasil*
[2] *Embrapa Instrumentação, São Carlos - São Paulo, Brasil*
[3] *Instituto Mato-grossense do Algodão, Primavera do Leste - Mato Grosso, Brasil*
***Corresponding author**: alexa-mat@hotmail.com
DOI: http://dx.doi.org/10.22456/2175-2745.108126 • **Received:** 14/05/2021 • **Accepted:** 22/06/2021
CC BY-NC-ND 4.0 - This work is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License.

1. Introdução

Dados recentes (ano 2021) mostram que o Brasil é o "quarto maior produtor mundial de algodão" [2][3], ficando atrás apenas da Índia, China e EUA, de acordo com a Associação Brasileira dos Produtores de Algodão (Abrapa) na (safra 19/20). O país é o"segundo maior exportador" [4] ficando atrás apenas dos Estados Unidos e o "primeiro em produtividade em sequeiro" [5], isto é, em áreas não irrigadas. Vale frisar que o trabalho descrito neste artigo foi de uma pesquisa de mestrado desenvolvida durante o periodo de 2015 a 2019.

A produção comercial do algodão no Brasil teve início nos estados do Nordeste, sendo o Maranhão o primeiro grande produtor e exportador da fibra. No início da década de 80 com aparecimento do bicudo-do-algodoeiro, praga de maior impacto dessa cultura, ocorreu o declínio das lavouras algodoeiras, fazendo com que o Brasil passasse de exportador a importador. Esta crise teve como uma de suas consequências o deslocamento do eixo de produção dos estados de São Paulo, Paraná e do Nordeste, para os cerrados do Centro-Oeste, mais precisamente para o Mato Grosso. Na década de 90, o Brasil passou da condição de importador para exportador de pluma, pois com o avanço das tecnologias para "explorar as grandes extensões de terras planas mecanizáveis e o clima favorável, possibilitou um rápido crescimento da produção de algodão no Centro-Oeste" [6]. Hoje a região do Cerrado responde por 97% da produção brasileira, sendo o Estado de Mato Grosso o maior produtor de algodão herbáceo (de fibra curta), ocupando a primeira posição em área cultivada e produção (66,2% do total) [7].

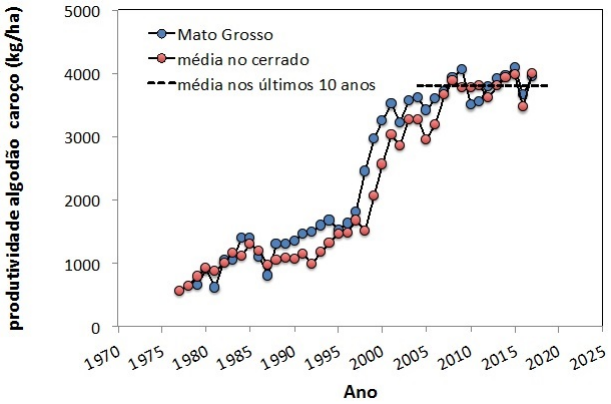

Figure 1. Evolução da produtividade do algodão no Estado do Mato Grosso e na média de todos os estados produtores na região do Cerrado (MT, MS, GO, BA, MA), evidenciando a tendência de estagnação nos últimos 10 anos (linha pontilhada)

A Empresa Brasileira de Pesquisa Agropecuária (EMBRAPA) vem atuando muito bem na área de manejo do algodão em parceria com Universidades, Institutos de Pesquisa e Empresas privadas realizando pesquisas para o aperfeiçoamento dos sistemas de produção, porém, um dos desafios atuais é o do aumento da produtividade desta cultura, uma vez que nos últimos anos houve uma estagnação da produtividade no cerrado, como pode ser observado na Figura 1. Verificou-se que a produtividade média está estagnada em torno de 4 toneladas por hectare (aproximadamente 272 @/ha), sendo que o potencial produtivo máximo do algodão nas condições de Cerrado é de cerca de 6 toneladas/ha ou 408 @/ha [8]. Assim, surge o seguinte questionamento: quais são os fatores que estão limitando o aumento da produtividade do algodoeiro no Mato Grosso e demais áreas produtoras do Cerrado?

Nesse contexto, aplicar análises Matemáticas para extrair modelos de bases de dados como as bases agricolas [9], diversas técnicas de MD e AM têm sido utilizadas para avaliar o efeito de diferentes atributos, manejáveis ou não na variável produtividade, buscando, assim, prever a produtividade das culturas agrícolas em diferentes cenários [10]. Portanto, o objetivo geral deste trabalho foi aplicar tais técnicas, mais especificamente a classificação baseada em regras em um BD de produção comercial de algodão no Estado do Mato Grosso, visando a compreensão dos fatores que impactam na produtividade. Assim, identificar padrões e tendências estatísticas contribuem na tomada de decisões. A aplicação da MD "no agronegócio, é extremamente útil para a agricultura de precisão e para a agroindústria, entre outros aspectos" [9].

Além da Introdução apresentada na **Seção 1**, a estrutura do trabalho está organizada no seguinte modo: A **Seção 2** apresenta os trabalhos relacionados com a atua pesquisa, fazendo uma descrição dos principais pontos. A **Seção 3** aborda sobre as ferramentas utilizadas para a geração do modelo de AM. A **Seção 4** apresenta os métodos de discretização baseados em entropia que foram indispensáveis para o trabalho. Na **Seção 5** veremos como foi montada a metodologia para o desenvolvimento do trabalho. A **Seção 6** descreve os resultados de cada etapa do desenvolvimento do trabalho e por fim a **Seção 7** apresenta as conclusões, observações e as possíveis recomendações para o desenvolvimento de trabalhos futuros.

2. Trabalhos Relacionados

Na literatura é escasso trabalhos com aplicações de MD na cultura do algodão, porém, há inúmeros trabalhos relacionados a outras culturas, como os de [11] , [12], [13], [14], [15], [16], [17], [18], [19] e [20].

No trabalho de [11] foram aplicadas técnicas de MD (modelos de regressão) em variáveis: climáticas, propriedades dos solos, irrigação e o local da produção. O objetivo foi identificar as condições mais adequadas de produção em termos de produtividade das culturas de arroz, batata e trigo em diferentes distritos de Bangladesh. A classificação/regressão foi realizada usando regressão linear, o algoritmo de k-Nearest Neighbour (vizinhos próximos) e Redes Neurais, para a predição do atributo (alvo) produtividade das culturas. Árvores de Decisão foram utilizadas no trabalho de [13] para identificar os fatores de maior influência na produtividade da cana-de-açúcar e a produção de etanol. O BD foi composto por atributos do solo, a variedade de cana utilizada e informações de manejo como os sistemas de plantio, irrigação e colheita, o

espaçamento das linhas de plantio, dentre outros, coletados em um período de 5 anos. O principal resultado obtido foi a identificação do atributo variedade como sendo o mais relevante para o atributo-alvo produtividade. Similar ao trabalho realizado e descrito neste artigo, no que tange à busca por regras explicativas, foi realizado por [14] utilizando Regras de Associação (RA) e o algoritmo APRIORI [21], um estudo para encontrar os principais fatores que influenciam a produtividade do açaí no estado do Amapá. Os resultados mostraram que a grande maioria dos produtores de açaí no Amapá utilizam mão-de-obra familiar e não têm acesso a financiamentos e assistência técnica. O produto é explorado em áreas pequenas e a maior parte da produção é comercializada diretamente no porto. A eficácia e consistência das regras de associação geradas foram comprovadas pelo analista de dados e o especialista no domínio da aplicação. Uma análise geral feita com as regras geradas, mostram que a produção do açaí no Amapá ainda é praticada de maneira "amadora", como meio de subsistência, uma vez que a maioria dos produtores possuem outra atividade principal. Em uma revisão recente, [15] descrevem algumas aplicações com técnicas de MD, como Redes Neurais Artificiais, Redes Bayesianas, Máquinas de Vetores de Suporte e Associação de Regras em uma variedade de bases agrícolas. Dentre elas, destacam-se a estimativa de preço de safras agrícolas, a previsão da produção agrícola baseada em dados climáticos e na fertilidade dos solos, a identificação de padrões de infestação de doenças e pragas agrícolas e a classificação de solos. Na cultura do arroz, que é altamente demandante de água, trabalhos foram realizados aplicando técnicas de MD para se obter associações entre a produtividade do arroz e parâmetros climáticos como precipitação, radiação solar e temperaturas mínimas e máximas, com dados da Índia [16], [17] e Colômbia [18]. Em cana-de-açúcar utilizou-se três técnicas de MD nas análises de BD de usinas de cana-de-açúcar no Estado de São Paulo com o objetivo de ordenar variáveis que condicionam a produtividade da cana-de-açúcar de acordo com sua importância, bem como o desenvolvimento de modelos matemáticos de produtividade dessa cultura [19].

Assim, a agricultura é uma atividade que depende de fatores que muitas vezes foge do controle do agricultor, como as variáveis climáticas (que não estão presentes no BD do presente estudo), o aparecimento de doenças e pragas inesperadas, e a oscilação de preços (variável também ausente no BD em estudo) tanto dos insumos como das culturas (grãos, frutas, fibras). Portanto, a busca por produtividades elevadas e a diminuição dos custos de produção são fatores preponderantes para a lucratividade e a competitividade do setor. Por esse motivo, a mineração de dados agrícolas tem focado na correlação dos fatores que influenciam a produtividade e geração de modelos de previsão ou estimativa do rendimento das culturas.

3. Ferramentas de Aprendizado de Máquina utilizadas

Para este trabalho utilizamos desde o tratamento dos dados até a descoberta de padrões a ferramenta WEKA *(Waikato Environment for Knowledge Analysis)*. "O Weka é uma coleção de algoritmos de Machine Learning e Data Mining escrita em Java para resolver problemas de MD do mundo real, foi desenvolvido na Universidade de Waikato, Nova Zelândia" [22] [23]. Frisando que a escolha dos algoritmos e *softwares* para a MD deve seguir os critérios necessários para atender as necessidades do usuário, ou seja, devem estar de acordo com o problema a ser resolvido. De acordo com a literatura, o *software* WEKA tem sido bastante utilizado para gerar modelos de AM no agronegócio, como por exemplo, indução de árvores de decisão [9], geração de regras de classificação, dentre outros, com um único objetivo, cotribuir para tomada de decisão.

3.1 Algoritmo PART

O algoritmo PART (*partial decision trees*) é uma variação do classificador J48 [24] e tem como finalidade gerar regras a partir de uma árvore de decisão baseada no conjunto de dados de treinamento rotulados [1]. É uma abordagem alternativa para a indução de regras que evita a otimização global e produz conjuntos de regras precisas e compactas. O processo de geração de regras de produção atua em dois estágios: gera uma lista de decisão e, assim como o J48, também usa a técnica de dividir-para-conquistar. O algoritmo constrói uma árvore de decisão C4.5 [24] parcial a cada iteração e coloca a melhor folha dentro de uma regra [1]. Regras são induzidas inicialmente de uma árvore e posteriormente são refinadas. Para cada regra criada é estimada a cobertura das instâncias da base. Isso ocorre repetidamente até que todas as instâncias estejam cobertas. As regras com coberturas mais altas são apresentadas para o usuário e as demais são descartadas [1].

3.2 Algoritmo JRip

O algoritmo JRip - RIPPER (*Repeated Incremental Pruning to Produce Error Reduction*) [25]: Poda Incremental Repetida para Produzir Redução de Erro é um dos algoritmos mais populares implementados no pacote WEKA. É baseado no algoritmo RIPPER que repete uma poda incremental para a redução de erro, proposto como uma versão aperfeiçoada de IREP (*Incremental Reduced Error Pruning*) [26], ou seja, com a poda incremental repetida, o algoritmo JRip produz uma redução de erro. É um algoritmo de aprendizado de máquina que aprende regras proposicionais a partir de exemplos. Ele implementa uma ordenação de classes seguindo a técnica dividir-para-conquistar, elencando linearmente o número de exemplos para treino (aprendizagem), realizando tal esquema para cada exemplo em sua base de regras [27]. Isto é sequencialmente repetido até que as chances de erro sejam as menores possíveis de serem detectadas pelo sistema. A regra produzida com menor incidência de erro é eleita para a classificação, ou seja, a classe que se sobressai é es-

colhida como padrão, auxiliando na determinação da classe minoritária. Ele foi proposto por Willian W. Cohen [25] em uma versão otimizada do IREP.

4. Métodos de discretização baseados em entropia

Em seu conceito mais simples, entropia [28] [29] é a medida da desordem de um conjunto de instâncias. A "desordem" não deve ser compreendida como "bagunça" e sim como a forma de organização de sistema. A entropia reflete a dificuldade de se calcular a probabilidade de um estado qualquer de um sistema. Em termos práticos, uma alta entropia significa a inexistência de regiões mais prováveis no espaço de estados possíveis. A discretização utilizando rótulos ou informações de classe empregam abordagens baseadas em entropia. A entropia de informação de classe utiliza um conceito similar, isto é, quando um sistema é "bem comportado", apresenta distribuições de probabilidades de suas partes constituintes que são modeláveis e relativamente previsíveis. Assim, é possível representar essas distribuições por meio de modelos Matemáticos com complexidades menores. Atributos com menor entropia terão maior ganho de informação, ou seja, é selecionado como atributo teste para o nó corrente, então, este atributo faz uma minimização da informação essencial para classificar as instâncias em partições resultantes e reflete a menor aleatoriedade ou "impureza" naquelas partições. A entropia de informação de classe mede a quantidade de informação que seria necessária para especificar a qual classe uma instância pertence [30].

4.1 Técnica de discretização MDL (descrição de comprimento mínimo)

O MDL (*Minimum Description Length*) ou descrição de comprimento mínimo [31] [32] [33] é uma teoria de inferência indutiva que pode ser aplicada a problemas em estatísticas, aprendizado de máquina e reconhecimento de padrões [34].

A ideia principal do princípio MDL é a busca de um modelo com a menor descrição dos dados observados. Isso é feito encontrando regularidades nos dados que são usados para compactá-los [35], isto é, para descrevê-los usando menos símbolos do que o número de símbolos necessários para que descreva os dados literalmente. Quanto mais regularidades existem, mais os dados podem ser comprimidos. Equiparando "aprendendo" com "encontrar regularidade", podemos, portanto, dizer que quanto mais somos capazes de comprimir os dados, mais aprendemos sobre eles. Em termos gerais, a melhor explicação para um determinado conjunto de dados é fornecida pela descrição mais curta desses dados [34]. "Regularidade" pode ser identificada como a "capacidade de comprimir". O MDL combina esses dois *insights*, vendo a aprendizagem como compressão de dados: por exemplo, para um determinado conjunto de hipóteses H e conjunto de dados D, devemos tentar encontrar a hipótese ou combinação de hipóteses em H que mais comprime D [33].

O desafio no AM é o desenvolvimento de métodos que evitem o superajustamento - *overfitting* que é a memorização dos dados de treinamento, isto é , o modelo descreve com precisão os exemplos usados para construí-lo mas falha em aprender a generalizá-los, e modelos com essa falha não podem ser usados para tarefas preditivas [35]. Assim, o procedimento MDL protege automática e inerentemente contra sobreajuste [33]. Além desta propriedade, o MDL não necessita de exemplos negativos na seleção do modelo [35].

O MDL é considerado como um método de discretização que utiliza o procedimento de minimização de entropia [36]. Esse método realiza a discretização de variáveis contínuas em classes, ou seja, estabelece os pontos de corte das classes utilizando a entropia de informação ou ganho de informação de classe de partições para selecionar limites na discretização. A Entropia de [28], ou Entropia da Informação é a medida de incerteza (ou desordem) de um sistema, está relacionada às probabilidades de encontrar um sistema em cada estado que ele pode assumir. Se um sistema tem poucos estados possíveis a probabilidade de prever em qual estado ele se encontra é maior do que se um sistema possui grande número de estados possíveis [37].

Em termos práticos, isso não significa que há falta de homogeneidade. O que existe são "adensamentos probabilísticos", ou seja, regiões de maior probabilidade, assim, uma alta entropia significa a inexistência de "regiões mais prováveis no espaço de estados possíveis", ou seja, maior será a necessidade de recursos para o canal de envio da informação. E quando a entropia é baixa, a quantidade de bits de informações para um canal de comunicação também o é.

A entropia é utilizada porque as aglomerações no espaço (n-dimensional), quando ocorrem, reduzem a entropia do sistema de dados. Assim, selecionar um atributo pela sua entropia pode refletir aglomerações imanentes às classes. A medida de entropia é muito usada para geração de árvores de decisão e para classificadores baseados em regras, sendo este ultimo, a técnica selecionada para o estudo do banco de dados do presente artigo.

O método MDL utiliza uma abordagem de cima para baixo, onde vários intervalos são criados para formar uma árvore por meio de múltiplas divisões do atributo numérico ao mesmo nó para produzir intervalos discretos [36]. Assim, considera todo o intervalo conhecido ou medido de uma variável contínua, calcula a entropia de todo o conjunto de dados e em seguida percorre todo o conjunto fazendo-se partições recursivamente em subintervalos para encontrar as divisões com maior ganho de informação. As divisões são encontradas classificando dois valores vizinhos da lista e esse procedimento vai se repetindo até satisfazer um critério de parada, isto é, quando o ganho de informação for menor de um certo limite (ruído intrínseco do conjunto). Existem várias formas de se calcular o limite mínimo, inclusive empiricamente. O critério de divisão MDL é conservador e se o conjunto de dados apresentar ruído moderado o MDL poderá encontrar poucas partições, se houver [38]. Os procedimentos MDL

protegem automática e inerentemente contra sobreajuste [33]. Na Figura 2 é apresentado um fluxograma ilustrando as etapas do algoritmo MDL.

4.1.1 Método de discretização *Better Encoding (BE)*

Em algumas situações, como é o caso de problemas multi-classes, os algoritmos convencionais baseados em entropia apresentam limitações, produzindo pontos de corte inadequados. Isso ocorre pelo fato do algoritmo minimizar a entropia média ponderada de 2 conjuntos na partição binária e o ponto de corte separar exemplos de uma mesma classe. Na Figura 3 essa situação é ilustrada utilizando o atributo pH do solo em água. Ao invés do ponto de corte (T) cair em um dos limites B1, B2 ou B3 do atributo, que seriam as melhores divisões, caiu entre o intervalo B2 e B3, de modo que a entropia média de ambos os lados seja minimizada. Isso seria indesejável, uma vez que de forma desnecessária separa exemplos da mesma classe, resultando em árvores maiores e de qualidade baixa [39], gerando consequentemente um número maior de regras [36].

Na indução descendente das árvores de decisão por exemplo, várias funções de impureza são usadas para estimar a qualidade dos atributos para selecionar o "melhor" para dividir [40], como o Índice Gini e a Entropia. Então para resolver o problema de pontos de cortes "ruins" na discretização, usa-se uma codificação de ponto de divisão mais eficiente para o MDL (a Técnica de discretização *Better Encoding*).

Almeja-se realizar divisões que melhorem a informação que recebemos de nossos dados. Consequentemente, para se ter melhores divisões é feito o uso do critério alternativo de divisão baseada em informação, que no caso seria a medida de impureza Entropia. O fato de que apenas os pontos limite são considerados faz com que a derivação do intervalo de cima para baixo seja viável (uma vez que o algoritmo nunca se compromete com um corte "ruim" na parte superior) e reduz o esforço computacional [36].

4.1.2 Método de discretização *Kononenko (KO)*

O método MDL baseado em entropia de *Kononenko (KO)* utiliza os mesmos conceitos do MDL original de [36], incluindo um ajuste para a discretização de múltiplos atributos [40]. Para isso, o algoritmo de (KO) apresenta uma correção para o viés que a medida da entropia apresenta para atributos com muitos valores [41].

Isso é obtido utilizando dois critérios de seleção, um baseado no princípio MDL e outro no algoritmo RELIEF [42]. A extensão incluída no algoritmo RELIEF permite tratar de forma mais eficiente dados com ruído, faltantes e problemas multi-classes [40]. Como resultado as medidas de seleção, melhoram com o aumento do número de atributos. O viés é estável e apresenta melhor resultado que outros critérios de seleção como a relevância, χ^2 (qui-quadrado), índice Gini dentre outros. As medidas de seleção introduzidas no método KO têm interpretação natural e mostra quando o atributo não é útil, ou seja, quando não é compressivo. Vale mencionar que uma das vantagens do método KO é reduzir a complexidade da aprendizagem [43].

5. Métodos

Neste trabalho foi utilizado um BD gerado por um projeto de pesquisa coordenado pelo Instituto Mato-Grossense do Algodão (IMAmt) em parceria com a Empresa Brasileira de Pesquisa Agropecuária (EMBRAPA), entre 2012 e 2015. O objetivo do projeto foi realizar um levantamento do nível de infestação de fitonematoides em áreas produtoras de algodão no Mato Grosso [44]. Fitonematoides são organismos vermiformes e microscópicos que habitam o solo e retiram nutrientes das raízes causando danos à planta hospedeira. Os fitonematoides estão entre os principais problemas fitossanitários da cultura do algodoeiro no Brasil [7].

5.1 O Banco de Dados

Juntamente com o levantamento da densidade populacional dos fitonematoides no solo e nas raízes, foram realizadas medidas de atributos químicos e físicos do solo e obtidas informações gerais do sistema de produção e a produtividade do algodoeiro. No total foram avaliados 1.799 talhões comerciais de produção de algodão (instâncias) em 431 fazendas do Mato Grosso, conforme ilustrado na Figura 4.

A Tabela 1 mostra a distribuição dos pontos (talhões) amostrados por safra agrícola, as épocas das amostragens de solo e raiz e obtenção de informações gerais, bem como os períodos de plantio do algodão. Os talhões são as unidades de plantio e manejo onde uma dada cultivar, manejo e tratos culturais são utilizados visando maximizar a produtividade. O ciclo do algodão é de 6 meses e cada produtor possui áreas muito grandes divididas em talhões. Os Talhões típicos nessas regiões possuem áreas entre 200 e 300 hectares. De um modo geral, o período de plantio foi de novembro ou dezembro indo até fevereiro ou março no máximo. A colheita inicia-se em junho e pode ir até agosto, dependendo de quando o talhão foi plantado. A coleta de dados concentraram-se nos meses de janeiro a maio nos 4 anos ou safras agrícolas. O procedimento para a coleta de informações é descrito a seguir.

Table 1. Períodos de coleta de dados e amostras de solo e raízes, períodos de plantio e número de talhões em cada safra agrícola para a construção do BD

Safra	Período de Coleta	Período de Plantio	DAP*	Nº de talhões
2011-2012	09/03 a 29/05/12	26/11/11 a 04/03/12	41 a 185	254
2012-2013	18/01 a 27/05/13	03/12/12 a 08/03/13	19 a 164	908
2013-2014	26/01 a 20/05/14	26/11/13 a 03/03/14	16 a 144	337
2014-2015	25/02 a 18/05/15	07/12/14 a 05/03/15	12 a 154	300

**DAP: dias após o plantio referente a data da coleta das amostras e dados*

A equipe de amostragem se reunia com o gerente da fazenda, que indicava os talhões a serem avaliados (variando entre 2 e 10, dependendo do tamanho da fazenda). O critério de seleção utilizado foi a escolha de talhões contrastantes de alta e baixa produtividade e talhões com e sem a presença visual de ataque das plantas de algodoeiro por fitonematoides. Nessa reunião inicial o gerente fornecia também informações

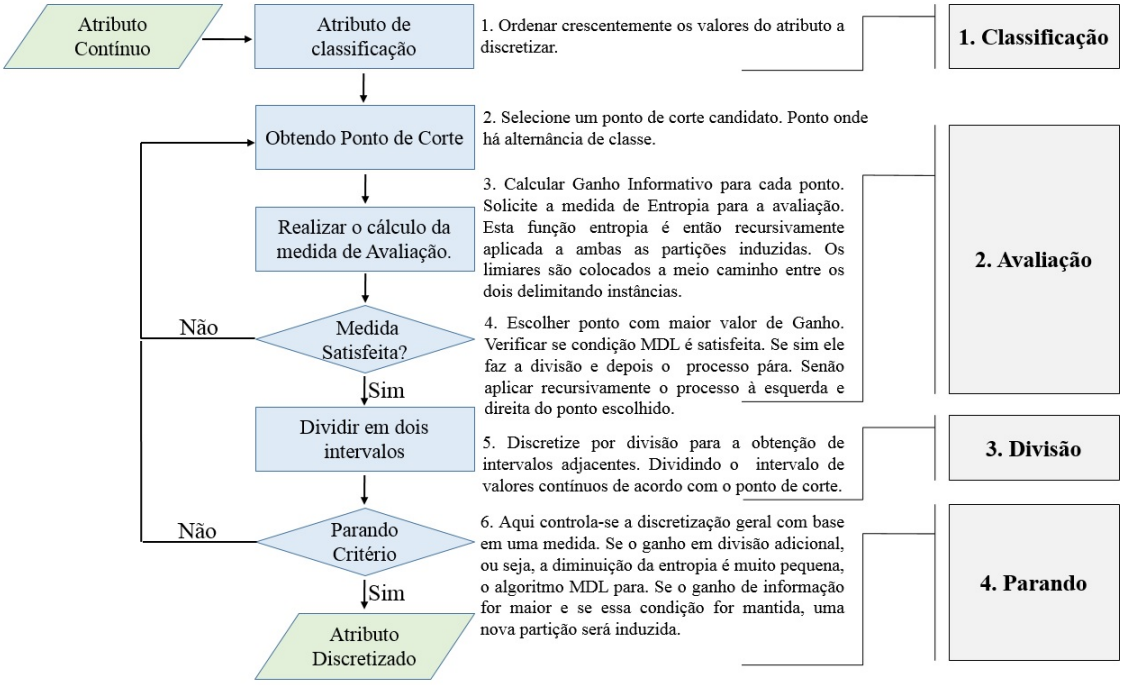

Figure 2. Fluxograma do algoritmo de discretização MDL

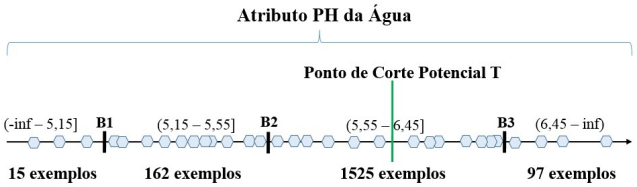

Figure 3. Exemplo de ponto de corte provisório para o atributo pH em água

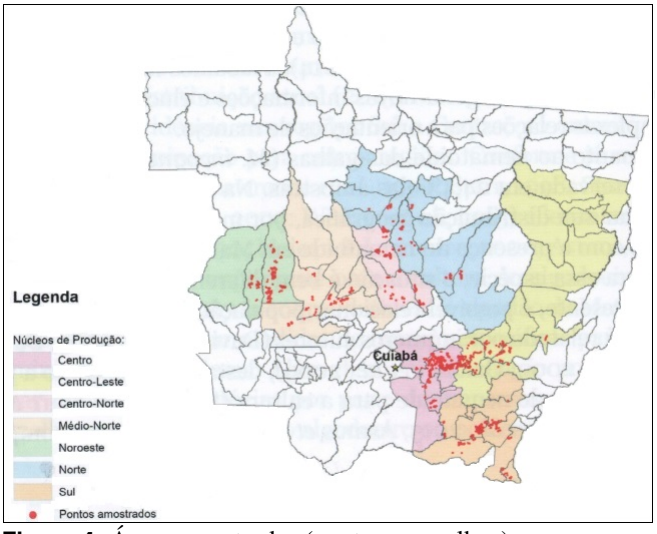

Figure 4. Áreas amostradas (pontos vermelhos) em diferentes núcleos de produção de fazendas de algodão no Estado de Mato Grosso

gerais do sistema de cultivo utilizado e demais informações constantes em um questionário aplicado de forma estruturada. Posteriormente, a equipe se dirigia ao campo, nos talhões indicados, e realizava a coleta de solo e raízes utilizando um protocolo pré-definido.As amostras coletadas eram enviadas a laboratórios credenciados para a determinação dos atributos químicos, físicos e biológicos e todos os dados coletados e analisados foram registrados em planilha do Excel. As análises físicas do solo foram realizadas na Embrapa Instrumentação, em São Carlos, SP, e as análises químicas do solo e de nematoides na Associação dos Produtores de Sementes do Mato Grosso (APROSMAT).

O espaçamento entre linhas de plantio (ESP) utilizado foi basicamente de 2 tipos, o Convencional (linhas espaçadas de 90, 80 ou 76 cm) e o Adensado (linhas espaçadas de 45 cm). Durante as 4 safras avaliadas apenas 12,8% dos talhões foram plantados no sistema adensado e dos 87,2% cultivados no sistema convencional, 50,9% foram no espaçamento 90 cm e 36,3% no espaçamento 76/80cm.

O valor médio do número de sementes por linha de plantio (Planta/m) foi de 10 sementes. Cerca de 60% do algodão foi de safrinha (com a soja como cultura de safra, na maioria das vezes) e 40% de algodão de safra. As cultivares mais utilizadas foram a FM 975 WS (43,9%), FM 951 LL (11%), FM 910 (7,6%), FM 933 (5,9%), FMT 701 (4,8%), TMG 81 WS (4,5%) e FMT 709 (4,4%).

Com relação ao sistema de plantio (SIST-PLANTIO) em cerca de 90% dos talhões foram utilizados o Plantio Direto na Palha (PDP) e 10% o Plantio Convencional (PC) com aração e gradagem no preparo do solo. Em cerca de 20% dos talhões foi realizada a subsolagem do solo (PREP-SOLO) para

descompactação na operação de preparo do solo anteriormente ao plantio do algodão.

Sobre as classes texturais dos solos utilizados, 33,4% foram da classe Muito Argilosa, 50,3% Argilosa, 13,5% Média e 2,8% Arenosa. Isso mostra que os solos com melhor qualidade física (com maiores teores de argila e menor de areia) e, por conseguinte, melhor fertilidade natural, são geralmente selecionados para o cultivo do algodoeiro no Mato Grosso.

O atributo alvo a ser avaliado foi a produtividade do algodoeiro, que apresenta dois tipos de informações, a produtividade na safra avaliada (PROD-SAFRA) e a produtividade histórica (PROD-HIST), que é a produtividade média dos talhões nos anos em que houve plantio de algodão em um determinado talhão.

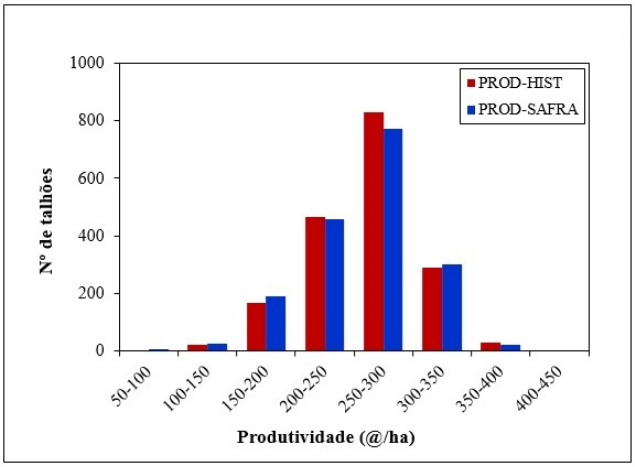

Figure 5. Histogramas de frequência das produtividades do algodoeiro na safra (PROD-SAFRA) e a produtividade histórica (PROD-HIST)

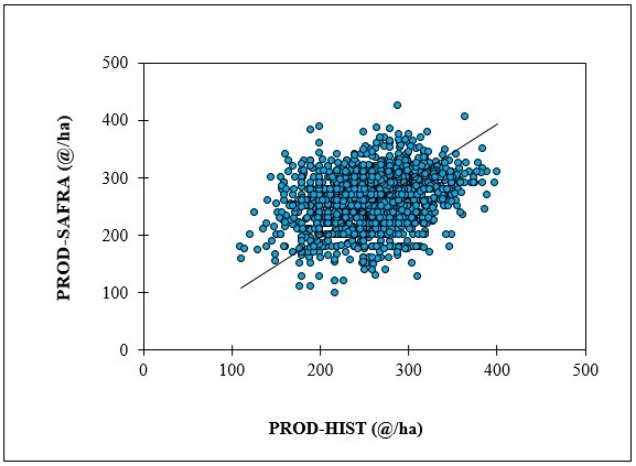

Figure 6. Correlação entre essas 2 produtividades

Vale ressaltar que ao longo dos 4 anos na coleta de dados não foi repetida em nenhum dos talhões a amostragem de um ano para outro, ou seja, cada talhão foi analisado uma única vez. A produtividade histórica (PROD-HIST) representa a média das produtividades durante todos os anos em que houve plantio de algodão no talhão. Por exemplo, se em determinado talhão foi reiniciado o plantio de algodão há 10 anos, a média de produção histórica será dos 10 anos, se o talhão começo a ser plantado em apenas dois anos, a média será destes dois últimos 2 anos. Já a produtividade da safra (PROD-SAFRA) é a produtividade do talhão no ano de amostragem. As distribuições de frequência dessas duas produtividades 5 e a correlação direta entre elas estão apresentadas respectivamente nas figuras 5 e 6.

5.2 Procedimento metodológico para criação do modelo de Regras de Classificação

O procedimento metodológico adotado para a obtenção das regras entre os atributos e informações do sistema de produção do algodão com a variável alvo Produtividade é sintetizado na Figura 8.

Na base de dados houve a necessidade de aplicar algumas técnicas de pré-processamento [45] para garantir a consistência dos dados [46]. Foram aplicadas técnicas de seleção de variáveis, transformação de variáveis de doenças (nematóides) e aplicação de técnica para detecção de *outliers* com ponderação do desvio interquartil de 1,5 conforme a Figura 7, esta análise foi realizada no *software* Excel diretamente na variável alvo (produtividade do algodoeiro), uma vez que em alguns talhões foram observados valores de produtividade da safra bastante distintos da produtividade histórica.

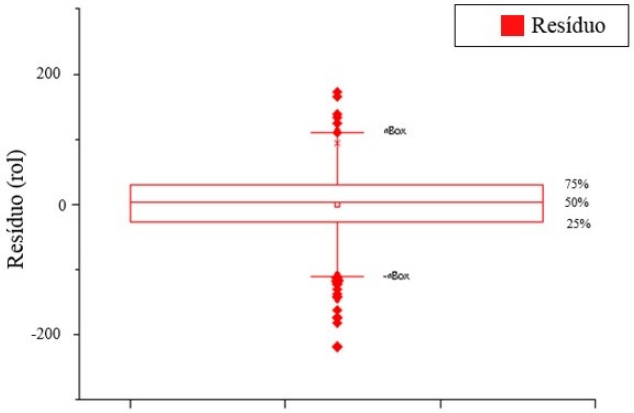

Figure 7. Gráfico de *box plot* dos resíduos entre as produtividades histórica e da safra

Após todo este processo foi necessário aplicar a discretização de dados, que não deixa de ser uma técnica de pré-processamento, já que os dados passam por uma transformação. Isto foi necessário pelo fato dos algoritmos de AM geradores de regras utilizados neste trabalho, aceitarem apenas dados discretos. Além da discretização transformar dados numéricos em dados nominais, este método reduz e simplifica o conjunto de dados, tornando o aprendizado mais rápido e os resultados mais compactos e fáceis de analisar [47]. Nesta etapa foram utilizadas 4 técnicas de discretização: três automáticas: (*Kononenko (KO)*, *Better Encoding (BE)*, combinação *Kononenko (KO)*

com *Better Encoding (BE)*, nomeado como (KO)+(BE), que foram realizadas com o *software* WEKA, e uma manual: feita com o auxílio do especialista. Os limites dos intervalos e o número de intervalos foram definidos manualmente para os atributos numéricos. O número de intervalos resultantes da discretização variou de 4 a 6 intervalos.

Em seguida ocorreu a etapa da MD propriamente dita, fazendo o uso dos algoritmos PART e JRip. Para as técnicas de mineração foram feitos testes para escolher o melhor classificador. As métricas precisão (P), recall (R), medida-F, número de regras geradas, custo (C) e o índice I_{FC} (criado a partir de (P), (R) e (C)) foram utilizadas para a avaliação do desempenho das técnicas de discretização e classificação. O modelo foi avaliado através da validação cruzada *k-fold* [46].

Um outro passo importante foi a aplicação da técnica da conversão binária (*Binary Split*) que tem por função transformar atributos discretos multi-rótulos em múltiplos problemas binários usando a técnica um-contra-todos. Uma das formas mais populares de lidar com atributos de múltiplos rótulos é distingui-los em outro conjunto de atributos binário. Os algoritmos de conversão trabalham apenas sobre os atributos, convertendo, por exemplo, um atributo de 4 rótulos em 4 novos atributos. Um exemplo dessa técnica é ilustrado na Figura 9 na qual é considerado o problema com atributo Classe Textural multirótulo do banco de dados do trabalho cujos os valores são: "Argiloso", "Arenoso", "Media" e "Muito Argilosa".

Por exemplo, o novo conjunto de dados criado para a categoria "Argiloso" terá apenas um rótulo. O rótulo será "True" para todos os pontos de dados que tivessem "True" para "Argiloso" no conjunto de dados original. Do mesmo modo, o conjunto de dados criado para "Arenosa" terá rótulo como "True" para todos os pontos de dados que tiveram o "Arenosa" como "True" no conjunto de dados original. Isso é feito para todas as categorias presentes no conjunto de dados.

$$Cobertura = \left(1 - \frac{8,0}{44}\right) * 100 = 81,81\% \tag{1}$$

O critério para a escolha das melhores regras foi feito através da avaliação de sua cobertura, onde as regras com coberturas superiores a 50% foram selecionadas para o estudo final da pesquisa. A cobertura é a porcentagem de tuplas que são cobertas pela regra [48]. Para aprender automaticamente, um classificador baseado em regras busca o melhor conjunto de pares (atributo, valor), ou seja, as premissas que determinam a decisão a ser tomada. Por exemplo na Figura 10 temos um exemplo de regra onde a cobertura é calculada utilizando os números entre parenteses, o cálculo está descrito em (1).

Uma avaliação geral das estatísticas dos atributos numéricos foi feita nas variáveis de doenças. Tal avaliação apresenta valores mínimos, máximos, médios, medianos, desvios padrão (DP), coeficientes de variação (CV), assimetria e curtose destes dados. Valores baixos de assimetria e curtose entre 0 e 3 [49] indicam uma tendência de distribuição normal dos dados. Para normalizar os dados, foi utilizada a Eq. (2):

$$P_i^* = Log(P_i + 1) \tag{2}$$

onde P_i é a população medida para o nematoide no talhão e P_i^* o valor transformado pelo logaritmo na base 10. O valor unitário somado a P_i deve-se ao fato de haverem talhões com contagem zero de nematoides. Vale ressaltar que em geral, não houve a necessidade de normalizar os dados, pois todos foram discretizados, até mesmo os atributos de doenças normalizados pelo algoritmo. A normalização de dados é recomendável quando os limites de valores de atributos distintos são muito diferentes, para evitar que um atributo predomine sobre outro [45], ou seja, pode existir uma coluna no banco de dados que está uma escala diferente e então este atributo dependendo do classificador, se ele atribuir uma distância, este atributo vai ser muito mais importante que o outro.

A discretização manual do atributo alvo produtividade foi realizada com o auxílio do especialista da área, onde considerou-se um limiar de lucratividade para a cultura entre 90 e 97 @ de pluma de algodão/ha. Isso em termos de produtividade do algodão em caroço representa cerca de 250@/ha.

Dentre os 68 atributos do BD, 2 deles (PROD-SAFRA e PROD-HIST) são atributos alvos. Esses atributos foram discretizados em 3 classes: BAIXA, MÉDIA e ALTA. Os outros 66 atributos foram utilizados para a obtenção das regras para os atributos alvo produtividade. Destes, 10 eram atributos nominais e 56 numéricos, havendo, portanto, a necessidade da transformação deles em atributos discretos utilizando-se para isso de técnicas de discretização presentes no *software* WEKA.

Medir adequadamente o desempenho de classificadores através do erro (ou precisão) tem papel importante na MD, uma vez que o objetivo é se obter classificadores com baixa taxa de erro quando aplicado a novos exemplos. Então, a matriz de custo foi utilizada para ponderar os erros apresentados na matriz de confusão visando a avaliação dos tipos de erros cometidos, especialmente quando a variável alvo (atributo *target*) é nominal ordinal (qualitativo ordinal), como é o caso da produtividade. Dessa forma, um classificador que erra uma produtividade ALTA como BAIXA, deve ser pior do que outro que erra a produtividade ALTA como MÉDIA.

A Tabela 2 (a) ilustra a matriz de custo de k classes, com pesos atribuídos às instâncias ou registros classificados corretamente ou erroneamente. A diagonal da matriz (n_0) indica as classes preditas corretamente (CP = CA). Pesos ou penalidades (n) são aplicados para os casos onde os registros são classificados erroneamente e no caso de acerto é considerado um peso igual a zero (n_0 = 0). Em geral, os indutores assumem que custo (CP_i, CA_j) = 1 para $i \neq j$, mas pode ser maior que 1 dependendo das características das classes do atributo alvo avaliado. Dessa forma, o custo é calculado multiplicando-se o número de instâncias classificadas corretamente ou erroneamente pelos pesos estabelecidos e quanto menor o custo melhor a predição (menor erro) do modelo. A definição dos pesos depende do problema específico estu-

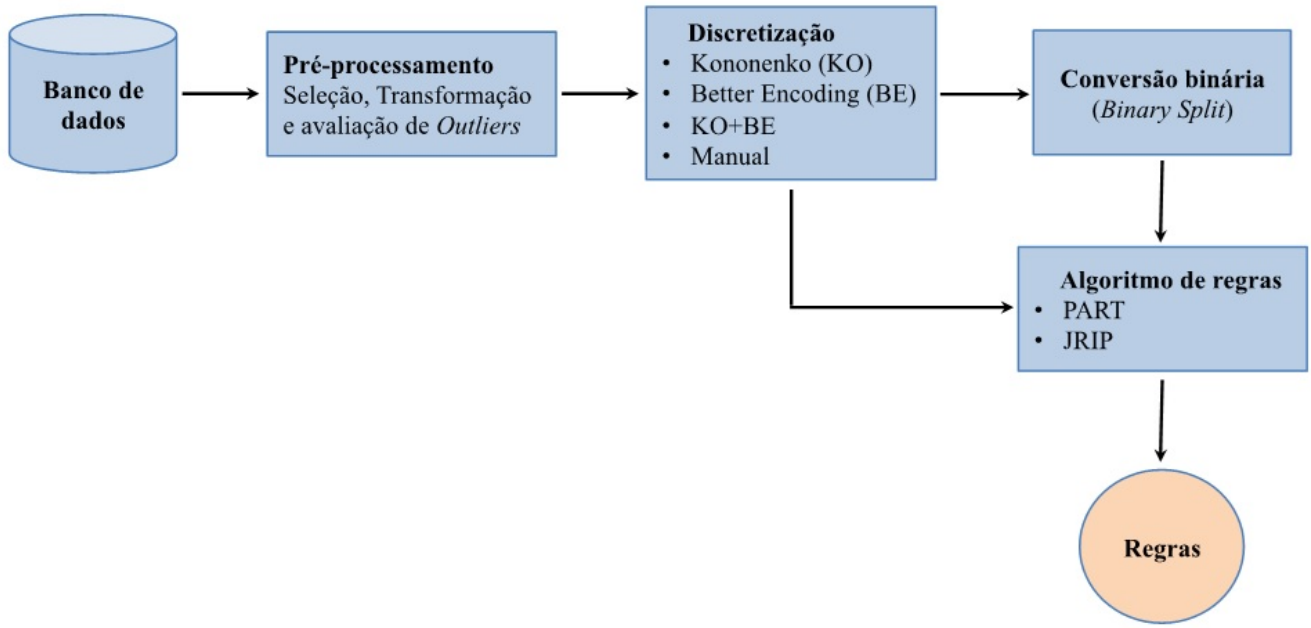

Figure 8. Esquema do procedimento utilizado para a geração de regras

Table 2. Ilustração de uma matriz de custo geral de ordem k com a indicação de pesos n_0 a n_k (a); dos pesos da matriz de custo adotada nesse trabalho (b) e um exemplo de matriz de custo obtida com o algoritmo PART aplicado aos dados desse trabalho (c)

a)

	CA_1	CA_2	...	CA_{k-1}	CA_k
CP_1	n_0	n_1	$\rightarrow$	n_{k-1}	n_k
CP_2	n_1	n_0	$\rightarrow$	$\rightarrow$	n_{k-1}
$\vdots$			n_0		
CP_{k-1}	n_{k-1}	$\leftarrow$	$\leftarrow$	n_0	n_1
CP_k	n_k	n_{k-1}	$\leftarrow$	n_1	n_0

b)

Alta	Média	Baixa	Classificação
0	1	2	**Alta**
1	0	1	**Média**
2	1	0	**Baixa**

c)

Alta	Média	Baixa	Classificação
374	113	134	**Alta**
195	139	215	**Média**
140	86	403	**Baixa**

dado. No caso deste trabalho, adotou-se as classes ALTA, MÉDIA e BAIXA para o atributo alvo produtividade do algodoeiro e considerou-se uma penalidade de 1 para as classes vizinhas (ALTA-MÉDIA; MÉDIA-BAIXA) e 2 (penalidade no custo de 200%) para classes extremas (ALTA-BAIXA), conforme mostrado na Tabela 2 (b). Isso porque instâncias classificadas como ALTA sendo BAIXA (ou vice-versa) são condição muito ruins de predição, enquanto que de ALTA para MÉDIA (ou vice-versa) e MÉDIA para BAIXA (ou vice-versa) não são tão graves quanto as anteriores.

O cálculo do custo total para o exemplo da matriz de custo apresentado da Tabela 2 é realizado conforme o cálculo em (3) descrito à seguir:

$$\begin{aligned} Custo = &(374\times 0)+(113\times 1)+(134\times 2)\\ &+(195\times 1)+(139\times 0)+(215\times 1)\\ &+(140\times 2)+(86\times 1)+(403\times 0)\\ &= 1157 \end{aligned} \tag{3}$$

Além do custo total, foram utilizadas outras métricas indicadoras do desempenho das regras de classificação geradas, como precisão (P), revocação (R) e medida F. Adicionalmente foi criado e introduzido um novo índice denominado de I_{FC}, que integra as métricas P, R, F e C, definido na Eq. (4) como:

$$I_{FC} = \frac{100\text{F}}{Log_2 C} \tag{4}$$

A Tabela 3 apresenta os cálculos dessas métricas para um exemplo real do BD utilizado neste trabalho. Os resultados de P, R e F são apresentados para as 3 classes individualmente e a média entre as 3 classes. Na tabela também é apresentado o resultado do cálculo do índice I_{FC}.

Para a avaliação do desempenho dos classificadores além das métricas descritas, o número de regras gerado na MD foi utilizado como um indicador importante na seleção do melhor classificador.

6. Resultados e Discussão

6.1 Resultado da avaliação geral das estatísticas dos atributos numéricos

Após uma avaliação geral das estatísticas dos atributos numéricos, observou-se uma anormalidade nos valores dos atributos de

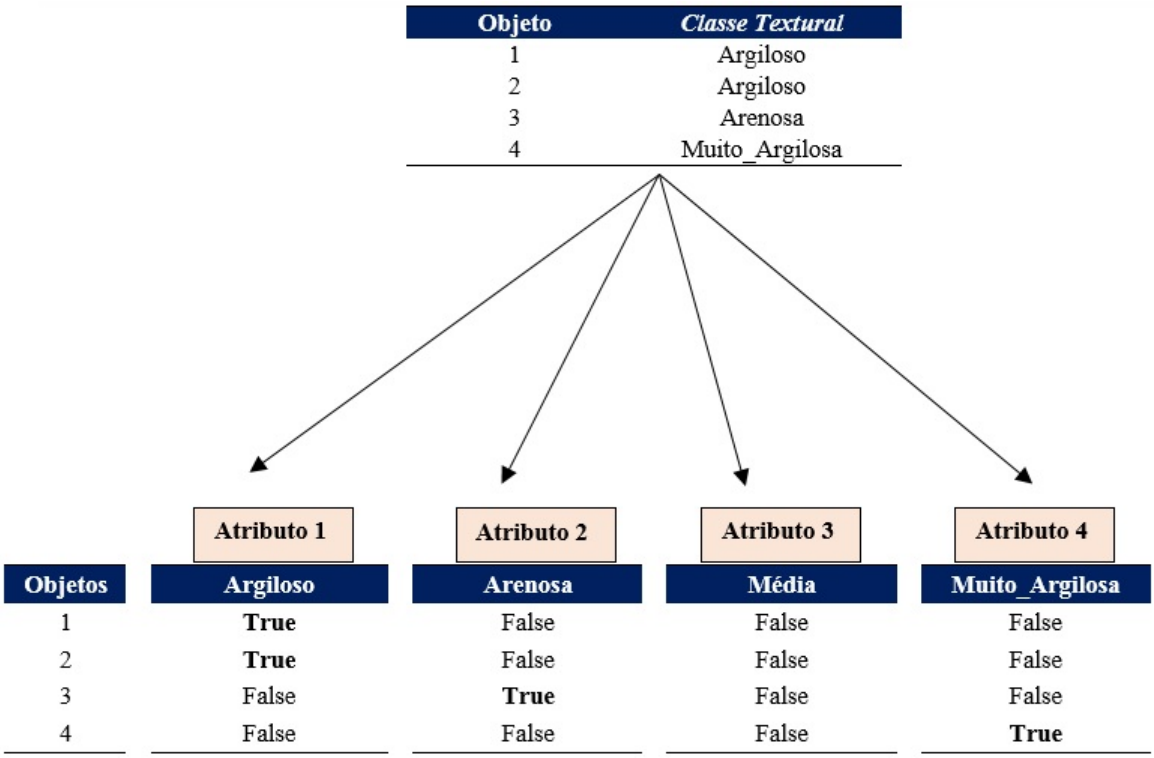

Figure 9. Transformação baseada em rótulos

doenças. Tais valores estavam fora dos intervalos normais de assimetria e curtose que seriam de 0 a 3 [49]. Assim, fica claro pelos dados da Tabela 4 que no caso das contagens populacionais das diversas espécies de nematoides em amostras de solos e raízes as distribuições não são normais (assimetria entre 3 e 19,9; curtose entre 11,3 e 347,7) e dessa forma foi realizada uma transformação dos dados de doenças (nematoides) antes da etapa de discretização.

6.2 Resultado da discretização do atributo alvo - Produtividade

Observa-se na Tabela 5 que a distribuição do número de instâncias em cada classe foi relativamente homogênea, ou seja, não foram observados problemas significativos de desbalanceamento das classes geradas não sendo necessário fazer um balanceamento de dados. Isso é importante, pois o desbalanceamento é um dos fatores que interferem negativamente no desempenho dos algoritmos de AM [50] . A existência de classes desbalanceadas ocorre quando o número de elementos entre as classes é desproporcional [45]. Nesse caso, exemplos da classe minoritária são geralmente classificados erroneamente.

6.3 Resultado da discretização manual dos atributos com informações fornecidas por especialistas da área

Os atributos numéricos passaram por uma discretização manual. Nesta discretização foram obtidos os limites dos intervalos e o número de intervalos que foram definidos para os atributos numéricos. O número de intervalos resultantes da discretização variou de 4 a 6 intervalos.

6.4 Análise de *outliers* dos atributos alvo (resíduo entre produtividades histórica e na safra)

Uma análise preliminar (Tabela 6) mostrou que em alguns talhões houve diferenças grandes entre as variáveis (PROD-SAFRA) e (PROD-HIST), o que pode ter sido causado por erros de medida no campo ou ocorrência de doenças e pragas no ano da coleta. Assim, foi necessário uma avaliação de *outliers* e a necessidade de uma análise para decidir excluir ou não estas instâncias.

A Tabela 6 apresenta a estatística do cálculo, indicando que apenas 28 instâncias (talhões) apresentaram resíduos fora dos limites inferiores e superiores estabelecidos pela ponderação do desvio interquartil de 1,5.

Em [51], destaca-se que para identificar *outliers*, faz-se uma análise com o coeficiente estatístico 1,5, ou seja, observações com afastamento superior a 1,5 desvio interquartílico para cima ou para baixo são consideradas atípicas. Ressaltamos que testes realizados com e sem a inclusão dessas 28 instâncias tiveram efeito pouco significativo na precisão da classificação e essas instâncias foram, portanto, mantidas no BD.

6.5 Análise comparativa dos classificadores PART e JRip

Resultados da validação cruzada (10-*folds*) são apresentados nas Tabelas 7, 8, 9 e 10 para todos os tipos de discretização.

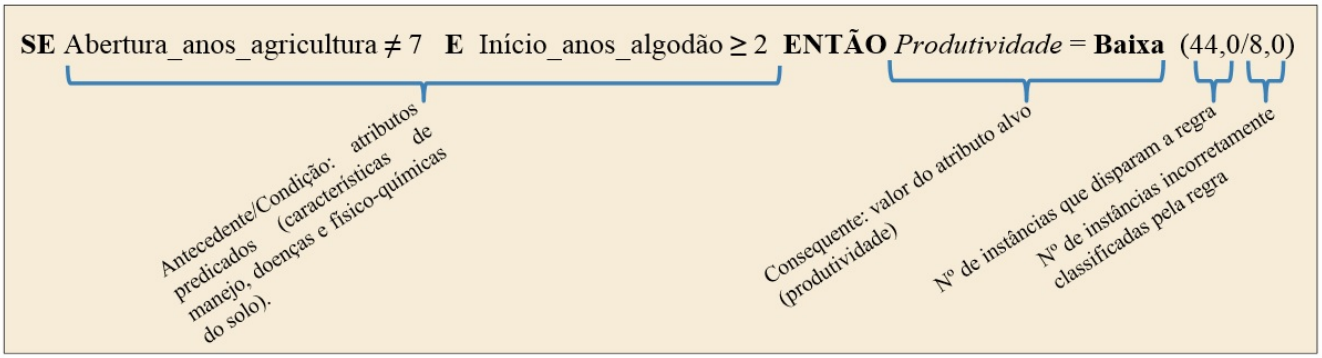

Figure 10. Estrutura de uma regra de classificação

Table 3. Cálculos da precisão (P), revocação (R), medida F e índice I_{FC} para o exemplo de classificação, utilizando o algoritmo PART

Métrica	Fórmula	Classe predita como	Cálculo	Média
Precisão	P = TP/(TP+FP)	Alta	374/(374+113+134) = 0,602	0,499
		Média	139/(139+195+215) = 0,253	
		Baixa	403/(403+140+86) = 0,641	
Revocação	R = TP/(TP+FN)	Alta	374/(374+195+140) = 0,528	0,491
		Média	139/(139+113+86) = 0,411	
		Baixa	403/(403+134+215) = 0,535	
Medida F	F = 2xPxR/(P+R)	Alta	2x0,602x0,528/(0,602+0,528) = 0,563	0,486
		Média	2x0,253x0,411/(0,253+0,411) = 0,313	
		Baixa	2x0,641x0,535/(0,641+0,535) = 0,583	
I_{FC}	$I_{FC} = 100.F/Log_2.C$			4,776

Os números de i/fla (instâncias por folha) que foram selecionados para essa avaliação foi de 20 a 50, de modo a evitar valores muito baixos e muito altos, os quais podem gerar tanto sobrestimação (*over fitting*) como subestimação (*under fitting*) da classificação.

Como podemos observar nas Tabelas 7, 8, 9 e 10 o classificador JRip apresenta número de regras bem menor que o PART, o que pode ser interessante por facilitar na interpretação. Entretanto, as demais métricas como custo, precisão e revocação são favoráveis ao classificador PART. Nota-se que apesar da revocação obtida para o JRip serem ligeiramente maiores para a classe de BAIXA produtividade (B) nas 4 técnicas, foi menor para classe de ALTA (A) e muito inferior para a classe de produtividade MÉDIA (M), tanto a revocação quanto a precisão, sendo na maioria das vezes zero.

Os gráficos na Figura 11 apresentam um panorama geral de comparação de desempenho dos classificadores PART e JRip, bem como do efeito do i/fla nas métricas avaliadas. Na maioria dos casos avaliados, tanto em termos de diferentes métodos de discretização como em i/fla, obtém-se menores custos e maiores precisões com o classificador PART. Por esses motivos e aliado ao fato do algoritmo JRip apresentar revocação e precisão quase nulos para a classificação da produtividade M, conclui-se que o algoritmo PART tem desempenho muito superior ao JRip para o caso desse banco de dados analisado, sendo, portanto, selecionado para as próximas etapas do trabalho.

6.6 Desempenho das técnicas de discretização, da conversão binária e do número i/fla

A discretização manual apresentou custos muito superiores que as 3 técnicas de discretizações automáticas. Assim, na análise comparativa das técnicas de discretização apresentadas nas Tabelas 11, 12 e 13 foi desconsiderada a discretização manual, pelo seu baixo desempenho.

De um modo geral, considerando a média de todos os valores de i/fla avaliados (últimas linhas nas Tabelas 11, 12 e 13), observou-se pequenas diferenças entre as 3 técnicas de discretização automática, mas com uma tendência de superioridade da técnica de KO+BE, conforme indicado pela maioria das métricas utilizadas. O indicador que melhor expressa o desempenho em termos de precisão, revocação (medida F) e o custo é o índice I_{FC}, pois quanto maior for a precisão e revocação e menor for o custo, maior será o I_{FC}.

As métricas de acurácia, precisão, medida F e número de regras diminuem e o custo aumenta com o aumento de i/fla. Considerando o melhor compromisso entre precisão e custo (representado por I_{FC}) e o número de regras gerado, selecionou-se i/fla de 30 (*bias trade-off*) como ideal e está indicado pela seta vermelha na vertical na Figura 13.

Essa escolha se justifica pelo conceito de *trade-off*. Trata-se de uma expressão em inglês que significa o ato de escolher

Table 4. Avaliação estatística dos atributos com características numéricas

	Atributos	Mínimo	Máximo	Média	Mediana	DP	CV(%)	Assimetria	Curtose
Informações Gerais	Planta/m	4	20	9,8	10	1,7	17	0,3	0,8
	Planta/ha	18000	153000	77635	81000	19408	25	-0,1	-0,2
	MÊS	1	12	5,3	2	5,2	98	0,5	-1,8
	ABERTURA	5	40	21,9	20	7,3	33	0	-0,7
	INI-ALGODAO	1	25	10	10	5	48	0,1	-0,8
	PROD-HIST	98	426	264,4	270	44,1	17	-0,4	0,5
	PROD-SAFRA	28	417	262,1	265	48,9	19	-0,4	0,9
Doenças	MELOI-SOLO	0	23600	391	0	1345	344	6,2	63,8
	MELOI-RAIZ	0	10880	105	0	563	538	10,8	159,5
	MELOI-TOTAL	0	34480	495	0	1757	355	7,2	92,9
	PRATY-SOLO	0	2130	56	20	159	285	7,1	64,4
	PRATY-RAIZ	0	2870	228	110	337	148	3,3	13,8
	PRATY-TOTAL	0	2990	283	140	381	134	3	11,3
	HETE-SOLO	0	1180	7	0	39	550	18,1	468,3
	ROTY-SOLO	0	13080	163	0	783	480	7,5	76,3
	ROTY-RAIZ	0	920	3	0	38	1130	19,9	436,7
	ROTY-TOTAL	0	13100	166	0	794	477	7,3	73,2
	OUTROS-SOLO	0	11350	263	50	498	190	7,8	140,3
	OUTROS-RAIZ	0	5490	49	0	200	407	15,4	347,7
	OUTROS-TOTAL	0	11470	312	70	557	179	6,5	97
Parâmetros Físicos do Solo	DS	0,64	1,71	1,25	1,23	0,17	13	0,3	-0,2
	UMID-VOL	2,7	45,5	26,7	28,1	7,5	28	-0,6	-0,1
	UMID-MAS	2,1	41,4	21,6	22,7	7,2	34	-0,4	-0,5
	DP	2,61	2,96	2,72	2,73	0,05	2	0	-0,2
	PT	35,9	76,8	53,9	54,9	6,7	12	-0,4	-0,3
	ARGILA	4	79	51	55	15	30	-0,9	0,1
	SILTE	0	32	8	6	6	76	1,2	1
	AREIA	7	91	41	38	18	44	0,8	-0,2
	AD	1	30	5	3	5	104	1,6	2,4
	CE	0,02	3,32	0,55	0,48	0,28	51	2,1	11,1
	RP10-40	0,74	9,02	2,12	1,84	1,11	52	2,2	6,2
Parâmetros Químicos do Solo	pH-AGUA	4,8	7,6	6	6	0,3	6	-0,1	1
	pH-CLORETO	4,2	7	5,2	5,2	0,3	7	0	1,3
	P	2,7	87,8	34,6	33,6	16,1	47	0,4	-0,6
	K	10	283	73,9	69	31,8	43	1,1	2,6
	Ca	1	8	4,02	4	0,98	24	0,4	0,6
	Mg	0,70	6,20	2,94	2,90	0,75	25	0,4	0,7
	Ca + Mg	0,30	1,90	1,09	1,10	0,24	22	0,2	0,6
	Al	0	0,70	0,03	0	0,10	324	3,4	11,3
	H	0,1	8,9	4	4	1,2	30	0,2	0,1
	MO	10,7	54,2	34,4	33,9	7	20	-0,1	-0,1
	SOMA-BASES	1,2	8,2	4,2	4,2	1	24	0,4	0,6
	CTC	3,7	14	8,3	8,3	1,4	17	0	-0,1
	SAT-BASES	13,1	98,4	51,5	51,6	10,6	21	0,1	1,3
	Ca/Mg	2	3,3	2,7	2,7	0,1	6	-0,2	0,2
	Ca/K	4,1	65,4	17,8	15,7	8,5	48	1,4	2,8
	Mg/K	1,7	25,5	6,6	5,8	3,1	47	1,4	2,6
	SAT-Ca	8	72,5	35,6	35,5	7,9	22	0,2	1,3
	SAT-Mg	3,7	23,9	13,3	13,3	2,6	20	0,1	1,5
	SAT-Al	0	37	1	0	3,6	350	4,6	26,6
	SAT-K	0,6	8,9	2,3	2,2	0,9	40	0,9	1,7
	SAT-H	1,6	80,4	48,1	48,4	10	21	-0,4	1,1
	Zn	0,7	19,2	6,4	6	3,1	49	0,9	1
	Cu	0,2	9,7	1,7	1,5	0,9	53	1,8	6,8
	Fe	22	212,3	80	76	27,7	35	1,3	2,5
	Mn	4,2	111,9	16,1	14,8	7,8	48	3,8	30,9
	S	5,8	60,6	13,8	12,2	5,7	41	2,4	9,6
	B	0,2	2,9	0,5	0,5	0,2	40	3	17,1
	P-res	1,2	97,2	38,8	36,5	17,5	45	0,7	0,1

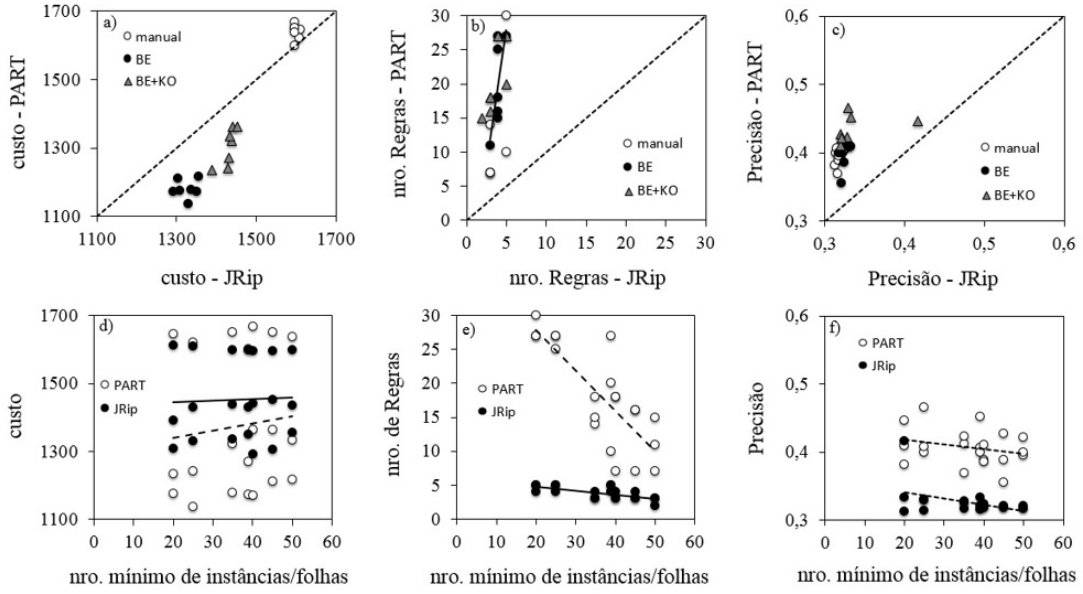

Figure 11. Comparação das métricas de custo, número de regras e precisão (a, b, c) e suas variações em função do número mínimo de instâncias por folha (d, e, f) para os classificadores PART e JRip (dados das Tabelas 7, 9 e 10).

Table 5. Discretização do atributo produtividade

Classe	Produtividade (@/há)	*instâncias (talhões)*	
		número	*percentual*
Baixa	inferior a 250	681	37,9
Média	251 a 280	489	27,2
Alta	acima de 280	629	35

Table 6. Estatística utilizando uma ponderação ao desvio interquartil de 1,5

Estatísticas - Cálculo de outliers	
Q_1	$-25,3713$
Q_3	$29,68973$
Desvio Interquartil	$84,091545$
Limite Inferior (outlier)	$-110,462845$
Limite Superior (outlier)	$113,781275$
Coef. Outliers	$1,5$
Total de outliers	28

uma coisa em detrimento de outra e muitas vezes é traduzida como "perde-e-ganha". *O trade-off* implica um conflito de escolha e uma consequente relação de compromisso (que no caso, o melhor foi entre precisão, custo e número de regras).

Entre as 3 técnicas de discretização [36], as diferenças nas métricas de avaliação de desempenho foram menos significativas do que a variação com i/fla. Entretanto, a técnica KO+BE apresentou-se, em média, levemente superior, além de apresentar uma menor dispersão dos dados quando avaliada em função de i/fla, como pode ser observado na Figura 12. Assim, este foi um dos motivos em considerar o desempenho da técnica de discretização KO+BE superior às demais.

Assim, para reforçar o fato de que a técnica KO+BE foi a melhor dentre as três técnicas de discretização, foi aplicado nestes resultados a técnica da conversão binária e observamos que as métricas de desempenho da classificação com e sem a aplicação da conversão binária (*binary split*) e utilizando as diferentes produtividades, ou seja, a real, a histórica e a média das duas, que estão apresentadas nas Tabelas 14 e 15 e na figura 14, que a conversão binária produz um aumento significativo do I_{FC} sem aumento do número de regras (há na realidade uma pequena diminuição no número de regras, em média). Na prática, com o uso da conversão binária obteve-se o melhor desempenho provavelmente devido a ocorrência de regras que operam com a não-linearidade do fenômeno. Quando o atributo foi transformado em binário, permitiu-se avaliar também intervalos no meio da escala, que muito provavelmente é decorrente da falta de linearidade.

6.7 Avaliação das regras de Classificação geradas

As 35 regras apresentadas nos quadros das figuras 16, 17 e 18, estão no formato SE-ENTÃO, e foram obtidas pelo classificador PART, discretização pelas técnicas de KO, BE e KO+BE

Table 7. Revocação, precisão, custo e número de regras geradas por validação cruzada (10-*folds*), com os classificadores PART e JRip, em função do número mínimo de instâncias por folha (i/fla), com Discretização Manual. Classes: B=baixa, M=média, A=alta produtividade

	Revocação						Precisão						Custo		Nº de Regras	
i/fla	PART			JRip			PART			JRip			PART	JRip	PART	JRip
	B	M	A	B	M	A	B	M	A	B	M	A				
20	0,615	0,08	0,515	0,834	0	0,378	0,454	0,235	0,456	0,426	0	0,511	1645	1612	30	5
25	0,614	0,106	0,515	0,844	0	0,366	0,458	0,274	0,466	0,425	0	0,517	1621	1608	27	5
30	0,667	0,078	0,514	0,833	0	0,394	0,457	0,281	0,482	0,43	0	0,516	1600	1598	10	5
35	0,74	0,008	0,493	0,846	0	0,375	0,457	0,19	0,46	0,428	0	0,521	1651	1597	14	3
40	0,731	0,022	0,467	0,858	0	0,356	0,448	0,262	0,456	0,425	0	0,528	1668	1596	7	3
45	0,721	0,033	0,471	0,858	0	0,356	0,448	0,254	0,463	0,425	0	0,528	1652	1596	7	3
50	0,721	0,033	0,486	0,855	0	0,362	0,454	0,267	0,465	0,427	0	0,523	1638	1598	7	3

Table 8. Revocação, precisão, custo e número de regras geradas por validação cruzada (10-*folds*), com os classificadores PART e JRip, em função do número mínimo de instâncias por folha (i/fla), com discretização pelo algoritmo Kononenko (KO). Classes: B=baixa, M=média, A=alta produtividade

	Revocação						Precisão						Custo		Nº de Regras	
i/fla	PART			JRip			PART			JRip			PART	JRip	PART	JRip
	B	M	A	B	M	A	B	M	A	B	M	A				
20	0,681	0,086	0,56	0,868	0,002	0,385	0,494	0,288	0,494	0,432	0,2	0,567	1331	1438	34	6
25	0,687	0,055	0,585	0,852	0	0,404	0,491	0,235	0,503	0,433	0	0,553	1322	1441	26	5
30	0,659	0,059	0,601	0,853	0	0,418	0,495	0,25	0,487	0,442	0	0,542	1339	1421	21	4
35	0,673	0,039	0,59	0,855	0	0,407	0,474	0,26	0,489	0,437	0	0,548	1378	1433	20	4
40	0,681	0,029	0,603	0,821	0	0,445	0,485	0,318	0,474	0,44	0	0,529	1379	1431	18	3
45	0,642	0,027	0,603	0,821	0	0,482	0,469	0,228	0,467	0,45	0	0,544	1420	1385	16	3
50	0,649	0,039	0,58	0,828	0	0,456	0,466	0,257	0,47	0,444	0	0,544	1421	1407	11	3

Table 9. Revocação, precisão, custo e número de regras geradas por validação cruzada (10-*folds*), com os classificadores PART e JRip, em função do número mínimo de instâncias por folha (i/fla), com discretização pelo algoritmo Better Encoding (BE). Classes: B=baixa, M=média, A=alta produtividade

	Revocação						Precisão						Custo		Nº de Regras	
i/fla	PART			JRip			PART			JRip			PART	JRip	PART	JRip
	B	M	A	B	M	A	B	M	A	B	M	A				
20	0,655	0,072	0,576	0,859	0	0,424	0,489	0,259	0,481	0,445	0	0,553	1175	1309	27	5
25	0,671	0,057	0,599	0,868	0	0,402	0,501	0,23	0,493	0,441	0	0,551	1137	1331	25	4
30	0,658	0,049	0,59	0,844	0	0,399	0,485	0,235	0,48	0,434	0	0,528	1173	1351	27	4
35	0,674	0,045	0,588	0,856	0	0,404	0,477	0,268	0,491	0,44	0	0,536	1179	1337	15	4
40	0,698	0,012	0,604	0,814	0	0,463	0,487	0,188	0,48	0,445	0	0,526	1171	1292	18	4
45	0,645	0,012	0,603	0,815	0	0,452	0,465	0,136	0,467	0,443	0	0,52	1211	1305	16	4
50	0,645	0,037	0,583	0,838	0	0,399	0,462	0,265	0,471	0,435	0	0,518	1217	1355	11	3

Table 10. Revocação, precisão, custo e número de regras geradas por validação cruzada (10-*folds*), com os classificadores PART e JRip, em função do número mínimo de instâncias por folha (i/fla), com discretização por Kononenko (KO) e Better Encoding (BE). Classes: B=baixa, M=média, A=alta produtividade

	Revocação						Precisão						Custo		Nº de Regras	
i/fla	PART			JRip			PART			JRip			PART	JRip	PART	JRip
	B	M	A	B	M	A	B	M	A	B	M	A				
20	0,692	0,106	0,598	0,862	0,008	0,42	0,514	0,289	0,535	0,445	0,235	0,57	1235	1390	27	4
25	0,699	0,11	0,606	0,853	0	0,412	0,51	0,353	0,535	0,438	0	0,55	1242	1429	27	5
30	0,67	0,086	0,636	0,862	0	0,401	0,507	0,333	0,517	0,434	0	0,564	1271	1431	20	5
35	0,656	0,041	0,652	0,849	0	0,409	0,501	0,274	0,492	0,434	0	0,549	1322	1439	18	3
40	0,655	0,025	0,642	0,816	0	0,442	0,5	0,261	0,469	0,44	0	0,52	1363	1441	18	3
45	0,658	0,07	0,59	0,83	0	0,418	0,492	0,315	0,475	0,437	0	0,521	1363	1453	16	3
50	0,711	0,033	0,596	0,821	0	0,442	0,502	0,281	0,482	0,442	0	0,52	1334	1435	15	2

Table 11. Métricas de avaliação da classificação por validação cruzada, com o classificador PART, em função de i/fla, com discretização pelo método de Kononenko (KO), utilizando a conversão binária. Classes: B=baixa, M=média, A=alta produtividade (média da real e histórica)

i/fla	Acurácia	Revocação			Precisão			Média	Custo	Nº de Regras	Medida F			Média	IFC
		B	M	A	B	M	A				B	M	A		
20	49,0%	0,59	0,235	0,615	0,535	0,364	0,509	0,469	1189	31	0,561	0,286	0,557	0,468	4,58
25	49,9%	0,645	0,182	0,631	0,547	0,339	0,514	0,467	1158	25	0,592	0,237	0,567	0,465	4,57
30	49,7%	0,595	0,25	0,617	0,534	0,387	0,514	0,478	1181	22	0,563	0,304	0,561	0,476	4,66
35	48,8%	0,625	0,197	0,607	0,528	0,352	0,505	0,462	1202	19	0,572	0,253	0,551	0,459	4,48
40	49,6%	0,644	0,188	0,62	0,526	0,359	0,519	0,468	1182	15	0,579	0,247	0,565	0,464	4,54
45	49,9%	0,615	0,202	0,643	0,52	0,368	0,53	0,473	1175	10	0,564	0,261	0,581	0,468	4,59
50	48,7%	0,604	0,182	0,639	0,509	0,376	0,505	0,463	1229	13	0,552	0,245	0,564	0,454	4,42
Média	49,4%							0,469	1188	19				0,465	4,55

Table 12. Métricas de avaliação da classificação por validação cruzada, com o classificador PART, em função de i/fla, com discretização pelo método de Better Encoding (BE), utilizando a conversão binária. Classes: B=baixa, M=média, A=alta produtividade (média da real e histórica)

i/fla	Acurácia	Revocação			Precisão			Média	Custo	Nº de Regras	Medida F			Média	IFC
		B	M	A	B	M	A				B	M	A		
20	49,0%	0,579	0,231	0,628	0,518	0,357	0,527	0,467	1185	36	0,547	0,281	0,573	0,467	4,57
25	49,8%	0,617	0,248	0,599	0,532	0,378	0,525	0,478	1169	22	0,571	0,300	0,560	0,477	4,68
30	50,1%	0,625	0,224	0,62	0,522	0,414	0,514	0,483	1196	21	0,569	0,291	0,562	0,474	4,63
35	48,8%	0,603	0,242	0,589	0,514	0,39	0,508	0,471	1218	18	0,555	0,299	0,546	0,466	4,55
40	47,3%	0,595	0,175	0,614	0,503	0,332	0,497	0,444	1250	19	0,545	0,229	0,549	0,441	4,29
45	48,7%	0,63	0,189	0,605	0,509	0,364	0,512	0,462	1219	18	0,563	0,249	0,555	0,456	4,44
50	48,6%	0,564	0,188	0,671	0,524	0,356	0,501	0,460	1216	14	0,543	0,246	0,574	0,454	4,43
Média	48,9%							0,467	1208	21				0,462	4,51

Table 13. Métricas de avaliação da classificação por validação cruzada, com o classificador PART, em função de i/fla, com discretização pelo método de KO+BE, utilizando a conversão binária. Classes: B=baixa, M=média, A=alta produtividade (média da real e histórica)

i/fla	Acurácia	Revocação			Precisão			Média	Custo	Nº de Regras	Medida F			Média	IFC
		B	M	A	B	M	A				B	M	A		
20	50,0%	0,579	0,259	0,633	0,528	0,396	0,524	0,483	1176	36	0,552	0,313	0,573	0,480	4,70
25	49,8%	0,614	0,244	0,605	0,522	0,385	0,529	0,479	1177	24	0,564	0,299	0,564	0,476	4,66
30	49,7%	0,623	0,25	0,588	0,54	0,38	0,513	0,478	1174	24	0,579	0,302	0,548	0,476	4,67
35	49,2%	0,618	0,189	0,631	0,513	0,369	0,516	0,466	1205	20	0,561	0,250	0,568	0,459	4,49
40	48,8%	0,641	0,182	0,604	0,504	0,364	0,518	0,462	1218	17	0,564	0,243	0,558	0,455	4,44
45	49,1%	0,604	0,235	0,604	0,521	0,389	0,508	0,473	1207	17	0,559	0,293	0,552	0,468	4,57
50	48,3%	0,572	0,23	0,617	0,514	0,355	0,515	0,461	1208	14	0,541	0,279	0,561	0,461	4,50
Média	49,3%							0,472	1195	22				0,468	4,58

Table 14. Métricas de desempenho (valores médios para i/fla entre 20 e 50) para a classificação das produtividades do algodoeiro (Real, Histórica e a Média das duas), utilizando PART, com diferentes técnicas de discretização (KO, BE, KO+BE), sem conversão binária

Parâmetro	**KO**			**BE**			**KO + BE**		
	Tipo de Produtividade								
	Real	**Histórica**	**Média**	**Real**	**Histórica**	**Média**	**Real**	**Histórica**	**Média**
Revocação	0,435	0,393	0,465	0,432	0,4	0,466	0,454	0,405	0,467
Precisão	0,409	0,381	0,448	0,396	0,396	0,455	0,435	0,399	0,459
Medida F	0,39	0,359	0,443	0,385	0,377	0,453	0,412	0,38	0,454
Custo	1370	1526	1239	1377	1505	1230	1304	1480	1247
IFC	3,74	3,4	4,32	3,69	3,57	4,41	3,96	3,61	4,42
Nº Regras	21	22	24	20	21	23	20	25	25

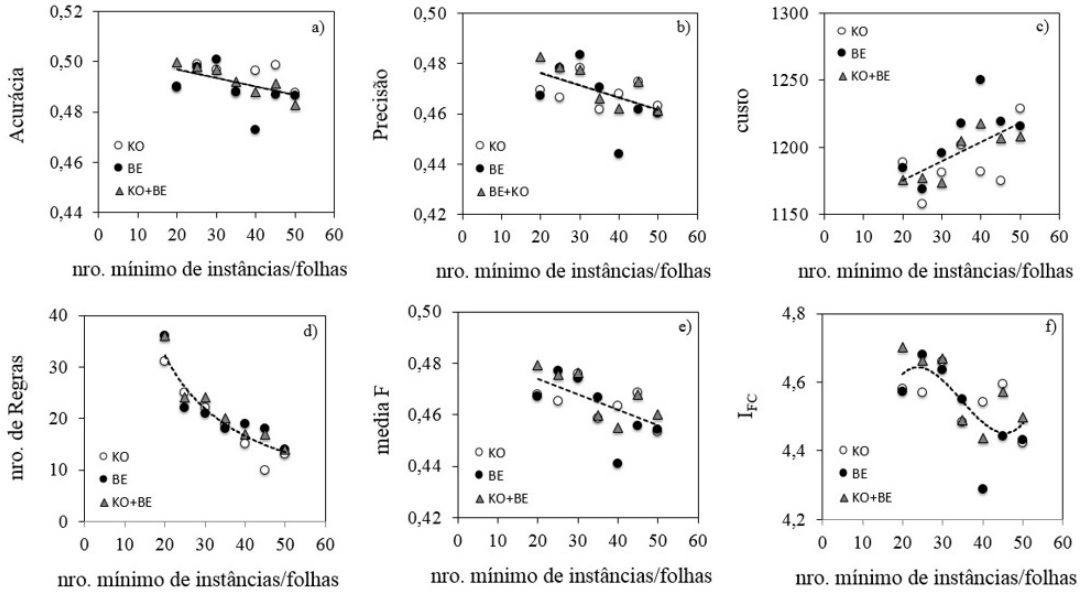

Figure 12. Variação das métricas de acurácia, precisão, custo, nro. de regras, medida F e I_{FC} com nro. mínimo de instâncias por folha (i/fla), obtidos por validação cruzada, com o classificador PART e as 3 técnicas de discretização automática (KO, BE e KO+BE).

Table 15. Métricas de desempenho (valores médios para i/fla entre 20 e 50) para a classificação das produtividades do algodoeiro (Real, Histórica e a Média das duas), utilizando PART, com diferentes técnicas de discretização (KO, BE, KO+BE) e com conversão binária

Parâmetro	KO			BE			KO + BE		
	Tipo de Produtividade								
	Real	**Histórica**	**Média**	**Real**	**Histórica**	**Média**	**Real**	**Histórica**	**Média**
Revocação	0,454	0,432	0,482	0,464	0,434	0,479	0,461	0,434	0,482
Precisão	0,444	0,427	0,469	0,458	0,432	0,467	0,451	0,431	0,471
Medida F	0,429	0,422	0,465	0,444	0,424	0,463	0,439	0,425	0,468
Custo	1295	1368	1188	1269	1376	1205	1268	1367	1193
IFC	4,15	4,05	4,51	4,31	4,07	4,52	4,26	4,08	4,58
Nº Regras	20	19	19	21	21	21	19	24	22

seguido da conversão binária e validação cruzada. O condicionante SE indica os atributos predicados (características de manejo, doenças e físico-químicas do solo) e o ENTÃO o valor do atributo alvo (produtividade).

Das 35 regras selecionadas para as 3 técnicas de discretização, 14 são para a classe de produtividade ALTA e 21 para a de BAIXA, sendo que nenhuma regra foi selecionada para a classe de produtividade MÉDIA, pois nesse caso a cobertura foi menor que 50%.

A Figura 15 apresenta uma análise quantitativa das regras geradas, do número de atributos por regra, do número de instâncias certas e do percentual de acertos, a partir das 35 regras apresentadas. Podemos observar regras com número de atributos variando de 1 a 12, mas com uma maior frequência entre 3 e 5 atributos (Fig. 15c) e coberturas (acertos) variando entre 51% a 88%, mas com a maioria na faixa de 50%-60% (Fig. 15d). Notamos que quando temos um número baixo de regras os erros são muito inferiores em relação ao número de instâncias recuperadas, por exemplo, para 8 regras com valores (44/8) de cobertura temos uma precisão de 81% pois, 44 é o número de instâncias que dispararam a regra e apenas 8 dessas 44 são o número de instâncias incorretamente classificadas pela regra, enquanto para um número superior de regras, no caso 12 regras, temos valores (66/30) de cobertura com precisão de 54% isso devido ao número de instâncias recuperadas erradas que no casso foram 30, isso equivale a quase 50% do número total de instâncias corretas. Logo, para este modelo, quanto maior o número de regras, maior o número de instâncias erradas recuperadas e consequentemente isto faz com que caia o número de acertos. Porém, vale ressaltar que este comportamento pode não ocorrer com outros tipos de problemas.

Em geral, verificou-se que o número de atributos em uma

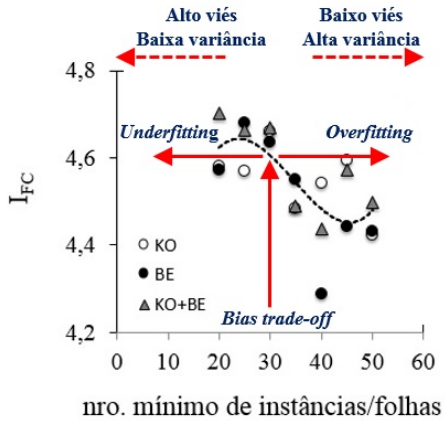

Figure 13. Melhor compromisso entre precisão, custo e o número de regras (seleção do I/fls)

regra influencia o percentual de instâncias certas (Fig. 15a), sendo que o percentual de acertos é maior quando o número de atributos envolvidos é muito baixo (por exemplo, 1 atributo) ou muito alto (por exemplo, 12 atributos). Entretanto, regras com 1 ou mais de 10 atributos "podem" ter pouco valor prático.

Outro aspecto interessante a ser observado é que o número total de instâncias certas está na faixa de 50 instâncias quando o número de atributos por regra varia de 4 a 6, aumentando para valores ao redor de 200 instâncias para número de atributos/regra maiores, entre 7 e 10 (Fig. 15b).

Uma avaliação geral das regras indica que algumas não apresentam sentido prático ou são contraditórias. Outras, entretanto, são coerentes e indicam efeitos sistemáticos dos parâmetros medidos na produtividade do algodoeiro.

Diante dos resultados com o uso da MD, temos abaixo as seguintes tendências que puderam ser observadas e algumas discursões sobre tais tendências:

- As cultivares de algodão utilizadas no plantio exercem influência na produtividade, pois apareceram como atributos predicativos em 15 regras, sendo 7 para a classe de BAIXA e 8 para a classe de ALTA produtividade. As cultivares que mais apareceram nas regras foram: FM966 LL, FM910, FMT 701, FM951 LL, FM975 WS e FM 993.

- O espaçamento entre linhas de plantio maior que 76 cm, ou seja, o espaçamento convencional está mais associado à classe de ALTA produtividade, porém, em estudos de [52] mostraram que as produtividades de algodão em caroço foram significativamente superiores no espaçamento adensado (0,45 m entre linhas de plantio).

- Solos com teores altos do micronutriente boro (aplicado entre 0,8 e 1,7 mg/dm3) apresentam ALTA produtividade, o que se deve, principalmente, ao fato de que a maior parte da área plantada está localizada na região dos cerrados, cuja maioria dos solos são naturalmente pobres em micronutrientes [53]. O boro (B) participa do transporte de carboidratos por meio da formação de complexos açúcar/borato, sendo importante na síntese de proteínas [54]. O algodoeiro é uma das plantas mais exigentes em B, acumulando de 170 g/ha a 680 g/ha [55]. O fornecimento regular desse nutriente favorece o florescimento e a frutificação, com reflexos positivos no aumento da produtividade e da qualidade das fibras. Durante o florescimento, a deficiência de B pode inviabilizar a germinação do grão de pólen, tornando os óvulos estéreis e impedindo a formação das sementes e das fibras advindas delas. A consequência final é a redução da produtividade [56]. É importante informar que em [57], com relação aos teores de Ca, Mg e a soma de Ca + Mg, quanto maiores os valores dessas variáveis, maiores foram os rendimentos de caroço de algodoeiro e que o teor de enxofre no solo também mostrou diferença entre os grupos de maior e de menor produção do algodoeiro, sendo que os maiores valores de produção corresponderam aos maiores valores de enxofre. Isso ressalta a importância da aplicação desse nutriente na cultura do algodoeiro.

- Plantio Direto na palha influenciou positivamente na produtividade, uma vez que 3 regras na classe de ALTA produtividade apresentaram esse atributo (SIST-PLANTIO=plantio direto), porém, quando dá para fazer plantio direto se faz, mas muita das vezes a palha não dá conta. Estudos mostram que uma das alternativas mais efetiva e eficiente de conservação do solo é o uso do plantio direto [58]. Este se fundamenta em programas de rotação de culturas, pelo cultivo em terreno coberto por palha e/ou plantas em crescimento e ausência de preparo do solo por tempo indeterminado [59]. Nos estudos de [60] afirma que a estabilidade da produção é ampliada com este tipo de plantio em comparação aos métodos tradicionais de manejo de solo. Ressaltamos também que o cultivo do algodoeiro em sistema plantio direto aumenta o estoque de carbono no solo, incrementa o teor de nitrogênio e ainda faz aumentar a produtividade em comparação com o sistema de preparo convencional do solo. Foi o que demonstrou um estudo realizado ao longo de nove anos por cientistas da Embrapa Algodão de Campina Grande (PB) no Cerrado brasileiro [61] .

- A ausência do fungo *Fusarium* foi o único atributo presente em uma regra com cobertura maior que 50%, que nesse caso foi associado à classe de ALTA produtividade, sendo um exemplo claro de causa e efeito, teve impacto mais relevante que os nematoides. Segundo [62] "a disseminação da doença ocorre principalmente através de solos infectados e as perdas são muito maiores quando os solos estão infestados por nematoides". Em [7] é salientado que são vários os fatores que afetam negativamente a cultura do algodoeiro, dentre os problemas fitossanitários, destacam-se os fitonematoides, causadores de danos econômicos, pois prejudicam a absorção de água e nutrientes pela planta, causando a diminuição da produtividade. Esses vermes ao fixarem o estilete nas plantas para se alimentarem, deixam uma "porta aberta" para entrada de outros patógenos. Assim, fungos e bactérias conseguem infectar as plantas com mais facilidade e consequentemente aumentam a área de infecção nos cultivos [62]. De acordo com [63], no cenário internacional a porcentagem de dano, e, consequentemente, de perdas de produção por nematóides, é

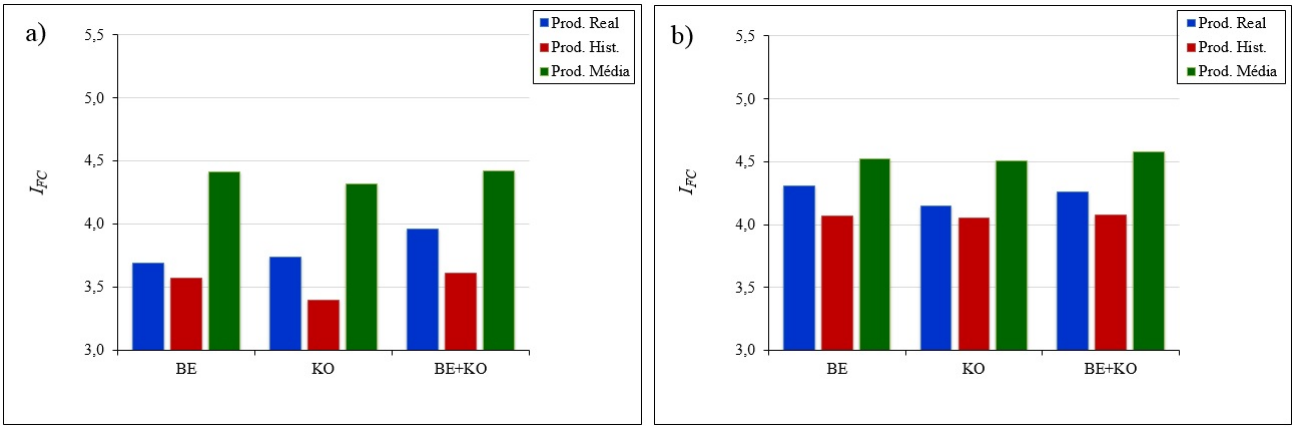

Figure 14. Valores médios de I_{FC} (100*F/Log_2.Custo) relativos aos desempenhos dos classificadores das produtividades Real, Histórica e Média das duas, utilizando PART, para as 3 técnicas de discretização avaliadas (KO, BE, KO+BE), com conversão binária (b) e sem (a)

mais elevada em condições de climas tropicais e subtropicais (14,6%), comparativamente com regiões de climas temperados (8,8%). No Brasil, as perdas causadas por nematoides são insignificantes em áreas de exploração agrícola recente [64]. Em estudos de [57], a presença de *murcha de fusarium*, causada por *Fusarium oxysporum f. sp. vasinfectum*, em plantas de algodoeiro foi confirmada em 24 talhões dos 1.162 amostrados, perfazendo cerca de 2% de toda a área amostrada. Isso demonstra sua grande concentração em determinada região, inclusive com danos expressivos à cultura, constatando-se que 100% das áreas com incidência do fungo apresentavam histórico de baixa produtividade.

Observamos valores de precisão entre 50% a 87% e de acertos até 90%. Verificamos que um mesmo atributo pode estar presente em uma regra de classificação para ALTA ou BAIXA produtividade, o que é plenamente aceitável e consistente. Entretanto, alguns atributos parecem influenciar mais fortemente a classe ALTA ou BAIXA. É o caso, por exemplo do micronutriente boro (B: 0,8 a 1,7 mg/dm3) e o sistema de plantio (SIST-PLANTIO=semeadura direta) que influenciam positivamente na produtividade do algodoeiro.

Uma síntese numérica dos dados observamos o número de vezes que cada atributo apareceram nas classes ALTA e BAIXA (e Total). Verificamos que o grupo de atributos de Manejo aparece com muito mais frequência (72 vezes), seguido do grupo de atributos Químicos (38 vezes) e Físicos do Solo (29 vezes) e por fim os de doenças (apenas 4 vezes).

Dentre os atributos de manejo, a ABERTURA, o tempo de cultivo de algodão (INI-ALGODÃO) na área e a CULTIVAR utilizada foram os mais frequentes, tendo, portanto, grande influência na produtividade do algodoeiro. Outros atributos como o espaçamento entre linhas (ESP), o sistema de plantio (SIST-PLANTIO) e o preparo do solo (PREP-SOLO), mostraram-se também de grande relevância. Em estudos de [57], o espaçamento entre linhas no sistema de cultivo em monocultura ou isolado varia de 0,76 a 0,90 m e o número de plantas por metro varia entre 7 a 10. As cultivares mais utilizadas são aquelas tolerantes a herbicidas e resistentes a lepidópteros.Um exemplo são as cultivares de algodão Bt que é o algodão que recebeu genes da bactéria de solo *Bacillus thuringiensis* [65] que produz proteínas tóxicas a determinados tipos de insetos, principalmente da ordem Lepidoptera [66]. Já no Sistema em sucessão de culturas o espaçamento entre fileiras de algodoeiro varia de 0,45 a 0,90 m, com predomínio do espaçamento de 0,76 m. Normalmente, no início do período de semeadura, utilizam-se espaçamentos maiores e, no final, espaçamentos menores. A frequência de uso do espaçamento de 0,45 m é baixa. O número de plantas por metro varia entre 7 a 10.

Dentre os atributos químicos do solo mais frequentes nas regras, os micronutrientes: zinco, ferro e boro, os macronutrintes: cálcio e magnésio são os únicos com cobertura maior que 50%. Embora exigidos em menores quantidades, os micronutrientes são tão importantes para a nutrição e o crescimento das plantas quanto os macronutrientes [56].

De um modo geral os atributos relacionados às doenças avaliadas (nematoides e fusarium) impactaram pouco nas regras de classificação. Isso se deve provavelmente ao intenso manejo fitossanitário do algodoeiro, com a aplicação de grandes volumes de inseticidas, nematicida e fungicidas, reduzindo bastante os efeitos dessas doenças na perda de produção, comparativamente com outros tipos de atributos. Em estudos de [57] devido ao modelo em uso do sistema de rotação de culturas, a pesar de ser pouco praticada em Mato Grosso, é comum se observar sinais claros de degradação dos atributos físicos, químicos e biológicos, levando à queda do potencial produtivo desses solos e, também, à elevação dos custos de produção devido ao maior uso de fertilizantes, inseticidas, fungicidas e herbicidas. A prática da rotação de culturas é indispensável para assegurar a sustentabilidade da produção de algodão.

Os atributos físicos apareceram com destaque também. Entretanto a maioria deles, como a umidade a base de massa (UMID-MAS) e a base de volume (UMID-VOL) e a den-

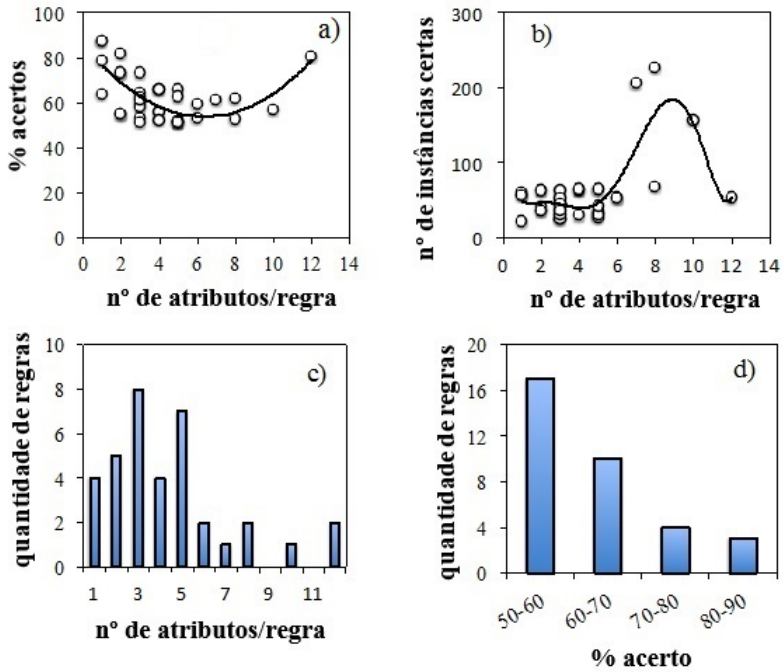

Figure 15. Correlação do número de atributos por regra com a porcentagem de acertos (a), o número de instâncias certas (b) e a quantidade de regras (c); e do percentual de acertos em função da quantidade de regras (d). Dados obtidos a partir dos quadros de 5 a 7.

sidade das partículas, são inerentes do tipo de solo, porém, constituem-se como atributos que podem ser modificados por manejo (apesar de não haver irrigação), que não é o caso dos atributos químicos e das doenças. Vale destacar também que os solos com melhor qualidade física, em termos de teor de argila, são em geral selecionados para o cultivo do algodoeiro, o que reduz o impacto desses atributos nas regras. Entretanto, a resistência à penetração (RP10-40), parâmetro relacionado a compactação do solo, e que pode ser modificada por manejo (subsolagem do solo), apareceu em algumas regras, com boa cobertura. Em estudos de [57], o parâmetro do solo que mais influenciou na produtividade foi a resistência à penetração (RP), que está diretamente relacionada à compactação do solo, ou seja, à limitação ao enraizamento da cultura. Para todas as classes de solo avaliadas houve aumento da RP no perfil para os pontos de BAIXA produtividade, indicando influência significativa da compactação na redução da produtividade. Os maiores valores de RP foram obtidos na camada de 20 a 30 cm de profundidade para todas as classes texturais e as classes que apresentaram os maiores valores de RP no perfil (maior compactação) foram as dos solos arenosos e argilosos.

7. Conclusões e trabalhos futuros

Apesar de ser um trabalho inicial, mostrou o potencial desse tipo de estudo e indica a necessidade de mais trabalhos com a aplicação de outras técnicas e algoritmos de MD. Algumas das conclusões e recomendações desse trabalho são relativamente óbvias para um especialista, ou seja, um hipotético especialista analisando os dados de forma convencional encontraria resultados parecidos (o que ocorreu nas discussões das tendências observadas). Por exemplo, um especialista observaria facilmente o mesmo resultado sobre a análise do boro no solo. Entretanto, deve-se destacar que a MD, permite a manipulação de um volume razoável de dados de uma forma rápida e evidencia algumas regras não óbvias mesmo para um especialista. Entretanto, sendo um trabalho pioneiro no estudo da produtividade do algodão no Brasil com o uso de técnicas de AM, empregou apenas uma das técnicas de MD, sendo possível se explorar muitos outros *softwares* e técnicas que poderão extrair mais conhecimentos do BD desse trabalho.

Algumas sugestões para os trabalhos futuros são o uso de remoção de atributos interdependentes, a utilização de 5 classes de intervalos para a produtividade (BAIXA, MÉDIA-BAIXA, MÉDIA, MÉDIA-ALTA, ALTA) e também ampliação da base de dados com informações de outros estados produtores de algodão como a Bahia e Goiás. Adicionalmente, a estratégia de discretização e conversão binária viabilizam a aplicação de algoritmos de mineração de regras gerais de associação como, por exemplo, o APRIORI. Variáveis como clima e flutuações de preço não fazem parte da base de dados disponibilizada para esse estudo. Entretanto são variáveis importantíssimas para ajudar a explicar a produtividade e deveriam ser incluídas em estudos futuros.

Deve-se destacar que a equação proposta neste trabalho para o índice I_{FC} é uma novidade para a área, pois neste formato integra vários indicadores. Utilizou-se aqui o logaritmo na base 2, porém, em trabalhos futuros poderiam ser comparados os desempenhos com o uso de outras bases logarítmicas.

R(1): **SE** Abertura_anos_agricultura6=7 **E** Início_anos_algodão > 2 **E** u_% > 8,43 **E** SOM_BASES > 2,65 **E** B_MG_DM3 = (0,82-1,67] **ENTÃO** Produtividade = **Alta** (97,0/33,0)

R(2): **SE** Início_anos_algodão 6=4 **E** Espaçamento > 85 **E** Início_anos_algodão 6=3 **E** Cultivar 6=FMT_701 **E** MF_incidência = Sem_Presença **E** Semeadura_direto_na_palha3_Semeadura = Semeadura_direto_na_palha **E** RP10_40cm_MPa≥2,68 **E** u_%6=(8,43-15,71] **ENTÃO** Produtividade = **Alta** (367,0/141,0)

R(3): **SE** Cultivar 6=FM_966_LL **E** Cultivar = FM_910 **E** Não_utilizado_Milho_cultura_cobertura = Soja **ENTÃO** Produtividade = **Alta** (39,0/14,0)

R(4): **SE** Cultivar 6=FM_966_LL **E** Início_anos_algodão 6=4 **E** Cultivar 6=FMT_701 **E** Silte_% > 11,0041 **E** FE_MG_DM36 = (75,95-155,15] **ENTÃO** Produtividade = **Alta** (127,0/62,0)

R(5): **SE** Cultivar6=FM_966_LL **E** Cultivar6=FM_951_LL **E** Silte_%≥11,0041 **E** Outros_Solo_mais_Raiz_log > 2,26 **E** FE_MG_DM3 = (57,75-75,95] **ENTÃO** Produtividade = **Alta** (53,0/26,0)

R(6): **SE** Abertura_anos_agricultura6=7 **E** Início_anos_algodão > 2 **E** u_%≥8,43 **ENTÃO** Produtividade = **Baixa** (86,0/23,0)

R(7): **SE** Abertura_anos_agricultura6=7 **E** Início_anos_algodão > 2 **E** SOM_BASES≥2,65 **ENTÃO** Produtividade = **Baixa** (74,0/28,0)

R(8): **SE** Abertura_anos_agricultura6=7 **E** Início_anos_algodão≥2 **ENTÃO** Produtividade= **Baixa** (44,0/8,0)

R(9): **SE** Início_anos_algodão6=4 **E** fi_%≥22,14 **E** Início_anos_algodão6=3 **E** Soja_Sorgo_Cultura_Anterior 6=Soja_Milho **E** Outros_Solo_mais_Raiz_log ≥2,26 **E** FE_MG_DM3 6=(57,75-75,95] **ENTÃO** Produtividade = **Baixa** (100,0/47,0)

R(10): **SE** Cultivar6=FM_966_LL **E** Início_anos_algodão = 4 **ENTÃO** Produtividade = **Baixa** (73,0/33,0)

R(11): **SE** Cultivar 6=FM_966_LL **E** Cultivar 6=FM_951_LL **E** Com_Subsolagem_preparo_de_solo 6=Sem_Subsolagem **E** u_%6 = (8,43014-15,717574] **ENTÃO** Produtividade = **Baixa** (54,0/24,0)

R(12): **SE** Cultivar6=FM_975_WS **E** u_%6=(8,43014-15,717574] **ENTÃO** Produtividade = **Baixa** (66,0/30,0)

Figure 16. Regras com cobertura maior que 50% para validação cruzada, classificador PART, conversão binária e discretização por BE

R(13): **SE** Início_anos_algodão > 2 **E** u_% > 8,43 **E** ZN_MG_DM3 < 13,45 **E** B_MG_DM3 = (0,825-1,675] **ENTÃO** Produtividade = **Alta** (96,0/33,0)

R(14): **SE** Dp_g_cm3≥2,68 **E** SATPOR_CA > 26,95 **E** Início_anos_algodão6=3 e 4 **ENTÃO** Produtividade = **Alta** (47,0/22,0)

R(15): **SE** Dp_g_cm3>2,68 **E** ZN_MG_DM3<13,45 **E** Cultivar6=FMT_701 **E** fi_%6=(14,088349-22,147452] **E** RP10_40cm_MPa ≥2,68 **E** Semeadura_direto_na_palha3_Semeadura = Semeadura_direto_na_palha **E** Espaçamento≥85 **ENTÃO** Produtividade = **Alta** (337,0/132,0)

R(16): **SE** ZN_MG_DM3 < 13,45 **E** CTC > 7,05 **E** Cultivar 6=FMT_701 **E** FE_MG_DM3 ≥57,75 **ENTÃO** Produtividade = **Alta** (123,0/59,0)

R(17): **SE** Início_anos_algodão > 2 **E** u_%≥8,43 **ENTÃO** Produtividade = **Baixa** (86,0/23,0)

R(18): **SE** Início_anos_algodão > 2 **E** ZN_MG_DM3 < 13,45 **E** SATPOR_CA≥26,95 **E** Silte_%≥11,0041 **E** Dp_g_cm3 > 2,68 **E** Plantas_ha≥89000 **ENTÃO** Produtividade = **Baixa** (86,0/35,0)

R(19): **SE** Início_anos_algodão≥2 **ENTÃO** Produtividade = **Baixa** (64,0/8,0)

R(20): **SE** ZN_MG_DM3 < 13,45 **E** Início_anos_algodão≥3 ou 4 **E** Dp_g_cm3 > 2,68 **E** Silte_%≥11,0041 **E** Espaçamento≥85 **ENTÃO** Produtividade = **Baixa** (60,0/29,0)

R(21): **SE** ZN_MG_DM3 < 13,45 **E** fi_% < 22,14 **E** Início_anos_algodão > 3 ou 4 **E** Cultivar6=FM_951_LL **E** RP10_40cm_MPa > 2,68 **ENTÃO** Produtividade = **Baixa** (68,0/33,0)

R(22): **SE** Dp_g_cm3≥2,68 **ENTÃO** Produtividade = **Baixa** (75,0/16,0)

R(23): **SE** ZN_MG_DM3 < 13,45 **E** Cultivar6=FM_993 **E** CTC≥7,05 **E** Com_Subsolagem_preparo_de_solo 6=Sem_Subsolagem **E** Soja_Sorgo_Cultura_Anterior6=Algodão **ENTÃO** Produtividade = **Baixa** (48,0/18,0)

Figure 17. Regras com cobertura maior que 50% para validação cruzada, classificador PART, conversão binária e discretização por KO

Agradecimentos

Agradecemos ao Instituto Mato-Grossense do Algodão (IMAmt) e ao Dr. Rafael Galbieri, pesquisador do IMAmt, por disponibilizar o Banco de Dados e fornecer todas as informações solicitadas, durante a discussão dos resultados.

Ao meu orientador Dr. Carlos Manoel Pedro Vaz pela imensa competência com que orientou-me neste trabalho. Sem suas orientações, apoio e confiança em todo o caminho percorrido até aqui, nada disso seria possível.

Ao meu coorientador Ednaldo José Ferreira pelas suas orientações a respeito das análises de cada resultado gerado.

À Embrapa - Empresa Brasileira de Pesquisa Agropecuária (Embrapa Instrumentação de São Carlos-SP), pela oportunidade de capacitação profissional.

Contribuição dos autores

- Alexandra Virgínia Valente da Silva: Fez pesquisa, explorou os dados no software de Aprendizado de Máquina (WEKA), elaborou gráficos e tabelas, análisou e interpretou os resultados, escreveu o artigo e fez uma revisão substancial.
- Carlos Manoel Pedro Vaz: Contribuiu do início ao fim com reuniões para analisarmos e debatermos cada resultado obtido, elaborou alguns gráficos e tabelas e fez uma revisão substancial.
- Ednaldo José Ferreira: Contribuiu do início ao fim com reuniões para debatermos e analisarmos cada resultado obtido e fez uma revisão substancial.
- Rafael Galbieri: Disponibilizou o Banco de Dados e forneceu informações importantíssimas para enriquecer os resultados.

R(24): **SE** Início_anos_algodão > 2 **E** u_% > 8,43 **E** MGCMOLCDM3 > 0,65 **E** B_MG_DM3 = (0,825-1,675] **ENTÃO** Produtividade = **Alta** (99,0/34,0)

R(25): **SE** Início_anos_algodão > 2 **E** MGCMOLCDM3 > 0,65 **E** Cultivar 6=FM_966_LL **E** ZN_MG_DM3 < 13,45 **E** Cultivar 6=FMT_701 **E** SATPOR_CA > 26,95 **E** Espaçamento > 85 **E** RP10_40cm_MPa ≥2,68 **E** Semeadura_direto_na_palha3_Semeadura = Semeadura_direto_na_palha **E** Outros_Solo_mais_Raiz_log 6=(1,181806-2,269356] **E** Classe_textural 6=Argiloso **E** Outros_Solo_mais_Raiz_log > 1,18 **ENTÃO** Produtividade = **Alta** (66,0/13,0)

R(26): **SE** Início_anos_algodão > 2 **E** MGCMOLCDM3 > 0,65 **E** ZN_MG_DM3 < 13,45 **E** Cultivar6=FM_966_LL **E** Cultivar 6=FMT_701 **E** CA_E_MG < 3,45 **E** RP10_40cm_MPa ≥2,68 **E** REL_CAMG > 2,55 **E** Cultivar6=FM_910 **E** Espaçamento > 85 **ENTÃO** Produtividade = **Alta** (275,0/119,0)

R(27): **SE** Início_anos_algodão > 2 **E** MGCMOLCDM3 > 0,65 **E** ZN_MG_DM3 < 13,45 **E** Cultivar6=FM_966_LL **E** Cultivar = FM_910 **ENTÃO** Produtividade = **Alta** (82,0/40,0)

R(28): **SE** Não_utilizado_Milho_cultura_cobertura6=Milheto **ENTÃO** Produtividade = **Alta** (33,0/12,0)

R(29): **SE** Início_anos_algodão > 2 **E** u_% > 8,43 **E** MGCMOLCDM3 > 0,65 **E** Início_anos_algodão > 4 **E** ZN_MG_DM3 < 13,45 **E** Cultivar6=FM_966_LL **E** fi_% < 22,14 **E** K_MG_DM3≥71,75 **ENTÃO** Produtividade = **Baixa** (130,0/62,0)

R(30): **SE** Início_anos_algodão > 2 **E** u_% > 8,43 **E** MGCMOLCDM3 > 0,65 **E** Início_anos_algodão ≥4 **ENTÃO** Produtividade = **Baixa** (88,0/36,0)

R(31): **SE** Início_anos_algodão > 2 **E** u_%≥8,43 **ENTÃO** Produtividade = **Baixa** (86,0/23,0)

R(32): **SE** Início_anos_algodão≥2 **ENTÃO** Produtividade = **Baixa** (68,0/8,0)

R(33): **SE** MGCMOLCDM3 > 0,65 **E** Cultivar6=FM_966_LL **E** SATPOR_CA≥26,95 **ENTÃO** Produtividade = **Baixa** (78,0/33,0)

R(34): **SE** Não_utilizado_Milho_cultura_cobertura 6=Milheto **E** HCMOLCDM3 = (2,45-6,55] **E** Com_Subsolagem_preparo_de_solo6=Sem_Subsolagem **ENTÃO** Produtividade = **Baixa** (59,0/23,0)

R(35): **SE** MGCMOLCDM3 < 0,95 **ENTÃO** Produtividade = **Baixa** (41,0/16,0)

Figure 18. Regras com cobertura maior que 50% para validação cruzada, classificador PART, conversão binária e discretização por KO+BE

Referências

[1] WITTEN, I.; FRANK, E. *Data Mining: Practical Machine Learning Tools and Techniques with Java Implementations*. Burlington / Massachusetts: Elsevier, 2009.

[2] ABRAPA. *Brasil finaliza colheita de uma nova safra recorde de algodão*. 2020. Disponível em: ⟨https://www.abrapa.com.br/Paginas/Not%C3%ADcias%20Abrapa.aspx?noticia=522⟩. Acesso em: 25 de abr. de 2021.

[3] GUIAJEANSWEAR. *Mato Grosso lidera produção de algodão no Brasil*. 2019. Disponível em: ⟨https://guiajeanswear.com.br/noticias/mato-grosso-lidera-producao-de-algodao-no-brasil-veja-ranking/⟩. Acesso em: 25 de abr. de 2021.

[4] NOTICIASAGRICOLAS. *Câmara Setorial do Algodão e Derivados (CSAD) aponta produção e exportações recordes*. 2019. Disponível em: ⟨https://www.noticiasagricolas.com.br/noticias/algodao/238097-camara-setorial-do-algodao-e-derivados-csad-aponta-producao-e-exportacoes-recordes.html#.YIXr6pBKg2w⟩. Acesso em: 25 de abr. de 2021.

[5] ABRAPA. *Algodão no Brasil*. 2021. Disponível em: ⟨https://www.abrapa.com.br/Paginas/dados/algodao-no-brasil.aspx⟩. Acesso em: 25 de abr. de 2021.

[6] EMBRAPA. *Série Desafios do Agronegócio Brasileiro (NT3) – Produto: Algodão – Parte 01: Caracterização e Desafios Tecnológicos*. 2019. Disponível em: ⟨https://www.alice.cnptia.embrapa.br/alice/bitstream/doc/1109655/1/SerieDesafiosAgronegocioBrasileiroNT3Algodao⟩. Acesso em: 25 de abr. de 2021.

[7] CULTIVAR, R. *Medidas de manejo contra nematoides em algodão*. 2020. Disponível em: ⟨https://www.grupocultivar.com.br/noticias/medidas-de-manejo-contra-nematoides-em-algodao#:~:text=S%C3%A3o%20v%C3%A1rios%20os%20fatores%20que,causando%20a%20diminui%C3%A7%C3%A3o%20da%20produtividade.⟩ Acesso em: 27 de abr. de 2021.

[8] EMBRAPA. *Algodão de alta produtividade e qualidade superior de fibra é apresentado na Tecnoshow Comigo*. 2018. Disponível em: ⟨https://www.embrapa.br/busca-de-noticias/-/noticia/33250842/algodao-de-alta-produtividade-e-qualidade-superior-de-fibra-e-apresentado-na-tecnoshow-comigo⟩. Acesso em: 09 de abr. de 2018.

[9] PAUTA, E. em. *Como a Mineração de Dados pode auxiliar o agronegócio?* 2017. Disponível em: ⟨https://excelenciaempauta.com.br/mineracao-de-dados-auxilia-o-agronegocio/#:~:text=Esse%20processo%20se%20denomina%20data,a%20agroind%C3%BAstria%2C%20entre%20outros%20aspectos.⟩ Acesso em: 25 de abr. de 2021.

[10] NEELAVENI, N.; RAJESWARI, S. Data mining in agriculture - a survey. *International Journal of Modern Computer Science* — Revista da Faculdade de Serviço Social da UERJ, Rio de Janeiro, v. 4, n. 4, p. 104–107, 2016. ISSN 2320-7868. Disponível em: ⟨http://www.ijmcs.info/current_issue/IJMCS160835.pdf⟩. Acesso em: 28 nov. 2018.

[11] AHAMED, A. T. M. S. et al. Applying data mining techniques to predict annual yield of major crops and recommend planting different crops in different districts in bangladesh. *International Conference on Software Engineering, Artificial Intelligence, Networking and Parallel/Distributed Computing (SNPD)*, IEEE/ACIS 16th, Takamatsu, Japan, p. 1–6, 2015. Disponível em: ⟨http://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=

7176185&isnumber=7176160⟩. Acesso em: 03 de jul. de 2018.

[12] VRIESMANN, L. M. et al. Análise de resultados obtidos por técnicas de inteligência artificial na mineração de dados de produtividade do solo. *Revista Brasileira de Agrocomputação*, Ponta Grossa-PR, DEINFO/UEPG, v. 2, n. 1, p. 11–18, 2004. Disponível em: ⟨http://www.agrocomputacao.deinfo.uepg.br/junho_2004/Arquivos/RBAC_Artigo_02.pdf⟩. Acesso em: 03 de jul. de 2018.

[13] GARCIA, E.; JÚNIOR, L. C. Classificação de fatores que mais impactam a produtividade da cana-de-açúcar usando mineração de dados. XSBIAGRO - X Congresso Brasileiro de Agroinformática, 2015. Disponível em: ⟨http://eventos.uepg.br/sbiagro/2015/anais/SBIAgro2015/pdf_resumos/16/16_ederson_garcia_73.pdf⟩. Acesso em: 30 de nov. de 2017.

[14] SILVA, C. F.; RODRIGUES, C. T.; MONTEIRO, M. V. B. Inteligência artificial - uso de regras de associação para descoberta de conhecimento na produtividade de açaí no estado do amapá. SULCOMP - Congresso Sul Brasileiro de Computação, 2010. ISSN 2359-2656. Disponível em: ⟨http://periodicos.unesc.net/sulcomp/article/view/297/304⟩. Acesso em: 30 de nov. de 2018.

[15] GANDHI, N.; ARMSTRONG, L. *A review of the application of data mining techniques for decision making in agriculture*. 2016a. Disponível em: ⟨https://ieeexplore.ieee.org/document/7917925/⟩. Acesso em: 30 de nov. de 2018.

[16] GANDHI, N.; ARMSTRONG, L. *Rice Crop Yield Forecasting of Tropical Wet and Dry Climatic Zone of India Using Data Mining Techniques*. 2016b. 357-363 p. Acesso em: 30 de nov. de 2018.

[17] DEY, U. K.; MASUD, A. H.; UDDIN, M. N. Rice yield prediction model using data mining. *ECCE: International Conference on Electrical, Computer and Communication Engineering, Cox's Bazar*, p. 321–326, 2017. Disponível em: ⟨http://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=7912925&isnumber=7912861⟩. Acesso em: 30 de nov. de 2018.

[18] DELERCE, S. et al. Assessing weather-yield relationships in rice at local scale using data mining approaches. *PLoS ONE 11(8):e0161620. doi:10.1371/journal.pone.0161620*, August 25, 2016, 2016. Disponível em: ⟨http://journals.plos.org/plosone/article?id=10.1371/journal.pone.0161620⟩. Acesso em: 30 de nov. de 2018.

[19] HAMMER, R. G. *Modelagem da Produtividade da Cultura da cana-de-açucar por meio do uso de técnicas de mineração de dados*. Dissertação (Mestrado) — Universidade de São Paulo - Campus São Carlos, São Carlos-SP, 2016.

[20] PATEL, H.; PATEL, D. A brief survey of data mining techniques applied to agricultural data. *International Journal of Computer Applications*, v. 95, n. 9, p. 6–8, 2014. ISSN 0975 – 8887. Disponível em: ⟨https://www.researchgate.net/publication/269669148_A_Brief_survey_of_Data_Mining_Techniques_Applied_to_Agricultural_Data⟩. Acesso em: 03 de jul. de 2018.

[21] AGRAWAL, R.; IMIELINSKI, T.; SWAMI, A. Mining association rules between sets of items in large databases. ACM Sigmod Conference, IBM Almaden Research Center / 650 Harry Road, San Jose, CA 95120, EUA, 1993. Disponível em: ⟨http://www.rakesh.agrawal-family.com/papers/sigmod93assoc.pdf⟩. Acesso em: 27 de abr. de 2021.

[22] DAMACENO, L. K. *Introdução a Machine Learning utilizando o Weka*. 2017. Disponível em: ⟨https://medium.com/cwi-software/introdu%C3%A7%C3%A3o-a-machine-learning-utilizando-o-weka-c38388514c40⟩. Acesso em: 28 de abr. de 2021.

[23] GARCIA, M. *Weka Data Mining – Software Open Source em Java*. 2020. Disponível em: ⟨https://www.cetax.com.br/blog/weka-data-mining-open-source/⟩. Acesso em: 28 de abr. de 2021.

[24] QUINLAN, R. C4.5: Programs for machine learning. Kluwer Academic Publishers, Boston. Manufactured in The Netherlands, p. 235–240, 1993. Disponível em: ⟨http://server3.eca.ir/isi/forum/Programs%20for%20Machine%20Learning.pdf⟩. Acesso em: 30 de nov. de 2018.

[25] COHEN, W. W. Fast effective rule induction. *Machine Learning Proceedings 1995: Proceedings of the Twelfth International Conference on Machine Learning*, Tahoe City, California, July 9–12, 1995, p. 115–123, 1995. Disponível em: ⟨http://www.cs.utsa.edu/~bylander/cs6243/cohen95ripper.pdf⟩. Acesso em: 30 de nov. de 2018.

[26] FURNKRANZ, J.; WIDMER, G. Incremental reduced error pruning. *Machine Learning Proceedings 1994: Proceedings of the Eleventh International Conference*, Machine Learning Proceedings, Rutgers University, New Brunswick, NJ, July 10–13, 1994, p. 70–77, 1994. Disponível em: ⟨https://www.sciencedirect.com/science/article/pii/B9781558603356500179⟩. Acesso em: 03 de jul. de 2018.

[27] MELO, I. R. S. et al. Avaliação do desempenho do algoritmo jrip na classificação do diagnóstico de doenças cardíacas. *IV Congresso Nacional de Pesquisa e Ensino em Ciências - CONAPESC*, Campina Grande - PB de 22 a 24 de agosto, 2019. ISSN 2525-6696. Disponível em: ⟨http://editorarealize.com.br/artigo/visualizar/56596⟩. Acesso em: 26 de abr. de 2021.

[28] SHANNON, C. E. A mathematical theory of communication. In: ____. July, October, 1948, 1948. v. 27, p. 379–423, 623–656. Disponível em: ⟨http://math.harvard.edu/~ctm/home/text/others/shannon/entropy/entropy.pdf⟩. Acesso em: 27 de set. de 2018.

[29] PAVIOTTI, J. R. *Considerações sobre o conceito de entropia na teoria da informação*. 118 p. Dissertação

(Mestrado) — Universidade Estadual de Campinas/Faculdade de Tecnologia, Limeira-SP, 2019.

[30] HOSOKAWA, E. O. *Técnica de Árvore de Decisão em Mineração de Dados*. 2011. Disponível em: ⟨http://www.fatecsp.br/dti/tcc/tcc0003.pdf⟩. Acesso em: 28 de out. de 2018.

[31] GRUNWALD, P. *The Minimum Description Length Principle*. Cambridge, MA: MIT Press (14 março 2016), 2016.

[32] RISSANEN, J. Modeling by the shortest data description. *Automatica*, v. 14, p. 465–471, 1978. Disponível em: ⟨https://msol.people.uic.edu/ECE531/papers/Modeling%20By%20Shortest%20Data%20Description.pdf⟩. Acesso em: 26 de abr. de 2021.

[33] GRUNWALD, P. Introducing the minimum description length principle. Amsterdam, v. 2, p. 20, 2007. ISSN 9780262256292. Disponível em: ⟨file:///C:/Users/Lenovo/Downloads/The_Minimum_Description_Length_Principle.pdf⟩. Acesso em: 26 de abr. de 2021.

[34] GRUNWALD, P.; ROOS, T. Minimum description length revisited. Cornell University - arXiv.org ¿ stat ¿ arXiv: 1908.08484v2, v. 2, p. 38, 2019. Disponível em: ⟨https://arxiv.org/abs/1908.08484⟩. Acesso em: 26 de abr. de 2021.

[35] RODRIGUES, E. S. C. *Teoria da informação e adaptatividade na modelagem de distribuição de espécies*. 137 p. Tese (Doutorado) — Escola Politécnica da Universidade de São Paulo, São Paulo, 2012.

[36] FAYYAD, U. M.; IRANI, K. B. Multi-interval discretization of continuous-valued attributes for classification learning. *IJCAI*, 1993. Disponível em: ⟨https://www.semanticscholar.org/paper/Multi-Interval-Discretization-of-Continuous-Valued-Fayyad-Irani/1dc53b91327cab503acc0ca5afb9155882b717a5⟩. Acesso em: 27 de out. de 2018.

[37] BARBOSA, A. A. V. *Entropia de Shannon e propriedades topológicas de redes funcionais do cérebro humano sob efeito de Ayahuasca*. Tese (Doutorado) — Universidade Federal do Rio Grande do Norte, 2015. Disponível em: ⟨https://repositorio.ufrn.br/jspui/handle/123456789/21549⟩. Acesso em: 02 de nov. de 2018.

[38] OTTERBACH, J. *MDLP and Conditional Inference - Strategies to prevent Decision Trees from overfitting*. 2016. Disponível em: ⟨http://jotterbach.github.io/2016/12/10/RecursivePartitioning/⟩. Acesso em: 07 de nov. de 2018.

[39] FAYYAD, U. M. *On the Induction of Decision Trees for Multiple Concept Learning (Machine Learning)*. Digitalizado em 25 mar. 2011, 263. p. Tese (Doutorado) — Universidade de Michigan, 1991. Disponível em: ⟨https://dl.acm.org/citation.cfm?id=144532⟩. Acesso em: 30 de nov. de 2018.

[40] KONONENKO, I. On biases in estimating multi-valued attributes. *IJCAI'95 Proceedings of the 14th international joint conference on Artificial intelligence*, Morgan Kaufmann Publishers Inc. San Francisco, CA, USA ©1995, Montreal, Quebec, Canada — August 20 - 25, 1995, v. 2, p. 1034–1040, 1995. Disponível em: ⟨https://dl.acm.org/citation.cfm?id=1643034⟩. Acesso em: 03 de ago. de 2018.

[41] ISMAIL, M. K.; CIESIELSKI, V. n empirical investigation of the impact of discretization on common data distributions. *Design and application of hybrid intelligent systems*, IOS Press Amsterdam, The Netherlands, The Netherlands ©2003, Department of Computer Science RMIT University- Melbourne, VIC 3001, Australia, p. 692–701, 2003. Disponível em: ⟨https://dl.acm.org/citation.cfm?id=998117⟩. Acesso em: 04 de set. de 2018.

[42] KIRA, K.; RENDELL, L. A. A practical approach to feature selection. *ML92 Proceedings of the ninth international workshop on Machine learning*, Aberdeen, Scotland, United Kingdom, p. 249–256, 1992. Disponível em: ⟨http://sci2s.ugr.es/keel/pdf/algorithm/congreso/kira1992.pdf⟩. Acesso em: 03 de ago. de 2018.

[43] PETER, C.; BEALE, R. *Affect and Emotion in Human-Computer Interaction: From Theory to Applications*. Springer, Berlin, Heidelberg, 1ª edição: Alta Books, 2008.

[44] GALBIERI, R. et al. *Nematoides fitoparasitas do algodoeiro nos cerrados brasileiros: Biologia e medidas de controle*. Cuiabá (MT): Boletim de P&D - nº3, maio, 2016.

[45] CARVALHO, A. C. P. L. F. de et al. *Inteligência Artificial: Uma abordagem de Aprendizagem de Máquina*. Rio de Janeiro: LTC, 2021.

[46] REZENDE, S. O. *Sistemas Inteligentes: fundamentos e aplicações*. Barueri-SP: MANOLE, 2003.

[47] GARCIA, S. et al. A survey of discretization techniques: Taxonomy and empirical analysis in supervised learning. *IEEE Transactions on Knowledge and Data Engineering*, v. 25, n. 4, p. 734–750, 2013. ISSN 1041-4347. Disponível em: ⟨http://doi.ieeecomputersociety.org/10.1109/TKDE.2012.35⟩. Acesso em: 03 de ago. de 2018.

[48] HAN, J.; KAMBER, M. *Data Mining: Concepts and Techniques*. San Francisco-CA/EUA: Morgan Kaufmann, 2006.

[49] FERREIRA, D. F. *Estatística Básica*. [S.l.]: UFLA, 2009. 663 p.

[50] PRATI, R. C.; BATISTA, G. E. A. P. A.; MONARD, M. C. Curvas roc para avaliação de classificadores. 2006. Disponível em: ⟨http://conteudo.icmc.usp.br/pessoas/gbatista/files/ieee_la2008.pdf⟩. Acesso em: 04 de set. de 2018.

[51] MEDRI, W. *Análise exploratória de dados*. Londrina-PR, 2011.

[52] ASMUS, G. L.; GALBIERI, R. Densidade populacional e distribuição espacial de meloidogyne incognita e rotylenchulus reniformis em algodoeiro em sistema de plantio adensado. Editado por Cláudia R. Dias-Arieira, Instituto Mato-Grossense do Algodão, Av. Rubens de Mendonça, 157, Sala 100, 78008-000 Cuiabá (MT) Brasil, v. 37(3-4), p. 6, 2013. Disponível em: ⟨https://ainfo.cnptia.embrapa.br/digital/bitstream/item/98644/1/Guilherme-Nematologia-Brasileira-2013.pdf⟩. Acesso em: 27 de jul. de 2020.

[53] LOPES, A. S. *Solos sob cerrados: características, propriedades e manejo*. 2. ed. Piracicaba: Associação Brasileira para Pesquisa da Potassa e do Fosfato, 1984. 162 p.

[54] MARSCHNER, H. *Mineral Nutrition of Higher Plants*. [S.l.: s.n.], 1995. 889 p.

[55] ROCHESTER, I. J. Nutrient uptake and export from an australian cotton field. *Nutrient Cycling in Agroecosystems*, Springer Holanda, v. 77, n. 3, p. 213–223, 2007. ISSN 1385-1314. Disponível em: ⟨https://www.infona.pl/resource/bwmeta1.element.springer-7e28f543-80d1-350d-b510-abf17c09ba88⟩. Acesso em: 25 de jun. de 2018.

[56] CARVALHO, M. S. *Nutrição e Adubação do Algodoeiro com Micronutrientes*. Campina Grande, PB, 2007. 1-17 p.

[57] GALBIERI, R. et al. *Áreas de produção de algodão em Mato Grosso: nematoides, murcha de fusarium, sistemas de cultivo, fertilidade e física de solo.* Embrapa Agropecuária Oeste, 2014. 1-16 p.

[58] FAGERIA, N. K.; STONE, L. F. Produtividade de feijão no sistema plantio direto com aplicação de calcário e zinco. *Pesquisa Agropecuária Brasileira (PAB)*, Embrapa Sede, Secretaria de Pesquisa e Desenvolvimento (SPD), Brasil, Embrapa Arroz e Feijão, Santo Antônio de Goiás, GO, v. 39, n. 1, p. 73–78, 2004. ISSN 1678-3921. Disponível em: ⟨https://www.scielo.br/pdf/pab/v39n1/19587.pdf⟩. Acesso em: 12 de set. de 2018.

[59] HERNANI L. C.; SALTON, J. C. *Manejo e conservação do solo*. Embrapa Agropecuária Oeste, 1998. 26-50 p.

[60] CRUZ, J. C. et al. Plantio direto e sustentabilidade do sistema agrícola. Informe Agropecuário, Belo Horizonte, v. 22, n. 208, p. 13–24, 2001. ISSN 0100-3364. Disponível em: ⟨file:///C:/Users/Lenovo/Downloads/Plantio-direto-3.pdf⟩. Acesso em: 10 de nov. de 2018.

[61] EMBRAPA. *Plantio direto do algodoeiro aumenta estoque de carbono no solo em 20%*. 2019. Disponível em: ⟨https://www.embrapa.br/busca-de-noticias/-/noticia/45923956/plantio-direto-do-algodoeiro-aumenta-estoque-de-carbono-no-solo-em-20⟩. Acesso em: 23 de mar. de 2019.

[62] INOVADORES, A. *Fusarium e nematoides: qual a sua relação?* 2019. Disponível em: ⟨https://agro.genica.com.br/2019/11/28/fusarium-e-nematoides-qual-a-sua-relacao/⟩. Acesso em: 28 de abr. de 2021.

[63] NICOL, J. M. et al. Current nematode threats to world agriculture. In: ____. 1. ed. [S.l.]: Springer Holanda, 2011. cap. 2, p. 21–43.

[64] ASMUS, G. L.; GALBIERI, R. Principais espécies de nematoides do algodoeiro no brasil. In: ____. 3. ed. editores técnicos: Rafael Galbieri e Jean Louis Belot - Cuiabá (MT): Boletim de P&D, 2016. cap. 1, p. 11–36.

[65] BALDIN, E. L. L. et al. O uso de plantas transgênicas resistentes a insetos no brasil. In: *Inovações em manejo fitossanitário*. Botucatu - SP — Brasil: FEPAF, 2017. p. 62–79.

[66] AGRONOMICAS, B. P. *Algodão Bt*. 2020. Disponível em: ⟨https://boaspraticasagronomicas.com.br/noticias/algodao-bt/⟩. Acesso em: 28 de abr. de 2021.