UNIVERSIDADE DE SÃO PAULO

Instituto de Ciências Matemáticas e de Computação

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

Alexandra Virginia Valente da Silva

Dissertação de Mestrado do Programa de Mestrado Profissional em
Matemática, Estatística e Computação Aplicadas à Indústria (MECAI)

SERVIÇO DE PÓS-GRADUAÇÃO DO ICMC-USP

Data de Depósito:

Assinatura: _____

Alexandra Virginia Valente da Silva

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

Dissertação apresentada ao Instituto de Ciências Matemáticas e de Computação – ICMC-USP, como parte dos requisitos para obtenção do título de Mestra – Mestrado Profissional em Matemática, Estatística e Computação Aplicadas à Indústria. *EXEMPLAR DE DEFESA*

Área de Concentração: Matemática, Estatística e Computação

Orientador: Dr. Carlos Manoel Pedro Vaz

USP – São Carlos
Dezembro de 2018

Ficha catalográfica elaborada pela Biblioteca Prof. Achille Bassi
e Seção Técnica de Informática, ICMC/USP,
com os dados inseridos pelo(a) autor(a)

S586c Silva, Alexandra Virginia Valente da
Classificação baseada em regras para estudo da
produtividade do algodão no estado do Mato Grosso /
Alexandra Virginia Valente da Silva; orientador
Carlos Manoel Pedro Vaz. -- São Carlos, 2018.
95 p.

Dissertação (Mestrado - Programa de Pós-Graduação
em Mestrado Profissional em Matemática, Estatística
e Computação Aplicadas à Indústria) -- Instituto de
Ciências Matemáticas e de Computação, Universidade
de São Paulo, 2018.

1. . I. Vaz, Carlos Manoel Pedro, orient. II.
Título.

Alexandra Virginia Valente da Silva

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

Dissertation submitted to the Institute of Mathematics
and Computer Sciences – ICMC-USP – in accordance
with the requirements of the Professional Master's
Program in Mathematics Statistics and Computing
Applied to Industry, for the degree of Master in Science.
EXAMINATION BOARD PRESENTATION COPY

Concentration Area: Mathematics, Statistics and
Computing

Advisor: Dr. Carlos Manoel Pedro Vaz

**USP – São Carlos
December 2018**

Este trabalho é dedicado aos meus amados pais, Raimunda Cáritas Valente da Silva e Francisco Soares da Silva, pela imensa dedicação em criar eu e meus irmãos, por todos os ensinamentos, pela confiança e amor, vocês sempre foram exemplo de força, coragem e honestidade na minha vida.

AGRADECIMENTOS

Primeiramente agradeço a Deus pela vida e por todas as graças a mim concedidas, dando-me força suficiente para concluir mais esta etapa de minha carreira acadêmica.

Agradeço a Rainha do Brasil, Nossa Senhora Aparecida, a quem nunca me desaparaou.

Aos meus pais Raimunda e Francisco, aos meus irmãos Cássia, Silvia, Francisco Jr. e Ricardo. Aos meus avós Neide Picanço e José, a minha tia querida Dirce Guarany e a memória de minha amada e inesquecível vó Virgília Soares, que viveu por muitos anos com a gente, sempre transmitindo luz e amor, mas teve que partir deixando um enorme vazio. À todos vocês meus familiares, sou eternamente grata por tudo que sou, pelas conquistas e pela felicidade. Obrigada pelo amor e incentivo sempre.

Ao professor Dr. José Camilo Ramos de Souza e a professora Dr^a Lucélida de Fátima Maia da Costa pela contribuição do meu ingresso no mestrado no Estado de São Paulo.

Agradeço aos meus professores do mestrado do ICMC/USP pela compreensão quando mais precisei, pois algumas das vezes eu não conseguia estar presente nas aulas por conta da distância que eu me encontrava, pois tinha que cumprir meu trabalho na cidade de Manaus-AM e ao mesmo tempo me deslocar até o Estado de São Paulo para estudar, agradeço a meus mestres por todos os seus ensinamentos, fica aqui meu reconhecimento ao lado de quem transpira sabedoria, meu respeito e admiração pela capacidade e o dom no ensino da Ciência, inibindo sempre a vaidade em prol da simplicidade e eficiência, muito obrigada: Dr. Antônio Castelo Filho, Dr. Francisco Louzada Neto, Dr. Fabrício Simeoni de Sousa, Dr. Seiji Isotani, Dr. Francisco Aparecido Rodrigues, Dr. Cláudio Fabiano Motta Toledo, Dr^a Ellen Francine Barbosa, Dr^a Maristela Oliveira dos Santos e Dr. André Carlos Ponce de Leon Ferreira de Carvalho.

Ao meu orientador Dr. Carlos Manoel Pedro Vaz e ao meu coorientador Dr. Ednaldo José Ferreira pela imensa competência com que orientaram este trabalho. Sem suas orientações, apoio e confiança em todo o caminho percorrido até aqui, nada disso seria possível.

Aos colegas de mestrado do ICMC/USP que já concluíram ou que ainda vão concluir, pela ajuda neste árduo percurso acadêmico, dentre eles em especial: Luís Carlos Otte Júnior (amigão), Rodrigo Rossetto Gati, Edilson da Silva Guimarães e Sama Rouhani.

Aos professores da banca examinadora, pelas sugestões e correções que muito contribuíram para a versão final desta dissertação.

Ao meu grande amigo parintinense João filho por me dar dicas na primeira versão do meu projeto de pesquisa e por sempre tirar minhas dúvidas em relação a redação deste trabalho.

À minha maravilhosa amiga parintinese Geovanna Gomes pelas suas palavras de incentivo sempre.

As escolas onde trabalhei durante o mestrado, pela compreensão em me liberar quando eu precisava me ausentar para fazer provas.

Aos meus queridos alunos pela compreensão das minhas ausências quando era necessário.

Aos meus cunhados Jackyson Vieira e Wallace pela amizade, incentivo e pelas caronas de carro para o aeroporto de Manaus nas madrugadas de viagens com destino a São Carlos-SP.

Aos amigos que conquistei nesse período no Instituto de Ciências Matemáticas e Computação, pela amizade, apoio, carinho e momentos de descontração: Alan Vinicius, Mágda, Victor, João Carlos Batista, João Cavalcanti Costa, João Paulo Guardabaxo Siqueira, Fabiana Goes, Gabriela Volpi e Bruna Soares.

Às minha queridas Claudia Colletti e Laura Andreolli, amigas sancarlense que conquistei nos últimos meses de 2017, obrigada pela amizade meninas.

Enfim, a todos que de alguma forma contribuíram para a realização deste trabalho.

*“Seja quem você for,
seja qualquer posição social que você tenha na vida,
no nível mais alto ao mais baixo,
tenha sempre como meta muita força,
muita determinação e sempre faça tudo com muito amor e com muita fé em Deus,
que um dia você chega lá.
De alguma maneira você chega lá.”
(Ayrton Senna da Silva)*

RESUMO

SILVA, A. V. V. DA. **Classificação baseada em regras para estudo da produtividade do algodão no estado do Mato Grosso**. 2018. 95 p. Dissertação (Mestrado – Mestrado Profissional em Matemática, Estatística e Computação Aplicadas à Indústria) – Instituto de Ciências Matemáticas e de Computação, Universidade de São Paulo, São Carlos – SP, 2018.

Contexto: O Brasil tem grande experiência no desenvolvimento de tecnologias para a cadeia produtiva do algodão, sendo um player mundial importante nesse segmento atualmente. 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 atualmente. Embora a cadeia do algodão tenha atingido um alto nível tecnológico, há ainda muitos desafios e problemas a serem vencidos, uma vez que nos últimos anos tem-se verificado uma estagnação na produtividade, apesar do aporte cada vez maior de tecnologias como cultivares transgênicas, máquinas, insumos e gestão de dados mais eficientes. Nesse contexto as técnicas de mineração de dados oferecem excelente oportunidade de se avaliar esse problema, uma vez os dados desse setor são bem organizados, como é o caso do Banco de Dados organizado pelo Instituto Matogrossense do Algodão (IMAmt) e que foi disponibilizado para esse estudo. **Problema:** O problema alvo desse trabalho é entender as causas da estagnação da produtividade do algodoeiro, ou seja, dos fatores que estão influenciando negativamente na produção de algodão no cerrado no estado de Mato Grosso. **Solução:** A solução proposta é se utilizar técnicas de mineração de dados em um BD complexo, contendo informações da produtividade e fatores do sistema de produção que influenciam na produção, para a extração de conhecimento e geração de recomendações para os cotonicultores do Estado. **Objetivo:** Aplicar técnicas de Mineração de Dados, mais especificamente classificação baseada em regras em um BD de produção de algodão no estado de Mato Grosso para identificar os principais fatores, como as práticas de manejo, as variáveis do solo e a incidência de doenças, que estão afetando a produtividade do algodão e consequentemente limitando o aumento da produtividade. **Estado-da-arte e Método proposto:** Diferentemente de outros trabalhos baseados em classificação de regras de associação que tem foco no resultado preditivo (acurácia), nesse trabalho a ênfase é da extração de regras de associação que tenham relevância prática e que possam auxiliar o agricultor na tomada de decisão. Na etapa de pré-processamento realizou-se a seleção e transformação de atributos e identificação de *outliers*. Os atributos numéricos foram discretizados utilizando 4 métodos, sendo 3 automáticos (técnica de Kononenko, Better Encoding e a combinação das duas) e um manual. Na etapa de modelagem de classificação, os algoritmos de regras utilizados foram o PART e o JRip e a conversão

binária de atributos foi também avaliada. O desempenho foi avaliado pela precisão, revocação, custo e a combinação delas no índice IFC ($100 * F / \log_2 \text{custo}$). **Resultados:** A avaliação das métricas indicou o melhor desempenho para o classificador PART, com a discretização pela técnica de Kononenko combinada com Better Encoding, seguida pela conversão binária. A análise das regras, com auxílio de especialistas da área, possibilitou a seleção daquelas mais relevantes e dos atributos que mais impactam na produtividade. Dentre esses atributos, destacam-se: a cultivar de algodão, o tempo da abertura e de início do cultivo de algodão na área, o sistema de plantio e de preparo do solo. Os principais atributos físico-químicos do solo foram os micronutrientes zinco, ferro e boro e os macronutrientes cálcio e magnésio e a resistência do solo à penetração. A ausência do fungo *fusarium* impactou positivamente na produtividade. Em geral, os atributos relacionados ao Manejo do algodoeiro apresentaram maior relevância na composição das regras de classificação, seguidos dos atributos Químicos e Físicos do Solo e dos atributos relacionados à doenças.

Palavras-chave: Produtividade do algodão, Mineração de dados, Classificação baseada em regras..

ABSTRACT

SILVA, A. V. V. DA. **Classification based on rules for the study of productivity in the state of Mato Grosso** . 2018. 95 p. Dissertação (Mestrado – Mestrado Profissional em Matemática, Estatística e Computação Aplicadas à Indústria) – Instituto de Ciências Matemáticas e de Computação, Universidade de São Paulo, São Carlos – SP, 2018.

Context: Brazil has great experience in developing technologies for the cotton sector, becoming an important world player presently. The cotton production has spread in the Brazilian cerrado stimulating and making viable a highly technical, efficient and profitable production, and leading the country from a condition of importer to a major exporter, currently. Although the cotton productive chain has reached a high technological level, there are many challenges and problems to be overcome, since the productivity is stagnant despite the use of more efficient transgenic cultivars, machines, agrochemicals and data management. In this context data mining offers excellent opportunities to evaluate this problem, since data in this sector is very well organized, as is the case of the Data Base (DB) of the Mato Grosso Cotton Institute, which is available for this study. **Problem:** The target problem of this study is to understand the causes of such productivity stagnation, that is, the factors influencing negatively the Mato Grosso cerrado cotton production. **Solution:** The proposed solution is to apply data mining to a complex DB, having productivity information and the factors that influences production, for extracting knowledge and generate recommendations for the cotton farmers in the State. **Purpose:** Apply data mining technique, more specifically the classification based on rules, in a DB of cotton production in the State of Mato Grosso to identify the main factors, as management practices, soil variables and diseases that are affecting cotton productivity and limiting the productivity increasing. **State-of-Art and Proposed Method:** Unlike other works based in classification based on rules with focus in prediction (accuracy) this work has emphasis in association rules with practical relevance to help farmers in decision-making. In the pre-processing step, it was performed selection, transformation and *outliers* identification. Numerical attributes were discretized using 4 methods, 3 automatic (techniques of Kononenko, Better Encoding and the combination of both) and one manual. In the classification step, the rule algorithms used were PART and JRip and binary split was also evaluated. The performance was evaluated by precision, recall, cost and the combination of them in an index called IFC ($100 * F / \log_2 \text{cost}$). **Results:** The evaluation of these metrics has indicated the best performance for the classifier PART, with discretization by the technique of Kononenko combined with Better Encoding, and using binary split. The rules were analyzed with the aid of specialists for the selection of the most relevant ones and select the attributes that mostly impact the productivity. Among

these attributes are: the cotton cultivar, time of opening the area and beginning of cotton cultivation, soil and growing system used. The main soil physico-chemical attributes were the micronutrients zinc, iron, bore, the macronutrients calcium and magnesium, and soil penetration resistance. The absence of the fungi *fusarium* has impacted positively in productivity. In general, attributes related to cotton management were more relevant in the classification rules composition, followed by soil Chemical and Physical attributes and lastly the ones related to diseases.

Keywords: Cotton productivity, Data mining, Classification based on rules..

LISTA DE ILUSTRAÇÕES

Figura 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, MG, MA), evidenciando a tendência de estagnação nos últimos 10 anos (linha pontilhada)	27
Figura 2 – Esquema do processo de descoberta de conhecimento em bases de dados (KDD)	33
Figura 3 – Exemplo ilustrativo da tarefa de Classificação de dados, utilizando parte dos dados da base de algodão, avaliado nesse trabalho	37
Figura 4 – Diagrama ilustrativo da estratégia <i>hold out</i> para avaliação de desempenho do modelo de classificação baseado em regras	42
Figura 5 – Ilustração do conceito de ponto de corte	46
Figura 6 – Exemplo de discretização do atributo idade	47
Figura 7 – Fluxograma geral de um método típico de discretização	48
Figura 8 – Fluxograma do algoritmo de discretização MDL	50
Figura 9 – Características do método MDL	50
Figura 10 – Exemplo de ponto de corte provisório para o atributo pH em água . . .	51
Figura 11 – Transformação baseada em rótulos	53
Figura 12 – Áreas amostradas (pontos vermelhos) em diferentes núcleos de produção de fazendas de algodão no Estado de Mato Grosso	58
Figura 13 – Histogramas de frequência das produtividades do algodoeiro na safra (PROD-SAFRA) e a produtividade histórica (PROD-HIST) (a) e a correlação entre essas 2 produtividades	61
Figura 14 – Esquema do procedimento utilizado para a geração de regras de associação	62
Figura 15 – Tela inicial do software WEKA para mineração de dados	64
Figura 16 – Tela do software WEKA mostrando as abas necessárias para a realização da etapa de discretização	66
Figura 17 – Janela do WEKA para alterar os pesos da matriz de custo	69
Figura 18 – Gráfico de <i>box plot</i> dos resíduos entre as produtividades histórica e da safra (a) e a estatística (b), utilizando uma ponderação ao desvio interquartil de 1,5	72

Figura 19 – 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 11, 12 e 13).	74
Figura 20 – Variação das métricas de acurácia, precisão, custo, nro. de regras, medida F e I_{FC} com nro. mínimo de instâncias por folha (I/F), obtidos por validação cruzada, com o classificador PART e as 3 técnicas de discretização automática (KO, BE e KO+BE).	76
Figura 21 – Valores médios de I_{FC} ($100 \cdot F / \log_2(\text{Custo})$) 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).	78
Figura 22 – 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.	81

LISTA DE QUADROS

Quadro 1 – Etapas da implementação do algoritmo PART (partial decision trees)	40
Quadro 2 – Etapas do processo de discretização	48
Quadro 3 – Síntese dos trabalhos sobre utilização de técnicas de MD com foco na avaliação dos fatores que mais influenciam na produtividade das culturas agrícolas	54
Quadro 4 – Melhores condições de classificação, considerando as 3 classes de produtividade (Baixa, Média e Alta).	78
Quadro 5 – Regras com cobertura maior que 50% para validação cruzada, classificador PART, conversão binária e discretização por BE.	78
Quadro 6 – Regras com cobertura maior que 50% para validação cruzada, classificador PART, conversão binária e discretização por KO.	79
Quadro 7 – Regras com cobertura maior que 50% para validação cruzada, classificador PART, conversão binária e discretização por KO+BE.	79

LISTA DE TABELAS

Tabela 1	– Matriz de confusão de um classificador	43
Tabela 2	– Matriz de confusão para um problema com duas classes	44
Tabela 3	– 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 Banco de Dados	59
Tabela 4	– Descrição de parâmetros físicos e químicos do solo utilizados no trabalho	60
Tabela 5	– Descrição de atributos gerais e de doenças do algodoeiro utilizados no trabalho	61
Tabela 6	– Avaliação estatística dos atributos com características numéricas	63
Tabela 7	– Avaliação estatística dos atributos com características numéricas	64
Tabela 8	– Limites de classes da discretização manual dos atributos numéricos realizada com informações fornecidas por especialistas da área	65
Tabela 9	– 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)	68
Tabela 10	– Cálculos da precisão (P), revocação (R), medida F e índice I_{FC} para o exemplo de classificação apresentado na Tabela 10c, utilizando o algoritmo PART	70
Tabela 11	– Revocação, precisão, custo e número de regras geradas por validação cruzada (10- <i>folds</i>), com os classificadores PART e JRip, em função do número mínimo de instâncias por folha (I/F), com discretização manual. Classes: B=baixa, M=média, A=alta produtividade.	72
Tabela 12	– Revocação, precisão, custo e número de regras geradas por validação cruzada (10- <i>folds</i>), com os classificadores PART e JRip, em função do número mínimo de instâncias por folha (I/F), com discretização pelo algoritmo Better Encoding (BE). Classes: B=baixa, M=média, A=alta produtividade.	73
Tabela 13	– Revocação, precisão, custo e número de regras geradas por validação cruzada (10- <i>folds</i>), com os classificadores PART e JRip, em função do número mínimo de instâncias por folha (I/F), com discretização por Kononenko (KO) e Better Encoding (BE). Classes: B=baixa, M=média, A=alta produtividade.	73

Tabela 14 – Métricas de avaliação da classificação por validação cruzada, com o classificador PART, em função de I/F, com discretização pelo método de Kononenko (KO), utilizando a conversão binária. Classes: B=baixa, M=média, A=alta produtividade (média da real e histórica).	75
Tabela 15 – Métricas de avaliação da classificação por validação cruzada, com o classificador PART, em função de I/F, 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).	75
Tabela 16 – Métricas de avaliação da classificação por validação cruzada, com o classificador PART, em função de I/F, com discretização pelo método de KO+BE, utilizando a conversão binária. Classes: B=baixa, M=média, A=alta produtividade (média da real e histórica).	75
Tabela 17 – Métricas de desempenho (valores médios para I/F 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.	77
Tabela 18 – Métricas de desempenho (valores médios para I/F 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.	77
Tabela 19 – Descrição dos atributos das regras geradas por validação cruzada, discretização por BE e uso da conversão binária, para a classe ALTA de produtividade do algodoeiro. Números entre parênteses nos atributos indicam o número de vezes que aparece na regra.	81
Tabela 20 – Descrição dos atributos das regras geradas por validação cruzada, discretização por BE e uso da conversão binária, para a classe BAIXA de produtividade do algodoeiro. Números entre parênteses nos atributos indicam o número de vezes que aparece na regra.	82
Tabela 21 – Descrição dos atributos das regras geradas por validação cruzada, discretização por KO e uso da conversão binária, para a classe ALTA de produtividade do algodoeiro.	82
Tabela 22 – Descrição dos atributos das regras geradas por validação cruzada, discretização por KO e uso da conversão binária, para a classe BAIXA de produtividade do algodoeiro.	83
Tabela 23 – Descrição dos atributos das regras geradas por validação cruzada, discretização por KO+BE e uso da conversão binária, para a classe ALTA de produtividade do algodoeiro. Números entre parênteses nos atributos indicam o número de vezes que aparece na regra.	83

Tabela 24 – Descrição dos atributos das regras geradas por validação cruzada, discretização por KO+BE e uso da conversão binária, para a classe BAIXA de produtividade do algodoeiro. Números entre parênteses nos atributos indicam o número de vezes que aparece na regra.	84
Tabela 25 – Síntese do número de vezes que determinado atributo apareceu nas 36 regras apresentadas nos Quadros 5, 6 e 7, nas classes de Alta e Baixa produtividade, separados por tipo de atributo (Físicos e Químicos do solo, Doenças e Manejo).	85

SUMÁRIO

1	INTRODUÇÃO	25
1.1	Contexto e Motivação	25
1.2	Objetivos	30
1.3	Estrutura do Trabalho	30
2	REVISÃO DA LITERATURA	31
2.1	Fundamentos de Mineração de Dados	31
2.1.1	<i>O processo de busca de conhecimento em banco de dados</i>	31
2.1.2	<i>Tarefas da Mineração de Dados</i>	35
2.1.3	<i>Classificador baseado em Regras</i>	38
2.1.3.1	<i>Algoritmo PART</i>	39
2.1.3.2	<i>Algoritmo JRip</i>	40
2.1.4	Estratégias para treinamento, validação e teste	41
2.1.4.1	<i>Holdout</i>	41
2.1.4.2	<i>Método de validação cruzada (10-folds-cross validation)</i>	42
2.1.4.3	<i>Matriz de confusão</i>	43
2.1.5	Técnica de pré-processamento: Discretização	46
2.1.5.1	<i>Métodos de discretização baseados em entropia</i>	48
2.1.5.1.1	<i>Técnica de discretização MDL (descrição de comprimento mínimo)</i>	49
2.1.5.1.2	<i>Técnica de discretização Better Encoding (BE)</i>	50
2.1.5.1.3	<i>Método de discretização Kononenko (KO)</i>	51
2.1.5.2	<i>Estratégia para gerar regras com conversão binária (Binary Split)</i>	52
2.2	Trabalhos relacionados	53
3	METODOLOGIA	57
3.1	Material e Métodos	57
3.1.1	O Banco de Dados	57
3.1.2	Procedimento metodológico utilizado	61
3.1.2.1	<i>Ferramenta utilizada</i>	64
3.1.2.2	<i>Discretização dos atributos</i>	64
3.1.2.3	<i>Algoritmos para extração de regras de classificação</i>	66
3.1.2.4	<i>Abordagem para treinamento e teste dos modelos de classificação</i>	67
3.1.2.5	<i>Avaliação sensível ao custo (Cost-sensitive evaluation)</i>	67

3.1.2.6	<i>Outros indicadores de desempenho utilizados</i>	69
4	RESULTADOS E DISCUSSÕES	71
4.0.1	<i>Análise de outliers dos atributos alvo (resíduo entre produtividades histórica e na safra)</i>	71
4.0.2	<i>Análise comparativa dos classificadores PART e JRip</i>	72
4.0.3	<i>Desempenho das técnicas de discretização, da conversão binária e do número I/F</i>	74
4.0.4	<i>Avaliação das regras de associação geradas</i>	77
5	CONCLUSÕES	87
	REFERÊNCIAS	89

INTRODUÇÃO

1.1 Contexto e Motivação

O Brasil é o quinto maior produtor mundial de algodão, terceiro maior exportador e o primeiro em produtividade em áreas não irrigadas, tendo uma história de pioneirismo, empreendedorismo e pesquisas tecnológicas para chegar a essas posições no panorama internacional da cotonicultura. Em termos de produção, o Brasil, com cerca de 6,4% da produção mundial, fica atrás apenas da Índia (25,6%), China (20,2%), EUA (15,6%) e Paquistão (8,5%), de acordo com dados da International Cotton Advisory Committee publicados no site da Associação Brasileira dos Produtores de algodão, na safra 2016/2017 ([ABRAPA, 2017](#)).

Por possuir uma enorme quantidade de terras cultiváveis e um clima favorável, a atividade agrícola sempre teve um papel relevante no cenário nacional, em especial a cotonicultura que é uma atividade extremamente importante para o país, gerando emprego, renda e desenvolvimento.

A produção comercial do algodão no Brasil teve início nos estados do Nordeste, sendo o Maranhão o primeiro grande produtor, que em 1970 exportou para a Europa as primeiras sacas do produto, principalmente para a Inglaterra. Mais tarde, 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 de algodão. A crise da cotonicultura brasileira 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 Mato Grosso.

A área plantada de algodão no Brasil diminuiu de cerca de 4 milhões de hectares na década de 70 para uma área oscilando em torno de 1 milhão de hectares nos últimos 15 anos. Entretanto, a partir do final da década de 90 houve mudanças no padrão tecnológico

do sistema de produção, proporcionando incrementos significativos à produtividade, de modo que o rendimento atual de fibra é cerca de três vezes maior do que da década de 70, mesmo com uma área plantada quatro vezes menor. Assim, o avanço da cultura no cerrado brasileiro viabilizou a produção em larga escala e a mecanização intensa, o que combinado com o uso de cultivares adaptadas, fertilizantes, corretivos e o controle químico de doenças e pragas tem feito com que o cerrado detenha as mais altas produtividades na cultura do algodoeiro no Brasil e no mundo, em áreas não irrigadas.

O algodão encontra no estado do Mato Grosso condições favoráveis para o seu desenvolvimento, como região plana, alta intensidade de luz e clima bem definido, com período seco na época da colheita, o que contribui para uma melhor qualidade de pluma. Com o avanço da cotonicultura em áreas do Cerrado, o Brasil passou da condição de importador para exportador de pluma (PENA, 2017). Atualmente a região do Cerrado (Mato Grosso, Mato Grosso do Sul, Goiás, Bahia e Maranhão) responde por 97% da produção brasileira de algodão, sendo o Estado de Mato Grosso o maior produtor de algodão herbáceo (de fibra curta), ocupando a primeira posição em área cultivada e produção (70% do total). Esse cenário é o resultado do esforço dos produtores, técnicos, pesquisadores e governos por meio de associações de produtores, das instituições públicas e empresas privadas na geração e transferência de novas tecnologias visando aperfeiçoar o sistema produtivo.

A Empresa Brasileira de Pesquisa Agropecuária (Embrapa) vem participando decisivamente da evolução da cotonicultura do Cerrado por meio de geração e transferência de tecnologias. A Embrapa, em parceria com Universidades, Institutos de Pesquisa e Empresas Privadas realizam pesquisas para o aperfeiçoamento dos sistemas de produção e o desenvolvimento de novas cultivares com as características de produtividade, qualidade de fibra e resistência às doenças demandadas pelo setor produtivo. Além das plantas geneticamente modificadas que são resistentes a pragas, outras tecnologias como mecanização e adubação também fazem a diferença gerando bons resultados no incremento da produtividade do algodoeiro.

Embora a cadeia do algodão tenha atingido um alto nível tecnológico, com competitividade e lucratividade, há ainda muitos desafios e problemas a serem vencidos e possibilidades de melhorias no sistema de produção. Um dos desafios atuais é o do aumento da produtividade, uma vez que nos últimos anos houve uma estagnação da produtividade do algodoeiro produzido no cerrado. Essa estagnação vem ocorrendo a despeito da constante introdução de novas tecnologias no sistema de produção, como pode ser verificado na Figura 1. Portanto, entender as causas dessa estagnação auxiliará na definição de estratégias por parte dos produtores para a reversão dessa tendência.

Figura 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, MG, MA), evidenciando a tendência de estagnação nos últimos 10 anos (linha pontilhada)

Fonte: Elaborado pela autora

Dado esse contexto, propõem-se 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?

Ao identificar tal problema e possuindo informações detalhadas de produtividade e fatores limitantes como dados de manejo, tipos e características químicas e físicas dos solos, pragas e doenças e cultivares utilizados, o Instituto Mato-Grossense do Algodão (IMAmt) em parceria com a EMBRAPA, organizou um banco de dados, contemplando quatro safras agrícolas (2011/2012 a 2014/2015), o qual pode auxiliar na compreensão das relações de causa-efeito e, por conseguinte, ajudar a elucidar a questão da estagnação da produtividade do algodoeiro nos últimos 10 anos.

O referido banco de dados gerado é bastante completo, contendo informações da produtividade de cada área analisada, atributos do solo como os indicadores químicos do solo (fertilidade) e físicos do solo (compactação, tipo de solo), diagnóstico de doenças e tipo de manejo empregado (plantio direto, convencional, algodão de safra ou safrinha, espaçamento da linha de plantio, cultivar utilizado, etc.).

Conhecer os fenômenos que estão explícitos em tais variáveis será de grande importância para repassar conhecimentos aos produtores de algodão para que haja melhoramento nas técnicas de cultivo do algodoeiro, visando o aumento da produtividade das áreas de menor potencial produtivo.

A mineração de dados têm sido aplicada na área agrícola em diversas culturas como a de cana-de-açúcar, arroz, trigo, açaí, dentre outras, principalmente na modelagem da produtividade, identificação de padrões de expansão de doenças, classificação de solos

agrícolas, dentre outros estudos ([NEELAVENI; RAJESWARI, 2016](#)). Nesse contexto, as técnicas de mineração de dados podem ser utilizadas na análise, interpretação e extração de conhecimentos de bancos de dados no processo de tomada de decisões eficientes.

Diversas técnicas da ciência da computação, como a mineração de dados e aprendizagem de máquinas têm sido utilizadas para avaliar o efeito de diferentes atributos (físicos, químicos, biológicos, climáticos, de manejo e outros), manejáveis ou não, na variável produtividade, buscando, assim, prever a produtividade das culturas agrícolas em diferentes cenários.

A mineração de dados é o processo de identificar padrões escondidos em dados complexos e volumosos ([KLOSGEN; ZYTKOW, 2002](#)) e pode, portanto, contribuir para identificar regras relevantes para a tomada de decisão em problemas complexos como é o caso da produção agrícola altamente tecnificada, que incorpora conhecimentos e tecnologias de ponta em todas as etapas da produção.

Existem vários trabalhos de aplicações de técnicas de mineração de dados no campo da agricultura. Por exemplo, em uma revisão recente ([GANDHI; ARMSTRONG, 2016a](#)) descrevem algumas dessas aplicações como técnicas de 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.

A agricultura é uma atividade que depende de fatores que muitas vezes foge do controle do agricultor, como as variáveis climáticas, o aparecimento de doenças e pragas inesperadas e a oscilação de preços tanto dos insumos como das culturas (grãos, frutas, fibras). Assim, 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 têm focado na correlação dos fatores que influenciam a produtividade e geração de modelos de previsão ou estimativa do rendimento das culturas.

Na cultura do arroz, que é altamente demandante de água, trabalhos foram realizados aplicando técnicas de mineração de dados 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 ([GANDHI; ARMSTRONG, 2016b](#)), ([DEY; MASUD; UDDIN, 2017](#)) e Colômbia ([DELERCE *et al.*, 2016](#)).

Em bases de dados de produção de cana-de-açúcar no Brasil e na Argentina, estudos foram realizados para se identificar os atributos do solo, parâmetros climáticos, cultiva- res, dentre outros fatores, com maior impacto na produtividade, de modo a orientar ações futuras para a tomada de decisão no manejo dessa cultura ([GANDHI; ARMSTRONG,](#)

2016b), (DEY; MASUD; UDDIN, 2017) e (DELERCE *et al.*, 2016).

Na cultura do açaí foram realizados estudos em (SILVA; RODRIGUES; MONTEIRO, 2010) na área de descoberta de conhecimento em base de dados, mostrando a implantação deste processo na análise e descoberta de associações que influenciam a produtividade do açaí no estado do Amapá. Em cana-de-açúcar utilizou-se três técnicas de mineração de dados nas análises de banco de dados 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 (HAMMER, 2016). Outro trabalho sobre cana (GARCIA; JÚNIOR, 2015) apresenta, através do processo de descoberta de conhecimento em bases de dados, um método sistemático para classificação dos fatores que impactam na produtividade de cana-de-açúcar utilizando a técnica de árvore de decisão, que mostra os fatores com maior influência na produtividade agrícola. Em arroz utilizaram-se técnicas de floresta de inferência condicional e agrupamento (DELERCE *et al.*, 2016) para avaliar as relações entre os fatores climáticos e a variabilidade do rendimento das culturas de arroz na escala local para cultivares e estádios de crescimento específicos. Em outro trabalho com dados de arroz, foram aplicadas técnicas de visualização de dados (GANDHI; ARMSTRONG, 2016b) para encontrar correlações entre os fatores climáticos e a produção. Esse último estudo utilizou também técnicas de mineração de dados para extrair o conhecimento do conjunto histórico de dados agrícolas para prever o rendimento das culturas de arroz em zonas tropicais úmidas e regiões mais secas da Índia.

De um modo geral, os trabalhos de aplicação de técnicas de mineração de dados descritos acima apresentam correlações e associações entre a produtividade e outros atributos, mas não avançam na interpretação das regras geradas. Verifica-se que em geral o principal objetivo é a obtenção da maior acurácia possível, sem preocupações ou compromissos com a consistência das associações geradas. Nesse trabalho, a preocupação principal não é apenas a obtenção de altas acurácias, mas principalmente que se possa explicar por meio das regras geradas, os principais fatores que estão causando as baixas produtividades do algodoeiro, de modo que se possa recomendar ações corretivas de manejo da cultura. Outra questão importante é que não foram identificados trabalhos de mineração de dados na cultura do algodoeiro em bases de dados do Brasil, o que abre uma oportunidade de contribuir com recomendações práticas aos produtores de algodão, visando o aumento da produtividade dessa importante cultura para a economia brasileira.

1.2 Objetivos

Este trabalho visa aplicar técnicas de mineração de dados, mais especificamente classificação baseada em regras em um banco de dados de produção comercial de algodão no Estado do Mato Grosso, visando a compreensão dos principais fatores (variáveis do solo e doenças das plantas) que afetam a produtividade do algodoeiro.

Os objetivos específicos do trabalho são:

- Testar diferentes técnicas de pré-processamento dos dados;
- Avaliar diferentes algoritmos de classificação baseados em regras para obter relações entre os atributos levantados com a produtividade do algodoeiro;
- Gerar recomendações técnicas para os produtores de algodão do Mato Grosso, baseados nos resultados da mineração.

O trabalho baseia-se principalmente na aplicação de algoritmos para indução de regras de classificação, utilizando os algoritmos JRip ([COHEN, 1995](#)) e PART ([WITTEN; FRANK, 1998](#)) para classificação baseada em regras de decisão. A análise da coerência das regras é então checada em conjunto com especialistas da área de produção de algodão.

1.3 Estrutura do Trabalho

Além da Introdução apresentada no capítulo anterior, a estrutura do trabalho está organizada nos seguintes capítulos:

Capítulo 2 – Revisão da Literatura: Apresenta duas subseções, a de Fundamentos de Mineração de Dados, e das Aplicações das técnicas de Mineração de Dados nas áreas agrícolas.

Capítulo 3 – Metodologia: Apresenta uma descrição do banco de dados, sua estrutura e atributos. Em seguida, descreve a metodologia da pesquisa, introduzindo as ferramentas e algoritmos utilizados para a obtenção dos resultados, e uma justificativa da escolha das técnicas selecionadas. É introduzida a ferramenta WEKA, as técnicas de discretização *Kononenko* e *Better Encoding*, precisão, recall, a matriz de avaliação de custos e uma análise do desempenho dos algoritmos de classificação JRip e PART.

Capítulo 4 – Resultados e Discussões: Apresenta os resultados da pesquisa através de regras de associação, bem como uma análise sobre dos resultados encontrados considerando a visão de especialistas na área agrícola.

Capítulo 5 – Conclusões: Apresenta as conclusões, observações, considerações finais e os possíveis desenvolvimentos futuros.

REVISÃO DA LITERATURA

No Capítulo 1 apresentou-se um breve histórico da cotonicultura e sua importância para o país, bem como uma introdução dos trabalhos sobre mineração de dados aplicados na agricultura, sejam eles com ou sem classificação baseada em regras. Nesse capítulo são apresentadas duas seções, uma sobre os fundamentos da mineração de dados e outra sobre os principais trabalhos realizados com aplicações da mineração de dados em bancos de dados agrícolas.

2.1 Fundamentos de Mineração de Dados

Nessa seção são descritos os processos de descoberta de conhecimento, denominado KDD (do inglês: *Knowledge Discovery in Databases*), a mineração de dados, e as principais tarefas de mineração com ênfase na classificação, bem como as técnicas de pré-processamento e análise utilizadas.

2.1.1 O processo de busca de conhecimento em banco de dados

A descoberta de conhecimento em bases de dados e a mineração ou prospecção de dados pode ser definida como um conjunto de procedimentos computacionais, constituídos de ferramentas e técnicas de aprendizagem ou classificação, para encontrar padrões de associação entre variáveis relevantes em bancos de dados complexos.

O objetivo principal da mineração dos dados é a descoberta de novos conhecimentos e que podem ser utilizados para tomar decisões, solucionar problemas, fazer diagnósticos, detectar fraudes, prevenir problemas, gerar novos modelos, dentre outras aplicações. A KDD é o processo de extração de conhecimento em banco de dados, que o torna útil, a partir da identificação de padrões e associações entre variáveis do banco de dados (FAYYAD;

PIATETSKY-SHAPIRO; SMYTH, 1996), ou seja, o ato da descoberta do conhecimento se manifesta pela mineração dos dados.

As organizações têm se mostrado bem eficientes em capturar, organizar e armazenar enormes quantidades de dados obtidos de suas operações. Entretanto, a maioria não utiliza adequadamente essa imensa quantidade de dados para transformá-la em conhecimentos que possam ser utilizados em suas próprias atividades.

Em diversas situações, muitas empresas não exploram adequadamente os dados gerados nos seus processos, não percebendo que têm em mãos uma fonte valiosíssima de conhecimento escondido nos seus dados, os quais podem ser "minerados" e utilizados para seu próprio benefício nos seus processos internos e externos. Um dos principais motivos para isso é o desconhecimento das ferramentas e técnicas, falta de equipes qualificadas e disposição da empresa em investir nessa área. De acordo com [Silva, Peres e Boscarioli \(2016\)](#), minerar dados para descobrir conhecimento não é uma tarefa trivial. É necessário conhecer os dados, o processo de análise e descoberta, as tarefas e técnicas de mineração e as ferramentas matemáticas e computacionais que se aplicam nesse contexto. É preciso ainda conhecer o ambiente em que os dados são produzidos e que tipo de conhecimento esse ambiente necessita e espera receber. Enfim, minerar dados exige conhecimento técnico, tempo e dedicação.

Descobrir conhecimento é um processo complexo em que é necessário conhecer a origem dos dados para se ter certeza de que tipo de conhecimento se deseja. Descobrimento tem como o objetivo encontrar padrões camuflados com o uso de técnicas de mineração apropriadas, sem que necessariamente exista uma ideia ou hipótese clara previamente estabelecida.

A mineração de dados (MD) faz parte de um processo de KDD, a qual possui um procedimento de preparação e exploração dos dados, interpretação de seus resultados e absorção dos conhecimentos minerados, como apresentado a seguir.

Deve-se observar que a MD se tornou mais conhecida até que o próprio processo de KDD por ser a etapa onde são aplicadas as técnicas de busca de conhecimentos. Considerando que as bases de dados são geralmente volumosas e que o conhecimento pode estar implícito, faz-se necessário um trabalho minucioso de busca para encontrar padrões de relevância para o problema avaliado. A área de MD é interdisciplinar, pois combina inteligência artificial, gerenciamento de banco de dados, visualização de dados, aprendizagem de máquina, algoritmos matemáticos e técnicas estatísticas ([HAN; KAMBER, 2006](#)).

Existem várias abordagens para a estratificação do processo de extração de conhecimento de bases de dados. Inicialmente, foi proposta uma divisão do processo em nove etapas ([FAYYAD; PIATETSKY-SHAPIRO; SMYTH, 1996b](#)), mas isso não é um consenso, sendo que alguns autores descrevem o processo formado por apenas quatro etapas

(WEISS; INDURKHYA, 1997). De modo geral, o processo de KDD contempla uma sequência iterativa contendo as seguintes etapas: *Limpeza, Integração, Seleção, Transformação e Mineração dos dados, Avaliação dos padrões, Apresentação e Assimilação dos conhecimentos gerados* (HAN; KAMBER, 2001). No entanto, neste trabalho utilizaremos uma sequência mais simplificada do processo, constituído de quatro etapas: Obtenção dos dados, Pré-processamento, Mineração dos dados (extração de padrões) e Pós-processamento.

Assim, tendo o objetivo de gerar conhecimentos úteis e que possam ser aplicados na tomada de decisão, serão apresentados nessa seção as diversas etapas do processo KDD. Conforme ilustrado na Figura 2, o processo de descoberta de conhecimento em bases de dados tem início com a obtenção dos dados e a organização das bases de dados contendo as informações da área de interesse e na qual se deseja descobrir informações relevantes para o processo.

Figura 2 – Esquema do processo de descoberta de conhecimento em bases de dados (KDD)

Fonte: adaptado de ((SILVA; PERES; BOSCARIOLI, 2016)

A fase de Pré-processamento inicia-se com a limpeza dos dados, que é um procedimento de tratamento para eliminação de instâncias repetidas e ou valores discrepantes, ou seja, remove-se dados inconsistentes e fora de padrões. Procedimentos de seleção de dados permitem que se escolha apenas os atributos e/ou exemplos relevantes nas suas diversas fontes, pois a escolha errada dos dados pode levar a resultados errôneos após aplicação da mineração de dados, prejudicando assim a tomada de decisão. Problemas como redundância e inconsistência estão relacionados, na maioria das vezes, ao uso de um número excessivo de variáveis e nesses casos recomenda-se o uso de técnicas de seleção de atributos.

Inúmeros algoritmos de aprendizado de máquina (AM) ao passar pela fase de mineração dos dados, têm dificuldades em processar os dados no seu formato original. Nesses casos, é necessário transformar os dados para que eles possam ser adequadamente utiliza-

dos na etapa de mineração. Na etapa de Transformação de dados, pode-se transformar ou consolidar os dados em formatos apropriados por meio de operações do tipo sumarização ou agregação, discretização e normalização, de modo que os dados possam ser avaliados em uma mesma escala e considerados com a mesma relevância pelos algoritmos de análise (FACELI *et al.*, 2011). As técnicas de pré-processamento com ênfase na transformação dos dados são importantes não apenas porque minimizam ou eliminam problemas nos dados, mas também porque podem torná-los mais adequados para utilização por um determinado algoritmo de aprendizado de máquina (A M). As técnicas de pré-processamento são utilizadas para melhorar a qualidade dos dados, levando a modelos mais ajustados à distribuição real dos dados, reduzindo sua complexidade computacional, tornando os ajustes dos parâmetros do modelo mais fáceis e rápidos, bem como a sua utilização posterior, facilitando também a interpretação dos padrões extraídos pelo modelo (FACELI *et al.*, 2011).

Após o pré-processamento, inicia-se a etapa de Mineração dos dados propriamente dita, onde as técnicas são aplicadas para análise e extração de padrões, envolvendo a resolução de tarefas como predição, agrupamento ou associação (HAN; KAMBER, 2001). Nesta etapa é que são descobertas as informações mais valiosas, como tendências e padrões que poderiam ser usados para melhorar as tomadas de decisões dos processos avaliados, além de outras aplicações (REZENDE, 2003).

A última etapa é a Avaliação de padrões, onde os resultados obtidos são avaliados e validados e formatados em gráficos, tabelas e relatórios estruturados. É nessa fase que se faz uso de técnicas de visualização e representação das informações geradas para expor o conhecimento adquirido aos usuários/clientes, bem como introduzi-los no âmbito estudado (HAN; KAMBER, 2001). Contudo, o importante nesta etapa é a validação e valoração do conhecimento extraído com o especialista do domínio, capaz de fazer corretamente a interpretação dos resultados gerados.

O processo como um todo (Figura 2) pode ser iterativo e interativo, ou seja, cada etapa pode ser executada diversas vezes, na sequência que foi descrita ou em uma outra, a depender do conjunto de dados original e/ou de decisões tomadas pelo especialista da área (SILVA; PERES; BOSCARIOLI, 2016). É importante mencionar também que não há uma ordem fixa para a aplicação das diferentes técnicas de pré-processamento (FACELI *et al.*, 2011).

Os usuários do processo podem ser divididos em três categorias (REZENDE, 2003), conforme descrito a seguir.

Especialista do Domínio: Usuário que deve possuir amplo conhecimento do domínio da aplicação e deve fornecer apoio para a execução do processo de extração de conhecimento.

Analista: Usuário especialista no processo de extração de conhecimento e responsável por sua execução. Este usuário deve conhecer profundamente as etapas que compõem o processo.

Usuário Final: Representa a classe de usuários que utiliza o conhecimento extraído no processo para auxiliá-lo em um processo de tomada de decisão. Não é necessário que esse usuário possua conhecimento profundo do domínio da aplicação.

Existem situações em que o especialista do domínio é também o usuário final, auxiliando e executando funções relacionadas ao analista (REZENDE, 2003).

2.1.2 Tarefas da Mineração de Dados

A mineração de dados (MD) é comumente classificada pela sua habilidade em realizar determinadas tarefas (LAROSE, 2005). Existem vários tipos de tarefas realizadas no processo de mineração de dados, dentre elas a classificação, agrupamento (clusterização) e associação. Essas tarefas serão conceitualmente apresentadas a seguir, com maior ênfase na etapa de classificação.

O Agrupamento tem como objetivo identificar e aproximar registros semelhantes, ou seja é uma coleção de registros similares, porém diferentes daqueles nos demais agrupamentos (CAMILO; SILVA, 2009). Sendo assim, o agrupamento ou clusterização é uma tarefa onde se procura determinar um conjunto finito de conjuntos ou grupos para descrever os dados (FAYYAD; PIATETSKY-SHAPIRO; SMYTH, 1996). Essa tarefa difere da classificação, pois não necessita que os registros sejam previamente categorizados, ou seja, é um aprendizado não-supervisionado. O agrupamento não tem a pretensão de classificar, estimar ou prever o valor de uma variável, mas sim identificar os grupos de dados parecidos (CAMILO; SILVA, 2009).

A Associação busca identificar regras para quantificar relações entre dois ou mais atributos, ou seja, busca identificar quais atributos estão relacionados. Esse tarefa apresenta configurações do tipo "SE atributo X ENTÃO atributo Y", sendo uma das tarefas mais conhecidas devido aos bons resultados obtidos, principalmente na análise conhecida como "Cestas de Compras" (*Market Basket*), onde identifica-se produtos que são adquiridos pelos consumidores (LAROSE, 2005). A Classificação é considerada como uma das tarefas de mineração de dados mais importantes e conhecidas na área de aprendizado de máquina. Ela visa identificar a qual classe um determinado registro pertence. Ou seja, analisa o conjunto de registros fornecidos, conhecendo-se previamente a classe a qual o registro pertence, afim de "aprender" como classificar um novo registro. É portanto um aprendizado supervisionado (SCOSS, 2006) e (CAMILO; SILVA, 2009). Se o domínio das classes é um conjunto finito e não ordenado de valores, o problema é denominado de Classificação e o estimador de Classificador (COSTA; BERNARDINI; FILHO, 2014).

Na tarefa de classificação existe um atributo chamado classe e é comum aplicar aprendizado de máquina para prever a classe. A classificação funciona com dados de eventos reais ocorridos, de modo óbvio já classificados, ou seja, por exemplo a base de dados algodão é uma base real com dados já classificados. Para melhor compreender a tarefa de classificação, a Figura 3 apresenta uma pequena parte do banco de dados utilizado nesse trabalho, com dados históricos da produtividade de algodão no Estado do Mato Grosso, o qual contem na totalidade 1799 instâncias e 72 atributos. Na Figura 3 são mostrados apenas quatro atributos e quinze instâncias, classificadas como Baixa, Média e Alta produtividade determinadas pelo algoritmo de classificação para construir o modelo. Após a construção do modelo, os dados reais de produtividade não são mais necessários. A cada nova instância com novos dados, ou seja, dados ainda não classificados em que não se sabe se a produtividade é Baixa, Média ou Alta, estas são aplicadas ao modelo que vai prever, com uma margem de erro, se aquela produtividade é Baixa, Média ou Alta.

No exemplo da Figura 3, o modelo criado tem formato de regras de associação para classificação. Um dado modelo gerado é dependente do tipo de algoritmo utilizado. No caso do algoritmo de árvore de decisão o modelo gerado tem formato de árvore com nós e folhas. Para a construção de um modelo que tem objetivo de classificar novas instâncias, os dados históricos são normalmente divididos em dois conjuntos, um para treinar o modelo e outro para testá-lo. Ou seja, o primeiro conjunto de dados é o que de fato vai ser utilizado para se obter o modelo e o segundo é para verificar a eficiência do processo de aprendizado, ou seja, o quão eficiente foi o modelo construído ([AMARAL, 2016](#)).

Figura 3 – Exemplo ilustrativo da tarefa de Classificação de dados, utilizando parte dos dados da base de algodão, avaliado nesse trabalho

Fonte: Elaborado pela autora

A divisão dos dados em conjuntos de treino e de teste para a validação do modelo pode ser feita utilizando-se diferentes métodos de validação, como *K-fold*, *Hold out* e o *Leave one out*. Uma das mais utilizadas é o *hold out*, que é um método de validação cruzada utilizado para avaliar o desempenho de um algoritmo ante os dados disponíveis (base de dados). Tal método possibilita comparar algoritmos de indução de modelos com relação ao se seus desempenhos e de acordo com alguma métrica (*F-measure*, ROC área, Recall, etc). Com o método *hold out*, os dados são divididos aleatoriamente sem substituição, em 70% para treino e 30% para teste, ou seja, ao ser selecionado uma instância para teste, ela não poderá ser selecionada para o grupo de treino, e vice-versa. O motivo de ser aleatório é para que as instâncias tenham as mesmas chances de serem selecionadas para teste ou para treino. Esse método será descrito em mais detalhes adiante.

Uma vez que os dados de treino do algoritmo são selecionados, o modelo é construído e sua eficiência deve ser avaliada. Aplicam-se então os dados de treino para que se faça a classificação destes dados. Observa-se na Figura 3 que os dados de treino desse banco de dados já estão previamente classificados, mas o algoritmo do modelo faz a sua própria classificação, independente da classificação prévia, atribuindo a sua própria classe. Quando esta etapa é concluída, testa-se a efetividade do modelo, fazendo-se uma comparação dos valores da classe dos dados de treino com os valores previstos, gerando com esse processo a matriz de confusão, a qual será descrita adiante.

A tarefa de classificação pode ser utilizada em várias situações e aplicações, como

por exemplo: identificar em uma Universidade, qual o curso mais adequado para um determinado candidato; localizar uma determinada enfermidade; identificar fraudes em transações de cartão de crédito; reconhecer comportamentos de ser ameaça para a segurança; identificar fatores limitantes da produtividade agrícola; identificar quais alunos podem desistir dos estudos, dentre outras.

Existem duas tarefas que são muito semelhantes à de Classificação que são a Regressão e a Predição. A regressão é uma tarefa usada quando o registro é identificado por um valor numérico e não um categórico, assim, pode-se estimar o valor de uma determinada variável analisando-se os valores das demais. Um exemplo prático para entendimento seria através de um conjunto de registros contendo os valores mensais gastos por diversos tipos de consumidoras de cosméticos e de acordo com os hábitos de cada uma, após ter analisado os dados, o modelo é capaz de dizer qual será o valor gasto por uma nova consumidora. A tarefa de regressão pode ser usada, por exemplo, para estimar a quantidade a ser gasta por um grupo de pessoas durante a volta às aulas e também estimar a massa corpórea de uma pessoa baseando-se no peso e na altura, prever a demanda futura de um novo produto ou estimar a expectativa de vida média dos brasileiros (CAMILO; SILVA, 2009). Já a Predição é também bastante similar às tarefas de classificação, porém ela visa descobrir o valor futuro de um determinado atributo. Tal tarefa tem por objetivo encontrar padrões para prever o futuro utilizando características de algumas entidades (FAYYAD; PIATETSKY-SHAPIO; SMYTH, 1996). Pode ser utilizada, por exemplo, para prever o valor de uma ação no futuro, o percentual de aumento de tráfego na internet caso a velocidade aumente, o time campeão de um determinado campeonato baseando-se na checagem das estatísticas dos times, etc.

2.1.3 Classificador baseado em Regras

Muitos sistemas inspiram-se na ideia de que a tomada de decisão humana pode ser moldada por meio de regras de condição do tipo SE (é uma lista de condições a serem satisfeitas) e ENTÃO (lista de conclusões e ações) (GOLDSCHMID, 2010). Sendo assim, um classificador baseado em regra é um modelo composto por um conjunto de regras lógicas com a estrutura SE/ENTÃO, cujo conjunto de premissas é composto por pares atributo/valor, em que o SE é a condição e o ENTÃO é a conclusão da regra (HAN; KAMBER, 2006). Portanto, as regras podem expressar relacionamentos lógicos e equivalentes de definições para simular o raciocínio humano. Um exemplo simples de regra (R1) é descrito a seguir.

R1: SE idade = jovem E aluno = sim, ENTÃO compra computador = sim.

O "SE" é a parte da regra conhecida como antecedente ou condição e o "ENTÃO" é a parte consequente da regra (HAN; KAMBER, 2006). No antecedente da regra, a condição consiste em um ou mais testes de atributos como idade (jovem) e aluno (sim) que são

logicamente combinados ($E =$). A regra consequente contém uma previsão de classe (neste caso, se um cliente irá comprar um computador). A regra R1 também pode ser escrita como R1: (idade = jovem) (aluno = sim) (compra computador = sim).

Um outro exemplo pode ser sobre um cenário de financiamento nas modalidades reembolsável e não reembolsável, com a seguinte regra: SE contratado = não, decisão-da-diretoria = sem decisão e arquivado = não, ENTÃO status = Em análise (CAMILO; SILVA, 2009).

Se as condições, (todos os testes de atributos em um antecedente de regra) forem válidas para uma dada tupla, dizemos que o antecedente da regra está satisfeito (ou simplesmente, que a regra está satisfeita) e que a regra abrange a tupla. Uma regra R pode ser avaliada pela sua cobertura e precisão, onde a cobertura é a porcentagem de tuplas que são cobertas pela regra e cujos valores de atributos são válidos para o antecedente da regra (HAN; KAMBER, 2006). Para a precisão da regra, avalia-se as tuplas que a cobrem e então analisa-se qual porcentagem delas pode ser classificada corretamente pela regra. 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 (CAMILO; SILVA, 2009). Portanto, regras de classificação consistem na identificação de qual classe pertence um elemento a partir de seus atributos. Além disso, ele também permite encontrar a função para a realização dessa classificação.

Sendo assim, as regras SE-ENTÃO são a base para algumas das linguagens de descrição de conceito mais populares na AM. Elas permitem que o "conhecimento" seja extraído de um banco de dados para ser representado de uma forma simples para as pessoas entenderem. Isso dá aos especialistas do domínio a chance de analisar a validade desse conhecimento e combiná-lo com fatos previamente conhecidos sobre o domínio. Há uma gama de algoritmos de classificação baseados em regra. Nesse estudo foram utilizadas duas técnicas contidas na ferramenta Weka para aplicação do processo de MD (WITTEN; FRANK, 2009), o PART que é um algoritmo baseado em árvores de decisão parcial (WITTEN; FRANK, 1998) e o JRip que implementa RIPPER, incluindo otimização global heurística do conjunto de regras (COHEN, 1995). Na seção à seguir são apresentadas essas duas técnicas de classificação.

2.1.3.1 Algoritmo PART

O algoritmo PART (*partial decision trees*) é um classificador desenvolvido a partir do algoritmo J48 e tem como finalidade gerar uma árvore de decisão baseada no conjunto de dados de treinamento rotulados (WITTEN; FRANK, 2009). É uma abordagem alternativa para a indução de regras que evita a otimização global e produz conjuntos de regras precisas e compactas. Sendo assim, o algoritmo PART constrói uma árvore J48 e depois extrai um caminho com a maior cobertura para formar a primeira regra. As

instâncias cobertas por esta regra são removidas dos dados de treinamento e o processo é então repetido para gerar a segunda regra, e assim, sucessivamente. Um conjunto de árvores J48 vai crescendo em subconjuntos menores da base de treinamento. Esta é uma classe para gerar lista de decisões PART. Ele usa uma abordagem de dividir e conquistar e constrói uma árvore de decisão parcial para cada iteração e percorre a “melhor” folha dentro de uma regra (DEVASENA, 2014). Tal abordagem divide um problema complexo em subproblemas mais simples, aplicando recursivamente a mesma estratégia a cada subproblema, dividindo o espaço definido pelos atributos em subespaços, associando-se a eles uma classe (WITTEN; FRANK, 2009).

Para criar uma regra, uma árvore de decisão podada é construída para o conjunto atual de instâncias, a folha com maior cobertura é transformada em uma regra e a árvore é descartada. Para cada regra criada, estima-se a cobertura das instâncias no banco de dados e isso é repetido diversas vezes até que todas as instâncias sejam cobertas. As regras com coberturas mais altas são mantidas e apresentadas para o usuário e as demais são descartadas (WITTEN; FRANK, 1998). Assim, a ideia principal do algoritmo PART é construir uma árvore de decisão parcial ao invés de uma geral. No Quadro 1 é descrito o funcionamento do algoritmo PART de acordo com (DEVASENA, 2014).

Quadro 1 – Etapas da implementação do algoritmo PART (partial decision trees)

1. Crie uma árvore de decisão parcial no conjunto atual de instâncias;   2. Crie uma regra da árvore de decisão; (A folha com maior cobertura é transformada em uma regra)   3. Rejeite a árvore de decisão;   4. Remova as instâncias cobertas pela regra;   5. Vá para a primeira etapa.

Fonte: (DEVASENA, 2014)

2.1.3.2 Algoritmo JRip

O algoritmo JRip (RIPPER: *Repeated Incremental Pruning to Produce Error Reduction* - Poda Incremental Repetida para Produzir Redução de Erro) é um dos algoritmos mais populares implementados no pacote WEKA. É baseado no algoritmo RIPPER (COHEN, 1995) que repete uma poda incremental para a redução de erro, proposto como uma versão aperfeiçoada de IREP (*Incremental Reduced Error Pruning*) (FURNKRANZ; WIDMER, 1994), ou seja, com a poda incremental repetida, o algoritmo JRip produz uma redução de erro. É um algoritmo de aprendizado de máquinas que aprende regras proposicionais a partir de exemplos e foi proposto por Willian W. Cohen (COHEN, 1995) em uma versão otimizada do IREP. Isso significa que um conjunto de exemplos da tarefa, representado na forma de atributo-valor, é fornecido ao algoritmo, que encontra um conjunto de regras com bom desempenho de acordo com o conjunto de exemplos e a medida

de performance. O desempenho é medido pelo ganho de informação obtido com a regra. O algoritmo RIPPER/JRip é composto por duas fases: uma de construção das regras e outra de otimização das regras. A fase de construção envolve uma etapa de crescimento da regra e uma posterior poda das mesmas, já a fase de otimização, cada regra é considerada por vez para ser otimizada, na ordem em que foram geradas (VELOSO; GOMES, 2007). O algoritmo JRip trabalha tratando todos os exemplos em particular julgamento dos dados de treinamento como uma classe, e encontra um conjunto de regras que cobre todos os membros da classe. Isso é executado da mesma forma para todas as classes em sequência até que todas elas tenham sido cobertas (RAJPUT *et al.*, 2011).

2.1.4 Estratégias para treinamento, validação e teste

Como estratégia para desenvolvimento e validação de um modelo é comum separar os dados em duas categorias: dados para treinamento (indução) e os dados para teste. Além dessa divisão, pode-se usar também uma subdivisão do conjunto de treinamento, criando um conjunto de validação, utilizado para verificar a eficiência da técnica quanto a sua capacidade de generalização durante o treinamento, e podendo ser empregado como critério de parada do treinamento. Em seguida, já com estes conjuntos definidos, eles são, geralmente, colocados em ordem aleatória para prevenção de tendências associadas à ordem de apresentação dos dados.

É importante ressaltar que uma das fases mais importantes em um trabalho de aprendizado de máquina (AM) é essa etapa de validação dos resultados. Várias das técnicas mais poderosas de aprendizado apresentam uma grande quantidade de parâmetros e quanto menos restrições colocarmos no nosso modelo maior a probabilidade de ocorrer um sobreajuste (do inglês *overfitting*).

O *Overfitting* é um problema que deve ser evitado. Ele ocorre quando o método de estimativa produz bons resultados apenas nos dados que eles foram treinados, apresentando baixa acurácia quando aplicado em novas observações. Ou seja, quando ocorre *overfitting* o método de aprendizado irá apresentar baixa capacidade de generalização (SILVA; PERES; BOSCAROLI, 2016). Para reduzir o *overfitting*, pode-se utilizar diferentes estratégias de geração e avaliação de modelos preditivos, como o *holdout* e a *validação cruzada*, os quais são descritos a seguir.

2.1.4.1 Holdout

O estimador *holdout* (conhecido também como *split-sample*) divide os dados em uma proporção para treinamento (p) e teste ($1 - p$), considerando normalmente $p > 1/2$. Valores típicos são $p = 2/3$ e $(1 - p) = 1/3$, embora não existam embasamentos teóricos sobre essa proporção (REZENDE, 2003). Os objetos a serem reservados em cada um dos subconjuntos devem ser escolhidos aleatoriamente.

Figura 4 – Diagrama ilustrativo da estratégia *hold out* para avaliação de desempenho do modelo de classificação baseado em regras

Fonte: Adaptado de (AMARAL, 2016)

No ambiente de aprendizado os dados são separados em treino e teste (Figura 4). O conjunto de dados de treino são submetidos ao classificador e produzem o modelo. O conjunto de dados de teste é então submetido ao modelo para que seja feita a previsão. Como são dados históricos que já possuem classificação, pode-se avaliar as classes real e a prevista, produzindo uma avaliação de desempenho. Caso o desempenho seja satisfatório o modelo pode ser colocado em produção para classificar novos dados.

2.1.4.2 Método de validação cruzada (*10-folds-cross validation*)

Na validação cruzada (*cross validation*) a amostra original é aleatorizada antes do teste e dividida em k partes com mesmo tamanho, onde $k-1$ partes são utilizadas para treino (aqui cada objeto participa o mesmo número de vezes do treinamento $k-1$ vezes) e 1 parte é utilizada para testes, onde cada objeto participa o mesmo número de vezes do teste (1 vez). Este processo é repetido k vezes, alternando em cada uma delas a parte utilizada para testes, sem repetição, de modo que todos os conjuntos de mensagens escolhidos para treino também servirão como teste e vice-versa. Ou seja, na estratégia de validação cruzada, todos os exemplares fazem parte em algum momento do conjunto de dados usado no teste do modelo preditivo.

Os subconjuntos são também conhecidos como *folds*. A estratégia de validação cruzada é comumente encontrada sob a nomenclatura de *k-fold cross validation*. Para minimizar a oscilação na medida, visto que o conjunto de dados para treino e para teste é aleatorizado e torna-se diferente a cada teste, utiliza-se normalmente 10 partições (número de *folds*). A avaliação final sobre a geração dos modelos preditivos pode ser feita ao menos de duas formas distintas: i) aplicando medidas estatísticas, como média, desvio-padrão e

intervalo de confiança ao conjunto de K avaliações obtidas; e ii) somando-se o desempenho obtidos pelos K modelos gerados e dividindo essa soma pelo número de exemplares do conjunto de dados original (REZENDE, 2003).

O principal objetivo de um modelo é prever com sucesso o valor de saída para novos exemplos. Diferentes medidas podem capturar diferentes aspectos do desempenho de classificadores. Sendo assim, existem várias medidas de “erro” e “acerto” que serão explicadas uma a uma a seguir, mas antes será apresentado uma tabela muito importante que serve de base para essas medidas, chamada de matriz de confusão também chamada de matriz de contingência, em que a avaliação de um modelo de classificação é baseada na análise de tal matriz.

2.1.4.3 Matriz de confusão

É uma espécie de tabela que permite a visualização do desempenho de um algoritmo de aprendizado, onde se conhece a saída desejada para cada exemplo (FACELI *et al.*, 2011). Com isso, o supervisor externo pode avaliar a capacidade da hipótese induzida de prever o valor de saída para novos exemplos. Cada coluna da matriz de confusão representa as instâncias de uma classe prevista, enquanto as linhas representam os casos de uma classe real. O nome é originado do fato de possibilitar verificar se o algoritmo está confundindo duas classes.

A matriz de confusão de uma hipótese h oferece uma medida efetiva do modelo de classificação ao mostrar o número de classificações corretas versus as classificações preditas para cada classe sobre um conjunto de exemplos T (REZENDE, 2003). Como ilustrado na Tabela 1, os resultados são totalizados em duas dimensões: classes verdadeiras e classes preditas, para k classes diferentes $C_1, C_2, \dots, C_k$. Cada elemento $M(C_i, C_j)$ da matriz, $i, j = 1, 2, \dots, k$, representa o número de exemplos de T que realmente pertencem à classe C_i , mas foram classificados como sendo da classe C_j .

Tabela 1 – Matriz de confusão de um classificador

Classe	Predita C_1	Predita C_2	...	Predita C_k
verdadeira C_1	$M(C_1, C_1)$	$M(C_1, C_2)$	...	$M(C_1, C_k)$
verdadeira C_2	$M(C_2, C_1)$	$M(C_2, C_2)$	...	$M(C_2, C_k)$
$\vdots$	$\vdots$	$\vdots$	...	$\vdots$
verdadeira C_k	$M(C_k, C_1)$	$M(C_k, C_2)$	...	$M(C_k, C_k)$

Fonte: De (REZENDE, 2003)

Para simplificar, consideremos um problema de apenas duas classes (Tabela 2). Com classes usualmente rotuladas como “+” (positiva) e “-” (negativa), as escolhas estão

estruturadas para prever a ocorrência ou não de um evento simples. Neste caso, os dois erros possíveis são denominados *falsos positivos* (FP) e *falsos negativos* (FN). A Tabela 2 ilustra a matriz de confusão para o exemplo com duas classes, onde TP é o número de exemplos positivos classificados corretamente e TN é o número de exemplos classificados corretamente do total de $n = (TP + FN + FP + TN)$ exemplos.

Tabela 2 – Matriz de confusão para um problema com duas classes

		Classe Predita	
		+	-
Classe Verdadeira	+	TP	FN
	-	FP	TN

Fonte: De (FACELI *et al.*, 2011)

Aparentemente existem vantagens ao se reduzir a matriz de confusão a uma única medida, pois através de um só valor é mais simples de fazer comparações de desempenho entre modelos, porém uma medida pode ser adequada para um certo problema, mas irrelevante para outros. É frequente também a avaliação de problemas de múltiplas situações, nas quais é possível definir várias medidas, sendo perfeitamente provável que um modelo fique melhor que outro para determinadas medidas, mas pior com relação a outras. Então, utilizar uma única medida para esses casos pode produzir a falsa impressão de que o desempenho pode ser avaliado utilizando-se apenas essa medida. Assim, a partir da matriz de confusão, uma série de outras medidas de desempenho podem ser utilizadas, como acurácia, erro, precisão, especificidade, recall, *F-measure* e classe, descritos a seguir.

Acurácia ou taxa de acerto

Essa métrica é a probabilidade do teste fornecer resultados corretos ou seja, é a probabilidade dos verdadeiros positivos e verdadeiros negativos como uma proporção de todos os resultados. Mede quão efetivo o sistema é do ponto de vista da classificação. A acurácia ou taxa de classificações corretas é a medida mais comum na avaliação de classificadores. Uma das limitações dessa métrica é que ela não reflete a realidade quando os dados não possuem a mesma quantidade de elementos por classe, ou seja, pode não ser adequada para classes desbalanceadas.

A acurácia (ac) é obtida pela soma dos valores da diagonal principal da matriz, dividida pela soma dos valores de todos os elementos da matriz (n), conforme equação

(2.1).

$$ac(f) = \frac{VP + VN}{n} \quad (2.1)$$

Taxa de erro total

A taxa de erro de classificação (*err*) é uma das medidas mais comumente utilizadas em aprendizado de máquina, que pode ser calculada a partir da matriz de confusão de acordo com a equação (2.2), ou seja, pela soma dos valores da diagonal secundária da matriz, dividida pela soma dos valores de todos os elementos da matriz.

$$err(f) = \frac{FP + FN}{n} \quad (2.2)$$

Existem várias situações em que a taxa de erro de classificação não é apropriada para a avaliação de modelos de classificação. Um caso comum se dá quando o número de exemplos em cada uma das classes é muito desbalanceado.

Precisão

Precisão ou preditividade positiva (*prec*) é a porcentagem de acertos ou verdadeiros positivos dentre todos os exemplos classificados como positivos, dada pela equação (2.3).

$$prec(f) = \frac{VP}{VP + FP} \quad (2.3)$$

Precisão ou também chamada de valor preditivo positivo é a fração de instâncias recuperadas que são relevantes.

Especificidade

A especificidade é a taxa com que se classifica como negativos todos os exemplos que são de fato negativos, ou seja, só considera os exemplos negativos. É chamada de taxa de verdadeiros negativos (TFP), a qual é calculada pela equação (2.4).

$$esp(f) = \frac{VN}{VN + FP} = 1 - TFP(f) \quad (2.4)$$

Recall, revocação ou taxa de verdadeiros positivos (VP)

A revocação ou *recall*, corresponde à taxa de acerto na classe positiva. Também é chamada de taxa de verdadeiros positivos (TVP), taxa com que classifica como positivos todos os exemplos que são de fato positivos. O Recall é a capacidade de recuperar instâncias úteis, ou seja, é a fração de instâncias relevantes que são recuperadas. O coeficiente de revocação de uma busca é a proporção de todos os itens relevantes em uma coleção

particular ou banco de dados que a busca é capaz de recuperar. É calculado pela equação (2.5).

$$sens(f) = rev(f) = TVP(f) = \frac{VP}{VP + FN} \quad (2.5)$$

F-Measure

A medida-F (*F-Measure*) é a média harmônica da precisão (eq. 3) e da revocação (eq. 5).

2.1.5 Técnica de pré-processamento: Discretização

Na Seção 2.1.1 foi apresentado um panorama geral da etapa de pré-processamento de dados e nesta seção apresenta-se um maior detalhamento dessa etapa.

A discretização como uma estratégia de redução de dados (DOUGHERTY; KOHAVI; SAHAMI, 1995) tem recebido atenção crescente dos pesquisadores nos últimos anos (LIU *et al.*, 2002), tornando-se uma das técnicas de pré-processamento mais empregadas em MD (GARCIA *et al.*, 2013).

Um aspecto importante no processo de discretização é o ponto de corte ou ponto de divisão, que é um valor real dentro do intervalo de valores contínuos que divide o intervalo em dois subintervalos, sendo um intervalo menor ou igual ao ponto de corte e o outro intervalo maior do que o ponto de corte (DOUGHERTY; KOHAVI; SAHAMI, 1995). A Figura 5 apresenta um exemplo ilustrativo do ponto de corte, onde um intervalo $[X, Y]$ é particionado em $[X, W]$ e $[W, Y]$, sendo W o ponto de corte.

Figura 5 – Ilustração do conceito de ponto de corte

Fonte: Elaborado pela autora

Embora muitas aplicações do mundo real envolvam atributos contínuos, determinados algoritmos de classificação como os baseados em regras lidam somente com atributos discretos. Assim, para viabilizar o uso desses algoritmos, faz-se necessária a discretização dos atributos contínuos. Alguns métodos de classificação permitem o uso de atributos contínuos, mas mesmo nesses casos a discretização proporciona geralmente em ganho de precisão (KURGAN; CIOU, 2004) e (CATLETT, 1991).

Com a discretização os dados numéricos podem ser tratados como dados nominais durante o processo de mineração e assim os métodos de discretização reduzem e sim-

plificam o conjunto de dados, tornando o aprendizado mais rápido e os resultados mais compactos (GARCIA *et al.*, 2013). Outra vantagem é os dados transformados em um conjunto de intervalos são mais relevantes cognitivamente para um ser humano interpretar (LIU *et al.*, 2002).

O processo de discretização permite uma melhor compreensão das regras descobertas (CATLETT, 1991) e (PFAHRINGER, 1995). A discretização pode ser considerada como um processo de abstração de dados, transformando dados de nível inferior em um nível superior (Figura 6). De acordo com Catlett (1991), geralmente no nível superior permite-se que o conceito alvo seja expresso de forma mais concisa, tornando o produto da aprendizagem mais preciso e com taxas de erro menores do que poderiam ser obtidas a partir da indução diretamente sobre os atributos brutos. Isso é ilustrado na Figura 6 onde os dados de idades de pessoas são discretizados.

Figura 6 – Exemplo de discretização do atributo idade

Fonte: De (FREITAS, 2002)

A discretização é realizada antes do processo de aprendizagem e pode ser dividida em duas tarefas. A primeira tarefa é encontrar o número de intervalos discretos. Apenas alguns algoritmos de discretização realizam isso; alguns algoritmos requerem a especificação do número de intervalos. A segunda tarefa é encontrar a largura, ou os limites (pontos de corte), dos intervalos dados o intervalo de valores de um atributo contínuo.

O especialista do domínio pode interagir nessa etapa de modo a ajustar os pontos de corte ao contexto técnico do estudo para o melhor resultado da discretização. No entanto, existem casos onde não há especialistas disponíveis, não há ainda conhecimento desenvolvido sobre o assunto ou o volume de dados é muito grande para ser abstraído pelo especialista. Nesses casos é que os métodos de discretização são mais interessantes (MULLENBACH; RAKOTOMALALA, 2005). O processo básico de discretização, consistindo de quatro etapas, é mostrado no Quadro 2 e ilustrado no fluxograma da Figura 7.

Quadro 2 – Etapas do processo de discretização

1. Ordenando os valores contínuos da característica a ser discretizada;
2. Avaliando um ponto de corte para intervalos de divisão (procedimento recursivo) ou adjacentes para a fusão;
3. De acordo com algum critério, divisão ou mesclando intervalos de valor contínuo;
4. Finalmente parando em algum ponto (se o critério de parada for satisfeito).

Fonte: Discretização segundo (DOUGHERTY; KOHAVI; SAHAMI, 1995)

Figura 7 – Fluxograma geral de um método típico de discretização

Fonte: Adaptado de (LIU *et al.*, 2002)

No presente trabalho são utilizados métodos de discretização baseados em entropia, sendo os principais descritos a seguir.

2.1.5.1 Métodos de discretização baseados em entropia

Entropia é a medida da desordem de um conjunto de instâncias. Ela 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.

A entropia de informação de classe mede a quantidade de informação que seria necessária para especificar a qual classe uma instância pertence (HOSOKAWA, 2011).

2.1.5.1.1 Técnica de discretização MDL (descrição de comprimento mínimo)

O MDL (Minimum Description Length ou descrição de comprimento mínimo) é um método de discretização que utiliza o procedimento de minimização de entropia (FAYYAD; IRANI, 1993). 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 Shannon (1948), 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 (BARBOSA, 2015).

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, que foi a técnica selecionada para a avaliação do banco de dados dessa dissertação.

O método MDL utiliza uma abordagem de cima para baixo, onde vários intervalos são criados para formar uma árvore por meio múltiplas divisões do atributo numérico ao mesmo nó para produzir intervalos discretos (FAYYAD; IRANI, 1993). 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, 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 (OTTERBACH, 2016). Na Figura 8 é apresentado um fluxograma ilustrando as etapas do algoritmo MDL.

Figura 8 – Fluxograma do algoritmo de discretização MDL

Fonte: Elaborado pela autora

A Figura 9 apresenta um esquema com as características do método de discretização MDL. O MDL é um algoritmo de discretização Estático, pois é executado antes do processo de aprendizagem (LIU *et al.*, 2002). É um método Incremental, pois o processo de discretização é realizado iterativamente até que um critério de parada seja alcançado (TAKATUZI, 2017). A discretização leva em consideração o atributo classe sendo assim um método Supervisionado. O método inicia a discretização de cima para baixo (Top-down) com uma lista vazia de pontos de corte no início e durante o processo novos pontos de divisão são inseridos dividindo valores em intervalos menores. Por fim, é Local, pois faz uso apenas de parte das informações para decidir as faixas de discretização (FAYYAD; IRANI, 1993).

Figura 9 – Características do método MDL

Fonte: Elaborado pela autora

2.1.5.1.2 Técnica 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 10 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 (FAYYAD, 1991), gerando consequentemente um número maior de regras (FAYYAD; IRANI, 1993).

Figura 10 – Exemplo de ponto de corte provisório para o atributo pH em água

Fonte: Elaborado pela autora

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 (KONONENKO, 1995), como o Índice Gini, Erro de Classificação 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 (*Técnica BetterEncoding*).

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 (FAYYAD; IRANI, 1993).

Vale frisar que não existe uma técnica ou abordagem perfeita para a escolha do ponto de corte, embora o *BetterEncoding* tenha uma estratégia melhor que outras para o problema em questão, isso não significa que ela é perfeita.

2.1.5.1.3 Método de discretização Kononenko (KO)

O método MDL baseado em entropia de Kononenko (KO) utiliza os mesmos conceitos do MDL original de Fayyad e Irani (1993), incluindo um ajuste para a discretização de múltiplos atributos (KONONENKO, 1995). Para isso, o algoritmo de (KO) apresenta

uma correção para o viés que a medida da entropia apresenta para atributos com muitos valores (ISMAIL; CIESIELSKI, 2003).

Isso é obtido utilizando dois critérios de seleção, um baseado no princípio MDL e outro no algoritmo RELIEF (KIRA; RENDELL, 1992). A extensão incluída no algoritmo RELIEF permite tratar de forma mais eficiente dados com ruído, faltantes e problemas multi-classes (KONONENKO, 1995). 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 (PETER; BEALE, 2008).

2.1.5.2 Estratégia para gerar regras com conversão binária (Binary Split)

Uma das formas mais populares de lidar com atributos de múltiplos rótulos é distingui-los em outro conjunto de atributos binário.

O algoritmos de conversão trabalha apenas sobre os atributos, convertendo um atributo de 4 rótulos em 4 novos atributos. Um exemplo dessa técnica é ilustrado na Figura 11 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.

Figura 11 – Transformação baseada em rótulos

Fonte: Elaborado pela autora

2.2 Trabalhos relacionados

O objetivo desta seção é dar uma visão geral dos trabalhos realizados por outros pesquisadores no campo da MD aplicada na área agrícola. O armazenamento e análise de bancos de dados é muito popular em áreas como de telecomunicações, comércio, indústria, bancos, saúde, bioinformática, dentre outras (KELES, 2017). Na agropecuária há um número relativamente pequeno de trabalhos com aplicação das técnicas de MD para geração de regras de classificação, principalmente para a avaliação dos fatores que impactam na produtividade. Esse tipo de avaliação auxilia em diversas etapas do processo de produção como o plantio, manejo, colheita e pós-colheita.

De um modo geral, trabalhos de aplicações de MD na agricultura têm sido realizados para as mais variadas finalidades, conforme descrito em diversas revisões de literatura, como é o caso de Gandhi e Armstrong (2016a); Neelaveni e Rajeswari (2016); Raorane e Kulkarni (2013); Kalpana, Shanthi e Arumugam (2014); Patel e Patel (2014); Geetha (2015); Veenadhari, Misra e Singh (2011). As aplicações mais comuns estão relacionadas à avaliação de doenças e pragas, qualidade de frutas, aplicação de pesticidas, simulação de variáveis climatológicas, classificação e mapeamento de solos, previsão de preço de produtos agrícolas, dentre outros.

O Quadro 3 apresenta uma síntese dos trabalhos encontrados na literatura que avaliam o efeito de atributos climáticos, propriedades químicas e físicas dos solos, características de manejo, de produção e pós-produção na produtividade/lucratividade de algumas culturas agrícolas.

Quadro 3 – Síntese dos trabalhos sobre utilização de técnicas de MD com foco na avaliação dos fatores que mais influenciam na produtividade das culturas agrícolas

Referência	Cultura	País	Alvo	Atributos	N°	N° de Instancias	Técnicas utilizadas	Software	Geração de Regras
				Tipos					
Ahamed <i>et al.</i> (2015)	Arroz Batata Trigo	Bangladesh	Produtividade	Climáticos Solo Outros	10	30	K-means clustering Regressão linear k-NN Redes neurais	RapidMiner Studio	Não
Vriesmann <i>et al.</i> (2004)	Soja	Brasil	Produtividade	Solo	17	2388	Árvores de decisão Algoritmos genéticos	Implementados em Prolog e Borland Delphi	Sim
Garcia e Camolesi, (2015)	Cana	Brasil	Produtividade	Solo Variedade Manejo	23	18.973	Árvore de decisão (J48)	Weka	Não
da Silva <i>et al.</i> (2010)	Açaí	Brasil	Escoamento Produtividade	Características de produção e comercialização	23	533	Regras de Associação Algoritmo <i>Apriori</i>	Weka	Sim
Souza <i>et al.</i> (2010)	Cana	Brasil	Produtividade	Solo	7	118	Árvores de decisão com redução de entropia e regras de parada	SAS Enterprise Miner Ferramenta <i>Tree</i>	Não

Fonte: Elaborado pela autora

No trabalho de [Ahamed *et al.* \(2015\)](#) foram aplicadas técnicas de MD (modelos de regressão) em variáveis climáticas, propriedades dos solos e outros atributos como o uso de irrigação e o local da produção. O objetivo foi de identificar as condições mais adequadas de produção em termos de produtividade das culturas de arroz, batata e trigo em diferentes distritos de Bangladesh. A técnica de *k-means* foi utilizada para identificar grupos (não hierárquicos) similares após o agrupamento por distritos de produção. 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. A acurácia da predição variou de 90 a 95%, mas devido ao pequeno número de dados utilizados (15 distritos com dados de 2 anos), algumas anomalias foram observadas. As técnicas de redes neurais e regressão linear apresentaram os melhores desempenhos.

O mapeamento de fatores climáticos, propriedades do solo e a produtividade de uma determinada cultura produz uma grande quantidade de dados que podem conter informações estratégicas para a tomada de decisão nos processos de produção no campo. Para isso, [Vriesmann *et al.* \(2004\)](#) utilizaram técnicas de Inteligência Artificial, como Árvore de Decisão (AD) e Algoritmos Genéticos (AGs) para avaliar os principais fatores do solo (atributos físicos e químicos) que influenciam a produção da soja. Como resultado foram geradas diversas regras de interesse, incluindo os atributos matéria orgânica do solo, argila, pH, fósforo, cobre e boro.

[Garcia e Júnior \(2015\)](#) utilizaram Árvore de Decisão para identificar os fatores de maior influência na produtividade da cana-de-açúcar e a produção de etanol. O banco de dados 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 nessa dissertação, no que tange à busca por regras explicativas, foi realizado por [Silva, Rodrigues e Monteiro \(2010\)](#) utilizando Regras de Associação (RA) e o algoritmo *Apriori*, 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.

A abordagem de indução de Árvore de decisão usada por [Souza *et al.* \(2004\)](#) teve por objetivo avaliar os atributos químicos do solo e a produtividade da cultura de cana-de-açúcar. A indução de árvores de decisão permitiu verificar que a altitude e o teor de potássio no solo foram os atributos de maior influência nos mapas de produtividade de cana-de-açúcar na área estudada. Os resultados de MD, juntamente com análises de geoestatística, foram utilizados para se definir zonas de manejo para aplicações espacialmente variada de insumos em Agricultura de Precisão.

METODOLOGIA

3.1 Material e Métodos

Neste capítulo são apresentados o banco de dados (BD) utilizado no trabalho, com a descrição das variáveis que o compõem, a ferramenta computacional utilizada e os procedimentos metodológicos empregados nas etapas de pré-processamento e na mineração dos dados propriamente dita. O pré-processamento consistiu na seleção dos atributos mais relevantes do BD, na avaliação da normalidade e transformação de atributos para a escala logarítmica e na identificação de outliers. Com relação aos procedimentos de MD são descritos os procedimentos e métodos utilizados.

3.1.1 O Banco de Dados

O banco de dados utilizado neste trabalho foi 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, ou seja durante quatro safras agrícolas. O objetivo do projeto foi de realizar um levantamento do nível de infestação de fitonematoides em áreas produtoras de algodão no Mato Grosso (GALBIERI; BELOT, 2016). 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 (PERINA *et al.*, 2015).

Juntamente com o levantamento da densidade populacional dos fitonematoides no solo e nas raízes do algodoeiro, 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 12.

Figura 12 – Áreas amostradas (pontos vermelhos) em diferentes núcleos de produção de fazendas de algodão no Estado de Mato Grosso

Fonte: GALBIERI et al. (2014)

A Tabela 3 apresenta 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. Talhões típicos nessas regiões possuem áreas entre 200 e 300 hectares. De um modo geral, o período de plantio foi de novembro a março e as coletas 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.

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 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 então enviadas a laboratórios credenciados para a determinação dos atributos químicos, físicos e biológicos e todos os dados coletados e analisados eram registrados em planilha 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).

A planilha original continha 99 atributos e após uma avaliação inicial decidiu-se

Tabela 3 – 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 Banco de Dados

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

Fonte: Elaborado pela autora

manter apenas 68, os quais formam considerados os mais relevantes para o trabalho de mineração de dados proposto. Uma descrição dos atributos utilizados é apresentada na Tabela 4 (parâmetros físicos e químicos do solo) e Tabela 5 (informações gerais e doenças).

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 (PD) na palha 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 é a produtividade do algodoeiro, que de acordo com a Tabela 6 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. A Figura 13 apresenta as distribuições de frequência dessas duas produtividades (13a) e a correlação direta entre elas (13b). O procedimento de MD será aplicado nos dois tipos de produtividades como atributos alvo,

Tabela 4 – Descrição de parâmetros físicos e químicos do solo utilizados no trabalho

	Atributos	Sigla	Unidade	Tipo	Valores distintos
Parâmetros Físicos do Solo	Densidade do solo	DS	g cm ⁻³	numérico	1651
	Umidade do solo a base de volume	UMID-VOL	%	numérico	1474
	Umidade do solo a base de massa	UMID-MAS	%	numérico	1799
	Densidade das partículas do solo	DP	g cm ⁻³	numérico	269
	Porosidade total do solo	PT	%	numérico	1799
	Teor de argila (partículas menores que 2 μ m)	ARGILA	%	numérico	992
	Teor de silte (partículas entre 2 e 50 μ m)	SILTE	%	numérico	752
	Teor de areia (partículas entre 50 μ m e 2mm)	AREIA	%	numérico	1173
	Classe Textural do solo (muito argilosa, argilosa, média, arenosa)	CLAS-TEXT	-	nominal	4
	Teor de água disponível no solo	AD	%	numérico	359
	Condutividade elétrica do solo	CE	dS m ⁻¹	numérico	1346
	Resistência do solo à penetração na camada de 10 a 40 cm de prof.	RP10-40	MPa	numérico	530
Parâmetros Químicos do Solo	pH (em água) do solo	pH-AGUA	-	numérico	29
	pH (em CaCl ₂) do solo	pH-CLORETO	-	numérico	28
	Teor de fósforo no solo	P	mg dm ³	numérico	371
	Teor de potássio no solo	K	mg dm ³	numérico	413
	Teor de cálcio no solo	Ca	cmol _c dm ⁻³	numérico	51
	Teor de magnésio no solo	Mg	cmol _c dm ⁻³	numérico	17
	Teor de cálcio + magnésio no solo	Ca+Mg	cmol _c dm ⁻³	numérico	64
	Teor de alumínio no solo	Al	cmol _c dm ⁻³	numérico	8
	Teor de hidrogênio no solo	H	cmol _c dm ⁻³	numérico	71
	Teor de matéria orgânica no solo	MO	g dm ³	numérico	47
	Soma de bases	SOMA-BASES	cmol _c dm ⁻³	numérico	63
	Capacidade de troca de cátions	CTC	cmol _c dm ⁻³	numérico	81
	Saturação por bases	SAT-BASES	%	numérico	458
	Relação cálcio-magnésio	Ca/Mg	-	numérico	13
	Relação cálcio-potássio	Ca/K	-	numérico	344
	Relação magnésio-potássio	Mg/K	-	numérico	151
	Saturação por cálcio	SAT-Ca	-	numérico	374
	Saturação por magnésio	SAT-Mg	%	numérico	162
	Saturação por alumínio	SAT-Al	%	numérico	113
	Saturação por potássio	SAT-K	%	numérico	55
	Saturação por hidrogênio	SAT-H	%	numérico	431
	Teor de zinco no solo	Zn	%	numérico	161
	Teor de cobre no solo	Cu	mg dm ³	numérico	57
	Teor de ferro no solo	Fe	mg dm ³	numérico	362
	Teor manganês no solo	Mn	mg dm ³	numérico	295
	Teor de enxofre no solo	S	mg dm ³	numérico	242
	Teor de boro no solo	B	mg dm ³	numérico	102
	Teor de fósforo no solo - método da resina	P-res	mg dm ³	numérico	525

Fonte: Elaborado pela autora

uma vez que a PROD-SAFRA pode estar relacionada com atributos que se modificam com o tempo como doenças e a fertilidade do solo (parâmetros químicos), dentre outros. Já a PROD-HIST deve estar mais relacionada com atributos mais estáveis como por exemplo o tipo de solo (teores de areia, silte e argila). Isso se justifica, pois conhecer regras de associação dos atributos da BD com esses 2 tipos de produtividades pode fornecer informações relevantes ao sistema de produção, podendo gerar recomendações técnicas aos produtores.

A média da PROD-HIST em todos os talhões foi de 264,4 @/ha e da PROD-SAFRA foi de 262,1 @/ha, ou seja há boa concordância geral dessas duas variáveis, mas em alguns casos verifica-se grandes diferenças entre o histórico e a safra atual. Isso indica a influência de fatores como a ocorrência de doenças, adubação e possivelmente variáveis climáticas (que não fazem parte desse banco de dados) no ano da safra avaliada (PROD-SAFRA). Dessa forma uma análise de outlier será realizada nos atributos de produtividades para avaliar a necessidade de ser excluir alguma instância, usando como critério a diferença entre as produtividades históricas e na safra.

Tabela 5 – Descrição de atributos gerais e de doenças do algodoeiro utilizados no trabalho

	Atributos	Sigla	Unidade	Tipo	Valores distintos
Informações Gerais	Espaçamento entre linhas de plantio	ESP	metro	nominal	4
	Número de plantas por metro na linha	Planta/m	número	numérico	15
	Número de plantas por hectare	Planta/há	número	numérico	42
	Sistema de produção (Plantio de Safra ou Safrinha)	Safra-Safrinha	-	nominal	2
	Mês de semeadura	MÊS	mês	numérico	4
	Cultivar plantada no talhão	CULTIVA	-	nominal	43
	Sistema de plantio (Plantio Direto ou Convencional)	SIST-PLANTIO	-	nominal	2
	Preparo do solo (com ou sem subsolagem)	PREP-SOLO	-	nominal	2
	Abertura da área para agricultura	ABERTURA	anos	numérico	32
	Início de plantio de algodão	INI-ALGODAO	anos	numérico	22
	Cultura anterior ao algodão na safra atual	ANTE-CULTURA	-	nominal	18
	Cultura de cobertura na entressafra	COBERTURA	-	nominal	10
	Produtividade histórica do algodão no talhão	PROD-HIST	arroba/hectare	numérico	191
	Produtividade do algodão no talhão na safra avaliada	PROD-SAFRA	arroba/hectare	numérico	248
Doenças	Ocorrência de murcha de <i>Fusarium</i>	FUSARIUM	-	nominal	2
	População do nematoide <i>Meloidogyne incognita</i> no solo	MELOI-SOLO	espécimes em 200 cm ³ solo	numérico	232
	População do nematoide <i>Meloidogyne incognita</i> na raiz	MELOI-RAIZ	espécimes em 5g raiz	numérico	119
	População do nematoide <i>Meloidogyne incognita</i> no solo + raiz	MELOI-TOTAL	espécimes em 200 cm ³ solo + 5g raiz	numérico	240
	População do nematoide <i>Pratylenchus brachyurus</i> no solo	PRATY-SOLO	espécimes em 200 cm ³ solo	numérico	77
	População do nematoide <i>Pratylenchus brachyurus</i> na raiz	PRATY-RAIZ	espécimes em 5g raiz	numérico	145
	População do nematoide <i>Pratylenchus brachyurus</i> no solo + raiz	PRATY-TOTAL	espécimes em 200 cm ³ solo + 5g raiz	numérico	163
	População do nematoide <i>Heterodera glycines</i> no solo	HETE-SOLO	espécimes em 200 cm ³ solo	numérico	25
	População do nematoide <i>Rotylenchulus reniformis</i> no solo	ROTY-SOLO	espécimes em 200 cm ³ solo	numérico	133
	População do nematoide <i>Rotylenchulus reniformis</i> na raiz	ROTY-RAIZ	espécimes em 5g raiz	numérico	19
	População do nematoide <i>Rotylenchulus reniformis</i> no solo + raiz	ROTY-TOTAL	espécimes em 200 cm ³ solo + 5g raiz	numérico	127
	População de outros nematoides no solo	OUTROS-SOLO	espécimes em 200 cm ³ solo	numérico	173
	População de outros nematoides na raiz	OUTROS-RAIZ	espécimes em 5g raiz	numérico	73
	População de outros nematoides no solo + raiz	OUTROS-TOTAL	espécimes em 200 cm ³ solo + 5g raiz	numérico	189

Fonte: Elaborado pela autora

Figura 13 – Histogramas de frequência das produtividades do algodoeiro na safra (PROD-SAFRA) e a produtividade histórica (PROD-HIST) (a) e a correlação entre essas 2 produtividades

Fonte: Elaborado pela autora

3.1.2 Procedimento metodológico utilizado

O procedimento metodológico adotado para a obtenção das regras de associação entre os atributos e informações do sistema de produção do algodão com a variável alvo Produtividade é sintetizado na Figura 14.

Figura 14 – Esquema do procedimento utilizado para a geração de regras de associação

Fonte: Elaborado pela autora

Uma avaliação geral das estatísticas dos atributos numéricos é apresentada na Tabela 6 com valores mínimos, máximos, médios, medianos, desvios padrão (DP), coeficientes de variação (CV), assimetria e curtose. Valores baixos de assimetria e curtose entre 0 e 3 indicam uma tendência de distribuição normal dos dados. Assim, fica claro pelos dados da Tabela 6 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 nematoides antes da etapa de discretização. A transformação utilizada é descrita a seguir.

$$P_i^* = \text{Log}(P_i + 1) \quad (3.1)$$

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.

Uma análise de *outlier* foi realizada com os dados de produtividade do algodoeiro, uma vez que em alguns talhões foram observados valores de produtividade da safra bastante distintos da produtividade histórica. Utilizou-se como critério para a identificação de *outliers* baseado no box-plot, com ponderação ao desvio interquartil de 1,5.

A discretização 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. Dessa forma estabeleceu-se as classes de produtivi-

dade Baixa, Média e Alta, onde a distribuição do número de instâncias em cada classe foi relativamente homogênea, conforme mostrado na Tabela 7.

Tabela 6 – 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
Parâmetros Químicos do Solo	RP10-40	0,74	9,02	2,12	1,84	1,11	52	2,2	6,2
	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

Fonte: Elaborado pela autora

Tabela 7 – Avaliação estatística dos atributos com características numéricas

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

Fonte: Elaborado pela autora

3.1.2.1 Ferramenta utilizada

A análise dos dados foi realizada com a ferramenta WEKA (*Waikato Environment for Knowledge Analysis*). O WEKA é um software de mineração de dados *open source* (código aberto), desenvolvido pelo Departamento de Ciência da Computação da Universidade de Waikato, Nova Zelândia, que é bastante utilizado mundialmente. É uma ferramenta simples, relativamente completa e de fácil utilização, que permite a aplicação de diferentes métodos de discretização, seleção de atributos e de balanceamento de classes.

Em geral, os trabalhos realizados no WEKA utilizam o modo *Explorer* indicado na página inicial do software e ilustrado na Figura 15. Nesse modo, pode-se visualizar e explorar o conjunto de dados, aplicar filtros, executar tarefas de aprendizado de máquina como classificação, agrupamentos, regras de associação e seleção de atributos (AMARAL, 2016).

Figura 15 – Tela inicial do software WEKA para mineração de dados

3.1.2.2 Discretização dos atributos

Dentre os 68 atributos do BD, 2 deles (PROD-SAFRA e PROD-HIST) são atributos alvo, os quais foram discretizados, com auxílio de especialistas da área, em 3 classes (baixa, média e alta, conforme mostrado na Tabela 7). Os outros 66 atributos são utilizados para a obtenção das regras de associação para os atributos alvo produtividade. Desses, 10 são atributos nominais e 56 são numéricos (Tabelas 4 e 5), havendo portanto

a necessidade da transformação deles em atributos discretos, utilizando as técnicas de discretização do Weka.

Neste trabalho foram utilizadas 4 abordagens para a discretização desses 56 atributos numéricos, sendo 3 automáticas (disponível no Weka) e uma manual. A discretização manual dos atributos foi realizada com informações fornecidas por especialistas da área. Os limites das classes definidas manualmente para os atributos numéricos são apresentados na Tabela 8. O número de classes variou de 4 a 6 classes.

Tabela 8 – Limites de classes da discretização manual dos atributos numéricos realizada com informações fornecidas por especialistas da área

Atributo	Classe 1		Classe 2		Classe 3		Classe 4		Classe 5		Classe 6	
	<i>Limites</i>	<i>Talhões</i>	<i>Limites</i>	<i>Talhões</i>	<i>Limites</i>	<i>Talhões</i>	<i>Limites</i>	<i>Talhões</i>	<i>Limites</i>	<i>Talhões</i>	<i>Limites</i>	<i>Talhões</i>
Planta/metro	4-7	145	8-9	655	10-11	716	12-20	283				
Planta/hectare	50-100 mil	471	100-150 mil	1014	150-200 mil	160	>200 mil	154				
ABERTURA	5-17	462	10-20	471	21-27	441	28-40	425				
INI-ALGODAO	1-6	501	7-9	529	10-14	405	15-25	364				
MELOI-SOLO	0 (ausente)	1401	10-100	101	100-1000	115	1000-3000	94	>3000	88		
MELOI-RAIZ	0 (ausente)	1446	10-30	113	30-200	100	200-1000	94	>1000	46		
PRATY-SOLO	0 (ausente)	517	10-50	946	50-80	131	80-200	108	>200	97		
PRATY-RAIZ	0 (ausente)	113	10-100	764	100-250	450	250-500	250	>500	221		
HETE-SOLO	0 (ausente)	1538	10-20	134	20-40	50	40-100	46	>100	31		
ROTY-SOLO	0 (ausente)	1577	10-40	57	40-500	54	500-2000	56	>2000	55		
OUTROS-SOLO	0 (ausente)	663	10-70	289	70-250	282	250-550	288	>550	277		
OUTROS-RAIZ	0 (ausente)	1103	10-20	174	20-50	194	50-130	176	>130	152		
DS	0,64-1,09	284	1,10-1,17	325	1,18-1,22	265	1,23-1,30	319	1,31-1,43	309	1,44-1,71	297
UMID-VOL	2,7-18,8	299	18,8-24,5	302	24,5-28	296	28,0-30,8	300	30,8-33,9	300	33,9-45,5	302
UMID-MAS	2,1-13,5	299	13,5-19,1	302	19,1-22,3	300	22,7-25,3	301	25,3-28,4	300	28,4-41,4	297
DP	2,61-2,67	340	2,68-2,70	316	2,71-2,72	230	2,73-2,74	299	2,75-2,77	364	2,78-2,96	250
ARGILA	4,3-35,0	293	35,1-47,5	303	47,6-54,5	302	54,6-60	302	60,1-65,0	308	65,1-79,1	291
SILTE	0-2,7	324	2,8-4,0	279	4,1-5,7	301	5,8-8,4	304	8,5-13,7	293	13,8-31,9	298
AREIA	7,0-24,0	296	24,1-30,3	304	30,4-37,5	302	37,6-45,7	301	45,8-60,6	299	60,7-91,2	297
CE	0,02-0,31	303	0,32-0,40	317	0,41-0,48	286	0,49-0,60	303	0,61-0,79	297	0,80-3,32	280
RP10-40	0,74-1,23	298	1,24-1,50	311	1,51-1,84	298	1,85-2,26	294	2,27-3,0	362	3,01-9,02	236
pH-AGUA	4,8-5,5	177	5,6-5,7	262	5,8-5,9	314	6,0-6,1	444	6,2-6,3	409	6,4-7,6	203
pH-CLORETO	4,2-4,8	255	4,9-5,0	188	5,1-5,2	500	5,3-5,4	398	5,5-5,6	334	5,7-7,0	124
P	2,7-17,6	299	17,7-21,9	308	25,0-33,6	298	33,7-41,8	295	42,0-51,7	307	52,1-87,8	292
K	10,0-43,8	301	44,0-56,0	311	56,2-69,0	297	69,1-83,7	292	84,0-103,0	301	103,9-283,0	297
Ca	0,7-2,2	295	2,3-2,6	343	2,7-2,9	339	3,0-3,2	273	3,3-3,6	258	3,7-6,2	291
Mg	0,3-0,8	258	0,9-1,0	552	1,1	333	1,2	353	1,3-1,4	128	1,5-1,9	175
Ca+Mg	1,0-3,1	297	3,2-3,6	341	3,7-4,0	345	4,1-4,4	267	4,5-4,9	261	5,0-8,1	288
Al	0 (ausente)	1622	0,1-0,2	45	0,3	75	0,4-0,7	57				
H	0,1-2,8	278	2,9-3,4	313	3,5-3,9	290	4,0-4,4	324	4,5-5,1	294	5,2-8,9	300
MO	10,7-24,8	174	25,6-29,5	284	29,5-33,9	446	34,8-37,8	359	38,9-41,0	264	42,2-54,2	271
SOMA-BASES	1,2-3,2	274	3,3-3,7	284	3,8-4,1	335	4,2-4,5	293	4,6-5,1	333	5,2-8,2	280
CTC	3,7-6,7	276	6,8-7,6	323	7,7-8,2	284	8,3-8,8	297	8,9-9,6	334	9,7-14,0	285
SAT-BASES	13,1-40,9	265	41,0-47,7	324	47,8-51,5	308	51,6-55,2	301	55,3-60,9	295	61,0-98,4	306
Ca/Mg	2,0-2,5	403	2,6	335	2,7	500	2,8	409	2,9-3,3	152		
Ca/K	4,1-10,2	288	10,3-12,9	309	13,0-15,7	309	15,8-19,1	281	19,2-25,1	310	25,2-65,4	302
Mg/K	1,7-3,8	264	3,9-4,8	326	4,9-5,8	311	5,9-7,1	287	7,2-9,3	314	9,4-25,5	297
SAT-Ca	8,0-28,1	278	28,2-32,7	315	32,8-35,5	309	35,6-38,4	306	38,5-42,4	290	42,5-72,5	301
SAT-Mg	3,7-10,8	274	10,9-12,3	299	12,4-13,2	307	13,3-14,3	360	14,4-15,6	276	15,7-23,9	282
SAT-Al	0 (ausente)	1622	0-7,3	56	7,4-11,0	62	11,1-37,0	59				
SAT-K	0,6-1,3	262	1,4-1,7	282	1,8-1,9	159	2,0-2,7	534	2,8-3,2	278	3,3-8,9	286
SAT-H	1,6-38,0	256	38,1-44,2	317	44,3-48,3	315	48,4-52,2	322	52,3-59,0	325	59,1-80,4	264
Zn	0,7-3,2	265	3,3-4,5	295	4,6-5,7	281	5,8-7,3	364	7,4-9,3	314	9,4-19,2	280
Cu	0,2-0,9	300	1,0-1,2	370	1,3-1,5	295	1,6-1,8	234	1,9-2,4	317	2,5-9,7	283
Fe	22,0-54,4	277	55,0-64,0	318	64,2-76,0	330	76,3-86,0	273	86,3-102,8	309	103,0-212,3	293
Mn	4,2-9,9	269	10,0-12,4	335	12,5-14,7	292	14,8-17,4	303	17,5-21,5	297	21,6-111,9	303
S	5,8-9,6	298	9,7-10,8	313	10,9-12,2	303	12,3-14,4	292	14,5-17,8	294	17,9-60,6	299
B	0,22-0,35	302	0,36-0,40	377	0,40-0,47	244	0,48-0,52	292	0,53-0,65	304	0,66-2,86	280
P-res	1,2-22,0	286	22,2-28,1	292	28,2-36,3	318	36,4-43,8	303	43,9-56,1	296	56,2-97,2	304

Fonte: Elaborado pela autora

Além da discretização manual, foram utilizadas 3 técnicas de discretização automática disponíveis no WEKA, baseadas no método MDL de Fayyad, denominados de Kononenko (KO), Better Encoding (BE) e a combinação das duas técnicas (KO+BE). O desempenho dessas 4 estratégias de discretização (3 automáticas e uma manual) serão

avaliadas neste trabalho.

A seguir são descritas as ações no WEKA para a execução da discretização automática dos atributos (ilustrado também na Figura 16). Na aba de pré-processamento seleciona-se a opção de filtro *Supervised => Attribute => Discretize*. Ao abrir a janela de configuração de discretização, seleciona-se a opção *Kononenko (KO)*, *Better Encoding (BE)* ou ambas (*KO + BE*) para posteriormente aplicar-se os algoritmos de aprendizado.

Figura 16 – Tela do software WEKA mostrando as abas necessárias para a realização da etapa de discretização

É importante mencionar que não foram observados problemas significativos de desbalanceamento das classes geradas. Isso é importante, pois o desbalanceamento é um dos fatores que interferem negativamente no desempenho dos algoritmos de aprendizado de máquina (PRATI; BATISTA; MONARD, 2006). A existência de classes desbalanceadas ocorre quando o número de elementos entre as classes é desproporcional. Nesse caso, exemplos da classe minoritária são geralmente classificados erroneamente.

3.1.2.3 Algoritmos para extração de regras de classificação

Após o pré-processamento, que inclui a transformação dos atributos de doenças (basicamente nematoides) para a escala logarítmica, a avaliação de *outliers* do atributo produtividade e a discretização dos atributos numéricos realizou-se a etapa de classificação, aplicando-se os algoritmos JRip e PART (HALL *et al.*, 2009). Maiores detalhes desses métodos podem ser obtidos em (COHEN, 1995) e (WITTEN; FRANK, 1998).

A classificação é a etapa mais importante da KDD, pois consiste na busca por relações que permitem associar cada instância do banco de dados aos rótulos denominados classes (GOLDSCHMIDT; PASSOS, 2005). Como é uma tarefa preditiva, essas relações podem ser aplicadas a novos registros para prever as classes em que tais registros se

enquadram. Feita a classificação, é gerada uma matriz denominada Matriz de Confusão que apresenta a quantidade de registros rotulada corretamente e incorretamente pelo algoritmo empregado. O desempenho dessas duas técnicas será comparado neste trabalho.

Para a geração das regras de classificação foi utilizada também a estratégia da conversão binária (*binary split*), que faz a transformação de atributos discretos multi-rótulos do banco de dados em múltiplos problemas binários usando a técnica *um-contratodos* descrita na seção sobre discretização.

3.1.2.4 Abordagem para treinamento e teste dos modelos de classificação

Em tarefas de classificação é necessário avaliar o desempenho do classificador. Há várias abordagens para avaliar a qualidade e as características de um modelo de mineração de dados. Entre elas inclui-se o uso de medidas estatísticas, como precisão, revocação (*recall*) e custo. Essas abordagens são úteis no processo de mineração de dados e são usadas iterativamente ao criar, testar e refinar modelos para responder a um problema específico.

Nesta etapa do trabalho é possível utilizar uma das três abordagens descritas a seguir para a avaliação do desempenho dos classificadores: 1) *Percent Split*: também conhecido como *hold out*, onde utiliza-se normalmente 70% dos registros para treino e 30% para teste; 2) *User training set*: aplica o modelo sobre o mesmo conjunto de dados em que foi treinado, ou seja, todos os registros são utilizados para a construção do modelo e também para avaliar o desempenho do modelo construído; e 3) *k-fold cross-validation* (validação cruzada): divide o conjunto de dados em K partes e separa uma parte para teste e realiza o treinamento com as demais partes, sendo que o procedimento é repetido para todas as partes. A acurácia final do modelo é a média das acurácias parciais calculadas para cada uma das partes.

Neste trabalho adotou-se a validação cruzada com o padrão de 10 *folds*, ou seja, com 10 testes sobre o modelo. A escolha por esse último método de validação deve-se ao fato de ser um dos mais utilizados e por fornecer bons resultados mesmo quando o conjunto de dados é pequeno.

3.1.2.5 Avaliação sensível ao custo (*Cost-sensitive evaluation*)

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 em novos exemplos.

Através da matriz de confusão é apresentado os resultados obtidos pelo algoritmo de classificação para cada uma das classes (de produtividade). Então, a matriz de custo é usada para ponderar os erros, apresentados na matriz de confusão, visando a avaliação

dos tipos de erros cometidos, especialmente quando a variável-target (atributo-target) é nominal ordinal (qualitativo ordinal), como é o caso da produtividade. Dessa forma, um classificador que erra uma produtividade ALTA como BAIXA, dever ser pior do que outro que erra a produtividade ALTA como MÉDIA. Para criar um sistema de avaliação desses erros, utiliza-se uma matriz de custo com pesos para esse tipo de erro.

Um dos principais indicadores da taxa de acerto da classificação é dado pela análise de custo, também denominada de matriz de custo, com ordem $k \times k$ (k : número de classes do atributo alvo). Como convenção adota-se que as linhas da matriz de custo correspondem a classes preditas (CP), enquanto as colunas correspondem às classes atuais ou reais (CA). A Tabela 9a ilustra uma 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 ($n_0=0$). Assim, custo (CP_i, CA_i) = 0, uma vez que não se constitui um erro, e custo(CP_i, CA_j) > 0, quando $i \neq j$. 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 estudado. No caso desse trabalho, onde adotou-se as classes Alta, Média e Baixa para o atributo alvo produtividade do algodoieiro, considerou-se uma penalidade de 1 para as classes vizinhas (Alta-Média; Média-Baixa) e 2 para classes extremas (Alta-Baixa), conforme mostrado na Tabela 9b. 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.

Tabela 9 – 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	→	n_{k-1}	n_k
CP ₂	n_1	n_0	→	→	n_{k-1}
⋮			n_0		
CP _{k-1}	n_{k-1}	←	←	n_0	n_1
CP _k	n_k	n_{k-1}	←	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

Fonte: Elaborado pela autora

A seguir são apresentadas as ações necessárias para se alterar os pesos da matriz de custo no WEKA. Na aba *Classify* selecione *More options* e depois *Classifier evaluation*

option (Figura 17), marcando-se a opção *Cost-sensitive evaluation* e clicando-se posteriormente no botão *Set* (indicado na Figura 17 com uma seta preta). Na opção *Classes* da janela *Cost Matriz Editor*, deve-se digitar o número de classes do atributo alvo, que no caso da produtividade do algodoieiro é 3 (Alta, Média e Baixa) e em seguida *Resize*. Com isso aparece a matriz com os pesos padrão de zero na diagonal e 1 nos demais campos, os quais podem ser alterados conforme interesse, com exceção dos valores zero da diagonal da matriz. Nesse trabalho optou-se por um valor de 2 (penalidade no custo de 200%) no caso onde as instância são classificadas como de Alta sendo de Baixa produtividade (ou vice versa).

Figura 17 – Janela do WEKA para alterar os pesos da matriz de custo

O cálculo do custo total (*Total cost*) para o exemplo da matriz de custo apresentado na Tabela 9c é realizado conforme descrito à seguir:

$$\text{Custo} = (374 \times 0) + (113 \times 1) + (134 \times 2) + (195 \times 1) + (139 \times 0) + (215 \times 1) + (140 \times 2) + (86 \times 1) + (403 \times 0) = 1157 \quad (3.2)$$

O valor obtido (custo=1157) traz um indicativo do grau de inadequação do classificador, medido por seu erros cometidos para uma variável-target ordinal (produtividade). Quanto menor o valor do custo total, mais adequado será o modelo está ao problema em questão.

3.1.2.6 Outros indicadores de desempenho utilizados

Além do custo total, foram utilizados outras métricas indicadoras do desempenho das regras de classificação geradas, como precisão (P), revocação (R) e medida F, definidas

no Capítulo 2. Adicionalmente foi introduzido um novo índice denominado de I_{FC} , que integra as métricas P, R, F e o custo (C), definido como:

$$I_{FC} = \frac{100.F}{\text{Log}_2.C} \quad (3.3)$$

A Tabela 10 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.

Tabela 10 – Cálculos da precisão (P), revocação (R), medida F e índice I_{FC} para o exemplo de classificação apresentado na Tabela 10c, 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}	$I_{FC} = 100xF/Log_2.C$			4,776

Fonte: Elaborado pela autora

Além das métricas descritas anteriormente, o número de regras gerado será utilizado como um indicador importante na seleção do melhor classificador.

RESULTADOS E DISCUSSÕES

Neste capítulo são apresentados e discutidos os resultados da mineração de dados para a classificação da produtividade do algodoeiro, bem como a avaliação das regras geradas. Inicialmente, faz-se uma análise de outliers baseada na diferença entre as produtividades histórica e da safra (resíduo). Dentre os aspectos avaliados na classificação, faz-se uma comparação e seleção do melhor classificador (PART, JRip), avalia-se o desempenho dos métodos de discretização (manual, automática pelo método de Kononenko, Better Encoding e a combinação dos dois) e por fim realiza-se a interpretação das regras geradas.

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

Conforme mencionado anteriormente, os atributos alvo definidos neste trabalho são as duas produtividades do algodoeiro disponíveis no banco de dados, que são as produtividades no ano da coleta (PROD-SAFRA) e a produtividade histórica da área (PROD-HIST), bem a média delas. Uma análise preliminar (Figura 13) mostrou que em alguns talhões houve diferenças grandes entre essas duas variáveis, o que pode ter sido causado por erros de medida no campo, ou variações climáticas entre diferentes safras (anos) e ocorrência de doenças e pragas no ano da coleta. Assim, faz-se necessário uma avaliação de *outliers* e da necessidade ou não da exclusão de instâncias.

A Figura 18 apresenta o gráfico de *box plot* e a estatística do cálculo dos *outliers*, indicando que apenas 28 instâncias (talhões) apresentaram resíduos fora dos limites inferiores e superiores estabelecidos pela ponderação do desvio interquartil de 1,5. Testes realizados com e sem a inclusão dessas 28 instâncias (*outliers*) tiveram efeito pouco significativo na precisão da classificação e essas instâncias foram, portanto, mantidos no BD.

Figura 18 – Gráfico de *box plot* dos resíduos entre as produtividades histórica e da safra (a) e a estatística (b), utilizando uma ponderação ao desvio interquartil de 1,5

Fonte: Elaborado pela autora

4.0.2 Análise comparativa dos classificadores PART e JRip

Uma avaliação comparativa do desempenho dos algoritmos de regras PART e JRip foi realizada considerando a discretização manual dos atributos (Tabela 8) e 2 discretizações automáticas, utilizando as técnicas de Better Encoding (BE) e a combinação Kono-nenko (KO) com BE, disponíveis no WEKA. Os resultados da validação cruzada (10-*folds*) são apresentados nas Tabelas 11 (para discretização manual), 12 (para discretização por BE) e 13 (para discretização por KO + BE) com as métricas de revocação, precisão, custo e número de regras geradas, em função do número mínimo de instâncias por folha na classificação (I/F). Os números de I/F foram selecionados para essa avaliação entre 20 e 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.

Tabela 11 – 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/F), com discretização manual. Classes: B=baixa, M=média, A=alta produtividade.

I/F	Revocação						Precisão						Custo		Nº de Regras	
	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

Fonte: Elaborado pela autora

Tabela 12 – 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/F), com discretização pelo algoritmo Better Encoding (BE). Classes: B=baixa, M=média, A=alta produtividade.

I/F	Revocação						Precisão						Custo		Nº de Regras	
	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

Fonte: Elaborado pela autora

Tabela 13 – 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/F), com discretização por Kononenko (KO) e Better Encoding (BE). Classes: B=baixa, M=média, A=alta produtividade.

I/F	Revocação						Precisão						Custo		Nº de Regras	
	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

Fonte: Elaborado pela autora

Como pode ser observado nas Tabelas 11, 12 e 13 o classificador JRip apresenta número de regras bem menor que o PART, o que pode ser interessante por facilitar a interpretação das associações geradas. Entretanto, as demais métricas como custo, precisão e revocação são muito mais favoráveis para o classificador PART. Nota-se que apesar da revocação e da precisão obtidas para o JRip serem ligeiramente maiores para a classe de baixa produtividade (B), foi menor para classe de alta (A) e muito inferior para a classe de produtividade média (M), sendo na maioria das vezes zero.

Os gráficos na Figura 19 apresentam um panorama geral de comparação de desempenho dos classificadores PART e JRip, bem como do efeito do número mínimo de instâncias por folha (I/F), nas métricas avaliadas. Na maioria dos casos avaliados tanto em termos de diferentes métodos de discretização como I/F, obtém-se menores custos, maiores números de regras 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 média (M), conclui-se que o algoritmo PART tem desempe-

no muito superior ao JRip para o caso desse banco de dados analisado, sendo portanto selecionado para as próximas etapas do trabalho.

Figura 19 – 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 11, 12 e 13).

Fonte: Elaborado pela autora

4.0.3 Desempenho das técnicas de discretização, da conversão binária e do número I/F

Conforme verificou-se na análise realizada na seção anterior para comparar os classificadores PART e JRip, a discretização manual (realizada pelo especialista) apresentou custos muito superiores que as 3 técnicas de discretização automática (KO, BE e KO+BE) disponíveis do WEKA. Assim, na análise comparativa das técnicas de discretização apresentadas na Tabela 14, 15 e 16 foi desconsiderada a discretização manual, pelo seu baixo desempenho, conforme já indicado anteriormente.

De um modo geral, considerando a média de todos os valores de I/F avaliados (últimas linhas nas Tabelas 14, 15 e 16), 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 e 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} .

Tabela 14 – Métricas de avaliação da classificação por validação cruzada, com o classificador PART, em função de I/F, 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/F	Acurácia	Revocação			Precisão			Média	Custo	Nº de Regras	Medida F			Média	I_{FC}
		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

Fonte: Elaborado pela autora

Tabela 15 – Métricas de avaliação da classificação por validação cruzada, com o classificador PART, em função de I/F, 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/F	Acurácia	Revocação			Precisão			Média	Custo	Nº de Regras	Medida F			Média	I_{FC}
		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

Fonte: Elaborado pela autora

Tabela 16 – Métricas de avaliação da classificação por validação cruzada, com o classificador PART, em função de I/F, 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/F	Acurácia	Revocação			Precisão			Média	Custo	Nº de Regras	Medida F			Média	I_{FC}
		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

Fonte: Elaborado pela autora

Figura 20 – Variação das métricas de acurácia, precisão, custo, nro. de regras, medida F e I_{FC} com nro. mínimo de instâncias por folha (I/F), obtidos por validação cruzada, com o classificador PART e as 3 técnicas de discretização automática (KO, BE e KO+BE).

Fonte: Elaborado pela autora

As métricas de acurácia, precisão, medida F e número de regras diminuem e o custo aumenta com o aumento de I/F. Considerando o melhor compromisso entre precisão e custo (representado por I_{FC}) e o número de regras gerado, selecionou-se I/F de 30 como ideal. Entre as 3 técnicas de discretização, as diferenças nas métricas de avaliação de desempenho foram menos significativas do que a variação com I/F. 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/F, como pode ser observado na Figura 20. Assim, considera-se que o desempenho da técnica de discretização KO+BE foi superior às demais.

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, são apresentadas nas Tabelas 17 e 18 e na Figura 21.

Verifica-se que a utilização da conversão binária produz um aumento significativo no I_{FC} , sem aumento do número de regras (há na realidade uma pequena diminuição no número de regras, em média). Na prática, a conversão binária inclui a opção "diferente de" além das opções convencionais de maior ou menor nas regras geradas. Essa opção de $\neq$ de é especialmente interessante para o BD de produtividade do algodoeiro, pois pode explicitar relações onde a resposta esperada de um determinado atributo na produtividade não é unidirecional, pois pode haver uma faixa ideal que maximiza ou minimiza a

Tabela 17 – Métricas de desempenho (valores médios para I/F 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
I _{FC}	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

Fonte: Elaborado pela autora

Tabela 18 – Métricas de desempenho (valores médios para I/F 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.

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

Fonte: Elaborado pela autora

produtividade. Com relação ao atributo alvo utilizado, a produtividade Média forneceu os melhores indicadores de desempenho, como pode ser verificado na Figura 21.

Em síntese, após a avaliação das métricas de desempenho considerando os algoritmos de classificação (PART e JRip), as técnicas de discretização (KO, BE, KO+BE e manual), os diferentes tipos de produtividade (safra, real, média), diferentes I/F e a conversão binária, obteve-se as melhores condições de classificação, considerando as 3 classes de produtividade (Baixa, Média e Alta), conforme descrito no Quadro 4 a seguir:

4.0.4 Avaliação das regras de associação geradas

Os Quadros de 5, 6 e 7 apresentam as regras, no formato SE-ENTÃO, obtidas por validação cruzada, com o classificador PART, discretização por KO, BE e KO+BE e a conversão binária. O condicionante SE são os atributos predicados (características de

Figura 21 – Valores médios de I_{FC} ($100 \cdot F / \log_2 \text{Custo}$) 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).

Fonte: Elaborado pela autora

Quadro 4 – Melhores condições de classificação, considerando as 3 classes de produtividade (Baixa, Média e Alta).

Classificador: PART	Nro. mínimo de instâncias por folha (I/F): 30
Técnica de discretização: combinação KO+BE	Uso da conversão binária: Sim
Tipo de produtividade: Média	

Fonte: Elaborado pela autora

manejo, doenças e físico-químicas do solo) e o ENTÃO o valor do atributo alvo (produtividade). São apresentadas apenas as regras com cobertura (taxa de acerto) maior que 50%.

Quadro 5 – Regras com cobertura maior que 50% para validação cruzada, classificador PART, conversão binária e discretização por BE.

ALTA
1. SE ABERTURA $\neq 7$ E INI-ALGODAO > 2 E UMID-MAS > 8 E SOMA-BASES $> 2,7$ E B = (0,8-1,7) ENTÃO PRODUTIVIDADE = Alta (97/33)
2. SE INI-ALGODAO $\neq 4$ E ESP > 76 E INI-ALGODAO $\neq 3$ E CULTIVAR $\neq$ FMT 701 E FUSARIUM = Sem_Presença E SIST-PLANTIO = Semeadura direta E RP10-40 $\geq 2,7$ E UMID-MAS $\neq$ (8-16) ENTÃO PRODUTIVIDADE = Alta (367/141)
3. SE CULTIVAR $\neq$ FM966 LL E CULTIVAR = FM910 E COBERTURA = Soja ENTÃO PRODUTIVIDADE = Alta (39/14)
4. SE CULTIVAR $\neq$ FM966 LL E INI-ALGODAO $\neq 4$ E CULTIVAR $\neq$ FMT 701 E SILTE > 11 E Fe $\neq$ (76-155) ENTÃO PRODUTIVIDADE = Alta (127/62)
5. SE CULTIVAR $\neq$ FM966 LL E CULTIVAR $\neq$ FM951 LL E SILTE ≥ 11 E Log_OUTROS-TOTAL $> 2,3$ E Fe = (58-76) ENTÃO PRODUTIVIDADE = Alta (53/26)
BAIXA
6. SE ABERTURA $\neq 7$ E INI-ALGODAO > 2 E UMID-MAS ≥ 8 ENTÃO PRODUTIVIDADE = Baixa (86/23)
7. SE ABERTURA $\neq 7$ E INI-ALGODAO > 2 E SOMA-BASES $\geq 2,7$ ENTÃO PRODUTIVIDADE = Baixa (74/28)
8. SE ABERTURA $\neq 7$ E INI-ALGODAO ≥ 2 ENTÃO PRODUTIVIDADE = Baixa (44/8)
9. SE INI-ALGODAO $\neq 4$ E UMID-VOL ≥ 22 E INI-ALGODAO $\neq 3$ E ANTE-CULTURA $\neq$ Soja_Milho E Log_OUTROS-TOTAL $\geq 2,3$ E Fe $\neq$ (58-76) ENTÃO PRODUTIVIDADE = Baixa (100/47)
11. SE CULTIVAR $\neq$ FM966 LL E INI-ALGODAO = 4 ENTÃO PRODUTIVIDADE = Baixa (73/33)
12. SE CULTIVAR $\neq$ FM966 LL E CULTIVAR $\neq$ FM951 LL E PREP-SOLO $\neq$ Sem_Subsolagem E UMID-MAS $\neq$ (8-16) ENTÃO PRODUTIVIDADE = Baixa (54/24)
13. SE CULTIVAR $\neq$ FM975 WS E UMID-MAS $\neq$ (8-16) ENTÃO PRODUTIVIDADE = Baixa (66/30)

Fonte: Elaborado pela autora

Quadro 6 – Regras com cobertura maior que 50% para validação cruzada, classificador PART, conversão binária e discretização por KO.

ALTA	
14. SE INI-ALGODAO > 2 E UMID-MAS > 8 E Zn < 13 E B = {0,8-1,7} ENTÃO PRODUTIVIDADE = Alta (96/33)	
15. SE DP ≥ 2,7 E SAT-Ca > 27 E INI-ALGODAO ≠ 3 e 4 ENTÃO PRODUTIVIDADE = Alta (47/22)	
14. SE DP > 2,7 E Zn < 13 E CULTIVAR ≠ FMT701 E UMID-VOL ≠ {14-22} E RP10-40 ≥ 2,7 E SIST-PLANTIO = Semeadura direta E ESP ≥ 76 ENTÃO PRODUTIVIDADE = Alta (337/132)	
16. SE Zn < 13 E CTC > 7 E CULTIVAR ≠ FMT701 E Fe ≥ 58 ENTÃO PRODUTIVIDADE = Alta (123/59)	
17. SE INI-ALGODAO > 3 ou 4 E CULTIVAR = FM975 WS E Fe = {58-76} ENTÃO PRODUTIVIDADE = Alta (59/29)	
BAIXA	
18. SE INI-ALGODAO > 2 E UMID-MAS ≥ 8 ENTÃO PRODUTIVIDADE = Baixa (86/23)	
19. SE INI-ALGODAO > 2 E Zn < 13 E SAT-Ca ≥ 27 E SILTE ≥ 11 E DP > 2,7 E Plantas/ha ≥ 89000 ENTÃO PRODUTIVIDADE = Baixa (86/35)	
20. SE INI-ALGODAO ≥ 2 ENTÃO PRODUTIVIDADE = Baixa (64/8)	
21. SE Zn < 13 E INI-ALGODAO ≥ 3 ou 4 E DP > 2,68 E SILTE ≥ 11 E ESP ≥ 76 ENTÃO PRODUTIVIDADE = Baixa (60/29)	
22. SE Zn < 13 E UMID-VOL < 22 E INI-ALGODAO > 3 ou 4 E CULTIVAR ≠ FM951LL E RP10-40 > 2,7 ENTÃO PRODUTIVIDADE = Baixa (68/33)	
23. SE DP ≥ 2,7 ENTÃO PRODUTIVIDADE = Baixa (75/16)	
24. SE Zn < 13 E CULTIVAR ≠ FM 993 E CTC ≥ 7 E PREP-SOLO ≠ Sem subsolagem E ANTE-CULTURA ≠ Algodão ENTÃO PRODUTIVIDADE = Baixa (48/18)	

Fonte: Elaborado pela autora

Quadro 7 – Regras com cobertura maior que 50% para validação cruzada, classificador PART, conversão binária e discretização por KO+BE.

ALTA	
25. SE INI-ALGODAO > 2 E UMID-MAS > 8 E Mg > 0,7 E B = {0,8-1,7} ENTÃO PRODUTIVIDADE = Alta (99/34)	
26. SE INI-ALGODAO > 2 E Mg > 0,7 E CULTIVAR ≠ FM966 LL E Zn < 13 E CULTIVAR ≠ FMT701 E SAT-Ca > 27 E ESP > 76 E RP10-40 ≥ 2,7 E SIST-PLANTIO = Semeadura direta E LogOUTROS-TOTAL ≠ {1,18-2,37} E CLAS-TEXT ≠ Argiloso E LogOUTROS-TOTAL > 1,18 ENTÃO PRODUTIVIDADE = Alta (66/13)	
27. SE INI-ALGODAO > 2 E Mg > 0,7 E Zn < 13 E CULTIVAR ≠ FM966 LL E CULTIVAR ≠ FMT701 E Ce+Mg < 3,5 E RP10-40 ≥ 2,7 E Ca/Mg > 2,6 E CULTIVAR ≠ FM910 E ESP > 76 ENTÃO PRODUTIVIDADE = Alta (275/119)	
28. SE INI-ALGODAO > 2 E Mg > 0,7 E Zn < 13 E CULTIVAR ≠ FM966 LL E CULTIVAR = FM910 ENTÃO PRODUTIVIDADE = Alta (82/40)	
29. SE COBERTURA ≠ Milheto ENTÃO PRODUTIVIDADE = Alta (33/12)	
BAIXA	
30. SE INI-ALGODAO > 2 E UMID-MAS > 8 E Mg > 0,7 E INI-ALGODAO > 4 E Zn < 13 E CULTIVAR ≠ FM966 LL E UMID-VOL < 22 E K ≥ 72 ENTÃO PRODUTIVIDADE = Baixa (130/62)	
31. SE INI-ALGODAO > 2 E UMID-MAS > 8 E Mg > 0,7 E INI-ALGODAO ≥ 4 ENTÃO PRODUTIVIDADE = Baixa (88/36)	
32. SE INI-ALGODAO > 2 E UMID-MAS ≥ 8 ENTÃO PRODUTIVIDADE = Baixa (86/23)	
33. SE INI-ALGODAO ≥ 2 ENTÃO PRODUTIVIDADE = Baixa (64/8)	
34. SE Mg > 0,7 E CULTIVAR ≠ FM966 LL E SAT-Ca ≥ 267 ENTÃO PRODUTIVIDADE = Baixa (78/33)	
35. SE COBERTURA ≠ Milheto E H = {2,5-6,5} E PREP-SOLO ≠ Sem subsolagem ENTÃO PRODUTIVIDADE = Baixa (59/23)	
36. SE Mg < 1 ENTÃO PRODUTIVIDADE = Baixa (41/16)	

Fonte: Elaborado pela autora

Das 36 regras selecionadas para as 3 técnicas de discretização, 15 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 22 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 dos dados dos Quadros 5, 6 e 7. Observam-se regras com número de atributos variando de 1 a 12, mas com uma maior frequência entre 3 e 5 atributos (Fig. 22c) e coberturas variando entre 51 e 88%, mas com a maioria na faixa de 50-60% (Fig. 22d).

Em geral, verifica-se que o número de atributos em uma regra influencia o percentual de instâncias certas (Fig. 22a), sendo que o percentual de acertos é maior quando o número de atributos envolvidos é muito baixo (por exemplo 1) ou muito alto (por exemplo 12). Entretanto, regras com um ou mais de 10 atributos são de 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. 22 b).

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.

Assim, as seguintes tendências podem ser observadas:

- As cultivares de algodão utilizadas no plantio exercem influência na produtividade, pois aparecem 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, ou seja o plantio adensado (45 cm de espaçamento entre linhas) deve ser evitado devido a sua tendência de baixa produtividade.

- Solos com teores altos do micronutriente boro (entre 0,8 e 1,7 mg/dm³) apresentam alta produtividade.

- Com relação ao sistema de plantio (SIST-PLANTIO) o 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).

- Com relação às doenças do algodoeiro 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.

As Tabelas 19 a 24 apresentam os atributos das regras dos quadros 5 a 7, separadas por tipo de atributo (Física e Química do Solo, Doenças e Manejo), incluindo também a precisão e a cobertura (acerto). Observam-se valores de precisão entre 50 e 87% e % de acertos de até 90%. Verifica-se 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.

Figura 22 – 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.

Fonte: Elaborado pela autora

Tabela 19 – Descrição dos atributos das regras geradas por validação cruzada, discretização por BE e uso da conversão binária, para a classe ALTA de produtividade do algodoeiro. Números entre parênteses nos atributos indicam o número de vezes que aparece na regra.

Regra	Precisão	Cobertura	Atributos			
			Física do Solo	Química do Solo	Doenças	Manejo
1	0,65	97/33	UMID_MAS	SOMA_BASES B		ABERTURA INI-ALGODÃO
2	0,61	367/141	RP10_40 UMID_MAS		FUSARUIM	INI-ALGODÃO (2) ESP CULTIVAR SIT_PLANTIO
3	0,64	39/14				CULTIVAR (2) COBERTURA
4	0,51	127/62	SILTE	Fe		CULTIVAR (2) INI_ALGODÃO
5	0,50	53/26	SILTE	Fe	Log_OUTROS_TOTAL	CULTIVAR (2)

Fonte: Elaborado pela autora

Tabela 20 – Descrição dos atributos das regras geradas por validação cruzada, discretização por BE e uso da conversão binária, para a classe BAIXA de produtividade do algodoeiro. Números entre parênteses nos atributos indicam o número de vezes que aparece na regra.

Regra	Precisão	Cobertura	Atributos			
			Física do Solo	Química do Solo	Doenças	Manejo
6	0,73	86/23	UMID_MAS			ABERTURA
7	0,62	74/28		SOMA_BASES		ABERTURA INI_ALGODÃO
8	0,81	39/14				ABERTURA INI_ALGODÃO
9	0,53	100/47	UMID_VOL	Fe	Log_OUTROS_TOTAL	INI_ALGODÃO (2) ANTE_CULTURA
10	0,54	73/33				CULTIVAR INI_ALGODÃO
11	0,55	54/24	UMID_MAS			CULTIVAR (2) PREP_SOLO
12	0,54	66/30	UMID_MAS			CULTIVAR

Fonte: Elaborado pela autora

Tabela 21 – Descrição dos atributos das regras geradas por validação cruzada, discretização por KO e uso da conversão binária, para a classe ALTA de produtividade do algodoeiro.

Regra	Precisão	Cobertura	Atributos			
			Física do Solo	Química do Solo	Doenças	Manejo
13		96/33	UMID_MAS	Zn B		INI_ALGODÃO
14		47/22	DP	SAT_Ca		INI_ALGODÃO
15		337/132	DP UMID_VOL RP10_40	Zn		CULTIVAR SIST_PLANTIO ESP
16		123/59		Zn CTC Fe		CULTIVAR
17		53/29		Fe		INI_ALGODÃO CULTIVAR

Fonte: Elaborado pela autora

Tabela 22 – Descrição dos atributos das regras geradas por validação cruzada, discretização por KO e uso da conversão binária, para a classe BAIXA de produtividade do algodoeiro.

Regra	Precisão	Cobertura	Atributos			
			Física do Solo	Química do Solo	Doenças	Manejo
18	0,73	86/23	UMID_MAS			INI_ALGODÃO
19	0,59	86/35	SILTE DP	Zn SAT_Ca		INI_ALGODÃO
20	0,87	64/8				INI_ALGODÃO
21	0,51	60/29	DP SILTE	Zn		INI_ALGODÃO ESP
22	0,51	68/33	UMID_VOL RP10_40	Zn		INI_ALGODÃO CULTIVAR
23	0,78	75/16	DP			
24	0,62	48/18		Zn CTC		CULTIVAR PREP_SOLO ANTE_CULTURA

Fonte: Elaborado pela autora

Tabela 23 – Descrição dos atributos das regras geradas por validação cruzada, discretização por KO+BE e uso da conversão binária, para a classe ALTA de produtividade do algodoeiro. Números entre parênteses nos atributos indicam o número de vezes que aparece na regra.

Regra	Precisão	Cobertura	Atributos			
			Física do Solo	Química do Solo	Doenças	Manejo
25	0,65	99/34	UMID_MAS	Mg B		INI_ALGODÃO
26	0,8	66/13	TEXT	Mg Zn SAT_Ca	Log_OUTROS_TOTAL	INI_ALGODÃO CULTIVAR (2) SIST_PLANTIO
27	0,56	275/119	RP10_40	Mg Zn Ca+Mg Ca/Mg		INI_ALGODÃO CULTIVAR (3) ESP
28	0,51	82/40		Mg Zn		INI_ALGODÃO CULTIVAR
29	0,63	33/12				COBERTURA

Fonte: Elaborado pela autora

Tabela 24 – Descrição dos atributos das regras geradas por validação cruzada, discretização por KO+BE e uso da conversão binária, para a classe BAIXA de produtividade do algodoeiro. Números entre parênteses nos atributos indicam o número de vezes que aparece na regra.

Regra	Precisão	Cobertura	Atributos			
			Física do Solo	Química do Solo	Doenças	Manejo
30	0,52	130/62	UMID_MAS UMID_VOL	INI_ALGODÃO Zn		INI_ALGODÃO (2) CULTIVAR
31	0,59	88/36	UMID_MAS	Mg		INI_ALGODÃO (2)
32	0,73	86/23	UMID_MAS			INI_ALGODÃO
33	0,87	68/4				INI_ALGODÃO
34	0,57	78/33		Mg SAT_Ca		CULTIVAR
35	0,61	59/23		H		COBERTURA PREP_SOLO
36	0,6	41/16		Mg		

Fonte: Elaborado pela autora

Uma síntese numérica desses dados é apresentada na Tabela 25, como o número de vezes que cada atributo aparece nas classes Alta e Baixa (e Total). Verifica-se 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.

Dentre os atributos químicos do solo mais frequentes nas regras, os micronutrientes zinco, ferro e boro, os macronutrientes cálcio e magnésio são os únicos com cobertura maior que 50

Tabela 25 – Síntese do número de vezes que determinado atributo apareceu nas 36 regras apresentadas nos Quadros 5, 6 e 7, nas classes de Alta e Baixa produtividade, separados por tipo de atributo (Físicos e Químicos do solo, Doenças e Manejo).

Atributo		Nº de vezes que aparece na classe		
Tipo	Sigla	Alta	Baixa	Total
Física do Solo	UMID-MAS	4	7	11
	DP	2	3	5
	RP10-40	3	1	4
	SILTE	2	2	4
	UMID-VOL	1	3	4
	TEXT	1	0	1
Química do Solo	Zn	6	5	11
	Mg	4	4	8
	Fe	4	0	4
	SAT_Ca	2	2	4
	B	3	0	3
	SOMA-BASES	1	1	2
	CTC	1	1	2
	Ca+Mg	1	0	1
	Ca/Mg	1	0	1
	K	0	1	1
	H	0	1	1
Doenças	Log_OUTROS-TOTAL	2	1	3
	FUSARIUM	1	0	1
Manejo	INI-ALGODÃO	11	17	28
	CULTIVAR	17	8	25
	ABERTURA	1	3	4
	ESP	3	1	4
	SIST-PLANTIO	3	0	3
	COBERTURA	2	1	3
	PREP-SOLO	0	3	3
	ANTE-CULTURA	0	2	2

Fonte: Elaborado pela autora

De um modo geral os atributos relacionados à doenças avaliadas (nematoides e *fusarium*) impactaram pouco nas regras de associação. Isso se deve provavelmente ao intenso manejo fitossanitário do algodoeiro, com a aplicação de grandes volumes de inseticidas, nematicida e fungicidas, reduzindo bastante o efeitos dessas doenças na perda de produção, comparativamente com outros tipos de atributos.

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 densidade das partículas, são inerentes do tipo de solo e não constituem-se como atributos que podem ser modificados por manejo (não há irrigação), como 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 relacionada à compactação do solo, e que pode ser modificada por manejo (subsolagem do solo), apareceu em algumas regras, com boa cobertura.

CONCLUSÕES

Neste trabalho foram aplicadas técnicas de mineração de dados para classificação baseada em regras, em um banco de dados de produção comercial de algodão no Mato Grosso, visando a compreensão dos principais fatores que afetam a produtividade do algodoeiro.

A etapa de pré-processamento dos dados consistiu na seleção e transformação de atributos e identificação de *outliers*. Os atributos numéricos foram discretizados (conversão numérica-nominal) utilizando 4 métodos, sendo 3 automáticos e um manual. A discretização dos atributos numéricos por intervalos ordinais (variável nominal-ordinal) foi utilizada visando a compreensibilidade das regras, assim como para limitar que o excesso de particionamentos recursivos pudesse ampliar a complexidade e interpretabilidade dos modelos. Na etapa de modelagem de classificação, os algoritmos de regras utilizados foram o PART e o JRip e a conversão binária de atributos (conversão nominal-binária) foi também avaliada.

O desempenho das técnicas de discretização e conversão binária associadas aos algoritmos de regras foi avaliado por diversas métricas como a precisão, revocação, custo e a combinação delas no índice IFC ($100 * F / \log_2 \text{custo}$). A avaliação dessa métrica indicou o melhor desempenho para o classificador PART, com a discretização pela técnica de Kononenko combinada com Better Encoding, seguida pela conversão binária.

O atributo alvo utilizado foi o indicador da produtividade, o qual possuía tipos de estimação: Histórica, da Safra ou por meio da Média de ambas. Os melhores resultados foram obtidos para a Média da produtividade, um indício consistente de que os contextos históricos (solo, manejos, matéria orgânica e outros) também impactam na produtividade e são importantes à compreensão e extração das regras. Foram obtidas regras com coberturas superiores a 50% para as classes de Alta (acima de 280@/ha) e de Baixa produtividades (abaixo de 250@/ha), ou seja, na classe de produtividade Média (entre 251

e 280@/ha) nenhuma regra gerada apresentou cobertura superior a 50%. A análise das regras, com auxílio de especialistas da área, possibilitou a seleção das regras mais relevantes e dos atributos que mais impactam na produtividade. Dentre esses atributos, destacam-se: a cultivar de algodão, o tempo da abertura e de início do cultivo de algodão na área, o sistema de plantio e de preparo do solo. Os principais atributos físico-químicos do solo foram os micronutrientes zinco, ferro e boro e os macronutrientes cálcio e magnésio e a resistência do solo à penetração. A ausência do fungo *fusarium* impactou positivamente na produtividade.

Em geral, os atributos relacionados ao Manejo do algodoeiro apresentaram maior relevância na composição das regras de classificação, seguidos dos atributos Químicos e Físicos do Solo e dos atributos relacionados à doenças.

Em função das regras geradas e da análise do especialista da área, pode-se sugerir as seguintes ações aos produtores:

- Utilizar do sistema de plantio direto (SIST-PLANTIO);
- Realizar o monitoramento da compactação do solo, por meio da medida da resistência do solo à penetração (RP) e a devida ação de descompactação (mecânica ou biológica), quando RP estiver muito alta, ou seja, acima de 3 MPa;
- Utilizar o espaçamento entre linha de plantio (ESP) de 76 ou 90 cm, evitando o sistema adensado (45 cm);
- Realizar análise da fertilidade do solo na recomendação de adubação, com especial atenção ao cálcio, magnésio, boro e zinco;
- Realizar o monitoramento da murcha de *fusarium* e o seu devido controle.

O trabalho realizado mostrou a importância e utilidade da mineração de dados para a extração de informações relevantes de um banco de dados complexo, com grande número de atributos e instâncias. Apesar de ser um trabalho relativamente simples e de utilizar um número restrito de algoritmos e técnicas de mineração, possibilitou a extração de informações relevantes aos produtores. Assim, é um trabalho inicial que mostra o potencial desse tipo de estudo e indica a necessidade de mais trabalhos com a aplicação de outras técnicas e algoritmos de mineração de dados.

Algumas sugestões para os trabalhos futuros são o uso de remoção de atributos interdependentes, a utilização de 5 classes de intervalos para a produtividade 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.

REFERÊNCIAS

ABRAPA. **Produção Mundial de Algodão**. 2017. Disponível em: <http://www.abrapa.com.br/Paginas/dados/producao-mundial-algodao_backup2017.aspx>. Acesso em: 27/11/2017. Citado na página 25.

AHAMED, A. T. M. S.; MAHMOOD, N. T.; HOSSAIN, N.; KABIR, M. T.; DAS, K.; RAHMAN, F.; RAHMAN, R. M. 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: 03/07/2017. Citado na página 54.

AMARAL, F. **Introdução a ciência de dados: mineração de dados e Big Data**. Rio de Janeiro: ALTA BOOKS, 2016. Citado nas páginas 36, 42 e 64.

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>>. Citado na página 49.

CAMILO, C. O.; SILVA, J. C. **Mineração de Dados: Conceitos, Tarefas, Métodos e Ferramentas**. Instituto de Informática - Universidade Federal de Goiás, 2009. Citado nas páginas 35, 38 e 39.

CATLETT, J. On changing continuous attributes into ordered discrete attributes. **Machine Learning — EWSL-91. Lecture Notes in Computer Science (Lecture Notes in Artificial Intelligence)**, Springer, Berlin, Heidelberg, v. 482, p. 164–178, 1991. Disponível em: <<https://link.springer.com/chapter/10.1007/BFb0017012>>. Acesso em: 03/07/2017. Citado nas páginas 46 e 47.

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>>. Citado nas páginas 30, 39, 40 e 66.

COSTA, J. J.; BERNARDINI, F. C.; FILHO, J. V. A mineração de dados e a qualidade de conhecimentos extraídos dos boletins de ocorrência das rodovias federais brasileiras. **AtoZ - Novas Práticas em Informação e Conhecimento**, v. 3, n. 3, p. 139–155, 2014. ISSN 2237-826X. Disponível em: <<file:///C:/Users/User/Downloads/1958-191-PB.pdf>>. Citado na página 35.

DELERCE, S.; DORADO, H.; GRILLON, A.; REBOLLEDO, M. C.; PRAGER, S. D.; NO, V. H. P.; VARÓN, G. G.; JIMÉNEZ, D. 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>>. Citado nas páginas 28 e 29.

DEVASENA, L. Proficiency comparison of zeror, ridor and part classifiers for intelligent heart disease prediction. **International Journal of Advances in Computer Science and Technology (IJACST)**, Special Issue of ICCAAC 2014 - Held on December 08,2014,Hyderabad, India, v. 3, n. 11, p. 12–18, 2014. ISSN 2320-2602. Disponível em: <<http://www.warse.org/pdfs/2014/iccaac2014sp03.pdf>>. Acesso em: 03/07/2017. Citado na página 40.

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>>. Citado nas páginas 28 e 29.

DOUGHERTY, J.; KOHAVI, R.; SAHAMI, M. Supervised and unsupervised discretization of continuous features. **Machine Learning Proceedings 1995: Proceedings of the Twelfth International Conference on Machine Learning**, Machine Learning Proceedings, Tahoe City, California, July 9–12, 1995, p. 194–202, 1995. Disponível em: <<https://www.sciencedirect.com/science/article/pii/B9781558603776500323>>. Acesso em: 03/07/2017. Citado nas páginas 46 e 48.

FACELI, K.; ; LORENA, A. C.; GAMA, J.; (AUTOR), A. C. P. L. F. de C. **Inteligência Artificial: Uma abordagem de Aprendizagem de Máquina**. Rio de Janeiro: LTC, 2011. Citado nas páginas 34, 43 e 44.

FAYYAD, U.; PIATETSKY-SHAPIO, G.; SMYTH, P. From data mining to knowledge discovery in databases. **AI Magazine**, (© AAAI) American Association for Artificial Intelligence, v. 17, n. 3, p. 37–54, 1996. Disponível em: <<https://www.aaai.org/ojs/index.php/aimagazine/article/viewFile/1230/1131>>. Citado nas páginas 32, 35 e 38.

_____. The kdd process for extracting useful knowledge from volumes of data. **COMMUNICATIONS OF THE ACM**, v. 39, n. 11, p. 27–34, 1996b. Disponível em: <http://sentra.ischool.utexas.edu/~i385q/readings/Fayyad_Piatetsky-1996-TheKDD.pdf>. Citado na página 32.

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>>. Citado na página 51.

FAYYAD, U. M.; IRANI, K. B. Multi-interval discretization of continuous-valued attributes for classification learning. **IJ-CAI**, 1993. Disponível em: <<https://www.semanticscholar.org/paper/Multi-Interval-Discretization-of-Continuous-Valued-Fayyad-Irani/1dc53b91327cab503acc0ca5afb9155882b717a5>>. Citado nas páginas 49, 50 e 51.

FREITAS, A. A. **Data Mining and Knowledge Discovery with Evolutionary Algorithms**. RSpringer-Verlag Berlin Heidelberg, 1ª edição: ALTA BOOKS, 2002. Citado na página 47.

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: 03/07/2017. Citado na página 40.

GALBIERI, R.; BELOT, J. L. **Nematoides fitoparasitas do algodoeiro nos cerrados brasileiros: Biologia e medidas de controle**. Cuiabá (MT): Boletim de P&D - n°3, maio, 2016. Citado na página 57.

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: 30/11/2017. Citado nas páginas 28 e 53.

_____. **Rice Crop Yield Forecasting of Tropical Wet and Dry Climatic Zone of India Using Data Mining Techniques**. 2016b. 357-363 p. Acesso em: 30/11/2017. Citado nas páginas 28 e 29.

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>. Citado nas páginas 29 e 54.

GARCIA, S.; LUENGO, J.; SáEZ, J. A.; LÓPEZ, V.; HERRERA, F. A survey of discretization techniques: Taxonomy and empirical analysis in supervised learning. **IEEE TRANSACTIONS ON KNOWLEDGE AND DATA ENGINEERING**, APRIL 2013, v. 25, n. 4, p. 734–750, 2013. ISSN 1041-4347. Disponível em: <<http://doi.ieeecomputersociety.org/10.1109/TKDE.2012.35>>. Acesso em: 03/07/2017. Citado nas páginas 46 e 47.

GEETHA, M. A survey on data mining techniques in agriculture. **International Journal of Innovative Research in Computer and Communication Engineering**, v. 3, n. 2, p. 887–892, 2015. ISSN 2320-9798. Disponível em: <<https://www.semanticscholar.org/paper/A-Survey-on-Data-Mining-Techniques-in-Agriculture-Geetha/dbfff0ee957017af0da440d54ef77327b6e264e8>>. Acesso em: 03/07/2017. Citado na página 53.

GOLDSCHMID, R. R. **Uma Introdução à Inteligência Computacional: Fundamentos, Ferramentas e Aplicações**. Rio de Janeiro: Série Livros Didáticos Digitais Gratuitos, 2010. Citado na página 38.

GOLDSCHMIDT, R.; PASSOS, E. **Data Mining: um guia prático**. Rio de Janeiro/RJ: Campus: Elsevier, 2005. Citado na página 66.

HALL, M.; FRANK, E.; HOLMES, G.; PFAHRINGER, B.; REUTEMANN, P.; WITTEN, I. H. The weka data mining software: An update. **ACM SIGKDD Explorations Newsletter**, ACM New York, NY, USA, v. 11, n. 1, p. 10–18, 2009. Disponível em: <<https://dl.acm.org/citation.cfm?id=1656278>>. Acesso em: 03/07/2017. Citado na página 66.

- HAMMER, R. G. **Modelagem da Produtividade da Cultura da cana-de-açúcar por meio do uso de técnicas de mineração de dados**. Dissertação (Mestrado) — Universidade de São Paulo - Campus São Carlos, São Carlos-SP, 2016. Citado na página 29.
- HAN, J.; KAMBER, M. **Data Mining: Concepts and Techniques**. San Francisco-CA/EUA: Morgan Kaufmann, 2001. Citado nas páginas 33 e 34.
- _____. **Data Mining: Concepts and Techniques**. San Francisco-CA/EUA: Morgan Kaufmann, 2006. Citado nas páginas 32, 38 e 39.
- HOSOKAWA, E. O. **Técnica de Árvore de Decisão em Mineração de Dados**. 2011. Disponível em: <<http://www.fatecsp.br/dti/tcc/tcc0003.pdf>>. Citado na página 49.
- 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: 03/07/2017. Citado na página 52.
- KALPANA, R.; SHANTHI, N.; ARUMUGAM, S. A survey on data mining techniques in agriculture. **Indian Journal of Innovations and Developments**, v. 5, n. 8, p. 1–6, 2014. ISSN 2277-5382, ISSN (online) : 2277-5390. Disponível em: <<file:///C:/Users/User/Downloads/110845-248030-1-SM.pdf>>. Acesso em: 03/07/2017. Citado na página 53.
- KELES, M. K. An overview: the impact of data mining applications on various sectors. **TECHNICAL JOURNAL**, v. 11, n. 3, p. 128–132, 2017. ISSN 1846-6168. Disponível em: <<https://hrcak.srce.hr/186659>>. Acesso em: 03/07/2017. Citado na página 53.
- 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: 03/07/2017. Citado na página 52.
- KLOSGEN, W.; ZYTKOW, J. **Handbook of data mining and knowledge discovery**. [S.l.]: Oxford University Press; 1 edition; June 15, 2002. Citado na página 28.
- 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: 03/07/2017. Citado nas páginas 51 e 52.
- KURGAN, L. A.; CIOU, K. J. Caim discretization algorithm. **IEEE Transactions on Knowledge and Data Engineering**, IEEE, v. 16, n. 2, p. 145–153, 2004. Disponível em: <<http://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=1269594&isnumber=28401>>. Acesso em: 03/07/2017. Citado na página 46.
- LAROSE, D. T. **Discovering Knowledge in Data: An Introduction to Data Mining**. Canada: John Wiley & Sons - WILEY, 2005. Citado na página 35.

- LIU, H.; HUSSAIN, F.; TAN, C. L.; DASH, M. Discretization: An enabling technique. **Data Mining and Knowledge Discovery**, Kluwer Academic Publishers, p. 393–423, 2002. ISSN 1573-756X. Disponível em: <<https://link.springer.com/article/10.1023/A:1016304305535>>. Acesso em: 03/07/2017. Citado nas páginas 46, 47, 48 e 50.
- MUHLENBACH, F.; RAKOTOMALALA, R. Discretization of continuous attributes. **John Wang. Encyclopedia of Data Warehousing and Mining**, Idea Group Reference, Hal Archives-Ouvertes.fr, v. 1, p. 397–402, 2005. Disponível em: <<https://hal.archives-ouvertes.fr/hal-00383757v1>>. Acesso em: 03/07/2017. Citado na página 47.
- 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. 2017. Citado nas páginas 28 e 53.
- 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: 30/11/2017. Citado na página 49.
- 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: 03/07/2017. Citado na página 53.
- PENA, C. **Importância Econômica da Cultura do Algodão**. Agricultura no Brasil: Agricultura Brasileira em foco: clima, solos, plantio, irrigação, tratamentos culturais, controle de pragas e doenças, colheita, comercialização e muito mais.: [s.n.], 2017. Disponível em: <<https://plantarcrecercolher.blogspot.com.br/2017/01/importancia-economica-da-cultura-do.html>>. Acesso em: 25/11/2017. Citado na página 26.
- PERINA, F. J.; BOGIANI, J. C.; RIBEIRO, G. C.; BRED, C. E.; FABRIS, A.; SANTOS, I. A. dos; SEIBEL, D. P. Levantamento e manejo de fitonematoides em algodoeiro no oeste da bahia, resultados safra 2016/17. Circular Técnica 03, n. 3, p. 8, 2015. Disponível em: <<file:///C:/Users/User/Downloads/Perina%20et%20al%202017.pdf>>. Acesso em: 03/09/2017. Citado na página 57.
- PETER, C.; BEALE, R. **Affect and Emotion in Human-Computer Interaction: From Theory to Applications**. Springer, Berlin, Heidelberg, 1ª edição: ALTA BOOKS, 2008. Citado na página 52.
- PFAHRINGER, B. Compression-based discretization of continuous attributes. **Machine Learning Proceedings 1995: Proceedings of the Twelfth International Conference on Machine Learning**, Tahoe City, California, July 9–12, 1995, p. 456–463, 1995. Disponível em: <<https://www.sciencedirect.com/science/article/pii/B9781558603776500633>>. Acesso em: 03/07/2017. Citado na página 47.
- 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: 03/09/2017. Citado na página 66.

RAJPUT, A.; AHARWAL, R. P.; DUBEY, M.; SAXENA, S. P.; RAGHUVANSHI, M. J48 and jrip rules for e-governance data. **International Journal of Computer Science and Security (IJCSS)**, v. 5, p. 201–207, 2011. ISSN 1985-1553. Disponível em: <<http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.232.4477&rep=rep1&type=pdf>>. Acesso em: 03/07/2017. Citado na página 41.

RAORANE, A.; KULKARNI, R. Review-role of data mining in agriculture. (**IJCSIT**) **International Journal of Computer Science and Information Technologies**, v. 4, n. 2, p. 270 – 272, 2013. ISSN 0975-9646. Disponível em: <<http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.300.533&rep=rep1&type=pdf>>. Acesso em: 03/07/2017. Citado na página 53.

REZENDE, S. O. **Sistemas Inteligentes: fundamentos e aplicações**. Barueri-SP: MANOLE, 2003. Citado nas páginas 34, 35, 41 e 43.

SCOSS, A. M. **A Clusterização e Classificação no Processo de Data Mining para Análise do Desempenho Docente no Ensino de Graduação**. Trabalho de Conclusão de Curso (Especialização) - Universidade do Extremo Sul Catarinense - UNESC, Criciúma, 2006. Citado na página 35.

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>>. Citado na página 49.

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>>. Citado nas páginas 29 e 55.

SILVA, L. A. da; PERES, S. M.; BOSCARIOLI, C. **Introdução à mineração de dados: com aplicações em R**. Rio de Janeiro: Elsevier, 2016. Citado nas páginas 32, 33, 34 e 41.

SOUZA, Z. M.; CERRI, D. G. P.; COLET, M. J.; RODRIGUES, L. H. A.; MAGALHÃES, P. S. G.; MANDONI, R. J. A. Análise dos atributos do solo e da produtividade da cultura de cana-de-açúcar com o uso da geoestatística e árvore de decisão. **Ciencia Rural**, v. 40, n. 4, p. 840–847, 2004. ISSN 0103-8478. Disponível em: <http://www.scielo.br/scielo.php?pid=S0103-84782010000400015&script=sci_abstract&tlng=pt>. Acesso em: 03/07/2017. Citado na página 55.

TAKATUZI, F. K. O. **Adaptive-IDT: algoritmo incremental para aprendizagem de árvores de decisão adaptativas**. Trabalho de Conclusão de Curso (Graduação) – Universidade Tecnológica Federal do Paraná, Guarapuava, 2017: [s.n.], 2017. Disponível em: <<http://repositorio.roca.utfpr.edu.br/jspui/handle/1/9426>>. Acesso em: 30/11/2017. Citado na página 50.

VEENADHARI, S.; MISRA, B.; SINGH, C. Data mining techniques for predicting crop productivity – a review article. **International Journal of Computer Science and Technology**, v. 2, n. 1, p. 98–100, 2011. ISSN 2229-4333. Disponível em: <https://www.researchgate.net/publication/305355715_Data_mining_Techniques_for_Predicting_Crop_Productivity_-_A_review_article>. Acesso em: 03/07/2017. Citado na página 53.

VELOSO, L. C.; GOMES, O. M. A. **Introduzindo Caminhadas Aleatórias no Aprendizado de Regras para Mineração de Dados**. 2007. Disponível em: <<http://www2.ic.uff.br/Graduacao/files/000004.pdf>>. Citado na página 41.

VRIESMANN, L. M.; GUIMARÃES, A. M.; CANTERI, M. G.; MOLIN, J. P.; CATA-NEO, A.; KOVALECHYN, D. 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: 03/07/2017. Citado na página 54.

WEISS, S.; INDURKHYA, N. **Predictive Data Mining: A Practical Guide**. San Francisco, CA/EUA: Morgan Kaufmann, 1997. Citado na página 33.

WITTEN, I.; FRANK, E. Generating accurate rule sets without global optimization. **Working paper 98/2**, New Zealand: University of Waikato, Department of Computer Science, 1998. ISSN 1170-487X. Disponível em: <<https://researchcommons.waikato.ac.nz/handle/10289/1047>>. Citado nas páginas 30, 39, 40 e 66.

_____. **Data Mining: Practical Machine Learning Tools and Techniques with Java Implementations**. Burlington / Massachusetts: ELSEVIER, 2009. Citado nas páginas 39 e 40.