

UNIVERSIDADE METODISTA DE PIRACICABA
FACULDADE DE ENGENHARIA ARQUITETURA E URBANISMO
PROGRAMA DE PÓS-GRADUAÇÃO EM ENGENHARIA DE PRODUÇÃO

**DESENVOLVIMENTO DE UM ROTEIRO
SISTEMATIZADO PARA O SUPORTE À TOMADA DE
DECISÃO NA GESTÃO DA PRODUÇÃO DE CANA-
DE-AÇÚCAR**

MARIA DAS GRAÇAS J.M. TOMAZELA

ORIENTADOR: PROF. DR. FERNANDO CELSO DE CAMPOS

SANTA BÁRBARA D'OESTE

2017

UNIVERSIDADE METODISTA DE PIRACICABA
FACULDADE DE ENGENHARIA ARQUITETURA E URBANISMO
PROGRAMA DE PÓS-GRADUAÇÃO EM ENGENHARIA DE PRODUÇÃO

**DESENVOLVIMENTO DE UM ROTEIRO
SISTEMATIZADO PARA O SUPORTE À TOMADA DE
DECISÃO NA GESTÃO DA PRODUÇÃO DE CANA-
DE-AÇÚCAR**

MARIA DAS GRAÇAS J.M. TOMAZELA

ORIENTADOR: PROF. DR. FERNANDO CELSO DE CAMPOS

Tese apresentada ao programa de Pós-graduação em Engenharia de Produção da Faculdade de Engenharia, Arquitetura e Urbanismo, da Universidade Metodista de Piracicaba - UNIMEP, como pré-requisito para obtenção do título de doutora em Engenharia de Produção.

SANTA BÁRBARA D'OESTE

2017

Dedico

Às minhas raízes e aos meus frutos.

AGRADECIMENTOS

A Deus, que segura minhas mãos e guia meus passos.

À minha família, minha maior bênção: meu marido Mauro, minhas filhas Renata e Priscila, meus genros Nuno e Wesley e meus netos Gabriel, Gustavo e Guilherme, as crianças lindas da vovó.

Aos meus pais, Vera (*in memoriam*) e Américo, pela dedicação e amor infinito.

Às minhas queridas tias Rita, Regina e Angelina pelo amor e carinho sem limites.

Aos amigos que colaboraram das mais diversas formas na elaboração deste trabalho: Mariana Vieira, Sérgio Clauss, Juliana Landgraf, Luciana Zaina, Angelina Melaré, Lilian Simão, Francisco Benedetti, Rosana Veroneze, Dilermundo Piva Jr., Michel Munhoz, Wellington Roque e Ivanete Bellucci. Quem tem amigos tem tudo! Sou muito grata a vocês pelos minutos, horas ou dias que dedicaram a me apoiar, deixando de lado suas próprias tarefas.

À Profª Rita Helena Junqueira, pela revisão gramatical e ortográfica.

Ao José Pedro Parizoto, pelas informações preciosas sobre a cultura da cana-de-açúcar.

Aos funcionários da Usina que forneceu os dados para a realização deste trabalho, pelas reuniões, entrevistas e todo o apoio para o entendimento dos processos referentes à produção da cana-de-açúcar.

Ao Prof. Dr. Tomaz Caetano C. Ripoli (*in memoriam*), pelos livros sobre cana-de-açúcar, que gentilmente emprestou para a realização deste trabalho.

Aos professores que fizeram parte da banca de qualificação pelas contribuições valiosas que me deram para a construção deste trabalho: Prof. Dr. Carlos Roberto Camello Lima, Prof. Dr. Aparecido dos Reis Coutinho, Prof. Dr. Eduardo Alves Portela e Prof. Dr. Marcos Milan.

À Marta Helena Bragaglia, secretária da pós-graduação, por sua dedicação e eficiência em seu trabalho e pelo carinho com que trata todos os alunos.

Ao meu orientando, Leonardo Moretto, pelo apoio no desenvolvimento do software desenvolvido para este trabalho.

Ao Prof. Dr. Luiz Antônio Daniel, pelas inúmeras orientações sobre os processos da cana-de-açúcar e também por propiciar contatos com outros profissionais.. Sua colaboração foi essencial para o desenvolvimento deste trabalho.

Ao meu amigo, Prof. Dr. Aldo Pontes, que me pegou pela mão, conduziu-me ao programa de doutorado da UNIMEP e não mediu esforços para que eu realizasse este grande sonho.

Ao meu orientador, Prof. Dr. Fernando Celso de Campos, pelo comprometimento, orientação precisa, profissionalismo e dedicação com que me conduziu na realização deste trabalho. Tem meu respeito e admiração, como profissional e como ser humano.

A todos dedico este poema, que aprecio muito, de Fernando Pessoa:

"Para ser grande, sé inteiro!
Nada teu exagera ou exclui.
Sê todo em cada coisa.
Põe quanto és no mínimo que fazes.
Assim, em cada lago,
a lua toda brilha,
porque alta vive!"

TOMAZELA, Maria das Graças Junqueira Machado. **Desenvolvimento de um roteiro sistematizado para o suporte à tomada de decisão na gestão da produção de cana-de-açúcar.** 2017.172 f. Tese de Doutorado em Engenharia de Produção— Faculdade de Engenharia Arquitetura e Urbanismo, Universidade Metodista de Piracicaba, Santa Bárbara d'Oeste.

RESUMO

Sistemas agrícolas são suscetíveis à variabilidade climática e biofísica, o que aumenta muito a complexidade do planejamento. Desta forma, o uso das tecnologias de informação pode contribuir para melhorar a eficiência da gestão desses sistemas. A utilização de técnicas de mineração de dados possibilita a manipulação de grandes conjuntos de dados e a identificação de padrões novos e úteis nesses conjuntos. Contudo, a tarefa de interpretar esses dados pode tornar o processo decisório mais complexo aos gestores agrícolas. Desta forma, prover mecanismos que auxiliem esses gestores durante a interpretação dos dados pode trazer benefícios diretos à tomada de decisão. Assim, o objetivo deste trabalho foi propor um roteiro sistematizado, a partir da aplicação de técnicas de mineração de dados, para dar suporte aos processos de tomada de decisão na gestão da produção de cana-de-açúcar. Para a realização da investigação, os seguintes procedimentos metodológicos foram adotados: *I)* revisão sistemática da literatura para identificação das contribuições e lacunas acadêmico-tecnológicas envolvendo esta temática; *II)* aplicação do processo de descoberta de conhecimento utilizando o modelo de referência CRISP-DM; *III)* elaboração de um roteiro sistematizado para melhoria da visualização dos resultados obtidos com o processo de descoberta de conhecimento; *IV)* implementação de uma ferramenta para exploração de cenários de produção da cana-de-açúcar. Os processos utilizados neste trabalho em conjunto com a ferramenta de visualização desenvolvida mostraram-se aptos a subsidiar atividades de planejamento e de tomada de decisão para a cultura da cana-de-açúcar.

Palavras-chave: Produtividade; Cana-de-açúcar; Mineração de dados; Árvore de Decisão; Sistematização; CRISP-DM.

TOMAZELA, Maria das Graças Junqueira Machado. **Development of a systematized road map to support decision-making in the management of sugarcane production.** 2017. 172 f. Doctoral Thesis in Production Engineering – School of Engineering, Architecture and Planning, Methodist University of Piracicaba, Santa Bárbara d'Oeste.

ABSTRACT

Agricultural systems are susceptible to climatic variability and biophysics, which greatly increases the complexity of planning. In this way, the use of information technologies can contribute to improve the efficiency of the management of these systems. The use of data mining techniques enables the manipulation of large data sets and the identification of new and useful patterns in these sets. However, the task of interpreting these data can make the decision-making process more complex to agricultural managers. Thus, providing mechanisms that assist these managers during the interpretation of data can bring direct benefits to decision making. Thus, the objective of this work was to propose a systematized roadmap, based on the application of data mining techniques, to support the decision-making processes in the management of sugarcane production. In order to carry out the research, the following methodological procedures were adopted: *i)* Systematic review of the literature to identify the contribution and academic-technological gaps enclosing this topic; *ii)* Application of the knowledge-discovery process making use of the CRISP-DM reference model; *iii)* Preparation of a systematized guide to improve the obtained results visualization with the KDD process; *iv)* Implementation of a tool to explore scenarios of the sugar cane production. The processes used on this work together with the developed visualization tool are able to subsidize the planning activities and the decision-making process to the sugar cane culture.

keywords: Productivity; Data Mining; Sugar Cane; KDD; Systematization; CRISP-DM.

SUMÁRIO

LISTAS DE FIGURAS	XI
LISTA DE QUADROS	XIV
LISTA DE TABELAS.....	XV
LISTA DE ABREVIATURAS E SIGLAS.....	XVI
1 INTRODUÇÃO	1
1.1 JUSTIFICATIVA E RELEVÂNCIA	2
1.2 PROBLEMA DE PESQUISA.....	4
1.3 OBJETIVOS	4
1.4 MÉTODO DA PESQUISA	5
1.5 ESTRUTURA DO TRABALHO	6
2 FUNDAMENTAÇÃO TEÓRICA	7
2.1 CARACTERÍSTICAS DA CULTURA DA CANA-DE-AÇÚCAR.....	7
2.2 TOMADA DE DECISÃO E PLANEJAMENTO PARA A INDÚSTRIA DA CANA.....	9
2.3 PRODUTIVIDADE DA CANA-DE-AÇÚCAR.....	12
2.3.1 Atributos Utilizados nas Aplicações de Produtividade da Cana-de-Açúcar	13
2.3.2 Produtividade de Cana-de-Açúcar e Aplicação de Técnicas	19
2.4 DESCOBERTA DE CONHECIMENTO EM BASE DE DADOS.....	35
2.4.1 Modelo de Referência de Processo para Realização do KDD.....	37
2.4.2 Tarefas e Técnicas de Mineração de Dados	39
2.4.3 Indução por Árvore de Decisão	41
2.4.4 Medidas de Desempenho dos Classificadores	43
2.4.5 Ferramentas de Mineração de Dados	47
2.4.6 Aplicação de Técnicas de Mineração de Dados em Produtividade de Cana-de-açúcar	49
3 ABORDAGEM METODOLÓGICA	55
3.1 DESCRIÇÃO DA REVISÃO SISTEMÁTICA DA LITERATURA (RSL)	56
3.1.1 Planejamento da revisão	58
3.2 PROCEDIMENTO TÉCNICO: MODELAGEM.....	62
4 DETALHAMENTO DO PROCEDIMENTO TÉCNICO: MODELAGEM – CRISP-DM	63
4.1 ENTENDIMENTO DO NEGÓCIO.....	63

4.1.1	Descrição da Unidade em Análise.....	64
4.2	ENTENDIMENTO DOS DADOS	70
4.2.1	Características dos Dados Disponibilizados.....	71
4.3	PREPARAÇÃO DOS DADOS	76
4.4	MODELAGEM DOS DADOS	82
4.4.1	Modelagem dos Dados com Conjunto Completo de Atributos.....	83
4.4.2	Modelagem dos Dados com Seleção de Atributos.....	88
4.5	AVALIAÇÃO	96
4.6	IMPLEMENTAÇÃO	100
4.7	SÍNTESE DOS PROCEDIMENTOS REALIZADOS NO PROCESSO DE KDD	103
5	PROPOSTA DO ROTEIRO SISTEMATIZADO.....	105
5.1	DESCRÍÇÃO DO ROTEIRO SISTEMATIZADO	105
5.2	DESENVOLVIMENTO E IMPLEMENTAÇÃO DA FERRAMENTA DE VISUALIZAÇÃO	109
5.2.1	Requisitos funcionais e não funcionais.....	112
5.2.2	Interações do Usuário e Funcionalidades do Sistema	113
5.2.3	Características da Estrutura de Armazenamento.....	115
5.2.4	Protótipo do Sistema.....	116
5.2.5	Atividades de Teste	121
5.2.6	Tecnologias.....	122
5.2.7	Resultados da Aplicação da Ferramenta.....	124
5.2.8	Considerações sobre a Ferramenta SECC	128
6	CONCLUSÃO.....	131
6.1	PROPOSTA PARA TRABALHOS FUTUROS	133
REFERÊNCIAS	135	
APÊNDICE A – DESCRIÇÃO DA APLICAÇÃO DO MÉTODO DELPHI.....	143	
APÊNDICE B – CONJUNTO DE VALORES POSSÍVEIS PARA OS ATRIBUTOS DA ÁRVORE DE DECISÃO	149	
APÊNDICE C – MAPA DAS ESTAÇÕES PLUVIOMÉTRICAS	153	
APÊNDICE D – HISTOGRAMAS DOS ATRIBUTOS UTILIZADOS NOS MODELOS DE PRODUTIVIDADE	157	
APÊNDICE E – DESCRIÇÃO DE MÉTODOS DE SELEÇÃO DE ATRIBUTOS DA FERRAMENTA WEKA	159	
APÊNDICE F – REGRAS DE DECISÃO REFERENTE À MODELAGEM	10161	

APÊNDICE G – PROTÓTIPO DAS TELAS PARA GERENCIAR USUÁRIOS DO SISTEMA.....	167
APÊNDICE H – TELAS PARA GERENCIAR USUÁRIOS DO SISTEMA.....	171

LISTA DE FIGURAS

Figura 1 - Outline da Pesquisa.....	5
Figura 2 - Etapas do KDD	37
Figura 3 - Fases do Modelo de Processo CRISP-DM.....	39
Figura 4 - Procedimentos Metodológicos a Serem Adotados neste Trabalho	56
Figura 5 - Fluxo de Atividades da RSL.....	57
Figura 6 - Fluxo das Atividades Executadas de Acordo com o Modelo CRISP-DM	62
Figura 7 - Itens Utilizados na Entrevista Semiestruturada	64
Figura 8 - Organograma da Área Qualidade Agrícola.....	65
Figura 9 - Distribuição dos Níveis de Produtividade da Cana em Função do Número de Cortes.....	74
Figura 10 - Precipitação Pluviométrica Acumulada por mês (2006 a 2009)....	75
Figura 11 - Média das Temperaturas Mensais (2006 a 2009)	76
Figura 12 - Fases Fenológicas da Cana	79
Figura 13 - Gráfico Método Correlation.....	91
Figura 14 - Gráfico Método GainRatio	92
Figura 15 - Gráfico Método InfoGain.....	92
Figura 16 - Gráfico OneR	93
Figura 17 - Gráfico Método ReliefF	93
Figura 18 - Gráfico Método Simmetrical Uncert	94
Figura 19 - Árvore de Decisão	102
Figura 20 - Regras de Decisão	103
Figura 21 - Fluxo de Execução do Roteiro Sistematizado	106
Figura 22 - Passos do Roteiro Implementados na Ferramenta SECC	109
Figura 23 - Processo de Desenvolvimento do SECC.....	111
Figura 24 - Diagrama de Casos de Uso	113
Figura 25 - Modelo Lógico do Banco de Dados	116
Figura 26 - Protótipo da Tela de Gerenciamento de Arquivos	117
Figura 27 - Protótipo da Tela de Identificação de Novo Arquivo	118
Figura 28 - Protótipo da Consulta a Partir dos Atributos de Produção.....	119
Figura 29 - Protótipo da Consulta a Partir do Nível de Produtividade	119
Figura 30 - Protótipo da Consulta dos Cenários Disponíveis.....	120
Figura 31 - Protótipo do Detalhamento dos Cenários Disponíveis.....	120
Figura 32 - Tela para Carregamento de Novas Árvores	124
Figura 33 - Características da Árvore Carregada.....	124
Figura 34 - Tela para Geração de Cenários.....	125
Figura 35 - Tela após Seleção dos Atributos	126
Figura 36 - Tela Escolher a Visualização de Cenários já Gravados	126

Figura 37 - Visualização de um Cenário Escolhido.....	127
Figura 38 - Busca por Nível de Produtividade.....	127
Figura 39 - Opção Salvar Cenário na Busca por Produtividade.....	128
Figura 40 - Protótipo da Tela para Gerenciar Usuários.....	167
Figura 41 - Protótipo da Tela para Cadastro de Usuário.....	168
Figura 42 - Protótipo da Tela de Login e Senha.....	169
Figura 43 - Tela para Inserção de Novo Usuário	171
Figura 44 - Tela para Cadastro de Usuários	171
Figura 45 - Tela de Login	172

LISTA DE QUADROS

Quadro 1 – Visão Geral dos Atributos e Autores	17
Quadro 2 - Categorias de Modelos De Produtividade e Publicações.....	20
Quadro 3 - Sumarização das Características dos Trabalhos da Categoria 1 ..	22
Quadro 4 - Sumarização das Características dos Trabalhos da Categoria 2 .	26
Quadro 5 - Sumarização Das Características Dos Trabalhos Da Categoria 3.	30
Quadro 6 - Sumarização das Características dos Trabalhos da Categoria 4 ..	33
Quadro 7 - Sumarização das Características dos Trabalhos da Categoria 5 ..	35
Quadro 8 - Principais Características das Ferramentas de Mineração De Dados.....	49
Quadro 9 - Tarefas de Mineração de Dados Utilizadas nos Trabalhos Revisados	52
Quadro 10 - Protocolo de Pesquisa	60
Quadro 11- Síntese das Características da Área de Qualidade Agrícola	69
Quadro 12 - Atributos Selecionados com o Método Delphi.....	71
Quadro 13 - Características dos atributos a serem utilizados na etapa de modelagem	81
Quadro 14 - Resultados das Parametrizações.....	86
Quadro 15 - Acurácia de Trabalhos que Utilizam Classificação.....	87
Quadro 16 - Resultados das Avaliações com Seleção de Atributos	95
Quadro 17 - Síntese dos Procedimentos realizados no Modelo CRISP-DM..	104
Quadro 18 - Casos de Teste	121
Quadro 19 - Caracterização dos especialistas.....	143
Quadro 20 - Resultado da primeira rodada do método Delphi	145
Quadro 21- Resultado da segunda rodada do método Delphi	146
Quadro 22 - Fertilidade	149
Quadro 23 - Textura	149
Quadro 24 - Ambiente de Produção.....	149
Quadro 25 - Fórmula do Adubo.....	150
Quadro 26 - Variedade.....	150
Quadro 27 - Tipo de Solo	151
Quadro 28 - Fazendas	153

Lista de Tabelas

Tabela 1 - Matriz de Confusão de um Classificador - para K classes	43
Tabela 2 - Matriz De Confusão - para 2 classes	44
Tabela 3- Resultado Quantitativo da Consulta às Bases	61
Tabela 4 - Quantidade de Linhas das Planilhas.....	72
Tabela 5 - Outliers dos Boxplots	75
Tabela 6 - Valores do Atributo Produtividade.....	80
Tabela 7 - Resultado do Ranqueamento Realizado pelos Métodos de Seleção De Atributos	89
Tabela 8 - Desempenho Geral do Classificador.....	98
Tabela 9 - Desempenho do Classificador por classe	99
Tabela 10 - Matriz de Confusão da Modelagem 10	100

LISTA DE ABREVIATURAS E SIGLAS

5W2H	<i>What, Why, Where, When, Who, How</i>
AC	Acurácia
CART	<i>Classification and Regression Trees</i>
CFS	<i>Correlation Feature Selection</i>
CLEARMINER	<i>CLimte and rEmote sensing Association patteRns Miner</i>
CRISP-DM	<i>CRoss Industry Standard Process for Data Mining</i>
DEA	<i>Data Envelopment Analysis</i>
EVI	<i>Enhanced Vegetation Index</i>
FN	Falso Negativo
FP	Falso Positivo
ISODATA	<i>Iterative Self-Organizing DATa Analysis</i>
KDD	<i>Knowledge Discovery in Databases</i>
MO – HIDS	<i>Multi-Objective Hybrid Intelligent Suite for Decision Support</i>
NDVI	<i>Normalized Difference Vegetation Index</i>
OBIA	<i>Object Based Image Analysis</i>
PCA	<i>Principal Component Analysis</i>
PVI	<i>Perpendicular Vegetation Index</i>
RF	<i>Random Forest</i>
RMSE	<i>Root Mean Square Error</i>
RVI	<i>Ratio Vegetation Index</i>
SAVI	<i>Soil Adjusted Vegetation Index</i>
SECC	Sistema dos Exploração dos Cenários da Cana
SVM	<i>Support Vector Machines</i>
TFN	Taxa de Falso Negativo
TFP	Taxa de Falso Positivo
TVP	Taxa de Verdadeiro Positivo
VN	Verdadeiro Negativo
VP	Verdadeiro Positivo
WRSI	<i>Water Requirement Satisfaction Index</i>

1 INTRODUÇÃO

Sistemas agrícolas são suscetíveis à variabilidade climática e biofísica (pragas, doenças, etc.), o que aumenta muito a complexidade do planejamento e da tomada de decisão subjacentes.

O clima é uma das ferramentas de gestão de risco que desempenham um papel importante na tomada de decisão. Previsões climáticas impactam em todos os setores da cadeia de valor da cana, assim devem ser utilizadas em uma perspectiva de cadeia de valor como um todo (EVERINGHAM *et al.*, 2002; MEINKE e STONE, 2005).

As diferentes necessidades dos decisores também é outro fator de complexidade para a tomada de decisão nos sistemas agrícolas (MEINKE e STONE, 2005). Nesse sentido, Ahumada e Villalobos (2009) e Jena e Poggi (2013) salientam a importância do planejamento agrícola ser realizado de forma hierárquica: estratégico, tático e operacional.

Meinke e Stone (2005) destacam a relevância da utilização modelos que possam traduzir a importância das projeções climáticas para as atividades de planejamento e tomada de decisão. Além de dados do clima, atributos como tipo do solo, variedade da cana, características da área, entre outros, podem ser utilizados para a modelagem da produtividade da cana.

Considerando essa complexidade envolvida na gestão da produção da cana-de-açúcar, o uso das tecnologias de informação pode contribuir para melhorar a eficiência da gestão desses sistemas. Nesse sentido, as técnicas de mineração de dados permitem a identificação de relacionamentos implícitos em grandes bancos de dados que envolvam um grande número de variáveis. Com isso, é possível descobrir novos padrões, dar maior precisão em padrões conhecidos e modelar fenômenos do mundo real (HAN e KAMBER, 2006).

.A mineração de dados faz parte de um processo denominado “Descoberta de Conhecimento em Bases de Dados”, conhecido como KDD (*Knowledge Discovery in Databases*). Esse processo de KDD é composto por atividades de

pré-processamento, que buscam dar qualidade aos dados que serão utilizados na etapa de mineração e também de atividades de pós-processamento, que visam a analisar os resultados da mineração de dados e apresentar os novos conhecimentos obtidos aos usuários do sistema.

A adequada apresentação dos conhecimentos obtidos é importante para que o processo de KDD seja realmente útil e contribua para as atividades de suporte à decisão e ao planejamento das empresas. Entretanto, ainda há carência de ferramentas que implementem os resultados da mineração e, com isso, contribuam para uma melhor exploração dos diversos cenários e padrões descobertos, subsidiando atividades de tomada de decisão.

1.1 JUSTIFICATIVA E RELEVÂNCIA

O agronegócio desempenha um importante papel na economia brasileira. O Brasil foi um dos países que mais cresceram no comércio internacional desse setor nas últimas décadas. O país é um dos líderes mundiais na produção e exportação de uma série de produtos agropecuários, entre eles os do setor sucroenergético. Além de referência mundial na produção de cana-de-açúcar, o Brasil é o primeiro do mundo na produção de açúcar e etanol, responsável por 53% da quantidade total de etanol vendido e por 61,8% das exportações de açúcar de cana (BRASIL, 2016).

Estimativas do Ministério da Agricultura indicam uma taxa média anual de crescimento de 3,3% na produção de açúcar, no período de 2013/2014 a 2023/2024. Para as exportações, a projeção é de um aumento de 3,7% ao ano nesse mesmo período. Para 2023/2024, é previsto um volume de exportação de 38,8 milhões de toneladas de açúcar (BRASIL, 2014).

As regiões produtoras de cana-de-açúcar concentram-se nos subsistemas regionais Centro-Sul e Norte-Nordeste. No Centro-Sul, destaca-se o Estado de São Paulo, que concentra mais de 50% da produção do país (ACOMPANHAMENTO..., 2015). As projeções do agronegócio para a safra 2023/2024 indicam que a produção de cana-de-açúcar do Estado de São Paulo deve ter um aumento de cerca de 24,6% na próxima década. As projeções

indicam ainda que apenas em Minas Gerais o aumento da produção se dará pelos ganhos em produtividade. Nos demais estados, o crescimento previsto da produção se fará, principalmente, pelo aumento de área plantada (BRASIL, 2014).

Além desses dados do Ministério da Agricultura ainda, de acordo com dados da Agência Paulista de Investimentos e Competitividade (2014), São Paulo é destaque, tanto no cultivo, como na produção de derivados de cana-de-açúcar. O Estado é líder mundial na produção de etanol a partir da cana-de-açúcar; além disso, é pioneiro em pesquisa e desenvolvimento nesse setor e detém uma das matrizes energéticas mais limpas do mundo. “Entre 2003 e 2012, a produção paulista de açúcar cresceu 73,8% e a de álcool 64,5%, impulsionada pelo mercado estadual de biocombustíveis. A economia do setor sucroenergético representa 44% de toda a agropecuária paulista” (SÃO PAULO, 2014).

Diante do cenário exposto, verifica-se que o setor sucroenergético tem grande relevância para geração de saldo positivo na balança comercial brasileira, e sua contínua modernização e adequação à realidade do mercado impactam favoravelmente no desenvolvimento econômico do país. Destaca-se também que esse setor é uma *commodity*, dessa maneira, o preço dos produtos é definido pelo mercado. Assim, aumentar os níveis de produtividade da cana-de-açúcar, tanto pelo aumento de produção, como pela redução de custos, é uma atividade imprescindível para a manutenção do país em sua posição de destaque nesse mercado.

Pela importância da cultura da cana-de-açúcar em termos representativos, tanto para a economia do Estado de São Paulo, quanto para a brasileira, e devido à complexidade inerente à gestão de sistemas agrícolas, justifica-se o estudo dos atributos que mais influenciam a produção dessa cultura; a categorização dos modelos de produtividade; a realização de um processo de descoberta de conhecimento; o desenvolvimento de um roteiro sistematizado e de uma ferramenta para dar apoio à gestão da cultura da cana-de-açúcar, oferecendo possibilidade de visualização dos inúmeros cenários da produção para a tomada de decisão.

1.2 PROBLEMA DE PESQUISA

Posto isto, é possível identificar o seguinte problema de pesquisa:

“Como dar suporte aos processos de tomada de decisão relacionados à gestão da produção de cana-de-açúcar por meio da utilização de técnicas de mineração de dados? ”

1.3 OBJETIVOS

O **objetivo geral** deste trabalho é:

- Propor um roteiro sistematizado, a partir da utilização de técnicas de mineração de dados, para dar suporte aos processos de tomada de decisão na gestão da produção de cana-de-açúcar.

São propostos cinco **objetivos específicos** para a realização desta investigação:

1. Categorizar os modelos de produtividade de cana-de-açúcar;
2. Identificar quais fatores produtivos (atributos) impactam na produtividade da cultura de cana-de-açúcar;
3. Sistematizar o processo de escolha de atributos para a elaboração de um modelo de produtividade;
4. Identificar as atividades de planejamento e tomadas de decisão da cultura da cana-de-açúcar;
5. Implementar uma ferramenta de apoio à decisão que apresente o conhecimento obtido a partir de um processo de KDD, com base em um roteiro sistematizado.

1.4 MÉTODO DA PESQUISA

Para atingir os objetivos desta pesquisa, optou-se pelo campo das pesquisas de natureza **explicativa**, que têm como preocupação central identificar os fatores que determinam ou que contribuem para a ocorrência dos fenômenos.

Neste campo, a referência será a abordagem **experimental**, que consiste em determinar um objeto de estudo, selecionar as variáveis que seriam capazes de influenciá-lo, definir as formas de controle e de observação dos efeitos que a variável produz no objeto (GIL, 2007).

Na Figura 1, é apresentado o *outline* da pesquisa, e os detalhes a respeito dos procedimentos metodológicos são apresentados no Capítulo 3, Abordagem Metodológica.

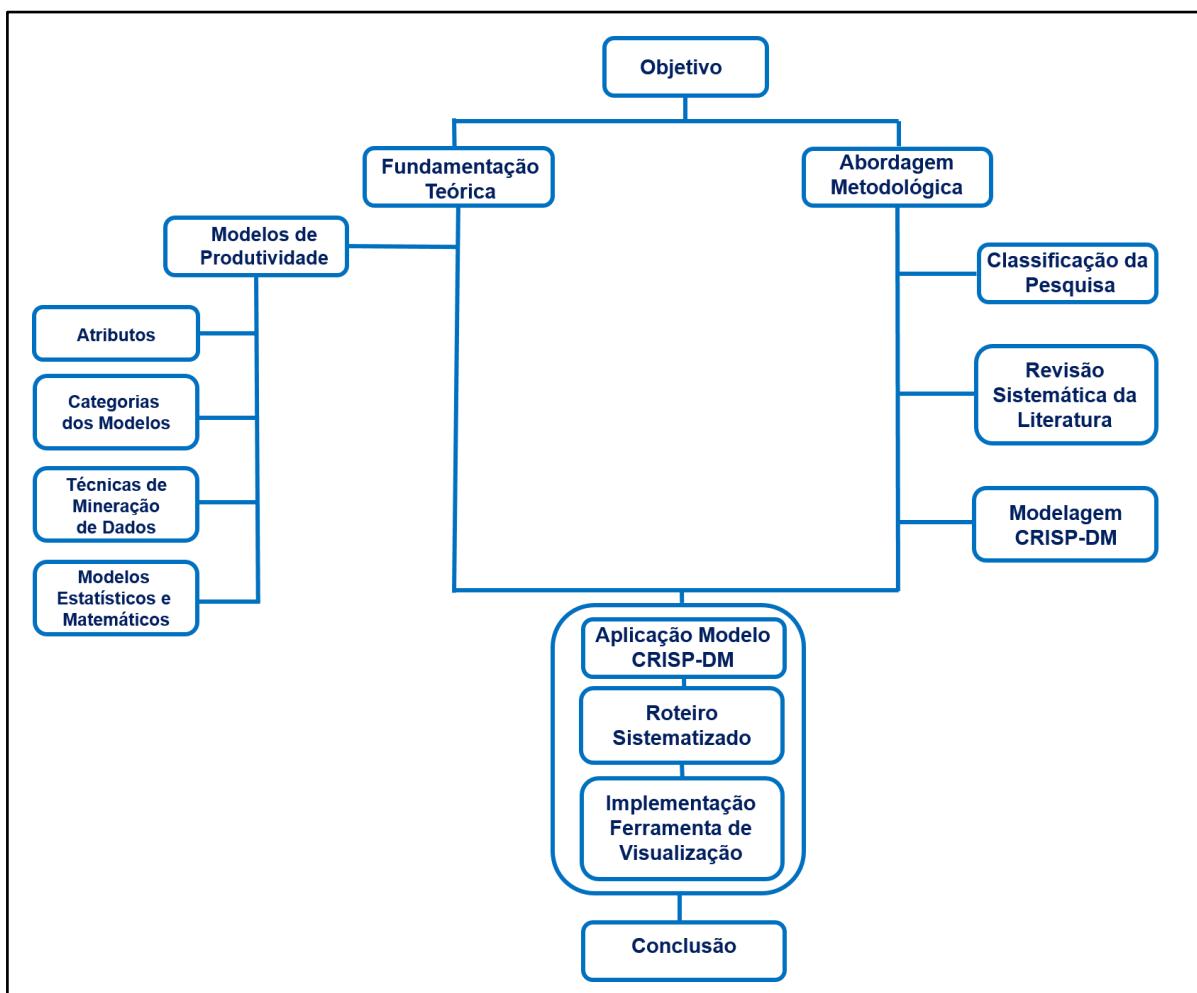


FIGURA 1 - OUTLINE DA PESQUISA

FONTE: A AUTORA

1.5 ESTRUTURA DO TRABALHO

O presente trabalho apresenta, no Capítulo 1, Introdução, contextualização, justificativa e relevância da pesquisa. Em seguida, é colocada a questão de pesquisa focada em modelos de produtividade de cana-de-açúcar e, a partir disso, os objetivos do trabalho e a visão geral do método da pesquisa, que será posteriormente detalhado no Capítulo 3.

No Capítulo 2, Fundamentação Teórica, são descritos os principais conceitos que embasam esta pesquisa, bem como os trabalhos relacionados obtidos a partir de uma Revisão Sistemática da Literatura, caracterizando o estado da arte da área em análise.

No Capítulo 3, Abordagem Metodológica, apresentam-se o planejamento e condução da Revisão Sistemática da Literatura e as etapas do modelo de referência utilizado para a realização do processo de KDD.

No Capítulo 4, Detalhamento do Procedimento Técnico: Modelagem – CRISP-DM, são descritas as atividades realizadas em cada etapa do modelo de referência CRISP-DM para o experimento deste trabalho.

No Capítulo 5, Proposta do Roteiro Sistematizado, são descritos os passos para a obtenção de uma ferramenta de apresentação do conhecimento obtido em processos de descoberta de conhecimento. Também são detalhadas as etapas do desenvolvimento dessa ferramenta.

No Capítulo 6, Conclusão, apresentam-se os principais resultados obtidos com a realização deste trabalho, assim como suas contribuições e trabalhos futuros.

2 FUNDAMENTAÇÃO TEÓRICA

Neste capítulo apresentam-se, inicialmente, os conceitos sobre cana-de-açúcar e tomadas de decisão no cultivo de cana-de-açúcar, bem como a análise dos resultados obtidos em uma Revisão Sistemática da Literatura (RSL). Nessa análise são destacados os principais atributos utilizados nos modelos de cana-de-açúcar, as características dos diversos modelos de produtividade, agrupados em cinco diferentes categorias, as técnicas e as ferramentas computacionais utilizadas em cada uma dessas pesquisas.

Em seguida, são apresentados os conceitos de descoberta de conhecimento em base de dados (KDD), o modelo de referência de processos para a realização de KDD, as principais tarefas, técnicas e ferramentas de mineração de dados e, por último, são destacados os trabalhos da RSL que utilizaram mineração de dados em seus modelos.

2.1 CARACTERÍSTICAS DA CULTURA DA CANA-DE-AÇÚCAR

A cana-de-açúcar (*Saccharum spp*) é uma gramínea semiperene, expressa um bom desenvolvimento em solos em que há boa aeração, boa drenagem e profundidade superior a um metro. Brunini (2008) destaca que essa cultura sofre influência das variáveis climáticas ao longo de todo o ciclo vegetativo e afirma que a temperatura do ar e a precipitação pluvial são os fatores determinantes para o sucesso da cultura e sua exploração econômica. Segundo o autor, o crescimento máximo da cana-de-açúcar se dá entre 30 e 34°C, o crescimento se torna lento em temperaturas abaixo de 25°C e acima de 35°C, e é praticamente nulo em temperaturas inferiores a 21°C e superiores a 38°C.

Ambiente de produção da cana-de-açúcar é o conjunto das interações dos atributos do solo com as condições climáticas locais. De acordo com Prado *et al.* (2008), o ambiente de produção da cana é definido em função das condições físicas, hídricas, morfológicas, químicas e mineralógicas dos solos sob manejo adequado da camada arável, associadas com as condições da subsuperfície do solo e ao clima (precipitação pluviométrica, temperatura, radiação solar,

evaporação). Os autores afirmam também que o ambiente de produção é a soma das interações dos atributos da superfície e da subsuperfície dos solos, considerando-se ainda o grau de declividade desses solos.

Percebe-se, portanto, que diversos atributos podem ser considerados para a elaboração de modelos de produtividade da cana-de-açúcar. Rossetto, Dias e Vitti (2008) salientam que a produtividade será otimizada quando os fatores de produção estiverem potencializados, com melhor solo, baixo déficit hídrico e variedade bem adaptada. Landell e Bressiani (2008) destacam que o grupo de solos favoráveis ao plantio de cana-de-açúcar é composto por aqueles de maior potencial químico e com maior potencial de armazenamento de água. De acordo com Silva, Marins e Dias (2015), o clima é a principal fonte de incerteza no agronegócio, com alto impacto nos custos e no comportamento da curva de maturação da cana-de-açúcar.

Um cultivar, ou variedade de cana-de-açúcar, deve reunir um conjunto de características favoráveis ao ambiente em que será cultivada. A produtividade superior de energia (açúcar, álcool e fibra), que está associada ao acúmulo de biomassa e ao teor de sacarose, é a principal qualidade de uma variedade, mas deve estar associada a inúmeras outras características como, por exemplo, resistência às doenças, acúmulo de sacarose nos períodos de colheita, tolerância à seca, etc. O manejo varietal em cana-de-açúcar tem como objetivo alocar diferentes cultivares comerciais no ambiente que proporcionem a melhor expressão produtiva no contexto considerado (LANDELL e BRESSIANI, 2008).

A cana-de-açúcar possui 2 tipos de ciclo de crescimento, dependendo do cultivar e da data de plantio: “cana-de-ano e meio”, também chamada “cana de 18 meses”, plantada nos meses de janeiro a abril, para ser colhida com mais de 12 meses de idade, e “cana-de-ano”, plantada nos meses de setembro e outubro para ser colhida com aproximadamente doze meses de idade (LANDELL e BRESSIANI, 2008; VIEIRA *et al.*, 2012). Landell e Bressiani (2008), entretanto, salientam que a cana é plantada quase o ano todo objetivando otimizar a estrutura do plantio e, assim, reduzir o custo de produção. Dessa forma, a categoria cana-de-ano está subdividida em: 1) cana-de-ano de inverno, com

plantio realizado nos meses de junho a agosto e; 2) cana-de-ano de primavera, com plantio realizado nos meses de setembro a novembro.

Após a primeira colheita (primeiro corte) a cana cresce novamente e é colhida em intervalos de 12 meses; cada ciclo anual é denominado “cana-soca”. As soqueiras são colhidas anualmente em um período que varia normalmente de 5 a 6 anos. As sucessivas colheitas anuais levam a perda gradual de produtividade até que a cultura não seja mais economicamente rentável, a partir disso o canavial é reformado e dá-se início a um novo ciclo (VIEIRA *et al.*,2012).

2.2 TOMADA DE DECISÃO E PLANEJAMENTO PARA A INDÚSTRIA DA CANA

Segundo Ahumada e Villalobos (2009), no contexto agrícola geral, podem ser identificadas quatro áreas funcionais principais: produção, colheita, armazenagem e distribuição. As decisões tomadas na produção incluem aquelas relacionadas ao cultivo, tais como a terra a ser alocada para cada cultura, tempo de semeadura e a determinação de recursos necessários para o cultivo das culturas. Entre as decisões relacionadas à colheita, estão a definição do cronograma da colheita e a determinação do nível de recursos necessários para realizar essa atividade. Algumas outras decisões tomadas no momento da colheita incluem a programação de equipamentos, mão de obra e equipamentos de transporte. Para a armazenagem é necessário o controle de inventário dos produtos agrícolas, caso necessitem ser armazenados antes ou durante a sua distribuição. Decisões relacionadas ao armazenamento também incluem a quantidade a se armazenar e vender em cada período de planejamento e como posicionar o inventário ao longo da cadeia de valor. As decisões associadas à distribuição incluem selecionar o modo de transporte, as rotas, e o cronograma de transporte do produto.

Para Everingham *et al.* (2002), a cadeia de valor das indústrias de cana é composta pelos setores: cultivo, colheita e transporte, moagem e *marketing/logística*. Os autores salientam o impacto do clima em cada um desses setores e identificam as decisões-chave da indústria que podem ser influenciadas por previsões climáticas sazonais. Ressalta que as previsões climáticas devem agregar valor às abordagens de tomada de decisão utilizadas.

Destacam também que, embora as decisões sejam feitas para os componentes específicos da cadeia de valor, é importante reconhecer que a cadeia representa um sistema integrado, assim, as decisões tomadas por um único setor podem afetar cada um dos outros setores da cadeia de valor. Alguns exemplos dessas decisões são listados a seguir. É possível perceber que o autor associa a previsão do tamanho da colheita a todos os setores.

- **Cultivo** - quando plantar, qual prática de plantio utilizar, qual herbicida aplicar, quando iniciar e quando parar de irrigar, quantidade de fertilizador utilizar e quando aplicar, qual o tamanho da colheita;
- **Colheita e transporte** - quando começar e quando terminar a colheita, quais blocos deveriam ser colhidos primeiro, qual o tamanho da colheita;
- **Moagem** - quando iniciar e quando terminar a moagem, tamanho da equipe, quanto de fibra na colheita, qual o tamanho da colheita;
- **Marketing/logística** - quanto de açúcar vender, quais são os requisitos ótimos de armazenagem e remessa, qual o tamanho da colheita.

Meinke e Stone (2005) afirmam que a previsão climática não é solução para todos os problemas na agricultura. Em vez disso, é uma das muitas ferramentas de gestão de risco que desempenham um papel importante nas tomadas de decisão. Segundo esses autores, a aplicação de informações climáticas requer a identificação dos principais pontos de decisão em sistemas de produção agrícola. Assim como Everingham *et al.* (2002), eles reconhecem que as previsões climáticas devem ser utilizadas em uma perspectiva de cadeia de valor como um todo, para garantir que os benefícios alcançados em um nível não sejam desfeitos no próximo nível. A utilização dessa abordagem pode possibilitar, entre outras coisas: *i*) a identificação das decisões-chave que influenciam a sustentabilidade e a rentabilidade que são impactados pelo clima; *ii*) as principais vulnerabilidades dentro da cadeia de valor em relação à variabilidade climática; *iii*) a avaliação dos benefícios e dos custos de tomada de decisão táticas com base em previsões climáticas em todos os diferentes componentes da cadeia de valor da indústria da cana.

As cadeias de valor mais tradicionais encontradas (automotiva, metal mecânica, entre outras) têm sistemas de baixa variabilidade e incerteza se comparadas aos

sistemas agrícolas. A cadeia de valor do açúcar, bem como a dos demais produtos derivados da cana, são sujeitas a substancial variabilidade e incerteza, tanto climática como biofísica. A complexidade é ainda maior para cadeias de valor nas quais a produção e a usina são controladas por diferentes agentes como, por exemplo, na Austrália, África do Sul e Tailândia (HIGGINS *et al.*, 2007). Everingham *et al.* (2002) destacam a interdependência dos diferentes setores da cadeia de valor e a diversidade dos decisores como fatores de complexidade para a cadeia de valor do açúcar. Meinke e Stone (2005) alertam ainda para o fato que é necessário um foco claro nas diferentes exigências e necessidades específicas dos decisores de forma a fornecer ferramentas mais adequadas para a tomada de decisão.

Ahumada e Villalobos (2009) afirmam que o planejamento na cadeia agrícola geralmente envolve vários níveis de decisões hierárquicas. Essas decisões são classificadas como estratégicas, táticas ou operacionais, dependendo de seus efeitos para a cadeia de valor como um todo. Os autores revisaram artigos que modelam aplicações agrícolas e os categorizaram de diversas formas, entre elas se o modelo era aplicável ao planejamento estratégico, tático ou operacional. De acordo com esses autores, os artigos revisados cobrem uma ampla gama de decisões estratégicas, como a seleção de equipamentos, seleção de tecnologia de agricultura, planejamento financeiro, planejamento de redes de fornecimento, gerenciamento de reservatórios, avaliação de culturas perenes e as estratégias de rotação de culturas. Os modelos táticos lidam com decisões de curto a médio prazo no planejamento agrícola, tais como planos de cultivo, colheita e políticas de plantio. A maioria dos modelos operacionais apresentados visam a determinar planos de colheita, programação de equipamento, alocação de água e preparação da terra. Os autores salientam que os trabalhos tratam mais da modelagem tática do que da operacional e que essa diferença pode refletir a importância do planejamento tático sobre o operacional para os produtos agrícolas.

Para Jena e Poggi (2013) a divisão em um planejamento tático e um planejamento operacional é uma abordagem eficaz para o planejamento da colheita. Inicialmente, realiza-se o planejamento tático para a época de colheita completa, sugerindo decisões para cada semana. Depois, o tempo de

planejamento total é dividido em períodos menores, de até 30 dias. O planejamento operacional é então realizado para cada um dos subperíodos usando as decisões do planejamento tático, com dados de entrada e de tomada de decisão para cada dia. Os autores destacam que é muito comum que a solução de planejamento operacional não possa ser exatamente executada na prática, tal como sugerido. Neste caso, haverá uma reformulação do planejamento operacional, ou até mesmo do planejamento tático.

É importante observar também que os modelos desenvolvidos em um país nem sempre são aplicáveis em outros, em razão das diferenças nas estruturas de negócios entre o plantio e colheita, do nível de infraestrutura mecânica na colheita e do tipo de sistema de transporte (HIGGINS *et al.*, 2007). Além disso, ressalta-se que os resultados econômicos são importantes, mas as decisões também são baseadas em diversos outros fatores, como consequências ambientais, o impacto de plantas daninhas e doenças, estilo de vida e do quadro político existente (MEINKE e STONE, 2005).

2.3 PRODUTIVIDADE DA CANA-DE-AÇÚCAR

De maneira geral, apesar das transformações que sofreu com o tempo, o conceito de produtividade hoje está intimamente ligado à área econômica, correspondendo à razão entre entradas e saídas. Assim, “a produtividade refere-se ao maior ou menor aproveitamento dos recursos no processo de produção, [...] a quanto se pode produzir partindo de uma certa quantidade de recursos” (MOREIRA, 2011, p. 606). Caracterizada dessa maneira, a produtividade assume grande relevância para os processos produtivos, pois seu aumento implica a diminuição dos custos desses processos à medida que serão desenvolvidos com uma quantidade menor de insumos (como capital, mão de obra e outros intermediários, como matéria-prima, combustível e energia elétrica) (MOREIRA, 2011).

Embora o conceito de produtividade, quando se trata de safras agrícolas, diga respeito à tonelada de cana produzida por hectare, neste trabalho entende-se a produtividade em uma perspectiva mais ampla. Dessa forma, entende-se por modelos de produtividade de cana-de-açúcar qualquer abordagem que vise, de

alguma forma, a aumentar a quantidade de cana colhida por hectare ou o teor de sacarose, ou à redução de custos, ou ainda à melhoria nos processos gerenciais.

Na próxima seção, os principais atributos utilizados nos modelos de produtividade de cana-de-açúcar encontrados na literatura são apresentados; na Seção 2.3.2, esses modelos de produtividade são categorizados e descritos.

2.3.1 ATRIBUTOS UTILIZADOS NAS APLICAÇÕES DE PRODUTIVIDADE DA CANA-DE-AÇÚCAR

Em razão da grande influência do clima na produtividade da cana-de-açúcar, muitos dos trabalhos utilizam dados agroclimáticos em seus modelos de produtividade, em especial índices pluviométricos e temperatura. Greenland (2005) utilizou 74 atributos climáticos, incluindo temperaturas diárias e mensais e índices pluviométricos diárias e mensais. Outro trabalho considerou sequências diárias de radiação, precipitação e temperatura entre os anos 1976 e 2003, que foram obtidos a partir de estações climáticas e utilizados como entrada para um sistema de suporte a decisão (EVERINGHAM, SMYTH e INMAN-BAMBER, 2009).

Ferraro, Rivero e Ghersa (2009) analisaram os atributos que mais influenciam a produtividade da cana-de-açúcar. Os autores utilizaram, entre outros atributos, os valores pluviométricos totais e dos meses de verão (dezembro, janeiro e fevereiro). Interessante notar que, nesse trabalho, as variáveis climáticas utilizadas não conseguiram explicar a variabilidade da produtividade. Cock *et al.* (2011) utilizaram dados climáticos em conjunto com dados espaciais, dados da produção, dados das características sociais dos proprietários e gestores de culturas da cana e do café na Colômbia.

Thuankaewsing, Pathumnakul e Piewthongngam (2011) utilizaram oito fatores que afetam o crescimento da cana para elaborar um modelo de previsão de produtividade. Os parâmetros climáticos utilizados nesse trabalho foram: chuva diária acumulada, média das temperaturas e umidade média entre a data da plantação e a data da colheita da cana. O uso de variáveis climáticas também

pode ser observado nos trabalhos de Carbonell e Osorio (2010) e Marin e Carvalho (2012).

O clima foi caracterizado em quatro períodos ao longo do ciclo de desenvolvimento da cana (brotação, perfilhamento, crescimento, maturação), para representar o efeito diferenciado nas diversas fases de crescimento da cana-de-açúcar, nos modelos de produtividade propostos por Bocca (2014). Para as diferentes fases foram calculados a precipitação total e média, bem como períodos de estiagem e número de veranicos (períodos de 10 dias consecutivos sem chuva). Para cada período, foram calculados também os valores médios da temperatura mínima, média e máxima diária.

Os avanços tecnológicos na área de sensoriamento remoto propiciaram seu uso em aplicações reais. Dessa forma, existe um grande volume de imagens de satélite que podem ser utilizadas na agricultura para melhorar o monitoramento das culturas, em especial aquelas cultivadas em grandes extensões, como a cana-de-açúcar (GONÇALVES *et al.*, 2011). Grande parte dos trabalhos que utilizam dados espectrais usam imagens NDVI (*Normalized Difference Vegetation Index*). Gonçalves *et al.* (2011) salientam que séries temporais NDVI, em conjunto com dados do clima, podem ser usadas no desenvolvimento de modelos para previsão de produtividade. O uso de imagens NDVI associados a dados do clima está presente, por exemplo, nos trabalhos de Romani *et al.* (2008), Hajj *et al.* (2009), Gonçalves *et al.* (2011), Fernandes, Rocha e Lamparelli (2011), Gonçalves *et al.* (2012) e Romani *et al.* (2013).

A utilização de imagens de sensoriamento remoto se dá também nos trabalhos de Everingham *et al.* (2007), Nascimento *et al.* (2009), Vintrou *et al.* (2013) e Nonato e Oliveira (2013), nesses casos, entretanto, sem a combinação com dados do clima.

Conforme mencionado na Seção 2.1, o tipo de solo é um dos fatores de grande influência na produtividade. Carbonell e Osorio (2010) listaram os tipos de solo e outras características para obtenção de máxima produtividade. Atributos químicos do solo são analisados, associados aos dados do relevo, no trabalho de Souza *et al.* (2010). Os autores mostraram que a altitude do terreno revelou

uma correlação positiva com a produtividade, e ainda que as variáveis altitude e potássio apresentaram os maiores valores de correlação com a produtividade.

Cerri e Magalhães (2012) avaliaram as correlações entre a produtividade da cana-de-açúcar e alguns atributos físicos e químicos do solo. Nesse trabalho essas correlações foram baixas e, dessa forma, não foram capazes de explicar a variação na produtividade da cana-de-açúcar, o que indica que, além das propriedades do solo, outras variáveis devem ser analisadas. Esse resultado está de acordo com o obtido em Souza *et al.* (2010).

No trabalho de Bocca (2014) foram utilizados atributos que descreviam os componentes químicos e as frações granulométricas do solo, assim como sua classificação. O autor utilizou ainda informações de manejo relativas à adubação, datas de plantio, colheita e colheita anterior, número de cortes, aplicação de torta de filtro e vinhaça. O atributo tipo de solo faz parte também dos modelos de Piewthongngam, Suksawat e Tenglolai (2007) e Thuankaewsing, Pathumnakul e Piewthongngam (2011).

Outro atributo importante em modelos de produtividade é a variedade da cana-de-açúcar. Utilizaram o atributo cultivar como parte de seus modelos os trabalhos de Jiao, Higgins e Prestwidge (2005), Everingham *et al.* (2007), Ferraro, Rivero e Ghersa (2009), Carbonell e Osorio (2010), Thuankaewsing, Pathumnakul e Piewthongngam (2011) e Silva, Marins e Dias (2015).

Como as sucessivas colheitas levam à perda de produtividade, alguns autores incluíram o número de corte em seus modelos. No trabalho de Ferraro, Rivero e Ghersa (2009), a produtividade da cana foi associada principalmente ao cultivar e ao número de cortes. Everingham *et al.* (2007) também utilizaram o número de cortes em seu modelo; fizeram uso de dados hiperespectrais referentes a nove ciclos da colheita da cana. No trabalho de Carbonell e Osorio (2010), as áreas de produtividade alta estavam relacionadas predominantemente ao primeiro, segundo e terceiro cortes, enquanto áreas com produtividade média e baixa estavam relacionadas ao primeiro, segundo, terceiro e quarto cortes.

Além dos atributos mais comuns, como dados do clima, imagens NDVI, tipo de solo, variedade e número de cortes, vale ressaltar a utilização de diversos outros

atributos nos diferentes modelos de produtividade, como, por exemplo: resultados de entrevistas com gestores (PIEWTHONGNGAM, SUKSAWAT e TENGLOLAI, 2007); traços de qualidade, morfologia e característica da biomassa (PACHECO, LUCAS e LIMA NETO(2008); dados das características do terreno (HAJJ *et al.*, 2009;FERRARO, RIVERO e GHERSA, 2009);dados das características sociais dos proprietários e gestores da cana (COCK *et al.*,2011); informações a respeito das práticas de gestão (fatores sociodemográficos e socioeconômicos, atributos biofísicos e fatores ambientais) (ANANTHARA, ARUNKUMAR e HEMAVATHY, 2013). Outra característica que se destaca é a variação de combinação de atributos nesses trabalhos, resultando em uma grande diversidade de modelos que objetivam o aumento da produtividade da cana-de-açúcar.

No Quadro 1, apresenta-se uma síntese dos principais atributos que foram referenciados pelos autores pesquisados.

QUADRO 1 – VISÃO GERAL DOS ATRIBUTOS E AUTORES

	Autores (ano)	
	Atributos	
Clima	X	Greenland (2005)
Imagens NDVI e dados do clima	X	Jiao, Higgins e Prestwich (2005)
Variedade	X	Everingham <i>et al.</i> (2007)
Características do Solo	X	Piewthongngam, Suksawat e Tengolai (2007)
Sensoriamento remoto	X	Pacheco, Lucas e Lima Neto (2008)
Número de cortes	X	Romani <i>et al.</i> (2008)
Características da área	X	Everingham, Smyth, e Imman-Bamber (2009)
	X	Hajji <i>et al.</i> (2009)
	X	Ferraro, Rivero e Ghersa (2009)
	X	Nascimento <i>et al.</i> (2009)
	X	Carbonell e Osorio (2010)
	X	Souza <i>et al.</i> (2010)
	X	Cock <i>et al.</i> (2011)
	X	Gonçalves <i>et al.</i> (2011)
	X	Fernandes Rocha e Lamparelli (2011)
	X	Romani <i>et al.</i> (2011)
	X	Thuankaewsing, Pathumnakul e Zhou <i>et al.</i> (2011)
	X	Ferraro, Ghersa e Rivero (2012)
	X	Gonçalves <i>et al.</i> (2012)
	X	Marin e Carvalho (2012)
	X	Santchurn <i>et al.</i> (2012)
	X	Vieira <i>et al.</i> (2012)
	X	Ananthara, Arunkumar e Hemavathy (2013)
	X	Barros, Oliveira e Oliveira (2013)
	X	Nawi <i>et al.</i> (2013)
	X	Nonato e Oliveira (2013)
	X	Romani <i>et al.</i> (2013)
	X	Vintrou <i>et al.</i> (2013)
	X	Bocca (2014)
	X	Mello, Atzberger e Formaggio (2014)
	X	Nawi <i>et al.</i> (2014)
	X	Silva, Marins e Dias (2015)

CONTINUAÇÃO QUADRO 1: VISÃO GERAL DOS ATRIBUTOS E AUTORES

		Autores (ano)
	Atributos	
Entrevistas com gestores		Greenland (2005)
Características sociais dos gestores da cana		Jiao, Higgins e Prestwidge (2005)
Práticas de gestão	X	Everingham <i>et al.</i> (2007)
Características morfológicas	X	Piewthongngam, Suksawat e Tenglolai (2007)
		Romani et al. (2008)
		Everingham, Smyth, e Inman-Bamber (2009)
		Hajji <i>et al.</i> (2009)
		Ferraro, Rivero e Ghersa (2009)
		Nascimento <i>et al.</i> (2009)
		Carbonell e Osorio (2010)
		Souza <i>et al.</i> (2010)
	X	Cock <i>et al.</i> (2011)
		Gonçalves <i>et al.</i> (2011)
		Fernandes Rocha e Lamparelli (2011)
		Romani et al. (2011)
		Thuankaewsing, Pathumnakul e Zhou <i>et al.</i> (2011)
		Ferraro, Ghersa e Rivero (2012)
		Gonçalves <i>et al.</i> (2012)
		Marin e Carvalho (2012)
		Santchurn <i>et al.</i> (2012)
		Vieira <i>et al.</i> (2012)
	X	Ananthara, Arunkumar e Hemavathy (2013)
		Barros, Oliveira e Oliveira (2013)
		Nawi <i>et al.</i> (2013), Nonato e Oliveira (2013)
		Romani <i>et al.</i> (2013)
		Vintrou <i>et al.</i> (2013)
		Bocca (2014)
		Mello, Atzberger e Formaggio (2014)
		Nawi <i>et al.</i> (2014)
		Silva, Marins e Dias (2015)

FONTE: A AUTORA

2.3.2 PRODUTIVIDADE DE CANA-DE-AÇÚCAR E APLICAÇÃO DE TÉCNICAS

A complexidade da gestão das culturas de cana-de-açúcar suscita a criação de diferentes modelos e ferramentas para apoio ao planejamento e à tomada de decisão. Esses modelos visam a dar suporte aos diversos setores da cadeia de valor da cana, para as decisões de nível estratégico, tático ou operacional. Assim, a partir da análise da literatura revisada, foi possível identificar vários tipos de propostas, que neste trabalho foram classificadas em 5 categorias.

Na categoria 1, estão os trabalhos cujo principal objetivo é subsidiar as atividades de programação da colheita; os modelos dessa categoria, em geral, estão ligados ao planejamento tático e/ou operacional. Na categoria 2, foram agrupados os trabalhos que visam à previsão de produtividade e se destinam ao planejamento estratégico e/ou tático. Foram classificados na categoria 3, aqueles trabalhos que, objetivando aumento de produtividade, buscam caracterizar áreas plantadas com cana-de-açúcar, geralmente em escala regional, cujos modelos podem ser aplicados ao planejamento estratégico e/ou tático. Na categoria 4, estão os modelos de apoio às práticas de gestão que, em sua maioria, levam em consideração a experiência dos gestores para a elaboração de seus modelos. Finalmente, na categoria 5, foram agrupados os diferentes trabalhos que não puderam ser classificados nas categorias anteriores, mas que também podem contribuir para o aumento da produtividade da cana-de-açúcar.

No Quadro 2, cada referência é associada à sua respectiva categoria de modelo de produtividade da cana-de-açúcar.

QUADRO 2 - CATEGORIAS DE MODELOS DE PRODUTIVIDADE E PUBLICAÇÕES

Categorias de Modelos de Produtividade	Autores/ ano da Pesquisa
1) Programação da colheita	Jiao, Higgins e Prestwidge (2005), Pacheco, Lucas e Lima Neto (2008), Thuankaewsing, Pathumnakul e Piewthongngam (2011), Silva, Marins e Dias (2015)
2) Previsão de produtividade	Greenland (2005), Romani et al. (2008), Everingham, Smyth, e Inman-Bamber (2009), Ferraro, Rivero e Ghersa (2009), Fernandes Rocha e Lamparelli (2011), Gonçalves et al. (2012), Ananthara, Arunkumar e Hemavathy (2013) Nawi et al. (2013), Mello, Atzberger e Formaggio (2014), Nawi et al. (2014) Bocca (2014)
3) Caracterização de áreas	Nascimento et al. (2009), Carbonell e Osorio (2010), Gonçalves et al. (2011), Romani et al. (2011), Santchurn et al. (2012), Vieira et al. (2012), Nonato e Oliveira (2013), Vintrou et al. (2013)
4) Apoio às Práticas de gestão	Everingham et al. (2007), Piewthongngam, Suksawat e Tenglolai (2007), Hajj et al. (2009), Souza et al. (2010), Cock et al. (2011), Romani et al. (2013)
5) Outros	Zhou et al. (2011), Ferraro, Ghersa e Rivero (2012) Barros, Oliveira e Oliveira (2013)

FONTE: A AUTORA

A seguir, são apresentados os trabalhos revisados, agrupados de acordo com essas 5 categorias. Destacam-se, para cada trabalho, as técnicas e ferramentas computacionais utilizadas, bem como seus objetivos principais.

Categoria 1: programação da colheita

Conforme já citado, a gestão de culturas de cana-de-açúcar é uma atividade complexa que demanda muito planejamento. Uma das tarefas críticas para o sucesso de todo o processo é a colheita. Um cronograma de colheita eficiente pode elevar o desempenho da cadeia de abastecimento, maximizando a

produtividade da cana-de-açúcar (THUANKAEWSING, PATHUMNAKUL e PIEWTHONNGAM, 2011).

Jiao, Higgins e Prestwidge (2005) desenvolveram um modelo polinomial de segunda ordem e incorporaram os parâmetros estimados em um modelo de programação linear com o objetivo de melhorar a programação da colheita, de forma a maximizar o ganho de conteúdo de açúcar da cana para três regiões de usinas pertencentes à indústria australiana de açúcar. Os autores também desenvolveram uma ferramenta, denominada Sugarmax, para aplicação desse modelo.

Pacheco, Lucas e Lima Neto (2008), investigaram como as preferências dos gestores, aplicadas às decisões de colheita, podem influenciar a produção final da cultura da cana-de-açúcar em Pernambuco. O modelo desenvolvido pelos autores utilizou um *framework* multiobjetivo (*Multi-Objective Hybrid Intelligent Suite for Decision Support - MO-HIDS*), que integrou técnicas inteligentes como redes neurais artificiais, lógica *fuzzy* e um algoritmo evolucionário multiobjetivo.

Thuankaewsing, Pathumnakul e Piewthongngam (2011) utilizaram um modelo de redes neurais para previsão de produtividade de cana-de-açúcar e, em seguida, usaram essa previsão em um modelo de programação linear a fim de obter uma solução ótima de programação de colheita para grupos de fazendeiros da Tailândia, de forma que o ganho em produtividade fosse equilibrado entre esses fazendeiros.

Silva, Marins e Dias (2015) elaboraram um modelo para auxiliar na tomada de decisão para o planejamento da safra de cana-de-açúcar de forma a otimizar o rendimento da produção de álcool e açúcar e reduzir os custos agroindustriais. Para a derivação do modelo foi utilizada programação orientada a objetivo, multiescolha revisada. Esse modelo foi implementado em um sistema de modelagem algébrica geral e testado em uma usina de açúcar e álcool de Minas Gerais.

No Quadro 3, são summarizadas as características dos trabalhos classificados na categoria 1. Destaca-se aqui o uso de programação linear em dois dos trabalhos e a combinação de duas diferentes técnicas nesses mesmos trabalhos. Quanto

aos atributos, percebe-se a utilização de características da área (talhões e *paddock*), presentes em dois dos trabalhos e a variedade (ou cultivar) da cana, utilizado em três dos trabalhos desse grupo. Interessante notar que as características climáticas foram utilizadas somente em um dos trabalhos. Destaca-se ainda a utilização dos dados da qualidade da cana, de sua morfologia e características de biomassa de cana.

QUADRO 3 - SUMARIZAÇÃO DAS CARACTERÍSTICAS DOS TRABALHOS DA CATEGORIA 1

Autor	Técnicas	Atributos	Ferramentas
Jiao, Higgins e Prestwidge (2005)	Modelo polinomial de segunda ordem Programação linear	Fazenda de origem do <i>paddock</i> , área do <i>paddock</i> , variedade, conteúdo de açúcar da cana, data da colheita, produtividade	Sugarmax
Pacheco, Lucas e Lima Neto (2008)	Framework multiobjetivo - redes neurais artificiais, lógica fuzzy e um algoritmo evolucionário multiobjetivo	Traços de qualidade da cana, traços de morfologia da Cana, e características de biomassa de cana	
Thuankaewsing, Pathumnakul e Piewthongngam (2011)	Redes neurais Programação linear	Tipo cana e cultivar, tipo de solo, chuva diária acumulada, temperatura média, umidade média e idade média de cana	Matlab
Silva, Marins e Dias (2015)	Programação orientada a objetivo, multi-escolha revisada.	Informações sobre talhões de terreno, a estratégia de corte, estado de cana de açúcar, condição de cana, e variedades de cana em diferentes fazendas.	

FONTE: A AUTORA

Categoria 2: previsão de produtividade

Everingham, Smyth e Inman-Bamber (2009) destacam a importância da realização de estimativas de produtividade argumentando que, se a safra já estiver negociada antes da colheita e a previsão for superestimada, os produtores terão de comprar a produção de um concorrente, provavelmente a preços mais altos. Por outro lado, se a safra for subestimada, a cultura pode não obter o preço máximo. Gonçalves *et al.* (2012) salientam a importância

estratégica para o país de informações corretas sobre as *commodities* agrícolas. Afirmam que o processo para a estimativa de produtividade das colheitas deveria ser preciso e objetivo. Afirmam ainda que no Brasil esse processo é subjetivo, baseado em entrevistas com profissionais do agronegócio e em informações sobre demanda de insumos na cadeia produtiva. Os trabalhos apresentados a seguir são propostas de modelos para apoiar a previsão de produtividade da cana.

Greenland (2005) realizou entrevistas com especialistas de cultura e um levantamento bibliográfico para identificar potenciais fatores climáticos que controlam a produtividade anual de cana na Louisiana. Realizou também uma análise estatística utilizando dados de produtividade da cana de todo o estado, de 1963 a 2002, e um banco de dados do clima. A partir disso foi construído um modelo para simular valores de produtividade.

Romani *et al.* (2008) propuseram uma técnica que combina a aplicação da teoria de fractais para determinar grupos de atributos correlacionados, e a mineração de dados para identificar padrões significativos no conjunto de dados de sensoriamento remoto da cana-de-açúcar e dados climáticos referentes a 10 municípios do Estado de São Paulo, a fim de melhorar a estimativa da produção da cana em nível nacional.

Para uma previsão de produtividade mais precisa, Everingham, Smyth e Inman-Bamber (2009) desenvolveram um modelo composto por diversos modelos (*ensemble*) utilizando a técnica Lasso de mineração de dados estatísticos. Como essa técnica é computacionalmente intensiva, os autores implementaram a estimativa “*stagewise forward*” como uma aproximação da solução Lasso mais eficiente. A ferramenta APSIM sugarcane foi utilizada para realizar simulações de produção da cultura de cana-de-açúcar, considerando vários tipos de opções ambientais e de gestão em Ayr, uma importante região de cana-de-açúcar da Austrália.

Ferraro, Rivero e Ghersa (2009), fizeram uso de mineração de dados a fim de determinar os fatores que mais influenciam na produtividade de cana-de-açúcar na Argentina. Os autores utilizaram os dados da produção para gerar modelos

que associam esses dados ao total de cana por hectare e ao total de açúcar por hectare. Embora o objetivo seja identificar os principais fatores que determinam a produtividade, esse trabalho foi classificado na categoria 2 (previsão de produtividade) porque utiliza uma técnica preditiva de mineração de dados, de forma que esse modelo poderia ser utilizado também na previsão de produtividade.

Fernandes, Rocha e Lamparelli (2011) utilizaram diversos algoritmos de seleção de atributos (*Qui-quadrado*, método *Wrapper* com algoritmo J48, CFS - *Correlation Feature Selection* - e uma combinação de *infoGain* e *GainRatio*) para selecionar os melhores atributos a serem usados em um modelo preditivo por meio do algoritmo de classificação denominado J48. O modelo foi utilizado para classificar o rendimento médio municipal dos 20 municípios selecionados para a pesquisa no Estado de São Paulo. A seleção de atributos e a classificação foram realizadas por meio da ferramenta Weka.

Gonçalves *et al.* (2012) utilizaram imagens de satélite (NOAA-AVHRR) para gerar índices NDVI para as regiões selecionadas: quatro cidades grandes produtoras de cana do Estado de São Paulo (Araras, Jaboticabal, Luís Antônio e Ribeirão Preto). Assim, usando dados meteorológicos, foi calculado o índice de satisfação do requisito de água (*Water Requirement Satisfaction Index* - WRSI). Em seguida, os autores aplicaram dois métodos estatísticos diferentes para séries temporais visando a analisar e gerar modelos de previsão de produtividade da cana-de-açúcar em uma escala regional.

Ananthara, Arunkumar e Hemavathy (2013), na Índia, propuseram um novo algoritmo de mineração de dados para ser utilizado na previsão de produtividade de diversos tipos de cultura, entre eles a cana-de-açúcar. Os autores utilizaram indicadores quantitativos da fazenda, dados estatísticos, agronômicos e do tipo da colheita extraídos de um banco de dados Oracle e informações a respeito das práticas de gestão. Foram utilizados algoritmos de seleção de atributos para melhor qualidade dos resultados da mineração, que foi realizada com a ferramenta Clementine.

Nawi *et al.* (2013), na Austrália, utilizaram dados espectrais, análise de componentes principais, modelo de mínimos quadrados parciais (PLS) e redes neurais artificiais para a previsão do conteúdo de açúcar da cana. Nawi *et al.* (2014) utilizaram dados espectrais, calibrados com o método de mínimos quadrados parciais, com base em valores de referência de *Brix*, teor de fibra e teor de umidade, com o objetivo de prever a qualidade da cana-de-açúcar.

O modelo de previsão de produtividade de Mello, Atzberger e Formaggio (2014) analisou dados de sensoriamento remoto e os comparou com dados históricos oficiais de produtividade da cana-de-açúcar do Estado de São Paulo. A qualidade dos resultados obtidos com as séries temporais de sensoriamento remoto em relação aos dados oficiais foi calculada utilizando raiz do erro quadrático médio (*Root Mean Square Error*, RMSE).

No modelo de previsão de produtividade proposto por Bocca (2014), foram utilizados dados de clima em conjunto com dados relacionados à produção de cana-de-açúcar de uma usina de cana-de-açúcar localizada no município de Teodoro Sampaio, no estado de São Paulo. O autor analisou o desempenho de diversas técnicas de mineração de dados em um contexto de seleção de atributos, para 75 diferentes atributos.

No Quadro 4 pode-se observar que, na categoria 2, grande parte dos trabalhos utilizam informações do clima, bem como dados de sensoriamento remoto, em especial imagens NDVI. O trabalho de Nawi *et al.* (2013), embora classificado nessa categoria, faz previsão de conteúdo de açúcar na cana e assim utiliza atributos diferente dos demais (amostras entrenós de diferentes variedades). Os atributos usados em Ananthara, Arunkumar e Hemavathy (2013) também são bastante distintos dos demais, envolvendo, entre outros, informações a respeito das práticas de gestão. Quanto às técnicas utilizadas, destaca-se o uso da mineração de dados em vários dos trabalhos, além de técnicas estatísticas.

QUADRO 4 - SUMARIZAÇÃO DAS CARACTERÍSTICAS DOS TRABALHOS DA CATEGORIA 2

Autor	Técnicas	Atributos	Ferramentas Computacionais
Greenland (2005)	Análise estatística	Variáveis climáticas críticas - temperatura média máxima de agosto de temperatura média mínima de fevereiro, excedente de água no solo entre abril e setembro e ocorrência de furacões de outono	
Romani <i>et al.</i> (2008)	Teoria dos fractais Mineração de dados (regras de associação)	Dados espectrais (NDVI e WRSI) e metereológicos (temperatura máxima e mínima e precipitação)	
Everingham, Smyth, e Inman-Bamber (2009)	Simulação Mineração de dados estatísticos – Lasso	Dados de produtividade, radiação, precipitação e temperatura	APSIM sugarcane
Ferraro, Rivero e Ghersa (2009) Fatores que mais influenciam na produtividade de cana-de-açúcar	Mineração de dados (<i>k-means</i> e CART)	Cultivar cana, código da fazenda, classe de cultura, mês da colheita, duração do cultivo, área de campo, regime de precipitação total e precipitação nos meses de verão	
Fernandes Rocha e Lamparelli (2011)	Diversos algoritmos de seleção de atributos Mineração de dados (árvore de decisão - algoritmo J48)	Imagens NDVI e dados climáticos (precipitação, radiação global e temperatura)	Weka
Gonçalves <i>et al.</i> (2012)	Métodos estatísticos	Dados espectrais (NDVI e WRSI) e metereológicos (temperatura e precipitação)	
Nawi <i>et al.</i> (2013) Prever conteúdo de açúcar na cana	Análise de componentes principais, Agricultura de precisão, Modelo de mínimos quadrados parciais (PLS) e redes neurais artificiais	Amostras entrenós representando três diferentes variedades de cana comerciais	
Ananthara, Arunkumar e Hemavathy (2013)	Mineração de dados (proposição algoritmo CRY)	Indicadores quantitativos da fazenda, dados estatísticos, dados agronômicos, dados do tipo da colheita e informações a respeito das práticas de gestão	Clementine

(CONT.) QUADRO 4: SUMARIZAÇÃO DAS CARACTERÍSTICAS DOS TRABALHOS DA CATEGORIA 2

Autor	Técnicas	Atributos	Ferramentas Computacionais
Nawi <i>et al.</i> (2014) Previsão da qualidade da cana	Método de mínimos quadrados parciais, com base em valores de referência de Brix, teor de fibra e teor de umidade	Dados espectrais	
Mello, Atzberger e Formaggio (2014)	Erro quadrático médio (<i>Root Mean Square Error</i> , RMSE).	Imagens NDVI e informações sobre o terreno	
Bocca (2014)	Mineração de dados: Diversas técnicas de classificação	Química e as frações granulométricas do solo, informações de manejo, precipitação total e média, estiagem e veranico para cada período de desenvolvimento da cana, médias das temperaturas mínima, média e máxima para cada período	Pacote R

FONTE: A AUTORA

Categoria 3: caracterização de áreas

Outro tema frequente nas pesquisas é o monitoramento agrícola como uma forma de identificar características comuns de áreas plantadas com cana-de-açúcar e, dessa forma, subsidiar os produtores e gestores nas tomadas de decisão nos diversos setores da cadeia de valor da cana. Nascimento *et al.* (2009) utilizaram séries temporais de imagens de satélite (12 imagens mensais de valor máximo de NDVI) e aplicaram dois métodos, análise harmônica e árvore de decisão, para identificar áreas com cana-de-açúcar no Estado de São Paulo. Os autores afirmam que a metodologia pode ser aplicada para dar suporte a sistemas oficiais de previsão de colheita. A ferramenta Weka foi utilizada para a indução da árvore de decisão.

No trabalho de Carbonell e Osorio (2010), para caracterizar áreas plantadas com cana-de açúcar e seus níveis de produtividade em uma região na Colômbia denominada Cauca River Valley, foram utilizados dois métodos de interpolação: *Ordinary Kriging* e *Voronoi polygons*, implementados no pacote de software

ARCGIS com módulos para análise geoestatística e análise espacial. Os dados foram obtidos de um banco de dados de produção de cana desde 1990, com um número aproximado de 311.000 registros e de um banco de dados geográficos.

Romani *et al.* (2011) desenvolveram um sistema para monitoramento da expansão da cana no Estado de São Paulo. Todo o processo, incluindo a extração de imagens NDVI multitemporais, métodos de *clusterização* e visualização geoespacial, foi implementado em um novo sistema chamado *SatImagExplore*.

O objetivo do trabalho de Gonçalves *et al.* (2011) foi avaliar a variação da produtividade da cana-de-açúcar em uma escala regional e assim melhorar a compreensão do desenvolvimento da cana e sua expansão para outras regiões do país. O estudo foi realizado em dezesseis cidades do Estado de São Paulo. Utilizaram-se as técnicas Análise de Componentes Principais (PCA – *Principal Component Analysis*) para reduzir a dimensão do conjunto de dados e a tarefa de mineração de dados *clusterização* para a caracterização dos grupos, por meio do software Minitab. O software ARCGIS foi usado para visualizar a distribuição espacial de cada *cluster* para todas as cidades e colheitas analisadas. Santchurn *et al.* (2012) também utilizaram PCA e *clusterização* para o agrupamento de variedades da cana com alta quantidade de biomassa, com dados obtidos do Instituto de Pesquisa da Indústria do Açúcar da República de Maurício.

Vieira *et al.* (2012) desenvolveram uma metodologia para contribuir na automatização de mapeamento de cana em grandes áreas. Para esse fim, foram combinadas duas técnicas principais: análise de imagens baseada em objetos (*Object Based Image Analysis - OBIA*), utilizando a ferramenta Definiens Developer®, e mineração de dados, por meio da ferramenta Weka.

O objetivo do estudo de Vintrou *et al.* (2013) foi testar uma abordagem de mineração de dados original para classificar e mapear a terra cultivada na África Ocidental usando imagens com resolução “grosseira” (*coarse resolution*) e comparar os resultados da classificação com aqueles obtidos a partir de uma abordagem clássica de classificação de imagens denominada ISODATA (*Iterative Self-Organizing DATa Analysis*).

Nonato e Oliveira (2013) utilizaram técnicas de mineração de dados para classificação e identificação de áreas cultivadas com cana-de-açúcar, em diferentes cidades produtoras no Estado de São Paulo (Ipuã, São Joaquim da Barra e Guará). Para essa identificação automatizada, foram utilizadas imagens do satélite Landsat 5/TM de áreas cultivadas com cana-de-açúcar em três fases fenológicas diferentes. Os autores utilizaram cinco índices de vegetação: *Normalized Difference Vegetation Index* (NDVI), *Enhanced Vegetation Index* (EVI), o *Perpendicular Vegetation Index* (PVI), o *Soil Adjusted Vegetation Index* (SAVI) e *Ratio Vegetation Index* (RVI). Segundo os autores, a introdução de atributos de textura contribuiu para melhorar a distinção de áreas cultivadas com cana-de-açúcar em meio a tipos diversos de cobertura do solo. Ressaltaram também que os índices de vegetação (como, por exemplo, o NDVI) mostraram-se relevantes na distinção da fase e do estado fenológico das culturas.

No Quadro 5 são sumarizados os trabalhos que tratam sobre caracterização de áreas plantadas com cana-de-açúcar. Em relação às técnicas utilizadas, assim como na categoria 2, a mineração de dados está presente em vários dos trabalhos, em alguns deles utilizada em conjunto com outras técnicas. Em dois dos trabalhos foi utilizada a técnica PCA para seleção de atributos. Quanto aos atributos, novamente as imagens NDVI são utilizadas na maioria dos trabalhos e dados climáticos são utilizados em apenas dois dos trabalhos dessa categoria. São utilizados ainda dados da produção e de produtividade.

QUADRO 5 - SUMARIZAÇÃO DAS CARACTERÍSTICAS DOS TRABALHOS DA CATEGORIA 3

Autor/objetivo	Técnicas	Atributos	Ferramentas Computacionais
Nascimento et al. (2009) Identificação de áreas com cana-de-açúcar	Análise harmônica Mineração de dados (árvore de decisão - algoritmo J48)	Dados de sensoriamento remoto (NDVI)	Weka
Carbonell e Osorio (2010) Áreas com máxima produtividade	Métodos de interpolação: <i>Ordinary Kriging</i> e <i>Voronoi polygons</i> ,	Ano de colheita, mês da colheita, tempo do corte em meses, variedade, solo, zona agroecológica, número de soqueiras, características climáticas	ARCGIS
Gonçalves et al. (2011) Variação de produtividade	Mineração de dados (Análise de componentes Principais, clusterização)	Terra cultivada, NDVI, precipitação, temperatura	Minitab - Mineração de dados - ARCGIS -visualizar a distribuição espacial de cada cluster
Romani et al. (2011) Monitoramento da expansão da cana	Mineração de dados (métodos de clusterização) e visualização geoespacial	Imagens NDVI, dados de produtividade	Desenvolvido um novo sistema- SatImagExplore.
Santchurn et al. (2012) Agrupamento de variedades da cana com alta quantidade de biomassa	Análise de componentes principais e clusterização	Traços da qualidade, da morfologia e da biomassa da cana	ASReml
Vieira et al. (2012) Mapeamento da cana em grandes áreas	Análise de imagens baseada em objetos (<i>Object Based Image Analysis</i> - OBIA)	Dados espectrais, espaciais, de textura e NDVI	Definiens Developer® - Análise de objetos Weka - mineração de dados
Vintrou et al. (2013) Classificação e mapeamento de áreas cultivadas	Clasificação ISODATA e mineração de dados	Dados de sensoriamento remoto	
Nonato e Oliveira (2013) Identificação de áreas cultivadas com cana-de-açúcar	Mineração de dados	Dados de sensoriamento remoto – imagens de cobertura do solo Índices de vegetação: NDVI, EVI, PVI, SAV, RVI	

FONTE: A AUTORA

Categoria 4: modelos de apoio às práticas de gestão

Os diversos modelos e sistemas de apoio à gestão permitem tomada de decisão de forma mais ágil e com maior eficiência, minimizando custos, otimizando os

recursos e as atividades produtivas e, dessa forma, acarretando ganho de produtividade e maximização dos lucros. Assim, algumas pesquisas objetivam associar as boas práticas dos gestores aos modelos propostos de forma a obter melhores resultados na gestão da cultura da cana-de-açúcar, como pode ser notado nos trabalhos a seguir.

Everingham *et al.* (2007) utilizaram dados hiperespectrais de sensoriamento remoto, técnicas estatísticas e de mineração de dados visando a classificar variedades de cana-de-açúcar e seu ciclo de desenvolvimento. O objetivo do trabalho foi subsidiar os gestores nas decisões táticas e estratégicas, em fazendas na Austrália. O pacote estatístico R foi utilizado para as análises.

O objetivo do trabalho de Piewthongngam, Suksawat e Tenglolai (2007) foi distinguir, entre os produtores de cana, os mais eficientes dos menos eficientes, de forma que as diferentes atividades agrícolas executadas pelos produtores eficientes possam servir como um *benchmark*. Os autores utilizaram análises qualitativas e quantitativas. Para a análise qualitativa, foram conduzidas várias entrevistas com gestores e suas equipes de serviços agronômicos e ainda vários grupos de fazendeiros, na Tailândia. Para a análise quantitativa, foi utilizada a Análise Envoltória de Dados (*Data Envelopment Analysis – DEA*).

Hajj *et al.* (2009) apresentaram uma abordagem para monitorar as práticas agrícolas, particularmente a colheita das culturas, usando séries temporais de imagens de satélite integradas com informações de modelagem de crescimento da cultura e conhecimento especializado. A área de estudo foi uma pequena ilha, *Reunion Island*, pertencente à França. Para a automatização do processo, foi desenvolvido um sistema de apoio à decisão baseado em lógica *fuzzy*.

O trabalho de Souza *et al.* (2010) auxilia a tomada de decisão sobre uso e manejo do solo. Os autores mapearam a produtividade da cana-de-açúcar em uma área de aproximadamente 23ha, em Catanduva, no Estado de São Paulo, por meio de um monitor de produtividade que permitiu a elaboração de um mapa digital que representa a superfície de produção para a área em estudo. No trabalho, foram utilizadas as técnicas de geoestatística e indução de árvore de decisão. A árvore de decisão foi induzida no programa SAS *Enterprise Miner*.

No trabalho de Cock *et al.* (2011), são descritos dois casos: café e cana-de-açúcar. Nos dois casos, a observação dos resultados obtidos pelos agricultores da Colômbia, com a variação natural no meio ambiente e as práticas de gestão distintas que eles aplicam, são usadas para determinar práticas de manejo da cultura local específica, por meio de pesquisa operacional.

Romani *et al.* (2013) desenvolveram um algoritmo não supervisionado, chamado CLEARMiner - *CLimte and rEmote sensing Association patteRns Miner*, para minerar regras de associação. Esse módulo foi integrado em um sistema de informação de sensoriamento remoto, Agri_remoto, desenvolvido para melhorar o acompanhamento dos canaviais. O estudo foi realizado com dados de cinco regiões produtoras do Estado de São Paulo.

No Quadro 6 são summarizadas as características dos trabalhos da categoria 4. Para cada trabalho é apresentada também a característica da atividade de apoio à gestão. Nessa categoria é possível notar a utilização de variadas técnicas, tais como: mineração de dados, técnicas estatísticas, lógica fuzzy, redes neurais e pesquisa operacional. Pode-se perceber também uma grande variedade de atributos utilizados. Além dos atributos climáticos e dados espectrais, destacam-se dados das características sociais dos proprietários e gestores da cana, conhecimento obtido com os produtores e gestores, dados da morfologia da cana, atributos físicos e químicos do solo e outros dados referentes à produção da cana.

QUADRO 6 - SUMARIZAÇÃO DAS CARACTERÍSTICAS DOS TRABALHOS DA CATEGORIA 4

Autor/objetivo	Técnicas	Atributos	Ferramentas Computacionais
Everingham et.al.(2007) Classificar variedades de cana-de-açúcar e seu ciclo de desenvolvimento	Técnicas estatísticas, Mineração de dados (random forest e support vector machine)	Dados hiperespectrais de nove variedades de cana, Número de cortes	Pacote estatístico R
Piewthongngam, Suksawat e Tenglolai (2007) Analisar os procedimentos eficientes dos gestores	Análise Envoltória de Dados –DEA	Dados do tipo de solo, irrigação, colheita, tamanho da área agrícola. Entrevista com produtores	
Hajj (2009) Monitoramento de práticas agrícolas, particularmente a colheita das culturas	Lógica fuzzy	Séries temporais de imagens de sensoriamento remoto, modelagem de crescimento da cultura e conhecimento especializado	Desenvolvido um sistema de apoio à decisão baseado em lógica fuzzy
Souza et al. (2010) Tomada de decisão sobre uso e manejo do solo	Técnicas de geo estatística e indução de árvore de decisão.	pH, cálcio, magnésio, potássio, fósforo, teor de matéria orgânica e saturação por bases de um Argissolo Vermelho-Amarelo	SAS Enterprise Miner
Cock et al. (2011) Melhorar práticas de gestão para melhorar produtividade	Pesquisa operacional de Agricultura Precisão	Dados espaciais, dados climáticos e meteorológicos, dados da produção, dados das características sociais dos proprietários e gestores da cana	
Romani et al. (2013) Acompanhamento de canaviais	Mineração de dados (regras de associação, algoritmo CLEARMiner integrado em um sistema de informação de sensoriamento remoto)	Dados espectrais (NDVI e WRSI) e meteorológicos (temperatura e precipitação)	Agri-remoto

FONTE: A AUTORA

Categoria 5: outros

Além dos trabalhos classificados, destacam-se aqui algumas pesquisas que não foram enquadradas nas quatro categorias citadas anteriormente, mas entende-se que também promovem o aumento da produtividade da cana. Zhou *et al.* (2011) desenvolveram um modelo de redes neurais, utilizando a ferramenta SAS *enterprise miner* para seleção de mudas de cana com maior rendimento. Ferraro, Ghersa e Rivero (2012) também utilizaram mineração de dados para determinar a hierarquia de fatores que influenciam na composição de ervas daninhas nas culturas de cana-de-açúcar. Barros, Oliveira e Oliveira (2013) desenvolveram um sistema de recomendação, a partir de técnicas de mineração de dados, para conteúdos relacionados à cultura da cana-de-açúcar. O sistema foi implantado no portal da Agência de Informação Embrapa, visando a organizar, tratar, armazenar e divulgar informações técnicas agrícolas para o setor sucroenergético.

No Quadro 7 são sumarizados esses modelos. Em relação às técnicas utilizadas, a mineração de dados foi utilizada em dois dos três trabalhos classificados nessa categoria. Para os atributos, ressalta-se ainda a utilização de número de colmo e dados do caule para o trabalho que visa a selecionar mudas com maior rendimento, bem como herbicidas aplicados, rendimento da biomassa, entre outros, utilizados no trabalho que objetiva identificar os fatores que influenciam a composição de ervas daninhas.

QUADRO 7 - SUMARIZAÇÃO DAS CARACTERÍSTICAS DOS TRABALHOS DA CATEGORIA 5

Autor/objetivo	Técnicas	Atributos	Ferramentas Computacionais
Zhou <i>et al.</i> (2011) Seleção de mudas com maior rendimento	Redes neurais	Número de colmos, altura e diâmetro do caule	SAS enterprise miner
Ferraro, Ghersa e Rivero(2012) Hierarquia de fatores que influenciam na composição de ervas daninhas	Mineração de dados (K-means e CART)	Área de campo, classe safra, herbicidas aplicados, rendimento biomassa, intervalo de tempo entre a amostra e colheita; tempo (mês) de amostra; quantidade de chuva durante 12 meses antes da pesquisa de plantas daninhas, genótipo da cana, qualidade do solo e campo da colheita com queima ou não queima de restos de colheita	
Barros, Oliveira e Oliveira (2013) Desenvolvimento de um sistema de recomendação	Mineração de dados (regras de associação)	Dados dos usuários e dados referentes ao acesso a <i>links</i> sobre cana-de-açúcar	Sistema de recomendação

FONTE: A AUTORA

2.4 DESCOBERTA DE CONHECIMENTO EM BASE DE DADOS

A informatização generalizada das mais diversas áreas, o aumento de capacidade e redução de custos dos dispositivos de armazenamento e a evolução dos dispositivos de coleta de dados, como leitores de código de barras e sensores, propiciaram o armazenamento de grandes quantidades de dados. Entretanto, a exploração desse grande volume de dados é muito difícil de ser realizada sem a utilização de técnicas e ferramentas que possibilitem a transformação dos dados em informação e conhecimento úteis.

A utilização de ferramentas de Mineração de Dados torna possível a análise dessas enormes quantias de dados (TSAI, 2013). De acordo com Tsai (2012) essa tecnologia fornece diversas metodologias para a tomada de decisão, resolução de problemas, análise, planejamento, diagnóstico, detecção, integração, prevenção, aprendizagem e inovação.

A mineração de dados é um campo interdisciplinar que combina inteligência artificial, gerenciamento de banco de dados, visualização de dados, aprendizagem de máquina, algoritmos matemáticos e técnicas estatísticas (HAN e KAMBER, 2006; TSAI, 2012; UNIVASO, ALE e GURLEKIAN, 2015), faz parte de um processo maior denominado descoberta de conhecimento em base de dados ou KDD, do inglês *knowledge Discovery in DataBase*. Para Fayyad, Piatetski-shapiro e Smyth (1996), KDD é um processo, de várias etapas, não trivial, interativo e iterativo, para identificação de padrões comprehensíveis, válidos, novos e potencialmente úteis a partir de grandes conjuntos de dados. Na realização de um processo de KDD é de grande importância a cooperação entre os responsáveis pela execução do processo e os especialistas do domínio, os quais devem possuir amplo conhecimento a respeito da área em análise.

O processo de KDD é constituído de várias etapas operacionais, normalmente definidas como: pré-processamento, mineração de dados e pós-processamento. Quanto melhor a qualidade dos dados, melhores serão os resultados na etapa de mineração de dados. Nesse sentido, as atividades de pré-processamento têm grande relevância no processo de descoberta de conhecimento. As atividades de pré-processamento podem ser divididas em: 1. **Limpeza** dos dados: etapa na qual são eliminados ruídos e dados inconsistentes; 2. **Integração** dos dados: etapa em que diferentes fontes de dados podem ser combinadas produzindo um único repositório de dados; 3. **Seleção**: etapa na qual são selecionados os atributos que interessam ao processo de descoberta de conhecimento; 4. **Transformação** dos dados: etapa em que os dados são transformados num formato apropriado para aplicação de algoritmos de mineração (por exemplo, por meio de operações de agregação) (AMO, 2004; HAN e KAMBER, 2006).

A mineração de dados, etapa essencial do processo de KDD, é composta por tarefas, definidas também como funcionalidades. A tarefa consiste na especificação do que se está buscando nos dados, que tipo de regularidades ou categoria de padrões tem-se interesse em encontrar, ou que tipo de padrões poderiam surpreender (por exemplo, um gasto de um cliente de cartão de crédito, fora dos padrões usuais de seus gastos) (AMO, 2004).

Ainda segundo Amo (2004) e Han e Kamber (2006), as atividades de pós-processamento são: 1. **Avaliação**: etapa em que são identificados os padrões interessantes; 2. **Visualização dos Resultados**: etapa na qual são utilizadas técnicas de representação do conhecimento para apresentar ao usuário o conhecimento minerado. A Figura 2 representa esquematicamente o processo de KDD.

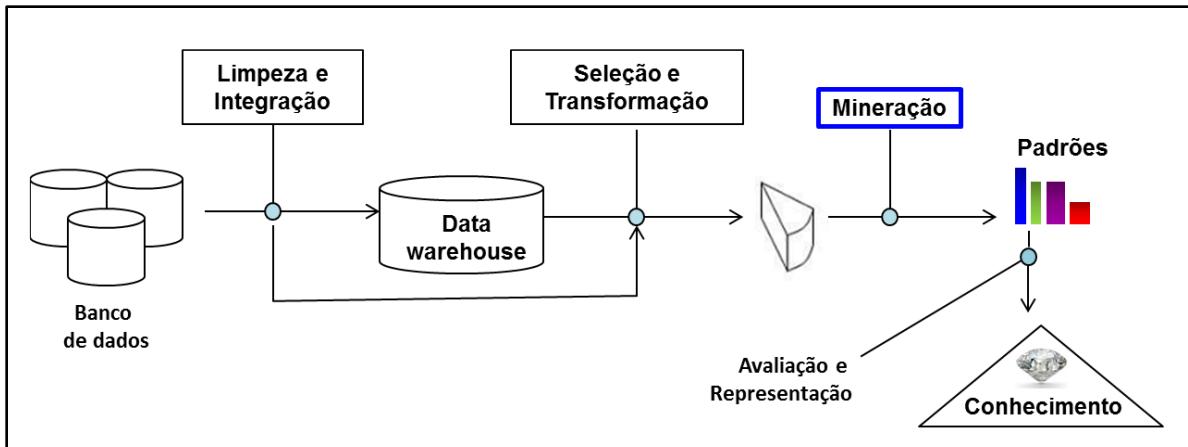


FIGURA 2 - ETAPAS DO KDD

FONTE:AMO (2004)

2.4.1 MODELO DE REFERÊNCIA DE PROCESSO PARA REALIZAÇÃO do KDD

O modelo de referência CRISP-DM - *CRoss Industry Standard Process for Data Mining* - (CHAPMAN *et al.*, 2000) é bastante utilizado em projetos de mineração de dados, tanto acadêmicos, como empresariais (BARROS, OLIVEIRA E OLIVEIRA, 2013). É composto por seis etapas, e reflete a natureza iterativa e interativa dos processos de descoberta de conhecimento:

- **Entendimento do negócio:** fase inicial que foca no entendimento, objetivos e requisitos necessários da perspectiva do negócio, convertendo, em seguida, esse conhecimento em um problema de mineração de dados e elaborando um plano preliminar para atingir os objetivos;
- **Entendimento dos dados:** essa fase começa com um conjunto inicial dos dados e prossegue com atividades que objetivam ganhar familiaridade com esses dados, para identificar problemas com sua qualidade, descobrir características preliminares ou detectar

subconjuntos interessantes para formar hipóteses a respeito de informações “escondidas”;

- **Preparação dos dados:** essa fase cobre todas as atividades que visem à construção do conjunto final de dados para a fase da mineração (dados que serão utilizados na ferramenta de modelagem) a partir dos dados brutos iniciais. Provavelmente, as atividades dessa fase serão executadas várias vezes e sem qualquer ordem pré-definida. Faz parte da preparação de dados: a seleção de atributos, a transformação e limpeza de dados para a ferramenta de modelagem;
- **Modelagem:** nessa fase várias técnicas de modelagem são selecionadas, aplicadas e os respectivos parâmetros “calibrados” para valores ótimos. Normalmente, existem várias técnicas para o mesmo tipo e problema de Mineração de Dados. Algumas técnicas têm requisitos específicos para o tipo e formato dos dados. Portanto, frequentemente é necessário voltar à fase de Preparação de Dados;
- **Avaliação:** nesse estágio, já foi construído um modelo (ou modelos) que parece ter boa qualidade sob a perspectiva da análise dos dados. Antes de prosseguir para a aplicação final do modelo, é importante que seja feita uma avaliação aprofundada desse modelo, juntamente com uma revisão dos passos executados para sua construção, para certificar-se de que ele atinge os objetivos do negócio. Um aspecto importante nesta fase é determinar se há algum item importante do negócio que não tenha sido suficientemente considerado. Ao final desta fase, decide-se se o modelo está pronto para ser usado ou deve ser revisto;
- **Implementação:** o conhecimento obtido e avaliado deverá ser organizado e apresentado de tal forma que o usuário possa utilizá-lo em um ambiente real da empresa. Dependendo do caso, essa fase pode ser tão simples quanto a geração de um relatório ou tão complexa quanto implementar o processo de mineração de dados por toda a empresa. É importante que o usuário compreenda que ações precisam ser realizadas para que o modelo criado seja efetivamente utilizado.

Na Figura 3, é possível perceber a natureza iterativa do modelo CRISP-DM.

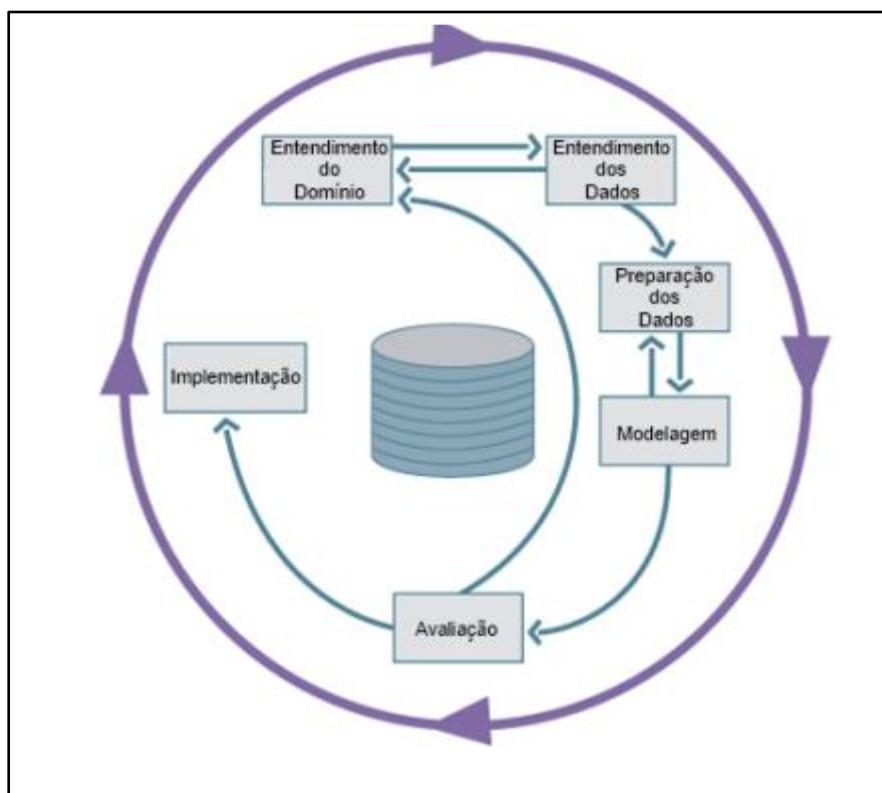


FIGURA 3 - FASES DO MODELO DE PROCESSO CRISP-DM

FONTE: ADAPTADO DE CHAPMAN ET AL., 2000

2.4.2 TAREFAS E TÉCNICAS DE MINERAÇÃO DE DADOS

Não há um consenso em relação à classificação das tarefas de mineração de dados. Colak, Sagiroglu e Yesilbudak (2012), por exemplo, classificam as tarefas de mineração de dados como caracterização e discriminação, classificação, *clusterização*, associação, análise de *outlier* e análise de evolução. Para Tsai (2013) as tarefas mais importantes são classificação, predição, associação, *clusterização* e sumarização. Essas tarefas são apresentadas resumidamente a seguir.

A **Classificação** é citada por Goldschmidt e Passos (2005) como uma das tarefas do KDD mais populares e importantes; consiste na busca por uma função que permita associar corretamente cada registro do banco de dados a uma classe. A tarefa de classificação é realizada em dois passos: 1) encontrar um modelo para o atributo alvo (também chamado de atributo meta ou classe) como

uma função dos valores dos outros atributos; 2) associar as instâncias com classes não conhecidas a uma determinada classe com a maior precisão possível. A tarefa de **predição (ou regressão)** é similar à classificação, entretanto o atributo para os quais os valores são previstos é um valor contínuo e não categórico como no caso da classificação.

A tarefa de **Associação** (ou regras de associação) consiste na busca por regras de associação frequentes e válidas, baseando-se em parâmetros de **suporte**, que dizem respeito à frequência da regra, e no nível de **confiança**, que expressa a força da regra. Esses parâmetros devem ser especificados pelo especialista em KDD, juntamente com o especialista no domínio da aplicação.

A tarefa de **Clusterização (ou agrupamento)** consiste em particionar os registros da base de dados em subconjuntos (ou *clusters*) de maneira que elementos presentes em um *cluster* compartilhem um conjunto de propriedades comuns e que os diferenciem dos elementos de outros *clusters*. Em geral, o conjunto de dados utilizados para efetuar a *clusterização* não possui uma classe pré-definida, por isso, essa tarefa é caracterizada como “não supervisionada”, enquanto a tarefa de classificação é definida como “supervisionada” (HAN e KAMBER, 2006).

A **sumarização** é uma estratégia para identificar as principais características em um conjunto de dados e apresentá-las de forma concisa e compreensível.

As tarefas de mineração de dados podem ser classificadas como atividades preditivas e descritivas. Nas tarefas preditivas de mineração de dados, constroem-se modelos, a partir de variáveis com valores conhecidos, para a previsão de um valor desconhecido de outra variável, o atributo meta. Nas tarefas descritivas, descrevem-se conceitos relevantes de forma concisa, informativa e discriminante. Os dois principais tipos de tarefas para predição são a classificação e a regressão, e para as descritivas são a associação e a *clusterização*.

As técnicas de mineração consistem na especificação de métodos que garantam como descobrir padrões potencialmente úteis em um conjunto de dados. Algumas das principais técnicas utilizadas em mineração de dados são: indução

por árvores de decisão, redes neurais artificiais, classificação bayseana, K-vizinhos mais próximos, algoritmos genéticos, lógica *fuzzy*, *support vector machine* (GOLDSCHMIDT e PASSOS, 2005; COLAK, SAGIROGLU e YESILBUDAK, 2012).

2.4.3 INDUÇÃO POR ÁRVORE DE DECISÃO

A Indução por Árvore de Decisão é uma das principais técnicas de mineração de dados utilizada para a tarefa de classificação (GOLDSCHMIDT e PASSOS, 2005). Essa importância se dá pela sua expressividade simbólica, pois, diferentemente de outras técnicas, é possível entender a estrutura preditiva do modelo, ou seja, quais atributos e seus respectivos valores são mais determinantes para a previsão da classe.

Uma árvore de decisão é um fluxograma com a estrutura de uma árvore. Cada nó interno representa um teste sobre um atributo, cada ramo representa o resultado do teste, os nós folhas representam as classes, o nó no topo da árvore é denominado nó raiz. A geração de uma árvore consiste de duas fases: 1) construção da árvore de decisão e 2) simplificação ou poda da árvore de decisão (GOLDSCHMIDT e PASSOS, 2005).

Na fase de **construção da árvore de decisão**, inicialmente, a raiz da árvore contém todas as instâncias de todas as classes; então, é utilizada uma heurística para selecionar o atributo que melhor discrimine as instâncias pelas suas classes. A esse atributo é associado um predicado que divide as instâncias em dois ou mais conjuntos (partições). Esse processo é repetido recursivamente até que o conjunto consista inteiramente ou predominantemente de instâncias de uma mesma classe, constituindo o nó folha (LOYOLA, MEDINA e GARCÍA, 2015).

Se o atributo associado ao predicado é um atributo discreto (categórico), um ramo é criado para cada valor do atributo. Se o atributo é contínuo, dois ramos são criados de acordo com o teste do valor do atributo em relação a um valor constante. Alternativamente, para os atributos discretos também pode ser gerado um teste binário.

O critério para a seleção dos atributos que melhor separam as instâncias em cada iteração é uma importante operação durante o processo de construção da árvore. Entre as principais medidas para seleção de atributos, estão: ganho de informação, razão do ganho de informação e o índice Gini (HAN e KAMBER, 2006).

A medida “**ganho de informação**” é baseada na teoria da informação. Para cada nó da árvore, calcula-se o atributo com maior ganho de informação, ou seja, aquele que minimiza a informação necessária para classificar uma instância na partição resultante e que reflete a menor “impureza” nessa partição. Essa medida é tendenciosa, pois seleciona atributos que tenham um grande número de valores. Assim, para superar esse problema, foi criada a medida “**razão do ganho de informação**”, que considera o número de instâncias para cada nó em relação ao número total de instâncias daquela partição. O algoritmo C4.5, um dos mais tradicionais algoritmos utilizados para a indução de árvores de decisão, utiliza a medida razão do ganho de informação.

O “**índice Gini**” é usado por outro importante algoritmo de indução de árvore de decisão, o CART (*Classification and Regression Trees*). Esse índice é calculado pela soma dos quadrados das probabilidades de uma instância em uma partição pertencer a uma classe específica. O índice Gini considera divisões binárias para cada atributo.

Na construção de uma árvore de decisão, alguns ramos podem refletir anomalias dos dados de treinamento em decorrência de ruídos ou *outliers*; assim, a árvore ficaria muito ajustada a esse conjunto de dados de treinamento. Esse tipo de problema é chamado de *overfitting*, e, para resolvê-lo, utilizam-se métodos de **poda**. Árvores podadas são menores e menos complexas e, por essa razão, mais fáceis de entender. Embora a poda seja um método bastante utilizado e eficaz na solução de *overfitting* é importante não podar demais a árvore, pois pode ocorrer o problema conhecido como *underfitting*, que ocorre quando o modelo de classificação não aprendeu o suficiente sobre os dados de treinamento e assim fica muito generalizado. Existem duas abordagens para a realização dessas podas: a pré-poda e a pós-poda (HAN e KAMBER, 2006).

Na pré-poda, a árvore é podada durante a sua construção, ou seja, utilizam-se medidas (significância estatística, ganho de informação, etc.) para avaliar se cada nó deve ser dividido ou não. Se a divisão do nó é avaliada abaixo de um determinado limiar, então a divisão não é realizada e esse nó torna-se nó folha.

Na pós-poda, as subárvoreas são removidas depois que a árvore foi construída. Uma subárvore de um determinado nó é podada removendo-se seus ramos e trocando-o por um nó folha, que será atribuído à classe mais frequente entre as subárvoreas podadas.

2.4.4 MEDIDAS DE DESEMPENHO DOS CLASSIFICADORES

Para medir o desempenho dos modelos de classificação, pode ser utilizada uma matriz de confusão, que resume as informações sobre as classes corretas e aquelas preditas por um sistema de classificação. Conforme apresentado na Tabela 1, os resultados são apresentados em duas dimensões: classes verdadeiras e classes preditas, para k classes distintas $\{C_1, C_2, \dots, C_k\}$. Cada elemento $M(C_i, C_j)$ da matriz, $i, j = 1, 2, \dots, K$, representa o número de exemplos da base de dados que pertencem à classe C_i , mas que foram classificados como sendo da classe C_j .

TABELA 1 - MATRIZ DE CONFUSÃO DE UM CLASSIFICADOR - PARA K CLASSES

Classes	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	$M(C_2, C_1)$	$M(C_2, C_2)$...	$M(C_2, C_k)$
...
Verdadeira C_k	$M(C_k, C_1)$	$M(C_k, C_2)$...	$M(C_k, C_k)$

FONTE: MONARD E BARANAUSKAS, 2003, P. 102

O número de acertos, para cada classe, está localizado na diagonal principal da matriz, $M(C_i, C_i)$. Os demais elementos $M(C_i, C_j)$, em que $i \neq j$, representam erros na classificação.

Várias medidas podem ser derivadas da matriz de confusão. Para exemplificar essas medidas, apresenta-se, na Tabela 2, por simplicidade, uma matriz para um problema com apenas duas classes, denominadas como C₊ (positiva) e C₋ (negativa). Nesse caso existem duas possibilidades de acerto: Verdadeiro Positivo (VP) e Verdadeiro Negativo (VN) e duas possibilidades de erro: Falso Positivo (FP) e Falso Negativo (FN).

TABELA 2 - MATRIZ DE CONFUSÃO - PARA 2 CLASSES

Classes	Predita C ₊	Predita C ₋
Verdadeira C ₊	VP	FN
Verdadeira C ₋	FP	VN

FONTE: ADAPTADO DE GOLDSCHMIDT E PASSOS, 2005

Assim, as entradas da matriz de confusão têm os seguintes significados:

- VP é o número de predições corretas da classe C₊;
- FN é o número de predições incorretas da classe C₊;
- FP é o número de predições incorretas da classe C₋;
- VN é o número de predições corretas da classe C₋.

As equações de 1 a 6 são medidas obtidas a partir da matriz de confusão:

- Acurácia (AC) é a proporção do número total de predições que foram corretas:

$$AC = (VP + VN) / (VP + VN + FP + FN) \quad (1)$$

- Sensitividade, Revocação, ou Taxa de Verdadeiro Positivo (TVP) é a proporção de casos positivos que foram identificados corretamente:

$$TVP = VP / (VP + FN) \dots \quad (2)$$

- Taxa de Falso Positivo (TFP) é proporção de casos negativos que foram classificados incorretamente como positivos:

$$TFP = FP / (FP + VN) \quad (3)$$

- Especificidade ou Taxa de Verdadeiro Negativo (TVN) é definida como a proporção de casos negativos que foram classificados corretamente:

$$TVN = VN / (VN + FP) \quad (4)$$

- Taxa de Falso Negativo (TFN) é a proporção de casos positivos que foram classificados incorretamente como negativos:

$$TFN = FN / (FN + VP) \quad (5)$$

- Precisão (P) é a proporção de casos positivos preditos que foram corretos:

$$P = VP / (VP + FP) \quad (6)$$

De acordo com Han e Kamber (2006) as taxas de verdadeiro positivo, verdadeiro negativo, falso positivo e falso negativo são úteis para avaliar os custos (ou riscos) e benefícios associados com o modelo de classificação.

Além das medidas citadas, é comum serem apresentados, na utilização de classificadores, a estatística *Kappa* e o erro médio absoluto:

- A estatística *Kappa* mede a concordância entre dois métodos de classificação, gerando, assim, um aspecto de confiabilidade e precisão dos dados classificados. Mede o grau de acurácia além do que seria esperado somente pelo acaso. Seus valores variam de zero a um. Quanto menor o valor de *Kappa*, menor a confiança de observação, o valor um implica a correlação perfeita.
- O erro médio absoluto é a média da diferença entre os valores reais e os preditos em todos os casos, é a média do erro da predição.

Existem diversos métodos para avaliar o desempenho de um classificador. Nesses métodos o conjunto é dividido em grupos de treinamento e teste. O modelo será construído a partir do grupo de treinamento e será testado pelo grupo de teste.

A divisão do conjunto de dados em grupos de treinamento e de teste pode ser feita de diversas formas. O método denominado *resubstituição*, consiste em construir o modelo e testar o seu desempenho no mesmo conjunto de dados, ou seja, o grupo de teste é igual ao grupo de treinamento. Esse método produz estimativas extremamente otimistas da precisão, pois o processo de classificação tenta maximizá-la. Dessa forma, o desempenho calculado com esse método possui um viés otimista, isto é, o bom desempenho do conjunto de treinamento em geral não se estende a conjuntos independentes de teste (MONARD e BARANAUSKAS, 2005).

Um método bastante utilizado é conhecido como *holdout* (HAN e KAMBER, 2006; WITTEN e FRANK, 2005). Nesse método o conjunto de dados é dividido em dois grupos distintos para o treinamento e teste. Uma proporção muito comum utilizada é a de 2/3 dos dados para treinamento e 1/3 restante para teste. Segundo Han e Kamber (2006), esse método produz uma estimativa pessimista porque somente uma porção inicial dos dados é usada para derivar o modelo.

Quando não há dados suficientes para partitionar em conjuntos distintos de treinamento e teste sem perder capacidade significativa de modelagem ou teste utiliza-se o método denominado validação cruzada (*cross-validation*) (WITTEN e FRANK, 2005). Nesse método o conjunto de dados é dividido em k subconjuntos mutuamente exclusivos de tamanhos aproximadamente iguais. O procedimento é executado k vezes, cada vez um subconjunto é utilizado para teste e os demais conjuntos para treinamento. A acurácia final é calculada pela média das acurácias obtidas em cada um dos subconjuntos de teste. No método denominado validação cruzada estratificada (*stratified cross-validation*) as partições são estratificadas de maneira que a distribuição das instâncias nas classes em cada partição é equivalente àquela do conjunto inicial (HAN e KAMBER, 2006).

De acordo com Witten e Frank (2005), testes extensivos em inúmeros conjuntos de dados mostraram que dez é o número de partições próximo do correto para se obter a melhor estimativa de acurácia pela validação cruzada, segundo esses autores algumas evidências teóricas apontam para esse número, entretanto ainda há muito debate sobre o assunto. Han e Kamber (2006) também afirmam

que a validação cruzada estratificada com dez partições tem viés e variância relativamente baixos.

2.4.5 FERRAMENTAS DE MINERAÇÃO DE DADOS

Existem diversas ferramentas disponíveis no mercado para auxílio ao processo de KDD, tanto “proprietárias”, como “livres”. Apresentam-se, a seguir, algumas dessas ferramentas, destacando-se as tarefas de mineração que implementam e as possibilidades de visualização dos resultados que oferecem.

SAS *Enterprise Miner*¹ contém um conjunto de ferramentas de análise com uma interface que pode ser usada para criação e comparação de vários modelos. Implementa as tarefas de classificação, clusterização, regras de associação, descoberta de padrão sequencial e análise de cesta de mercado. As ferramentas de visualização permitem análise de histogramas multidimensionais e comparação gráfica dos resultados da modelagem. O *Enterprise Miner* é projetado para PCs ou servidores que estão sendo executados nos sistemas operacionais Windows XP, UNIX, Linux ou versões posteriores desses ambientes.

O *Intelligent Miner*² é fabricado pela IBM, entretanto não é dependente do sistema da IBM, podendo ser utilizado com outros gerenciadores de Banco de dados relacionais. Implementa as seguintes tarefas de mineração de dados: clusterização, classificação, regressão, regras de associação, padrões sequenciais. Permite a utilização de algoritmos de mineração de dados de forma individual ou combinada. Disponibiliza diversos tipos de gráficos para análise visual dos resultados.

¹ Disponível em:
<http://support.sas.com/documentation/cdl/en/emgsj/62040/HTML/default/viewer.htm#a003307712.htm>

² Disponível em:
http://www.ibm.com/support/knowledgecenter/SSEPGG_9.5.0/com.ibm.im.overview.doc/c_ibm_db2_intelligent_miner_modeling.html

O *Oracle Data Mining*³ é o componente do Oracle para a realização de atividades de mineração de dados. As seguintes tarefas São implementadas: classificação, regressão, detecção de anomalias, regras de associação, *clusterização* e extração de características (cria novos atributos utilizando combinação linear dos atributos originais). O módulo *Oracle Data Miner* é a interface gráfica que disponibiliza meios para a preparação de dados, mineração e avaliação do modelo. Possui ainda a geração de histogramas, *Boxplot*, gráficos de linhas e de barras. Esse módulo possibilita também a geração de código para acesso a banco de dados (PL/SQL) para as atividades de mineração.

R⁴ é uma linguagem e ambiente para computação estatística. É uma ferramenta “livre” e de “código aberto”, que compila e roda em uma ampla variedade de plataformas UNIX e sistemas similares (incluindo FreeBSD e Linux), Windows e MacOS. R pode ser estendido facilmente via pacotes. Há cerca de oito pacotes fornecidos com a distribuição de R e muitos mais estão disponíveis por meio da família CRAN de sites na Internet (servidores Web que armazenam as versões atualizadas e documentações do R), que cobrem uma gama muito ampla de estatísticas modernas. R é um conjunto integrado de instalações de software para manipulação de dados, cálculo e visualização gráfica. Fornece uma grande variedade de técnicas estatísticas gráficas (modelagem não-linear testes estatísticos clássicos, análise de séries temporais linear, classificação, *clusterização*, etc).

O WEKA⁵ é um produto da Universidade de Waikato (Nova Zelândia); é uma ferramenta livre, de código aberto (*General Public License*) e possui uma coleção de algoritmos de aprendizado de máquina para as tarefas de mineração de dados. Os algoritmos podem ser aplicados diretamente da ferramenta, ou utilizados por programas Java. Implementa as tarefas de classificação, regressão, *clusterização* e regras de associação. Disponibiliza funcionalidades de pré-processamento e visualização. A visualização gráfica dos dados se dá

³ Disponível em: <http://www.oracle.com/technetwork/database/options/advanced-analytics/odm/overview/index.html>

⁴ Disponível em: www.r-project.org/

⁵ Disponível em: www.cs.waikato.ac.nz/ml/weka/

por meio de histogramas e, a apresentação dos resultados em árvores de decisão e diagramas de dispersão.

No Quadro 8, apresentam-se as tarefas de mineração de dados e as ferramentas de visualização disponíveis em cada uma das ferramentas citadas.

QUADRO 8 - PRINCIPAIS CARACTERÍSTICAS DAS FERRAMENTAS DE MINERAÇÃO DE DADOS

Ferramenta	Tarefas de KDD	Visualização	Empresa/grupo responsável
SAS Enterprise Miner	Classificação, clusterização, regras de Associação, descoberta de padrão sequencial e análise de cesta de mercado.	Histogramas multidimensionais e comparação gráfica dos resultados da modelagem	SAS Inc.
Intelligent Miner	clusterização, classificação, regressão, regras de associação, padrões sequenciais	Diversos tipos de gráficos para análise visual dos resultados	IBM Cor.
Oracle Data Mining	classificação, regressão, detecção de anomalias, regras de associação, clusterização e extração de características	Pode ser integrado ao R ou ao Microsoft Excel e utilizar suas capacidades gráficas	Oracle
Pacote R	Análise de séries temporais linear, classificação, clusterização.	Técnicas estatísticas gráficas	R Foundation
Weka	Classificação, Regressão, clusterização, regras de associação.	Histogramas a apresentação dos resultados em árvores de decisão e diagramas de dispersão	Machine Learning Group at the University of Waikato

FONTE: A AUTORA

2.4.6 APLICAÇÃO DE TÉCNICAS DE MINERAÇÃO DE DADOS EM PRODUTIVIDADE DE CANA-DE-AÇÚCAR

Segundo Chen, Hu e Zhang (2006), obter novos conhecimentos com as técnicas de mineração de dados é relativamente mais fácil do que com a utilização de abordagens estatísticas tradicionais. Os autores argumentam que a mineração de dados é sempre livre de hipóteses, enquanto técnicas estatísticas sempre requerem a definição de uma hipótese. Além disso, algoritmos de mineração de dados produzem um conjunto de tipos de dados muito mais amplos e fazem menos suposições, ou nenhuma suposição, a respeito de sua distribuição.

As técnicas de mineração de dados são apropriadas para processar grandes bases de dados e/ou bases com muitos atributos. Dessa forma, vários dos

trabalhos que utilizam dados de sensoriamento remoto das culturas, fazem uso de técnicas de mineração de dados para a construção de suas aplicações. Podem ser citados, como exemplo, os trabalhos de: Everingham *et al.* (2007), Gonçalves *et al.* (2011), Fernandes, Rocha e Lamparelli (2011), Romani *et al.* (2008), Romani *et al.* (2011), Vieira *et al.* (2012), Romani *et al.* (2013), Nonato e Oliveira (2013) e Vintrou *et al.* (2013).

A mineração de dados pode ser realizada por meio de diferentes tarefas e técnicas. Apesar das diversas possibilidades, observa-se nas pesquisas a recorrência das técnicas de classificação, *clusterização* e associação.

O trabalho de Romani *et al.* (2008) fez uso de regras de associação, enquanto que um novo algoritmo baseado em regras de associação foi proposto no trabalho de Romani *et al.* (2013). As regras de associação foram utilizadas também no desenvolvimento de um sistema de recomendação para conteúdos relacionados à cultura da cana-de-açúcar em Barros, Oliveira e Oliveira (2013). Os autores usaram dados de navegação do usuário em páginas *Web* para a aplicação das regras de associação. Vale ressaltar que esse é um uso típico dessa técnica de mineração de dados.

A tarefa de *clusterização*, com os algoritmos *K-means* e *K-medoids*, foi utilizada no trabalho de Romani *et al.* (2011) e Gonçalves *et al.* (2011), que fizeram uso do algoritmo de *clusterização K-means*. Um algoritmo de *clusterização*, baseado no comportamento das abelhas (*bee hive*), denominado CRY, foi apresentado e comparado com outros algoritmos no trabalho de Ananthara, Arunkumar e Hemavathy (2013).

Uma das tarefas de mineração de dados mais importantes e populares é a classificação (GOLDSCHMIDT e PASSOS, 2005). Vintrou *et al.* (2013) apresentaram um algoritmo de classificação original baseado em padrões seqüenciais. O trabalho de Everingham *et al.* (2007) utilizou as técnicas de classificação *Support Vector Machines* (SVM) e *Random Forest* (RF). As pesquisas de Nascimento *et al.* (2009), Souza *et al.* (2010), Fernandes, Rocha e Lamparelli (2011), Vieira *et al.* (2012) e Nonato e Oliveira (2013) também utilizaram a tarefa de classificação, mas a técnica utilizada nesses trabalhos foi

a indução por árvore de decisão, que é uma das principais técnicas de mineração de dados pela sua expressividade simbólica, conforme citado anteriormente. Bocca (2014), para o desenvolvimento do seu modelo de produtividade, comparou o desempenho de técnicas consideradas mais clássicas (árvore de regressão, redes neurais e regressão múltipla) e técnicas com potencial de melhor desempenho, dado seu desempenho em outros campos de aplicação (SVM, RF e árvore de modelos).

Goldschmidt e Passos (2005), ao descreverem os tipos de tarefas de mineração de dados, fazem referência à tarefa composta denominada “*Clusterização→Classificação*”. Para esses autores essa tarefa é aplicável quando as instâncias de dados não possuam classes pré-definidas. Desta forma os dados são agrupados em função de sua similaridade por meio de uma técnica de *clusterização*, em seguida um novo atributo é incluído ao conjunto de dados indicando a qual *cluster* cada instância pertence. Esse novo atributo passa a ser a classe dessas instâncias e, assim, as técnicas de classificação podem ser utilizadas para gerar modelos de conhecimento que possam prever a classificação de novas instâncias. Essa combinação de tarefas é utilizada nos trabalhos de Ferraro, Rivero e Ghersa (2009) e Ferraro, Ghersa e Rivero (2012), que fizeram uso da técnica de *clusterização k-means* e do algoritmo CART para indução de árvore de decisão.

No Quadro 9 são apresentados os trabalhos associados às tarefas de mineração de dados que eles utilizaram, destaca-se também a que categoria de produtividade foram relacionados. É possível notar que nenhum dos trabalhos da categoria 1 (programação da colheita) utilizou mineração de dados. Ressalta-se ainda que a maioria dos trabalhos (6) foram associados à categoria 3 (caracterização de áreas), seguidos por 5 trabalhos na categoria 2 (previsão de produtividade), 3 trabalhos na categoria 4 (apoio às práticas de gestão) e 2 trabalhos na categoria 5 (outros).

QUADRO 9 - TAREFAS DE MINERAÇÃO DE DADOS UTILIZADAS NOS TRABALHOS REVISADOS

Tarefa de Mineração de Dados	Categoria	Autor/ Ano da Publicação
Classificação	Categoria 2	Fernandes, Rocha e Lamparelli (2011), Bocca (2014)
	Categoria 3	Nascimento et al. (2009), Vieira et al., (2012), Nonato e Oliveira (2013), Vintro et al. (2013)
	Categoria 4	Everingham et al. (2007), Souza et al. (2010)
Clusterização	Categoria 2	Ananthara, Arunkumar e Hemavathy (2013)
	Categoria 3	Romani et al. (2011), Gonçalves et al. (2011)
Associação	Categoria 2	Romani et al. (2008)
	Categoria 4	Romani et al. (2013)
	Categoria 5	Barros, Oliveira e Oliveira (2013)
Clusterização→Classificação	Categoria 2	Ferraro, Rivero e Ghersa (2009), Ferraro e Ghersa e Rivero (2012)
	Categoria 5	

FONTE – A AUTORA

Conforme citado na Seção 2.4., uma das etapas do processo de KDD é o pós-processamento. Nessa etapa são realizadas atividades de avaliação para identificar os padrões interessantes, e os resultados sobre o conhecimento minerado devem ser apresentados ao usuário. Para que esse conhecimento possa, de fato, dar suporte à tomada de decisão, os resultados da mineração devem ser apresentados buscando facilitar a leitura e interpretação dos dados; ou seja, o grande conjunto de dados deve ser apresentado por meio de visualizações sumarizadas. A maioria das pesquisas revisadas e apresentadas, neste trabalho, faz a avaliação do conhecimento obtido na etapa de mineração de dados, mas não desenvolvem, ou utilizam, uma ferramenta que sintetize esse processo com foco em abreviar as tarefas dos gestores durante a tomada de decisão.

O desenvolvimento de um novo sistema para apoio à decisão resultante do processo de mineração de dados é realizado em Romani et al. (2011), por meio da visualização geoespacial de *clusters* formados nesse processo. A visualização espacial dos *clusters* utilizada em Gonçalves et al. (2011) é realizada por um pacote de software já existente, denominado ARCGIS. Um sistema de informação de sensoriamento remoto é utilizado para apresentar regras de associação em Romani et al. (2013) e Barros, Oliveira e Oliveira

(2013), que utilizam regras de associação no desenvolvimento de um sistema de recomendação.

No contexto analisado, isto é, modelos de produtividade de cana-de-açúcar, não foram encontradas ferramentas que apresentassem os resultados obtidos a partir da execução dos algoritmos de árvore de decisão com interfaces capazes de reduzir a complexidade inerente a esses modelos e, assim, subsidiar os gestores de maneira mais eficaz.

3 ABORDAGEM METODOLÓGICA

No processo de definição da **abordagem geral** desta pesquisa, optou-se pela articulação entre os campos das pesquisas **quantitativas e qualitativas**. Essa opção resulta de uma indicação de Minayo (1993), quando afirma que nenhuma das duas abordagens, por si só, é suficiente para a compreensão completa da realidade.

Para atender às especificidades dos **objetivos** propostos, dois tipos de abordagens aconteceram: exploratória e explicativa. A **pesquisa exploratória** deu-se pela pesquisa bibliográfica operacionalizada por uma Revisão Sistemática da Literatura (RSL). A **pesquisa explicativa** caracteriza-se pela busca da identificação de fatores que determinam ou contribuem para a ocorrência de fenômenos. Assim, tem como desafio o aprofundamento do conhecimento da realidade, implicando uma abordagem mais complexa (ANDRADE, 2002; GIL, 2007). No contexto da pesquisa explicativa, foi utilizada a **pesquisa experimental**, que implica em determinar um objeto de estudo, selecionar as variáveis que seriam capazes de influenciá-lo, definir as formas de controle e de observação dos efeitos que a variável produz no objeto (GIL, 2007; TRIVIÑOS, 1987).

Em relação aos procedimentos técnicos seguidos neste estudo experimental, a opção foi pela **Modelagem** que, de acordo com Berto e Nakano (2000), comprehende um processo que faz uso de técnicas matemáticas para descrever o funcionamento de um sistema ou parte de um sistema produtivo.

A Figura 4 representa esquematicamente os procedimentos metodológicos adotados.

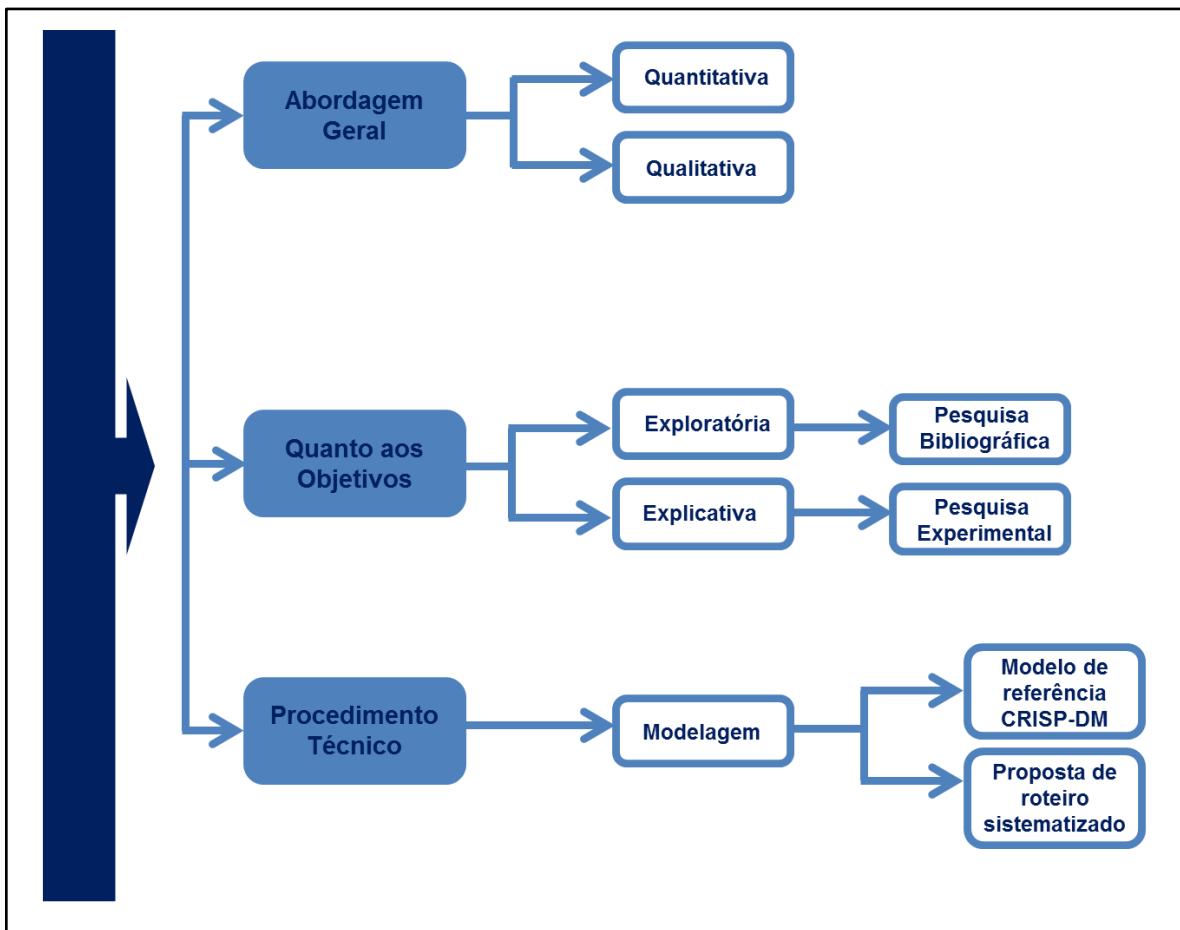


FIGURA 4 - PROCEDIMENTOS METODOLÓGICOS A SEREM ADOTADOS NESTE TRABALHO

FONTE: A AUTORA

Na seção 3.1 apresenta-se a descrição da RSL, em seguida, na seção 3.2, apresenta-se brevemente a pesquisa experimental, relatada em detalhes nos Capítulos 4 e 5.

3.1 DESCRIÇÃO DA REVISÃO SISTEMÁTICA DA LITERATURA (RSL)

Uma Revisão Sistemática da Literatura (RSL) é um meio de identificar, avaliar e interpretar todo material disponível e relevante sobre uma questão de pesquisa, um tópico ou um fenômeno de interesse. Deve ser conduzida de maneira formal, seguindo etapas bem definidas, de acordo com um protocolo previamente elaborado (BIOLOCHINI, 2005; KITCHENHAM e CHARTERS, 2007; STAPLES e NIAZI, 2008; VICENTE, DELAMARO e MALDONADO, 2009; GUESSI, OLIVEIRA e NAKAGAWA, 2011)

Uma revisão sistemática é composta por três fases principais:

- 1) **Planejamento** da revisão: definição dos objetivos da pesquisa e de um protocolo de revisão, que inclui as questões da pesquisa que será conduzida e os métodos que serão utilizados para executar a revisão;
- 2) **Condução** da revisão: identificação e avaliação dos estudos primários, segundo os critérios de inclusão e exclusão definidos, e seleção dos estudos;
- 3) **Análise** dos resultados: extração e síntese dos resultados.

Na Figura 5, é apresentado o fluxo de atividades realizadas na RSL.

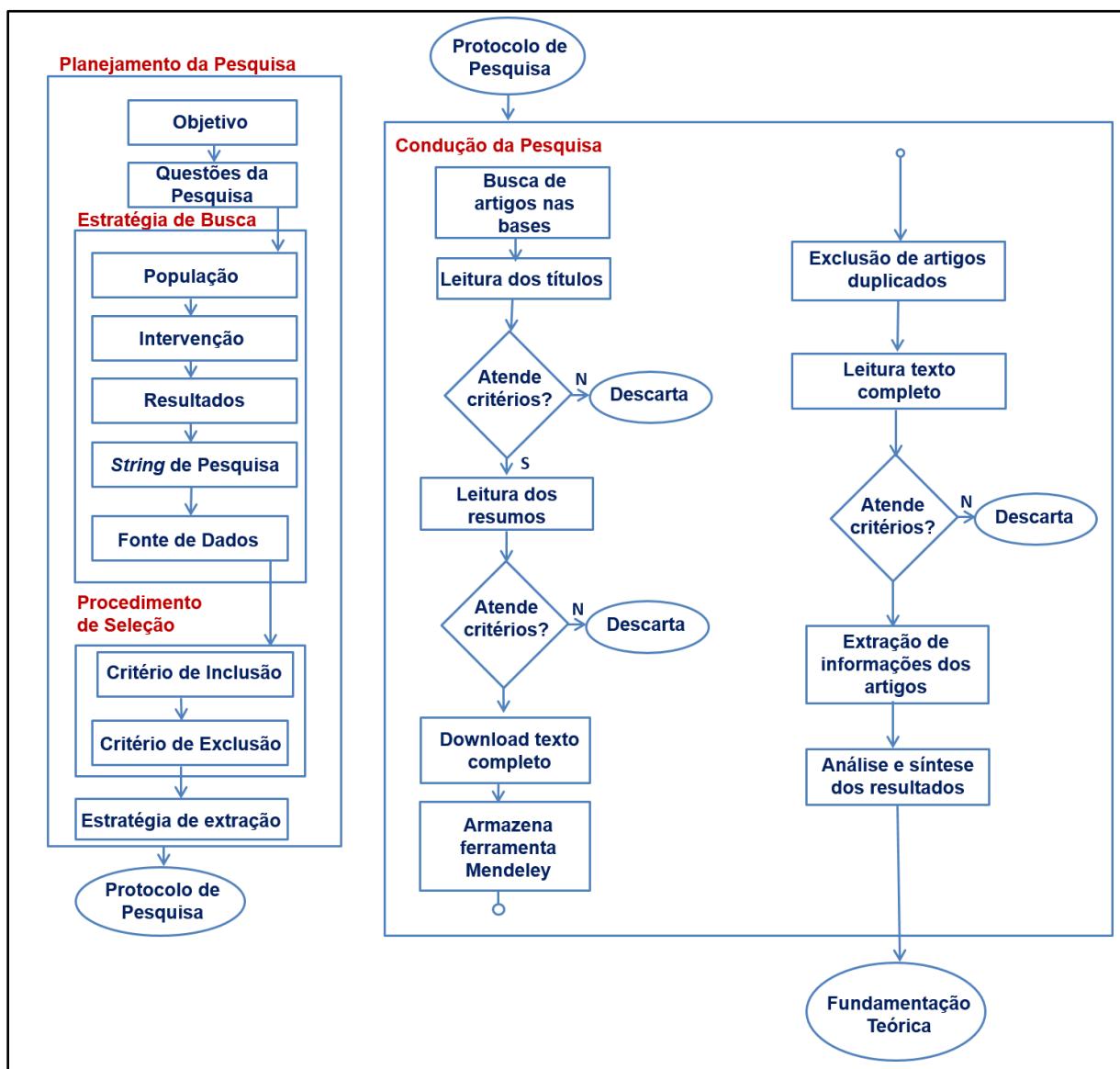


FIGURA 5 - FLUXO DE ATIVIDADES DA RSL

FONTE: A AUTORA

As atividades de planejamento da RSL que foram realizadas para a elaboração deste trabalho são descritas nas seções seguintes. A condução da RSL e análise dos resultados foram apresentadas no Capítulo 2, Fundamentação Teórica.

3.1.1 PLANEJAMENTO DA REVISÃO

Na fase de planejamento, foram definidos inicialmente os objetivos e as questões da pesquisa que norteariam a RSL. Em seguida, foram definidos a estratégia da busca, as fontes de dados e os procedimentos para a seleção dos artigos.

Assim, o **objetivo** definido para a pesquisa foi: avaliar a evolução das pesquisas que utilizam mineração de dados para otimizar a produtividade da cana-de-açúcar visando a identificar lacunas e tendências na área.

Para definir claramente o estado da arte das pesquisas na área alvo, ou seja, produtividade de cana-de-açúcar, quatro **questões** foram propostas:

1. Quais **estratégias e técnicas** estão sendo utilizadas para contribuir com o aumento da produtividade de cana-de-açúcar?
2. Quais os **principais atributos** que estão sendo utilizados para o aumento da produtividade da cana-de-açúcar?
3. Como os **modelos** de produtividade podem apoiar as atividades de gestão da cana-de-açúcar?
4. Quais as **características** (tarefas, técnicas e ferramentas) dos trabalhos relacionados com produtividade de cana-de-açúcar que utilizam mineração de dados?

Para a estratégia de busca, foram definidas: 1) a população, isto é, conjunto de objetos que se tem interesse porque estão relacionados às intervenções que serão avaliadas. 2) a intervenção, ou seja, a característica da população que se quer estudar e; 3) o resultado esperado dessa busca.

Para a **população**, foram selecionadas técnicas utilizadas para a construção de modelo de produtividade de cana-de-açúcar. As palavras-chave escolhidas fazem referência a técnicas de mineração de dados, a modelos estatísticos e matemáticos. Considerou-se como **intervenção** o tema produtividade de cana-de-açúcar; as palavras-chave desse item são sinônimos do termo “produtividade

de cana-de-açúcar". Os **resultados** esperados foram os modelos de produtividade de cana-de-açúcar. A partir da definição dos temas da pesquisa a seguinte *string* foi elaborada: (*data mining OR classification technique OR association rules OR clustering OR statistical method OR mathematical method*) *AND* (*sugarcane productivity OR sugarcane yield*).

As **fontes de dados** escolhidas para a realização da busca foram as seguintes bases de dados eletrônicas: *Web of Science*, *Scopus*, *Environmental Engineering*, *ScienceDirect*, *Emerald*, *Compendex*, *IEEE Xplore*. Vale destacar que as diferentes bases têm modos diferentes para a inserção das palavras-chave e execução da busca; dessa forma, algumas pequenas adaptações foram realizadas na *string* de pesquisa, de acordo com as regras de cada motor de busca. Foram selecionados artigos publicados em periódicos (*journals*) ou em anais de congressos (*proceedings*), no período de 2005 a 2015, redigidos em inglês ou português. Foram também definidos os seguintes critérios de inclusão e exclusão.

- **Inclusão:**

- ✓ Artigos que tratem sobre modelos de produtividade de cana-de-açúcar;
- ✓ Artigos que tratem sobre cana-de-açúcar e mineração de dados.

- **Exclusão:**

- ✓ Artigos que não tenham texto completo disponível;
- ✓ Editoriais, prefácios, resumos de artigos, entrevistas, notícias, análise (avaliações), correspondência, debates, comentários, cartas de leitores, resumos de tutoriais, *workshops*, painéis e sessões de pôsteres;

Na **estratégia de extração de informações**, foram selecionadas as características que identificassem os artigos e que pudessem responder às questões de pesquisa: ano da publicação, título do artigo, base de dados, técnicas e atributos utilizados para a elaboração do modelo, ferramentas computacionais, volume de dados processados, metodologia do estudo, objetivos, aspectos relevantes e contribuições do artigo.

O Quadro 10 resume o protocolo de pesquisa.

QUADRO 10 - PROTOCOLO DE PESQUISA

Objetivo:	Avaliar a evolução das pesquisas que utilizam mineração de dados para aumentar a produtividade da cana-de-açúcar visando a identificar lacunas e tendências na área.
Questões:	<ol style="list-style-type: none"> Quais estratégias e técnicas estão sendo utilizadas para contribuir com o aumento da produtividade de cana-de-açúcar? Quais os principais atributos que estão sendo utilizados para o aumento da produtividade da cana-de-açúcar? Como os modelos de produtividade podem apoiar as atividades de gestão da cana-de-açúcar? Quais as características (tarefas, técnicas e ferramentas) dos trabalhos relacionados com produtividade de cana-de-açúcar que utilizam mineração de dados?
Fontes de dados:	<i>Web of Science</i> <i>Scopus</i> <i>Environmental Engineering</i> <i>ScienceDirect</i> <i>Emerald</i> <i>Compendex</i> <i>IEEE Xplore</i>
Estratégia de busca:	
• População	<i>data mining; Classification technique; Association rules; Clustering; statistical method; mathematical method</i>
• Intervenção	<i>sugarcane productivity ; sugarcane yield</i>
• Resultados	Modelos de produtividade de cana-de-açúcar
• String de pesquisa	(<i>data mining or Classification technique or Association rules or Clustering or statistical method or mathematical method</i>) and (<i>sugarcane productivity or sugarcane yield</i>)
Critérios de inclusão	Artigos que tratem sobre modelos de produtividade de cana-de-açúcar. Artigos que tratem sobre cana-de-açúcar e mineração de dados.
Critérios de exclusão	Artigos que não tenham texto completo disponível, editoriais, prefácios, resumos de artigos, entrevistas, notícias, análise (avaliações), correspondência, debates, comentários, cartas de leitores, resumos de tutoriais, workshops, painéis, e sessões de pôsteres.
Período da pesquisa	2005 a 2015
Tipo de documento	Artigos de <i>journals</i> e <i>proceedings</i>
Idioma	Inglês

FONTE: A AUTORA

3.1.2 CONDUÇÃO DA RSL

A condução da RSL teve início com a busca de artigos nas bases selecionadas, de acordo com os parâmetros definidos na fase de planejamento. Para cada base, após inclusão dos critérios de busca, efetuou-se a leitura dos títulos dos artigos resultantes da pesquisa. Para os títulos aprovados, foi feita a leitura de seus resumos. Se o resumo se enquadrasse nos critérios de inclusão, era

realizado *download* do artigo completo. Para as bases que possuíam apenas os resumos, os trabalhos completos foram procurados em outras bases.

Os artigos selecionados foram armazenados na ferramenta *Mendeley*, que tem por objetivo gerenciar publicações, separados por pastas, uma para cada base de dados consultada. Nessa fase, foram verificados e retirados os artigos duplicados. Para os artigos resultantes, as informações definidas no planejamento da RSL foram extraídas e organizadas em planilha eletrônica.

Após a leitura do texto completo dos artigos, outras exclusões foram realizadas. Na Tabela 3, é apresentada a quantidade de artigos resultantes de cada fase, em cada base. As bases estão listadas de acordo com a ordem em que a pesquisa foi realizada.

TABELA 3- RESULTADO QUANTITATIVO DA CONSULTA ÀS BASES

Base	Data da consulta	Resultados	Seleção		
			Por título	Após leitura dos resumos	Após leitura dos artigos
<i>Science direct</i>	21/09/2015	555	24	18	10
<i>IEEE Xplore</i>	21/09/2015	70	15	15	12
			16		
<i>Web of Science</i>	24/09/2015	82	(5 - não encontrado texto completo)	8	6
<i>Scopus</i>	24/09/2015	8	5	Só duplicados	0
<i>Environmental Engineering</i>	25/09/2015	30	0	0	0
<i>Emerald</i>	25/09/2015	0	0	0	0
			4		
<i>Compendex</i>	25/09/2015	37	(1 - não encontrado texto completo)	3	3
Sistema de Bibliotecas da Unicamp	17/10/2015	5	3	1	1
Total		787	67	45	32

FONTE: A AUTORA

Os 32 artigos selecionados para análise foram agrupados em uma única pasta na ferramenta *Mendeley*, a fim de facilitar sua manipulação.

3.2 PROCEDIMENTO TÉCNICO: MODELAGEM

O procedimento técnico em relação à modelagem tem dois momentos: um passa pelas etapas do modelo de referência CRISP-DM, apresentado na Seção 2.4.1, e o outro chega na proposta de roteiro sistematizado. A Figura 6 representa o fluxo das tarefas executadas nessas etapas.

Apresentam-se, no Capítulo 4, as atividades que foram realizadas em cada uma das seis etapas do modelo CRISP-DM e, no Capítulo 5, o roteiro sistematizado e a documentação do desenvolvimento de uma ferramenta para visualização dos resultados obtidos no processo de mineração de dados.

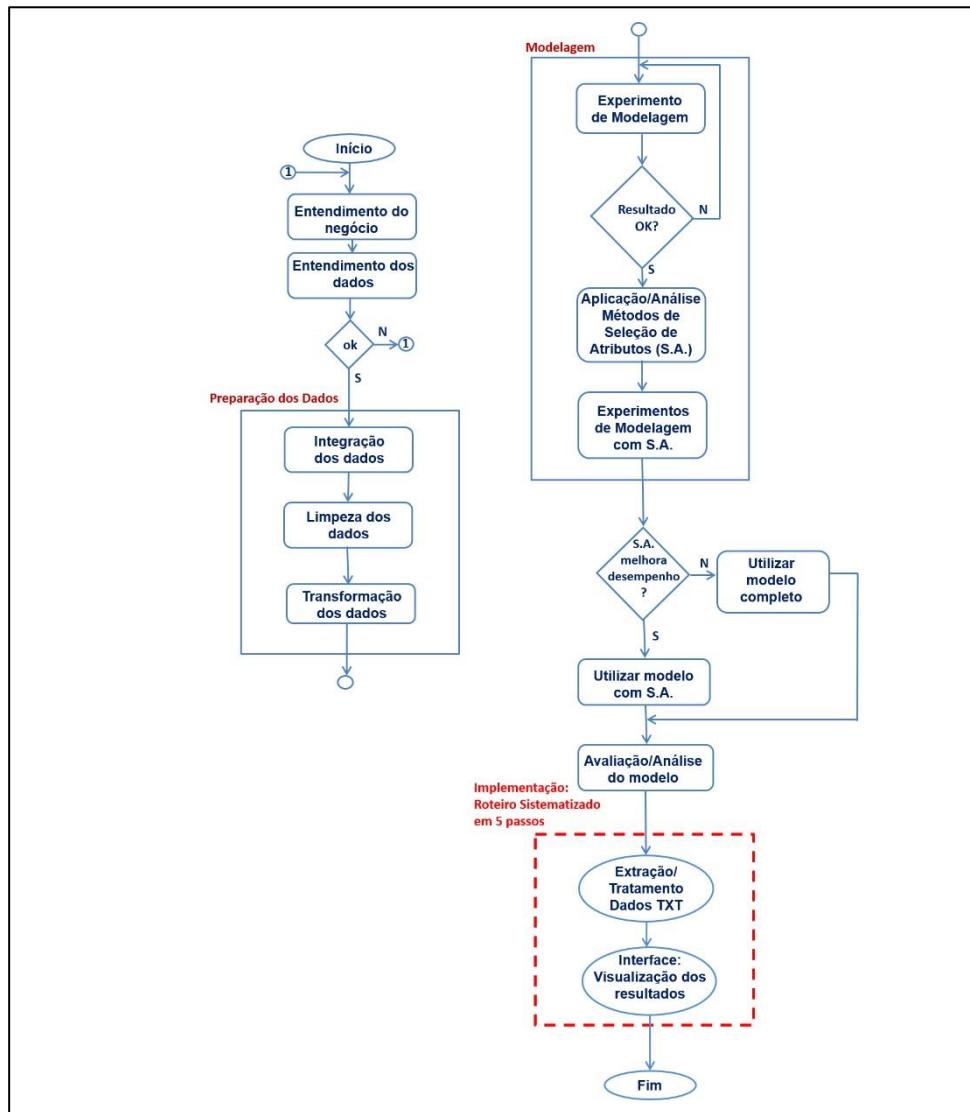


FIGURA 6 - FLUXO DAS ATIVIDADES EXECUTADAS DE ACORDO COM O MODELO CRISP-DM

FONTE: A AUTORA

4 DETALHAMENTO DO PROCEDIMENTO TÉCNICO: MODELAGEM – CRISP-DM

O modelo CRISP-DM, conforme citado na Seção 2.4.1, é composto por seis etapas: Entendimento do Negócio, Entendimento dos dados, Preparação dos dados, Modelagem, Avaliação e Implementação. Assim, para atender ao primeiro objetivo específico deste trabalho, nas seções seguintes são descritos os procedimentos realizados em cada uma dessas etapas.

4.1 ENTENDIMENTO DO NEGÓCIO

Neste trabalho, são utilizados os dados do censo varietal qualitativo referentes à cana-de-açúcar – 4 safras, 2006/2007 a 2009/2010, cedidos por um dos maiores grupos sucroenergéticos do Brasil, segundo a UNICA (União da Indústria de Cana-de-Açúcar) (UNICA, 2016), sediado no interior do Estado de São Paulo. O Grupo possui quatro usinas em operação: duas delas produzem açúcar e etanol; uma é dedicada à produção exclusiva de etanol; uma à produção de derivados de levedura. As usinas geram também energia elétrica a partir da queima do bagaço da cana (cogeração), garantindo autossuficiência e venda do excedente.

Segundo informações do site da empresa, o índice médio de mecanização da colheita do grupo é de 94%, chegando a 100% em uma das usinas, índices considerados como referência no setor. A empresa compra, cultiva, colhe e processa a principal matéria-prima usada na produção de açúcar e álcool. Na safra 2016/2017, foram processadas um total de 19,281 milhões de toneladas de cana que resultaram em 1.301 toneladas de açúcar e 667 mil m³ de etanol.

Manter-se em posição de destaque nesse setor requer utilização contínua de técnicas, tecnologias e ferramentas que deem suporte ao aumento da produção e/ou redução dos custos. Assim, foram realizadas reuniões com colaboradores do setor de qualidade da empresa para discutir como as técnicas de mineração de dados poderiam ser aplicadas nos dados da produção agrícola, de forma que

os diferentes cenários de produção pudessem ser investigados para a obtenção de melhor produtividade.

A exemplo do trabalho de Bocca (2014), foi realizada entrevista semiestruturada, adaptando-se a ferramenta 5W1H, para entendimento do funcionamento da área agrícola da empresa, por meio da caracterização das tomadas de decisões e atividades de planejamento. Assim, para cada atividade de planejamento da área agrícola foi associada a área funcional responsável pela elaboração do plano, quais decisões estão associadas a esse plano, a que nível hierárquico pertencem, quando devem ser tomadas e que ferramenta computacional dá suporte às atividades de planejamento. A Figura 7 representa os itens utilizados para nortear a entrevista que foi realizada com um profissional, suporte técnico, da área denominada Qualidade Agrícola.

Plano What	Área funcional Where	Decisões Why	Nível Organizacional Who	Quando When	Ferramenta how
-----------------------	-------------------------------------	-------------------------	-----------------------------------------	------------------------	---------------------------

FIGURA 7 - ITENS UTILIZADOS NA ENTREVISTA SEMIESTRUTURADA

FONTE: A AUTORA

4.1.1 DESCRIÇÃO DA UNIDADE EM ANÁLISE

Nesta seção apresenta-se, inicialmente, as características da área da Qualidade Agrícola, responsável pela produção agrícola. Em seguida são apresentadas as atividades referentes ao planejamento de plantio, colheita e tratos culturais.

A Qualidade Agrícola é suporte para todas as áreas de operação: planejamento de plantio, planejamento de colheita e tratos culturais. Faz parte da Qualidade Agrícola a área de controle de pragas, que possui uma equipe com esse conhecimento, que realiza o levantamento das principais e respectivo controle. Também compõe a Qualidade Agrícola a equipe de auditoria, cujo objetivo é auditar todos os processos para verificar se os indicadores de desempenho foram cumpridos, atuando diretamente na qualidade da operação.

Ligado à Qualidade Agrícola estão também os Serviços Agrícolas. Para uma área de aproximadamente 68 mil hectares existem 8 setores. Cada setor possui um administrador, responsável por manter a qualidade agrícola desse setor, e

uma pequena equipe para cuidar de plantas daninhas, dos carreadores, da limpeza de carreadores, e realizar algumas atividades de melhoria na área. Os administradores são responsáveis ainda por fazer estimativas de produtividade do seu setor.

Também dá suporte à Qualidade Agrícola a Manutenção Agrícola, responsável pela manutenção dos equipamentos. Influencia na qualidade da operação em relação à disponibilidade e rendimento dos equipamentos.

É importante ressaltar que as práticas agrícolas são todas interligadas. Assim, embora o planejamento tático do plantio, colheita e tratos culturais sejam responsabilidade da Qualidade Agrícola, esses planos são sempre realizados em conjunto com as equipes operacionais, para que se possa obter melhores resultados em termos de produtividade da cana-de-açúcar.

A Figura 8 representa a estrutura organizacional da área agrícola. Em seguida são apresentadas as principais características do planejamento do plantio, colheita e tratos culturais.

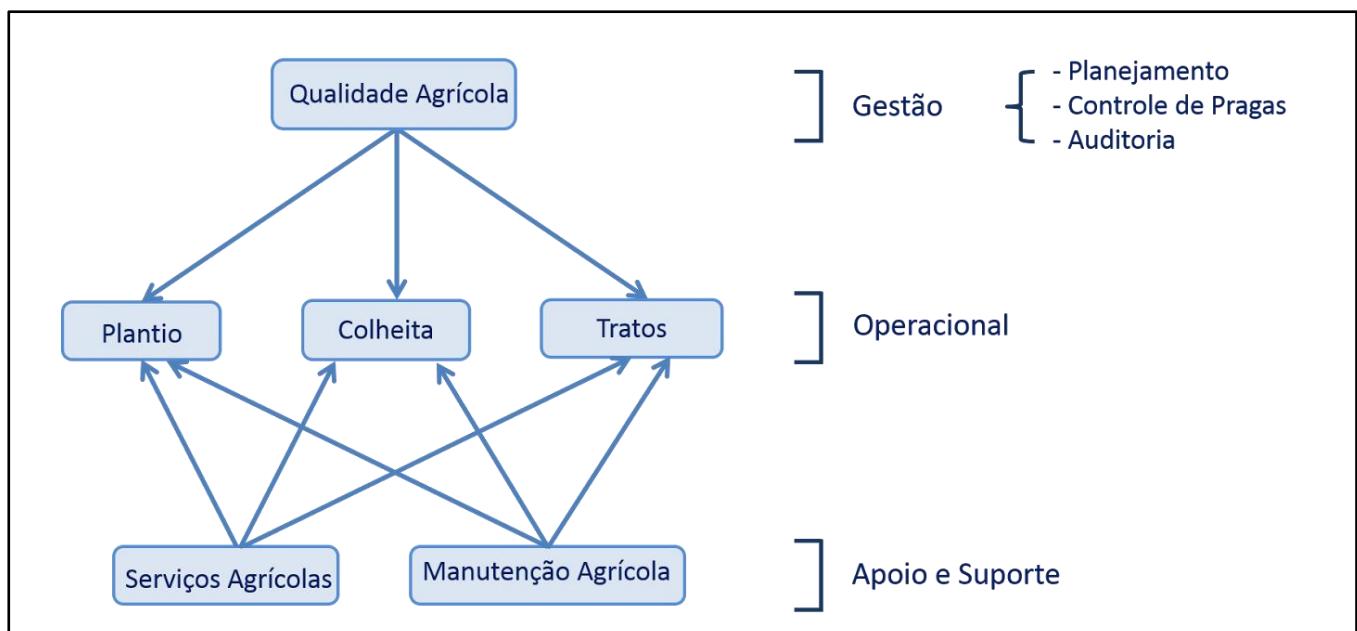


FIGURA 8 - ORGANOGRAMA DA ÁREA QUALIDADE AGRÍCOLA

FONTE: A AUTORA

Planejamento do plantio

Uma vez que a produtividade da cana decresce a cada corte, periodicamente é necessário replantar os canaviais. A decisão sobre a renovação das áreas é de responsabilidade da Qualidade Agrícola que elabora o planejamento de plantio, envolvendo as equipes operacionais de plantio e de tratos e tendo por base o planejamento da colheita.

De acordo com as características das áreas que serão liberadas e da época da colheita, são definidos os tipos de cana (cana de ano, inverno ou de 18 meses) e as variedades (precoce, média ou tardia) que serão plantadas.

Salienta-se, por exemplo, que o plantio da cana de ano não pode ser realizado em solos restritivos, apenas em solos com alta fertilidade. A cana de inverno requer áreas que têm possibilidade de irrigação, para suprir a necessidade hídrica da cana para a brotação. A cana de 18 meses deve ser plantada em áreas que serão colhidas mais tarde, áreas de plantio de rotação de cultura ou MEIOSI⁶.

Destaca-se também que solos de baixa fertilidade devem ter suas colheitas realizadas no início da safra. Assim, para solos desse tipo, devem-se plantar variedades de canas precoces, que são colhidas no início da safra, de forma a se obter maior ganho do potencial da área.

Para subsidiar o planejamento do plantio a Usina utiliza planilha eletrônica, contendo as características das áreas, identificação da fazenda, gleba e quadra, ambiente de produção e época de colheita.

O sistema denominado IREF, da empresa *Hexagon Agriculture*⁷, fornece subsídio para o planejamento de reforma ideal, a partir da curva de produtividade atual e a quanto se deseja obter de aumento a longo prazo. O sistema utiliza

⁶ Sigla que significa Método Inter-rotacional Ocorrendo Simultaneamente. O método consiste em intercalar culturas de interesse econômico e/ou agronômico com o canavial para reduzir custos de implantação, melhorar o sistema de logística e promover a melhora do local de cultivo.

Disponível em: <http://portalklff.com.br/publicacao/oldlink-1135>.

⁷ Disponível em: <http://www.hexagonagriculture.com/pt-Br>

simulação para sugerir a porcentagem de área de reforma e em quais ambientes de produção.

Planejamento de colheita

O planejamento da colheita visa a obter maior rentabilidade possível no que diz respeito à quantidade e qualidade da cana, bem como menores custos da operação. Para a elaboração desse planejamento são envolvidas, além da qualidade agrícola, a equipe operacional de plantio e de tratos.

A área de colheita tem em média um raio de 32km, com blocos de cana precoce, média e tardia. É necessário definir quais áreas, quando e que variedades serão colhidas, sempre com base na logística de colheita. Além disso, as áreas de reforma devem ser liberadas de forma escalonadas e distribuídas, de acordo com a capacidade de preparo da área, a ser realizada pela equipe de tratos.

Um sistema denominado ICOL, da empresa *Hexagon*, é utilizado para a elaboração do planejamento da colheita. O sistema recebe as informações e as restrições para a colheita e gera um plano que sofrerá um ajuste fino pelos profissionais responsáveis por esse planejamento.

Planejamento de tratos

Os tratos culturais têm como objetivos prover água e nutrientes, controlar pragas e plantas infestantes e nivelar as imperfeições da sulcação pós plantio.

Para a elaboração do plano de tratos leva-se em consideração tudo o que será colhido, em que época, qual o estágio de corte e o ambiente de produção. A partir desses dados é gerada a recomendação de adubação.

Nessa recomendação, é proposto o tipo de adubo, se mineral ou composto, dependendo do ambiente de produção dos blocos que serão tratados. Solos mais restritivos, por exemplo, precisam de adubação mais pesada para obter maior produtividade, diferentemente dos solos com mais alta fertilidade. A época da colheita impacta no tipo da adubação em decorrência das diferentes condições climáticas nas diferentes épocas de corte. Define-se ainda sobre a irrigação, aplicação de vinhaça e torta de filtro nesses blocos.

Por fim, o Quadro 11 resume as características da área da Qualidade Agrícola.

QUADRO 11- SÍNTSE DA CARACTERÍSTICAS DA ÁREA DE QUALIDADE AGRÍCOLA

Plano	Área funcional	Decisões	Nível Organizacional	Quando	Ferramenta
Plantio	Qualidade agrícola/ plantio – operação	<ul style="list-style-type: none"> • Área a ser plantada, • Quando será plantada, • Que tipo de cana • Qual variedade 	Tático/operacional	Planejamento de reforma – para os próximos 5 anos Plantio – de novembro a fevereiro, para a próxima safra	Planilha eletrônica IREF – Simulação de reforma (longo prazo)
Colheita	Qualidade agrícola/ colheita – operação	<ul style="list-style-type: none"> • Definir blocos de colheita <ul style="list-style-type: none"> ◦ Que área, ◦ Quando, ◦ Que variedade será colhida • Definir logística de colheita 	Tático/operacional	De novembro a fevereiro, para a próxima safra	ICOL
Tratos Culturais	Qualidade agrícola/ tratos culturais – operação	Definir recomendação de adubação de acordo com: <ul style="list-style-type: none"> • Época de colheita • Estágio de corte • Ambiente de produção. 	Tático/operacional	De novembro a fevereiro, para a próxima safra	

FONTE: A AUTORA

4.2 ENTENDIMENTO DOS DADOS

Conforme apresentado na Seção 2.3.1, diversos atributos, combinados de diferentes maneiras foram utilizados nos modelos de produtividade da cana-de-açúcar. Assim, visando a sistematizar o processo de seleção dos atributos para a elaboração de um modelo de produtividade, foi utilizado o método Delphi (DALKEY; HELMER, 1963 apud KAYO; SECURATO, 1997), conforme descrito no Apêndice A. O conjunto inicial de atributos foi definido com base naqueles utilizados em Ferraro, Rivero e Ghersa (2009) e Bocca (2014), que possuem características semelhantes a esta pesquisa.

Quatro especialistas participaram do processo de seleção, que foi executado em duas rodadas, nas quais esses especialistas foram solicitados a designar notas de zero a dez aos atributos que poderiam fazer parte de um modelo de produtividade de cana-de-açúcar. Na primeira rodada do método Delphi foram mantidos os atributos com frequência maior que dois em notas maiores ou iguais a sete. Ou seja, após a primeira rodada foram desconsiderados os atributos com nota atribuída inferior a sete.

Na segunda rodada, solicitou-se a atribuição de notas maiores ou iguais a oito para os atributos efetivamente relevantes para a produtividade da cana. Os atributos selecionados para elaboração de um modelo de produtividade de cana-de açúcar, resultantes da aplicação do método Delphi, são apresentados no Quadro 12.

QUADRO 12 - ATRIBUTOS SELECIONADOS COM O MÉTODO DELPHI

Atributos de solo	
Tipo do solo	
Fertilidade	
Textura	
Atributos de manejo	
Número de cortes	
Insumos	
Nitrogênio	
Potássio	
Fósforo	
Variedade	
Ambiente de manejo	
Atributos Climáticos	
Média das Temperaturas relacionadas a cada fase* do período de desenvolvimento (brotação, perfilhamento, crescimento e maturação)	
Precipitação Acumulada por fase*	
Sistema de produção	
Tipo de preparo do Solo	
Plantio Mecanizado	
Tratos Culturais	
Colheita Mecanizada	
Manejo de Biomassa	

FONTE: DA AUTORA

4.2.1 CARACTERÍSTICAS DOS DADOS DISPONIBILIZADOS

Os dados foram enviados pela empresa, descrita na Seção 4.1, por meio de planilhas eletrônicas, uma para cada ano-safra, referentes a quatro safras distintas, 2006/2007 a 2009/2010. Alguns dos atributos selecionados não puderam ser disponibilizados e, portanto, não foi possível sua utilizá-los: tipo de preparo do solo, plantio mecanizado e manejo da biomassa.

As planilhas, com os dados da produção, denominadas censo varietal, contêm, em cada instância, os seguintes atributos: Código da Fazenda, Código da Gleba (bloco), Código do Talhão, Tipo de Solo, Variedade da Cana, Datas, Estágio de Corte, Tipo de Corte, Condição de Corte, Fórmula do Adubo, Adubação, Ambiente de Produção, Fertilidade, Textura e Produtividade. Na Tabela 4, é apresentada a quantidade de linhas (instâncias) de cada planilha.

TABELA 4 - QUANTIDADE DE LINHAS DAS PLANILHAS

Planilha	Quantidade de linhas
Safra 2006/2007	6715
Safra 2007/2008	6630
Safra 2008/2009	7733
Safra 2009/2010	6735
Total	27813

FONTE: A AUTORA

Para realização do planejamento no setor sucroenergético, é necessário estimar a produtividade da cana-de-açúcar dos talhões que fornecerão matéria-prima para a unidade industrial. A identificação de cada talhão é feita a partir dos atributos: **Código da Fazenda; Código da Gleba; Código do Talhão.**

O **atributo Solo** contém o código referente à classificação do tipo do solo, de acordo com a classificação brasileira. A classificação traz informação do solo em vários níveis: o primeiro diz respeito à classe do solo, de acordo com a morfologia (latossolo, argissolo, etc); o segundo considera as cores no horizonte B (horizontes são camadas mais ou menos paralelas à superfície do terreno, diferenciadas pela cor, textura e estrutura); o terceiro considera as condições químicas do horizonte subsuperficial (eutrófico, distrófico, etc.). Detalhes dessa tipificação podem ser encontrados em Prado *et al.* (2008). A base de dados utilizada neste trabalho contém 39 tipos de solos distintos (vide Apêndice B).

O **atributo Variedade** diz respeito ao cultivar da cana-de-açúcar. São plantados 82 diferentes cultivares (vide Apêndice B).

As **datas** são divididas em: **Plantio, Primeiro Corte, Corte Anterior e Corte Atual.** Essas datas são necessárias para o cálculo da média da temperatura e precipitação acumulada por fase de crescimento da cana.

O **Estágio de Corte** é representado por um número que registra, na maioria das instâncias, duas informações. A primeira representa o total de vezes que a cana, de um talhão específico, foi cortada e, a segunda informa se esse talhão possui

“cana de ano” ou “cana de ano e meio”, por exemplo, o valor 3.12 indica terceiro corte de uma “cana de ano”. Para as “canas de inverno” foi utilizada também a letra i (por exemplo: “112i”). Para as canas socas com mais de 10 cortes o código “10>” foi utilizado.

O **atributo Tipo de Corte** informa se o corte da cana foi manual (MA, MANC ou MANQ) ou mecanizado (ME, MECC ou MECQ) e a **Condição de Corte** se a cana foi colhida após queima (Q) ou crua (C).

A **Formulação do Adubo** informa resumidamente a fórmula do adubo utilizado no talhão, com oito diferentes tipos de fórmulas (vide Apêndice B), e a **Adubação** diz respeito à quantidade desse adubo que foi aplicado, expresso em kg por hectare.

O atributo **Ambiente de Produção** mapeia os solos de acordo com suas características de textura, químicas e morfológicas. Possui cinco diferentes códigos: 1– Ambiente A; 2– Ambiente B; 3– Ambiente C; 4– Ambiente E e 5– Ambiente E. Detalhes a respeito dos ambientes de produção podem ser encontrados em Prado *et al.* (2008).

O **atributo Fertilidade do solo** é também representado por cinco códigos: 1– Alta; 2– Média Alta; 3– Média; 4– Média baixa e 5– Baixa. O atributo **textura** refere-se à proporção de argila, silte e areia do solo. Os seguintes códigos foram utilizados: 1– solo argiloso; 2– solo arenoso e 3– solo argiloso/arenoso.

O **atributo Produtividade** informa a quantidade de cana colhida no talhão, em toneladas por hectare. Na Figura 9, apresentam-se os *boxplots* com a distribuição dos níveis de produtividade, em tonelada de cana por hectare (TCH), de acordo com o número de cortes da cana. Os pontos assinalados em vermelho representam os *outliers*, nesse caso, os valores acima do limite superior ou abaixo do limite inferior. O cálculo desses limites é realizado da seguinte forma:

- Q1 (primeiro quartil), é o valor representado na parte inferior do retângulo;
- Q3 (terceiro quartil), é o valor representado na parte superior do retângulo;
- IQR = Q3 – Q1 (Amplitude Interquartil);

- **Limite Inferior** = $Q1 - 1.5 \times IQR$;
- **Limite Superior** = $Q3 + 1.5 \times IQR$.

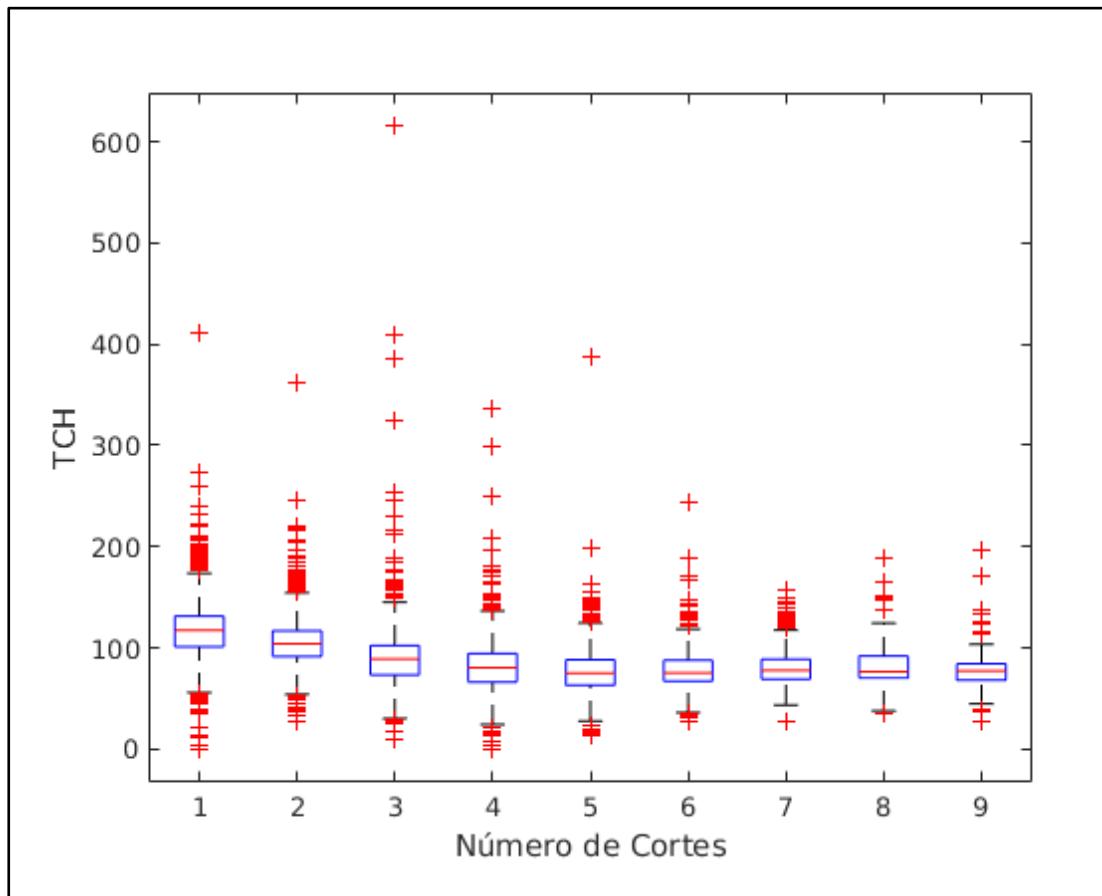


FIGURA 9 - DISTRIBUIÇÃO DOS NÍVEIS DE PRODUTIVIDADE DA CANA EM FUNÇÃO DO NÚMERO DE CORTES

FONTE: A AUTORA

O visual dos *outliers* nos *boxplots* da Figura 9 pode induzir a uma avaliação que aponte um excesso de dados discrepantes, o que não é o caso, uma vez que a maioria normal dos dados está compreendida entre os limites inferior e superior. Para embasar essa análise, a Tabela 5 apresenta os limites inferior e superior, o número total de instâncias, o número de *outliers* e a porcentagem de *outliers* para cada número de cortes da cana. Verifica-se que o maior percentual de *outliers* é 7.7%, quando o número de cortes é 9, e que para os demais *boxplots* esse percentual não ultrapassa 2.7%.

TABELA 5 - OUTLIERS DOS BOXPLOTS

Número de cortes	Limite inferior	Limite Superior	Total de instâncias	Total de outliers	% de outliers
1	55.7619	176.9099	3292	89	2.7%
2	53.5591	154.8015	4576	69	1.5%
3	29.50945	146.2043	5268	45	0.9%
4	23.965	136.725	4999	41	0.8%
5	25.477	126.0746	4666	109	2.3%
6	36.0865	118.7641	2484	56	2.3%
7	39.5775	118.4375	1101	29	2.6%
8	37.55625	124.6863	719	10	1.4%
9	43.81625	108.5463	705	54	7.7%

FONTE: A AUTORA

Além das planilhas do censo varietal, foram enviadas também planilhas contendo o **índice pluviométrico** e **temperatura** (mínima e máxima) diários, para cada mês, de cada ano do período em análise (2006 a 2009). Na Figura 10, apresenta-se a média mensal da precipitação acumulada por ano. Os índices pluviométricos são medidos em 16 estações meteorológicas. No Apêndice C, é apresentado um quadro que associa cada uma das fazendas à respectiva estação meteorológica.

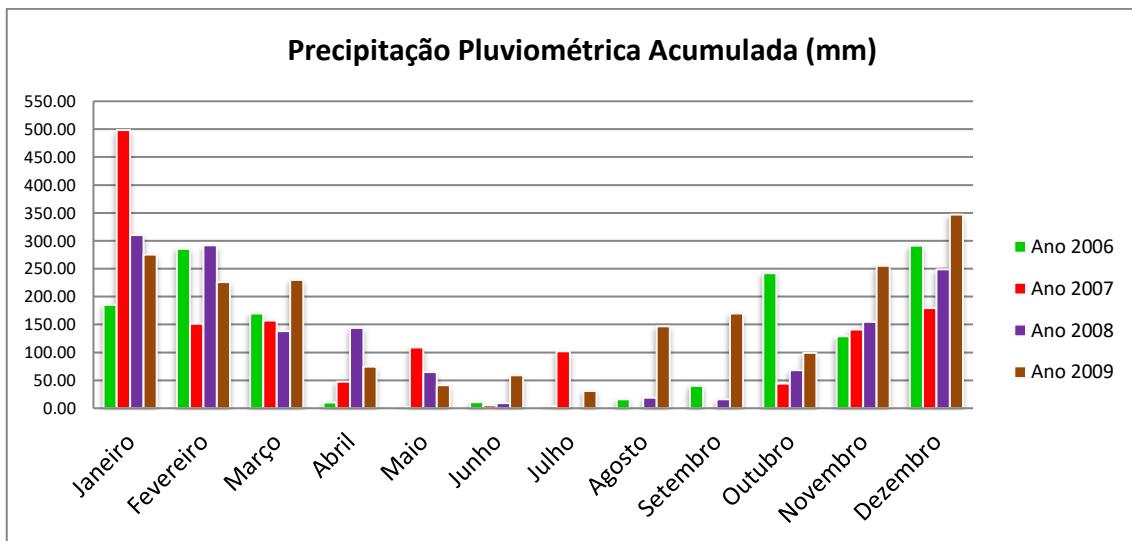


FIGURA 10 - PRECIPITAÇÃO PLUVIOMÉTRICA ACUMULADA POR MÊS (2006 A 2009)

FONTE: A AUTORA

Nota-se, no gráfico da Figura 10, que não há uma variação significativa no comportamento pluviométrico dos quatro anos observados, apesar de haver alguns meses de certos anos em que o índice pluviométrico foi acentuadamente maior do que o mesmo período em outros anos, como, por exemplo, outubro de 2006, janeiro e julho de 2007 e, agosto e setembro de 2009.

Na Figura 11, apresenta-se a média das temperaturas também para o período de 2006 a 2009. Nos gráficos, constata-se que a variabilidade das temperaturas dos quatro anos é relativamente parecida. O que se nota de diferença no período é o fato de que as médias nos anos de 2006 e de 2007 são maiores que as médias de 2008 e 2009.

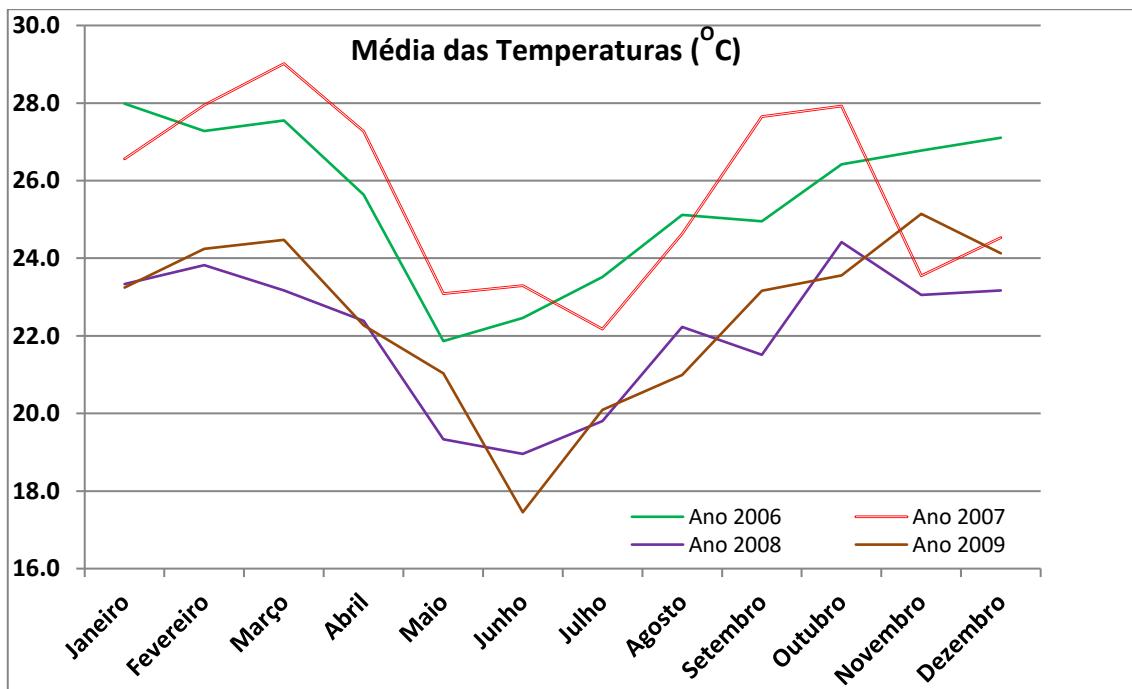


FIGURA 11 - MÉDIA DAS TEMPERATURAS MENSais (2006 A 2009)

FONTE: A AUTORA

4.3 PREPARAÇÃO DOS DADOS

As tarefas descritas a seguir, **limpeza, transformação e integração dos dados**, foram executadas para que fosse possível iniciar a etapa de modelagem.

Limpeza dos dados

Para encontrar possíveis inconsistências nos dados, foram realizadas inspeções visuais nas planilhas contendo os dados da produção. Para cada coluna verificou-se, por meio de filtros, se existiam valores inconsistentes ou vazios. Assim, foi possível verificar as seguintes inconsistências:

- Existia uma instância com vazios para a maioria dos campos e data de corte atual “16/11/2020”. Essa instância foi eliminada.
- Duas instâncias referentes à safra 2007/2008, que apresentavam valor 1 para o atributo estágio de corte, tinham como data de plantio o ano de 1999. Essas duas instâncias também foram eliminadas.
- Foram encontradas 163 instâncias de cana soca com o atributo **Data do Corte Anterior** contendo vazio. Para essas instâncias considerou-se a data de corte anterior como exatamente um ano antes da data do corte atual. A data do corte anterior é necessária para calcular os valores climáticos por fase de crescimento.
- O campo Quantidade de Adubo continha um traço (-), significando que não havia valor para esse atributo. Esse valor não significa um erro, mas foram substituídos por zero para que fosse possível calcular a quantidade de nutriente referente a Nitrogênio (N), Fósforo (P) e Potássio (K).

Transformação dos dados

Diversas tarefas para transformação dos dados foram realizadas, conforme relatadas a seguir.

Como era necessário obter as quantidades de N, P e K aplicadas em cada talhão, foram criados os atributos **InsumoN**, **InsumoP** e **InsumoK**. Os valores para esses atributos foram calculados a partir da **Fórmula do Adubo** e **Adubação**. Por exemplo, para a fórmula “Adubo 27-00-24”, e valor de adubação 400 kg, o valor de InsumoN é 108 ($0,27 \times 400$), do InsumoP é 0 (0×400) e do InsumoK é 96 ($0,24 \times 400$). O atributo **Estágio de Corte** ficou apenas com a informação numérica referente à quantidade de cortes da cana. A informação a

respeito do tipo da cana foi suprimida, uma vez que não fazia parte dos atributos selecionados para o processo de mineração.

O atributo **Tipo de Corte** possuía os códigos MA, significando corte manual; MANC, significando corte manual com cana crua; ou MANQ, significando corte manual com queima. Para os cortes mecanizados havia os seguintes códigos: ME, para corte mecanizado; MECC, para corte mecanizado com cana crua; ou MECQ, para corte mecanizado com queima. Como existe o atributo **Condição de Corte** (C ou Q), que já informa se a cana foi colhida crua ou com queima, foi possível transformar todos os códigos do atributo tipo de corte para MA (MA, MANC e MANQ) ou ME (ME, MECC e MECQ).

A exemplo dos trabalhos de Bocca (2014) e Fernandes, Rocha e Lamparelli (2011), o clima foi dividido em quatro períodos ao longo do ciclo de desenvolvimento da cana-de-açúcar: **brotação, perfilhamento, crescimento e maturação**. A média da temperatura para cada fase de desenvolvimento e a precipitação acumulada para essas fases foi calculada. O objetivo dessa caracterização do clima por fases é identificar seu efeito diferenciado nas diversas fases de crescimento da cana. Assim, foram criados os atributos: **TemperaturaB, TemperaturaP, TemperaturaC, TemperaturaM, PrecipitaçãoB, PrecipitaçãoP, PrecipitaçãoC e PrecipitaçãoM**. A Figura 12 representa o comportamento local típico da cultura da cana, conforme informado pelo profissional da Usina sob análise, para possibilitar o cálculo dos atributos climáticos.

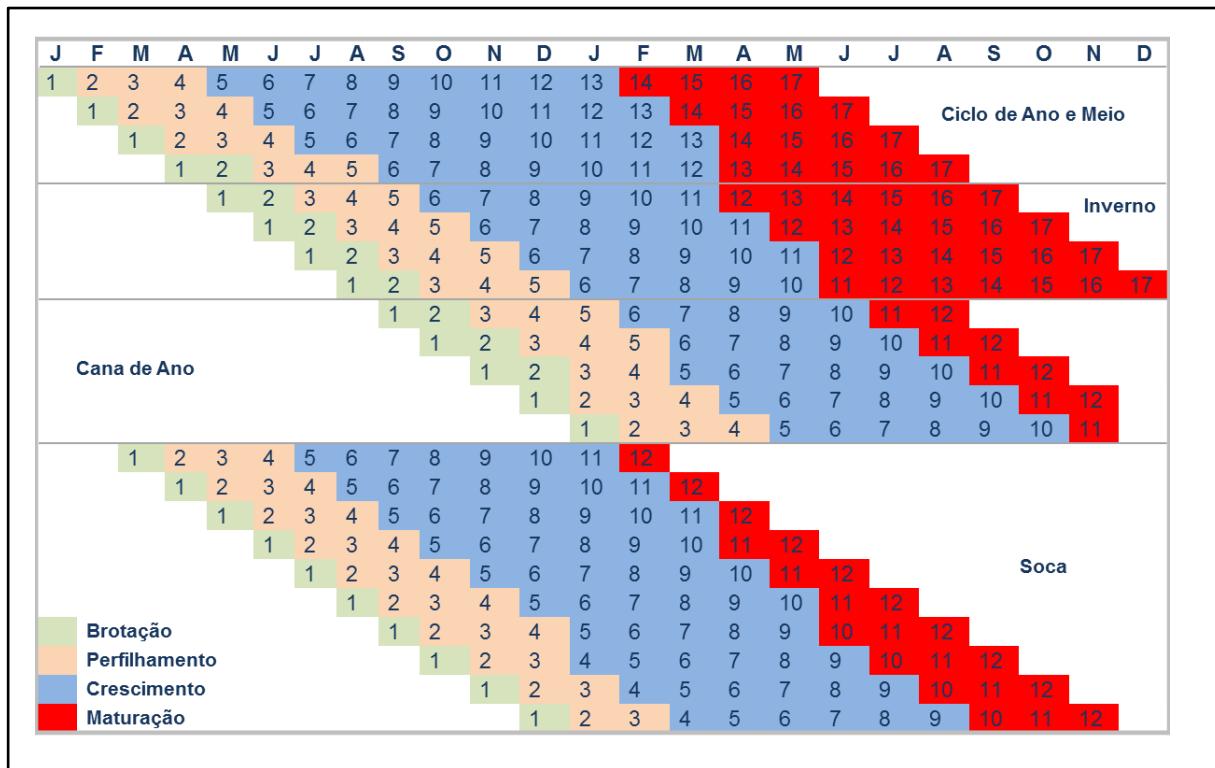


FIGURA 12 - FASES FENOLÓGICAS DA CANA

FONTE: ADAPTADO DE BOCCA (2014)

O cálculo dos valores para os atributos climáticos de acordo com as fases fenológicas foi realizado em várias etapas. Para os valores da precipitação pluviométrica, inicialmente, foram calculados os totais mensais, para cada ano do período (2006 a 2009), para cada uma das estações climáticas. Em seguida, foram utilizados filtros, na planilha com os dados da produção, para selecionar, por ano, cada uma das fazendas e, assim, identificar sua estação pluviométrica e quais seriam os valores a serem usados. Como os meses das fases fenológicas variam de acordo com o tipo da cana, conforme apresentado na Figura 12, foi selecionado também, para cada fazenda, o tipo da cana. Uma vez selecionados um ano, uma fazenda e um tipo específico de cana, com base em sua data de plantio para as canas-planta ou na data do corte anterior para as canas-soca, foram realizados os cálculos específicos para cada cenário.

O cálculo da temperatura foi realizado de maneira semelhante ao dos índices pluviométricos, entretanto essa etapa foi mais simples porque os valores da temperatura não estavam separados por estação meteorológica, dessa forma a etapa da seleção de valores por fazenda não foi necessária.

As atividades de transformação de dados descritas até aqui foram realizadas utilizando os recursos da planilha eletrônica MS-Excel. Outras transformações foram realizadas na ferramenta Weka, por meio dos filtros disponibilizados para atividades de pré-processamento, conforme descrito na sequência.

Os atributos **Ambiente**, **Fertilidade** e **Textura** possuem conteúdo numérico, entretanto esses números são códigos utilizados para representar uma informação categórica, não representam valores. Para transformar esses atributos de numéricos para nominais, foi aplicado o filtro “NumericToNominal” (Preprocess\filters\unsupervised\attribute\NumericToNominal).

O atributo **Produtividade**, definido como atributo meta, possui conteúdo numérico, mas a tarefa de classificação requer que esse tipo de atributo seja nominal. Dessa forma foi utilizado o filtro “Discretize” (Preprocess\filters\unsupervised\attribute\NumericToNominal). Para essa *discretização*, definiu-se em quantas partes o conjunto deveria ser dividido (bins igual a 3), uma vez que ficou definido, juntamente com o especialista da Usina, que o modelo teria três classes, porque é uma quantidade mais fácil para gerenciar. A opção “useEqualFrequency” foi configurada como *true*, indicando que cada um dos três conjuntos teria aproximadamente a mesma quantidade de elementos. Após a aplicação do filtro “Discretize” com essas configurações os valores de produtividade foram substituídos conforme apresentado na Tabela 6. Destaca-se que o especialista da Usina havia sugerido que os níveis produtividade fossem divididos em: *i*) menores ou iguais a 80; *ii*) maiores que 80 e menores ou iguais a 100 e, *iii*) maiores que 100, ou seja, muito próximo aos valores obtidos na *discretização* realizada pela ferramenta Weka.

TABELA 6 - VALORES DO ATRIBUTO PRODUTIVIDADE

Valor original da produtividade (numérico)	Valor transformado	Quantidade de instâncias por conjunto
<= 76,957	'(-inf-76.95741]'	9270
> 76,957 e <= 98,851	'(76.95741-98.851012]'	9270
98,851	'(98.851012-inf)'	9270

FONTE: A AUTORA

Integração dos dados

Todas as planilhas com dados da produção foram consolidadas em uma única planilha antes da realização das etapas de limpeza e transformação dos dados. Os dados referentes ao clima, precipitação acumulada e média da temperatura para cada fase fenológica, foram também inseridos nessa planilha.

Concluídas as tarefas de limpeza e transformação gerou-se um arquivo no formato CSV, uma das entradas permitidas pela ferramenta de mineração de dados Weka. Após a carga dessa planilha na ferramenta Weka, foram realizadas as transformações citadas anteriormente e removidos os atributos fazenda, gleba e talhão, que servem como identificação das instâncias, mas não devem ser utilizados na tarefa de classificação. A partir dessas alterações, um novo arquivo para ser utilizado nas atividades de modelagem foi gerado. Dessa forma, o conjunto de dados final contém 20 atributos para 27810 instâncias, sendo 19 variáveis independentes e o atributo meta. O Quadro 13 resume as características desses atributos.

QUADRO 13 - CARACTERÍSTICAS DOS ATRIBUTOS A SEREM UTILIZADOS NA ETAPA DE MODELAGEM

Nome do atributo	Tipo	Valores	Descrição
Produtividade	Categórico	1. '(-inf-76.95741]' 2. '(76.95741-98.851012]' 3. '(98.851012-inf)'	Atributo Meta- 3 níveis de produtividade
Solo	Categórico	Quadro 27 - Apêndice B	Sigla que descreve o tipo de solo
Variedade	Categórico	Quadro 26- Apêndice B	Sigla que descreve a variedade da cana
Corte	Numérico	1 a 10	Quantidade de vezes que a cana foi cortada
Tipo de corte	Categórico	MA ME	Especifica se o corte foi Manual ou Mecanizado
Condição de Corte	Categórico	C Q	Especifica se a cana foi cortada crua ou com queima
Ambiente	Categórico	1 a 5	Tipo do ambiente de manejo
Fertilidade	Categórico	1 a 5	Nível de fertilidade do solo
Textura	Categórico	1 a 3	Proporção de argila, silte e areia do solo
InsumoN	Numérico		Quantidade do insumo N em Kg por hectare

(CONT,) QUADRO 13 - CARACTERÍSTICAS DOS ATRIBUTOS A SEREM UTILIZADOS NA ETAPA DE MODELAGEM

Nome do atributo	Tipo	Valores	Descrição
InsumoP	Numérico		Quantidade do insumo P em Kg por hectare
InsumoK	Numérico		Quantidade do insumo K em Kg por hectare
TemperaturaB	Numérico		Média da temperatura na fase da brotação
TemperaturaP	Numérico		Média da temperatura na fase do perfilhamento
TemperaturaC	Numérico		Média da temperatura na fase do crescimento
TemperaturaM	Numérico		Média da temperatura na fase da maturação
PrecipitaçãoB	Numérico		Precipitação acumulada na fase da brotação
PrecipitaçãoP	Numérico		Precipitação acumulada na fase do perfilhamento
PrecipitaçãoC	Numérico		Precipitação acumulada na fase do crescimento
PrecipitaçãoM	Numérico		Precipitação acumulada na fase da maturação

FONTE: A AUTORA

No Apêndice D apresentam-se histogramas, gerados pela ferramenta Weka, de todos os atributos utilizados na etapa de modelagem.

4.4 MODELAGEM DOS DADOS

A ferramenta Weka, apresentada na Seção 2.4.5, foi utilizada neste trabalho porque contempla as principais tarefas de mineração de dados, é usada com frequência nas pesquisas que envolvem mineração de dados e também por ser gratuita.

Diversos experimentos de modelagem foram realizados visando a ajustar os parâmetros dos modelos com o objetivo de reduzir sua complexidade e manter níveis aceitáveis de acurácia. Na Seção 4.4.1 são apresentados os resultados dos experimentos realizados com o conjunto completo dos dados: todos os

atributos resultantes da aplicação do método Delphi, com todas as instâncias. Apresentam-se também os resultados obtidos após a exclusão das instâncias consideradas *outliers* e comparam-se esses resultados.

Na Seção 4.4.2 são apresentados os resultados obtidos com a aplicação de métodos de seleção de atributos. Esses resultados são analisados e, a partir disso, realizados os mesmos experimentos de modelagem relatados na Seção 4.4.1, entretanto com a retirada de alguns atributos. Esses procedimentos visam a obter melhor qualidade dos modelos.

4.4.1 MODELAGEM DOS DADOS COM CONJUNTO COMPLETO DE ATRIBUTOS

Para a realização da tarefa de classificação, foi utilizada a técnica indução de árvore de decisão. Uma árvore de decisão pode ser utilizada como um modelo para prever o atributo meta ou, como neste trabalho, para entender a estrutura preditiva do problema. Nesse último caso, o objetivo é compreender quais variáveis são mais relevantes na predição e como se dá a interação dessas variáveis. Optou-se pelo algoritmo J48, que é a implementação da ferramenta Weka do algoritmo C4.5 (QUINLAN, 1993), pois possui complexidade linear (o tempo de processamento cresce linearmente em relação à quantidade dos dados) e, assim, tornou factível o processamento dos dados em um pequeno intervalo de tempo.

Bocca (2014) utilizou vários algoritmos de classificação, entre eles o SVM e algoritmos de redes neurais, para criar um modelo de previsão de produtividade. Esse métodos, em geral, produzem modelos com valores de acurácia mais altos que as árvores de decisão, entretanto não têm a expressividade das árvores de decisão e, assim, não poderiam ser utilizados para o entendimento da estrutura preditiva do modelo, nem servir de base para a criação da ferramenta de visualização proposta neste trabalho. Ferraro, Rivero e Ghersa (2009), com o objetivo de definir os atributos que mais impactam na produtividade da cana, utilizaram o algoritmo CART (*Classification and Regression Tree*), para gerar árvores de decisão. O CART não foi utilizado no

presente trabalho porque, em mais de 48 horas de execução, o algoritmo não havia sido concluído e, dessa forma, optou-se por interromper sua execução.

A realização do processo de KDD é bastante empírica, assim vários experimentos devem ser realizados a fim de encontrar o modelo mais apropriado ao contexto analisado. No caso da classificação, podem ser verificados, por exemplo, o tempo de processamento, as diversas medidas de desempenho do modelo e a adequação da estrutura de predição ao problema que se está modelando. Por essa razão, foram elaborados diversos planos de modelagem, visando à calibração dos parâmetros, de forma a obter valores de desempenhos aceitáveis e, simultaneamente, uma estrutura de árvore com nível de complexidade possível de ser analisada e interpretada pelos usuários do sistema de visualização proposto neste trabalho.

Para as análises iniciais, os seguintes parâmetros foram ajustados: *i*) forma de construção da árvore (binária ou não binária); *ii*) número mínimo de instâncias nos nós e; *iii*) método de treinamento/teste do modelo. A seguir justifica-se a utilização desses parâmetros.

Na construção da árvore, o algoritmo J48 pode gerar árvores binárias e não binárias. Se a opção for por árvores não binárias, os atributos numéricos terão divisão binária na criação de novos nós e, para os atributos categóricos, haverá um novo nó para cada valor possível do atributo. Dessa forma, atributos categóricos com muitos valores distintos acarretam árvores muito específicas, com um número muito grande de nós. Para as árvores binárias, as divisões são binárias tanto para atributos numéricos como para atributos categóricos, e assim as árvores ficam mais generalizadas. Árvores binárias e não binárias foram avaliadas para as configurações propostas.

A configuração do número mínimo de instâncias por nó também altera o tamanho e, consequentemente, a complexidade da árvore gerada. Quanto maior for o número de instâncias requeridos em um nó antes que ele seja dividido em outros nós, menor será o tamanho da árvore. Assim, o objetivo do ajuste desse parâmetro é encontrar um valor que reduza o tamanho da árvore, de forma a não

gerar árvores muito específicas (*overfitting*), mas também não gerar árvores muito generalizadas (*underfitting*). Trata-se de um procedimento de pré-poda, como apresentado na Seção 2.4.3.

Para a seleção do método de treinamento/teste do modelo foram realizados experimentos com os métodos, explicados na Seção 2.4.4, *holdout* (opção *Percentage Split* na ferramenta Weka) e validação cruzada estratificada (*Stratified Cross-Validation* de 10 *folds*). A acurácia obtida com a utilização da validação cruzada é, em geral, mais alta que aquela obtida com o método *holdout*, entretanto o tempo gasto para geração do modelo e teste é bem maior na validação cruzada. Após testes com os dois métodos para as árvores não binária e binária, com número mínimo de instâncias por nó igual a 2, verificou-se que o tempo de construção do modelo na validação cruzada, apesar de maior, não foi impactante, não ultrapassando alguns segundos, assim optou-se por sua utilização nos demais experimentos de modelagem.

No Quadro 14 são apresentados os resultados da avaliação de cada uma das modelagens realizadas. Para cada opção de modelagem, também foram realizados testes em um conjunto reduzido, com 27308 instâncias, obtido pela eliminação das instâncias com valores de produtividade considerados *outliers* (502 instâncias), conforme apresentado na Seção 4.2.

Observando o Quadro 14 é possível verificar que à medida que aumenta o número mínimo de instâncias por nós, a árvore se torna menor e, portanto, menos complexa, mas, em contrapartida, a acurácia do modelo diminui. É verdade que as árvores geradas pelo processo de mineração de dados serão disponibilizadas para um sistema de visualização que facilitará a interpretação e análise dessas árvores, entretanto, mesmo com a utilização desse sistema, entende-se que árvores com muitos nós (acima de 200 nós) sejam de difícil análise.

QUADRO 14 - RESULTADOS DAS PARAMETRIZAÇÕES

Modelagem	Parametrização	Avaliação	Avaliação Conjunto Reduzido
Modelagem 1	Não binária MinNumObj: 2 <i>Cross-validation</i>	Acurácia: 86,850% NumFolhas: 7758 TamÁrvore: 8762	Acurácia: 87,663% NumFolhas: 6470 TamÁrvore: 7394
Modelagem 2	Binária MinNumObj: 2 <i>Cross-validation</i>	Acurácia :87,170% NumFolhas: 1225 TamÁrvore: 2449	Acurácia: 87,706% NumFolhas: 1254 TamÁrvore: 2507
Modelagem 3	Não Binária MinNumObj: 2 <i>Percentage Split</i>	Acurácia: 85,478% NumFolhas: 7758 TamÁrvore: 8762	Acurácia: 86,462% NumFolhas: 6470 TamÁrvore: 7394
Modelagem 4	Binária MinNumObj:2 <i>Percentage Split</i>	Acurácia: 86,060% NumFolhas:1225 TamÁrvore: 2449	Acurácia: 86,645% NumFolhas: 1254 TamÁrvore:2507
Modelagem 5	Não Binária MinNumObj: 10 <i>Cross-validation</i>	Acurácia: 82,391% NumFolhas: 3228 TamÁrvore: 3669	Acurácia: 83,250% NumFolhas: 3009 TamÁrvore: 3422
Modelagem 6	Binária MinNumObj: 10 <i>Cross-validation</i>	Acurácia: 83,322% NumFolhas: 636 TamÁrvore: 1271	Acurácia: 83,700% NumFolhas: 598 TamÁrvore: 1195
Modelagem 7	Não Binária MinNumObj: 50 <i>Cross-validation</i>	Acurácia: 74,617% NumFolhas:1233 TamÁrvore: 1350	Acurácia: 75,450% NumFolhas: 1264 TamÁrvore: 1374
Modelagem 8	Binária MinNumObj:50 <i>Cross-validation</i>	Acurácia: 75,246% NumFolhas: 227 TamÁrvore: 453	Acurácia: 76,193% NumFolhas: 201 TamÁrvore: 401
Modelagem 9	Não Binária MinNumObj:100 <i>Cross-validation</i>	Acurácia: 71,370% NumFolhas: 556 TamÁrvore: 621	Acurácia: 71,865% NumFolhas: 598 TamÁrvore: 663
Modelagem 10	Binária MinNumObj: 100 <i>Cross-validation</i>	Acurácia: 71,118% NumFolhas: 106 TamÁrvore: 211	Acurácia: 72,158% NumFolhas: 99 TamÁrvore: 197

FONTE: A AUTORA

No que diz respeito à acurácia, os níveis aceitáveis são dependentes da área de aplicação, como não existe um padrão mínimo de acurácia para a área agrícola, foram analisados os trabalhos levantados pela RSL que utilizaram a tarefa de classificação, objetivando-se definir o valor mínimo que seria utilizado neste trabalho. No Quadro 15 são apresentadas as acurácia obtidas em pesquisas sobre cana-de-açúcar.

QUADRO 15 - ACURÁCIA DE TRABALHOS QUE UTILIZAM CLASSIFICAÇÃO

Autor (ano)	Objetivo da aplicação da tarefa de classificação	Acurácia
Everingham <i>et al.</i> (2007)	Caracterizar culturas da cana-de-açúcar:	
	1) Modelos que classificaram o ciclo da cana (número de cortes)	72,5% a 89,5%
	2) Modelos que classificaram a variedade da cana	71,3% a 92,3%,
Nascimento <i>et al.</i> (2009)	Identificar campos com cana-de-açúcar:	
	1) Quando não havia cana-de-açúcar	88,0% (sensitividade)
	2) Quando a plantação era de cana-de-açúcar	92,0% (sensitividade)
Fernandes, Rocha e Lamparelli (2011)	Prever a produtividade da cana-de-açúcar	66,7% a 86,5%
Vintrou <i>et al.</i> (2013)	Mapear terras cultivadas com cana-de-açúcar	57,8%
Nonato e Oliveira (2013)	Classificar áreas cultivadas com cana-de-açúcar	94,9% e 97,2%.

FONTE: A AUTORA

Observa-se que Nascimento *et al.* (2009) não apresentaram os valores da acurácia geral, mas, sim, os valores de sensitividade, entretanto esses valores são, normalmente, próximos da acurácia geral do modelo. Com base nos valores utilizados nas pesquisas apresentadas no Quadro 15, definiu-se manter níveis de acurácia acima de 70% para este trabalho. Por essa razão não foram utilizadas configurações de modelagens com número mínimo de instâncias por nó superior a 100.

Da análise do Quadro 14, nota-se que os valores de acurácia para a modelagem do conjunto reduzido foi ligeiramente melhor que para o conjunto completo dos dados, em todas as opções de modelagem. Embora o número de nós tenha sido um pouco maior, no conjunto reduzido, nas modelagens 2, 4,7 e 9, acredita-se que a opção sem *outliers* é a melhor e, dessa forma, o processo de seleção de atributos, apresentado na próxima seção, foi realizado com o conjunto reduzido de instâncias.

4.4.2 MODELAGEM DOS DADOS COM SELEÇÃO DE ATRIBUTOS

O objetivo da seleção de atributos é remover atributos irrelevantes, isto é, aqueles que não contêm informações úteis para o modelo, e também os redundantes, que são aqueles altamente correlacionados com algum outro atributo e, por essa razão, não agregam informação na construção do modelo.

Os métodos de seleção de atributos são geralmente classificados como filtros e *wrappers* (GUYON e ELISSEEFF, 2003; HALL e HOLMES, 2003; PRATI, BATISTA E MONARD, 2008). *Wrappers* avaliam os atributos utilizando a acurácia obtida por um algoritmo de aprendizado especificado, o objetivo é encontrar um subconjunto de atributos que minimize o erro de predição. Filtros se baseiam nas características dos dados e trabalham de forma independente dos algoritmos de aprendizado. Os métodos do tipo filtro especificam um *ranking* para os atributos, já os métodos *wrappers* tem como resultado um subconjunto de atributos selecionado pelo método, sem que seja especificado um *ranking*. Os métodos *wrappers* não foram utilizados neste trabalho por restrições computacionais. Na Tabela 7, são apresentados os resultados dos ranqueamentos realizados a partir dos métodos de seleção de atributos do tipo filtro disponíveis no Weka (no Apêndice E, apresenta-se uma breve descrição desses métodos).

Da análise das classificações apresentadas na Tabela 7, algumas considerações podem ser feitas. Inicialmente é possível observar a importância do número de cortes, classificado em primeiro lugar em três dos seis métodos utilizados e em segundo lugar em um deles. A influência do número de cortes na distribuição da produtividade pode ser verificada também nos *boxplots* da Figura 9, apresentados na Seção 4.2.1. Esses resultados estão de acordo com os obtidos no trabalho Ferraro Rivero e Ghersa (2009) em que o atributo número de cortes foi o segundo mais importante, representando 80% de importância para a construção do modelo.

TABELA 7 - RESULTADO DO RANQUEAMENTO REALIZADO PELOS MÉTODOS DE SELEÇÃO DE ATRIBUTOS

Atributos	Métodos de Seleção de Atributos					
	Correlation	GainRatio	InfoGain	OneR	ReliefF	Symmetrical Uncert
Solo	14	14	13	13	4	13
Variedade	11	13	12	12	1	12
Corte	1	1	11	11	2	1
Tipo de Corte	19	19	19	19	14	19
Cond. Corte	18	18	18	18	16	18
Ambiente	12	16	14	14	19	14
Fertilidade	13	17	15	15	18	15
Textura	10	15	16	16	15	17
InsumoN	4	4	2	1	13	2
InsumoP	17	2	17	17	17	16
InsumoK	6	3	5	4	12	3
TemperaturaB	7	7	8	6	3	6
TemperaturaP	2	11	10	9	5	10
TemperaturaC	16	12	9	10	6	11
TemperaturaM	15	8	7	7	8	7
PrecipitaçãoB	8	9	6	8	9	9
PrecipitaçãoP	3	6	3	3	10	5
PrecipitaçãoC	5	5	1	2	7	4
PrecipitaçãoM	9	10	4	5	11	8

FONTE: A AUTORA

Os insumos N e K também se mostraram relevantes para a elaboração do modelo de predição de produtividade, exceto para o método ReliefF. Esses dois atributos ocuparam as primeiras posições do ranqueamento. O mesmo não aconteceu com o insumo P que ficou, na maioria dos métodos, nas posições finais. O nitrogênio (N) é um dos nutrientes essenciais absorvido em maior quantidade pela cana-de-açúcar, perdendo, em geral, apenas para o fósforo (P) e tem um papel fundamental no desenvolvimento da cana (VITTI *et al.*, 2008). O fósforo (P) é muito importante para a produtividade da cana-de-açúcar, mas existem controvérsias sobre a eficiência da adubação fosfatada nas soqueiras (ROSSETO *et al.*, 2008a). O potássio (K), também extraído em grande quantidade pela cana-de-açúcar, tem grande influência na produtividade da cana (ROSSETO *et al.*, 2008b). Verificando-se a planilha com os dados da produção, identificou-se que em, aproximadamente, 80% das instâncias foi aplicado o insumo N, em 60% foi aplicado o insumo K e em apenas 2% foi utilizado formulações que continham insumo P, portanto isso explica a classificação desse insumo.

Esses resultados estão em desacordo com aqueles obtidos em Souza et al. (2010), em que a correlação linear entre a produtividade da cana-de-açúcar e os atributos químicos do solo mostrou coeficientes de correlação baixos para todos os atributos do solo estudados, exceto para variável potássio, que apresentou significância.

Para cada tipo de solo está associado um ambiente, uma fertilidade e uma textura, essa correlação provavelmente levou o atributo solo a obter, exceto em um dos métodos, posições melhores que os atributos Ambiente, Fertilidade e Textura. Esses atributos ficaram em posições finais na maioria dos métodos.

O atributo Variedade foi classificado, exceto no método ReliefF, nas posições centrais, uma ou três posições antes do atributo solo. Em Ferraro, Rivero e Ghersa (2009), o atributo Variedade obteve resultados bem superiores, 100% de importância no modelo de produtividade medido em toneladas de açúcar por hectare (TSC – do inglês, *Ton of Sugar per Hectare*) e em torno de 70% para o modelo medido em tonelada de cana por hectare (TCH).

A maioria dos métodos classificou os atributos climáticos em posições iniciais, isso significa que eles foram capazes de capturar influência dos fatores climáticos para a produtividade da cana-de-açúcar, como já destacada na Seção 2.1.

No trabalho de Bocca (2014), em que foram utilizados métodos *wrappers* para a seleção de atributos, os resultados dessa seleção em relação ao solo e aos atributos climáticos foram equivalentes aos obtidos neste trabalho com os métodos do tipo filtro, uma vez que os atributos climáticos foram selecionados pela maioria dos métodos, mas o mesmo não ocorreu com os atributos referentes ao solo.

No trabalho de Fernandes, Rocha e Lamparelli (2011), cujo modelo associava atributos de sensoriamento remoto (NDVI) e atributos climáticos à produtividade da cana-de-açúcar, os métodos de seleção de atributos utilizados (*wrappers* e filtros) não selecionaram atributos climáticos. Também no trabalho de Ferraro,

Rivero e Ghersa (2009) os atributos climáticos colaboraram pouco para a criação do modelo de produtividade, menos de 20 % para a precipitação acumulada total e menos de 10 % para a precipitação acumulada nos meses de verão.

Finalmente, os atributos Condição de Corte (crua ou queima) e Tipo de Corte (manual ou mecanizado) obtiveram as duas últimas posições em cinco dos seis métodos utilizados, o que é bastante coerente, uma vez que os demais atributos, pela sua natureza, certamente têm maior influência na produtividade da cana.

Os gráficos das Figuras 13 a 18 exibem a contribuição de cada atributo para a identificação da classe.

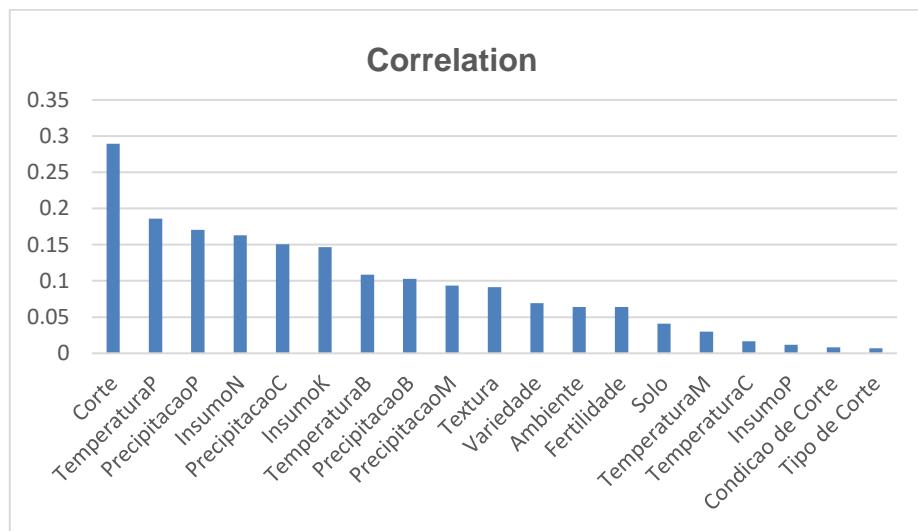


FIGURA 13 - GRÁFICO MÉTODO CORRELATION

FONTE: AUTORA

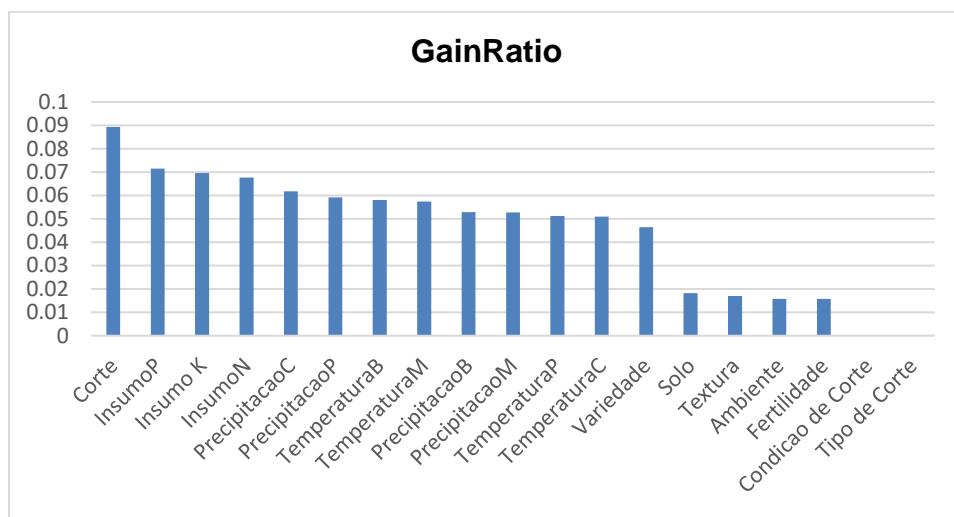


FIGURA 14 - GRÁFICO MÉTODO GAINRATIO

FONTE: AUTORA

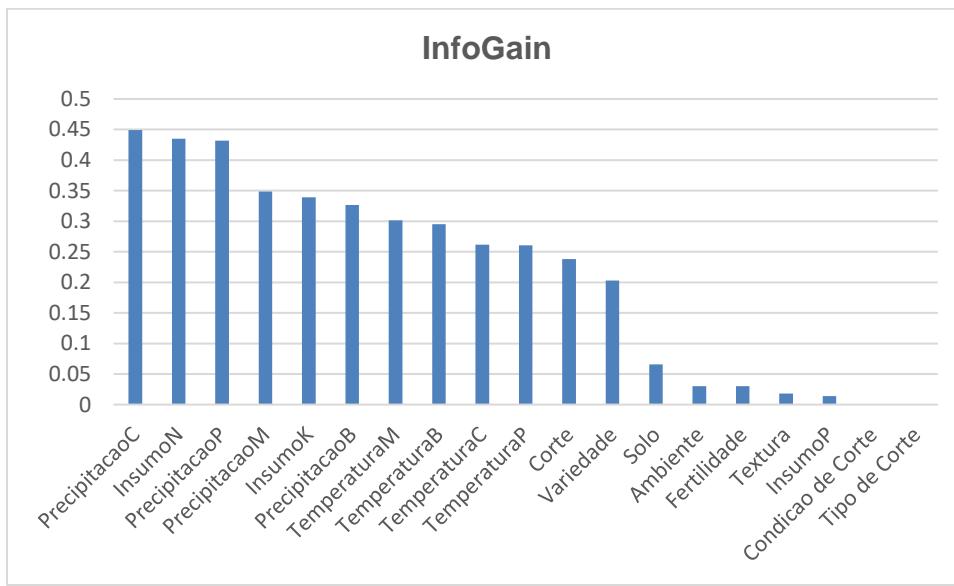


FIGURA 15 - GRÁFICO MÉTODO INFOGAIN

Fonte: Autora

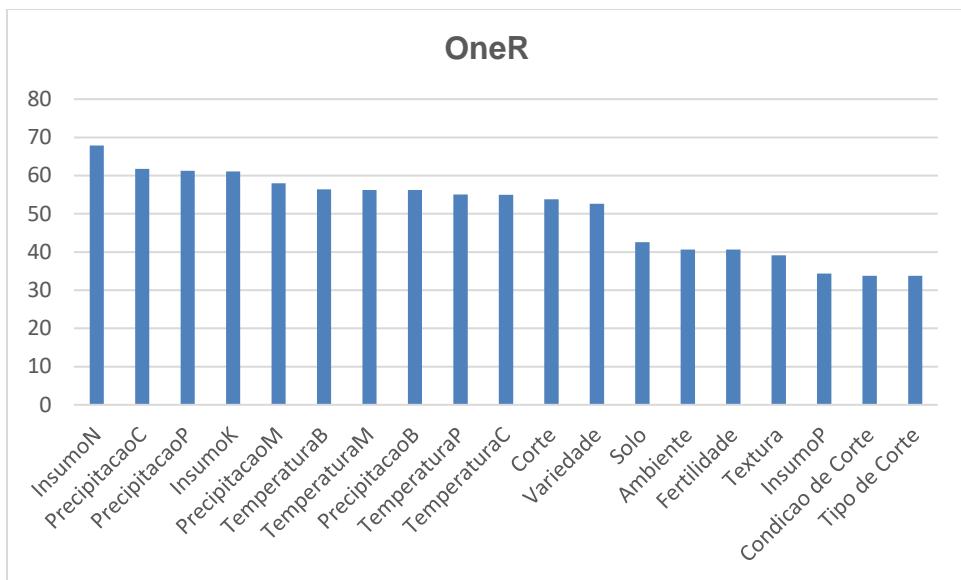


FIGURA 16 - GRÁFICO ONE R

FONTE: AUTORA

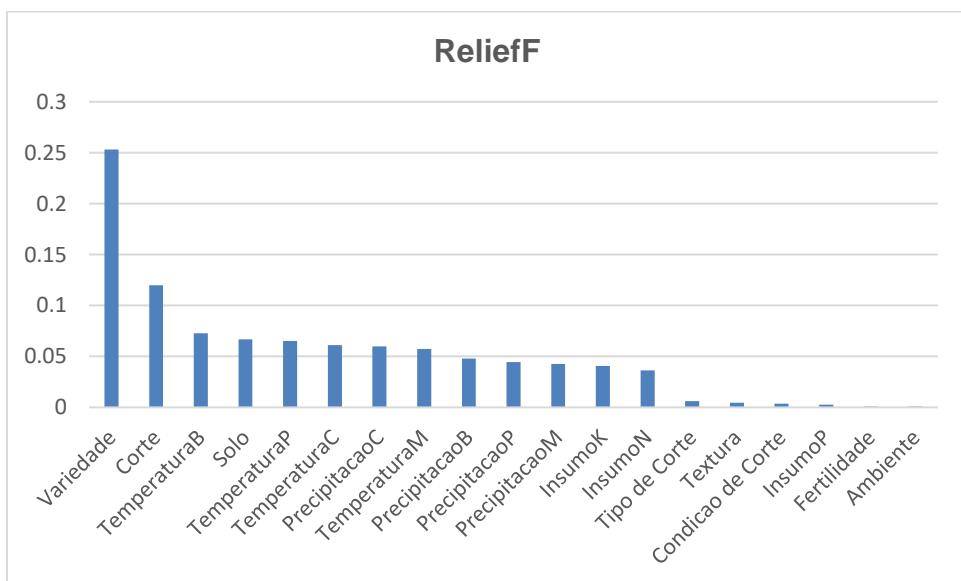


FIGURA 17 - GRÁFICO MÉTODO RELIEFF

FONTE: AUTORA

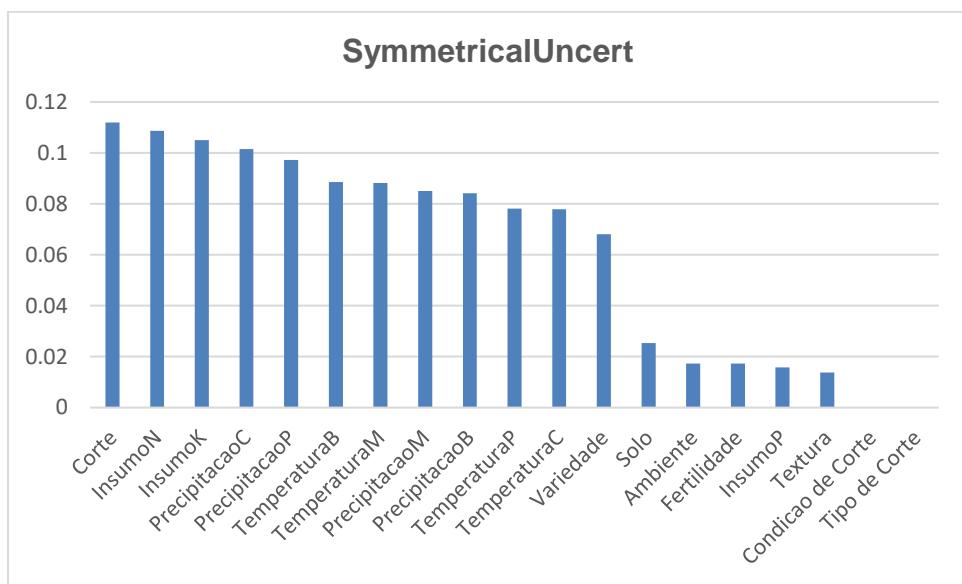


FIGURA 18 - GRÁFICO MÉTODO SIMMETRICAL UNCERT

FONTE: AUTORA

Pela visualização dos gráficos das Figuras 13 a 18, é possível perceber que os atributos Tipo de Corte e Condição de Corte tiveram nível de contribuição muito pequena para a construção dos modelos na maioria dos métodos, exceto no método OneR em que não houve muita variação na pontuação dada a cada atributo. Destaca-se ainda a contribuição do atributo Variedade no método ReliefF, bem como a queda acentuada de pontuação a partir do atributo Solo nos métodos Infogain, Gainration e SymmetricalUncert.

Dois experimentos de seleção de atributos foram realizados para cada uma das modelagens apresentadas no Quadro 14: *i)* retirada dos atributos tipo de corte e condição de corte, pela baixa pontuação desses atributos na maioria dos modelos e; *ii)* retirada dos atributos Ambiente, Fertilidade e Textura, porque o atributo Solo já contempla as informações de ambiente, fertilidade e textura. No Quadro 16 são apresentados os resultados obtidos pelos modelos com a seleção de atributos.

QUADRO 16 - RESULTADOS DAS AVALIAÇÕES COM SELEÇÃO DE ATRIBUTOS

Modelagem	Avaliação Conjunto Reduzido	Avaliação com Seleção de Atributos	Avaliação com Seleção de Atributos
	Todos os atributos	Retirados: tipo e condição de corte	Retirados: ambiente, fertilidade e textura
Modelagem 1	Acurácia: 87,663% NumFolhas: 6470 TamÁrvore: 7394	Acurácia: 86,798% NumFolhas: 4946 TamÁrvore: 5968	Acurácia: 87,732% NumFolhas: 7070 TamÁrvore: 7977
Modelagem 2	Acurácia: 87,706% NumFolhas: 1254 TamÁrvore: 2507	Acurácia: 86,428% NumFolhas: 1328 TamÁrvore: 2655	Acurácia: 87,604% NumFolhas: 1246 TamÁrvore: 2491
Modelagem 3	Acurácia: 86,462% NumFolhas: 6470 TamÁrvore: 7394	Acurácia: 84,792% NumFolhas: 4946 TamÁrvore: 5968	Acurácia: 87,732% NumFolhas: 7070 TamÁrvore: 7977
Modelagem 4	Acurácia: 86,645% NumFolhas: 1254 TamÁrvore: 2507	Acurácia: 84,771% NumFolhas: 1328 TamÁrvore: 2655	Acurácia: 87,604% NumFolhas: 1246 TamÁrvore: 2491
Modelagem 5	Acurácia: 83,250% NumFolhas: 3009 TamÁrvore: 3422	Acurácia: 82,648% NumFolhas: 2571 TamÁrvore: 3091	Acurácia: 83,546% NumFolhas: 3371 TamÁrvore: 3795
Modelagem 6	Acurácia: 83,700% NumFolhas: 598 TamÁrvore: 1195	Acurácia: 82,078% NumFolhas: 667 TamÁrvore: 1333	Acurácia: 83,568% NumFolhas: 606 TamÁrvore: 1211
Modelagem 7	Acurácia: 75,450% NumFolhas: 1264 TamÁrvore: 1374	Acurácia: 73,978% NumFolhas: 936 TamÁrvore: 1081	Acurácia: 75,585% NumFolhas: 1214 TamÁrvore: 1327
Modelagem 8	Acurácia: 76,193% NumFolhas: 201 TamÁrvore: 401	Acurácia: 73,191% NumFolhas: 228 TamÁrvore: 455	Acurácia: 75,413% NumFolhas: 194 TamÁrvore: 387
Modelagem 9	Acurácia: 71,865% NumFolhas: 598 TamÁrvore: 663	Acurácia: 69,635% NumFolhas: 424 TamÁrvore: 493	Acurácia: 72,341% NumFolhas: 654 TamÁrvore: 713
Modelagem 10	Acurácia: 72,158% NumFolhas: 99 TamÁrvore: 197	Acurácia: 68,866% NumFolhas: 114 TamÁrvore: 227	Acurácia: 71,814% NumFolhas: 103 TamÁrvore: 205

FONTE: A AUTORA

A realização do processo de seleção de atributos visa a:

- Melhorar o desempenho dos algoritmos de aprendizado de máquina;
- Simplificar os modelos de predição;
- Reduzir o custo computacional para executar esses modelos;
- Fornecer um estudo prévio sobre o relacionamento entre os atributos.

Verifica-se, no Quadro 16, que os modelos com menos atributos não foram capazes, muitas vezes, de melhorar o desempenho do algoritmo de aprendizado utilizado nas diversas configurações de modelagem propostas. Os modelos em que foram retirados os atributos condição de corte e tipo de corte obtiveram valor

de acurácia menor que os modelos com todos os atributos em todas as dez modelagens. Os modelos nos quais foram retirados os atributos: Ambiente, Fertilidade e Textura, obtiveram resultados melhores em relação à acurácia. Apenas nas modelagens 2 e 6 seus valores de acurácia foram inferiores aos dos modelos completos.

No que diz respeito à simplificação dos modelos, quando foram retirados os atributos: Condição de Corte e Tipo de Corte, as modelagens não binárias obtiveram valores melhores (menos nós nas árvores) que as modelagens com todos os atributos. Quando foram retirados os atributos: Ambiente, Fertilidade e Textura, o tamanho da árvore gerada foi menor no modelo com seleção de atributos nas modelagens 2, 4,7 e 8.

Em relação ao custo computacional, os tempos de construção dos modelos foram muito pequenos, não ultrapassando poucos segundos (em torno de 2 segundos), e por essa razão não foi necessário considerá-los na análise.

Verificando-se todos os resultados de forma conjunta, concluiu-se que, embora a experiência de seleção de atributos tenha sido efetiva no propósito de entender e validar a estrutura do modelo, a estrutura completa é a mais adequada, pois os modelos com menos atributos não foram capazes de obter resultados muito melhores nos quesitos acurácia e redução de complexidade. Além disso, a utilização do conjunto com todos os atributos contempla as escolhas realizadas pelos especialistas por meio do método Delphi. Assim, para a sequência do processo de KDD, na próxima seção será realizada a avaliação dos resultados obtidos com a modelagem utilizando todos os atributos.

4.5 AVALIAÇÃO

Como foram realizados vários experimentos de modelagem, nesta seção será realizada a análise da estrutura e dos parâmetros de desempenho da árvore resultante da Modelagem 10 com todos os atributos (as regras de decisão resultantes dessa modelagem se encontram no Apêndice F). Essa árvore possui 197 nós, com 99 nós folhas. O caminho mais longo, em profundidade, possui 23

nós e o mais curto possui 4. O atributo Corte está na raiz da árvore, indicando ser esse o atributo que melhor divide as instâncias da base de dados, ou, dito de outra forma, o atributo que traz maior ganho de informação para a classificação dessas instâncias.

A primeira divisão da árvore separa as instâncias com número de cortes “ ≤ 2 ” (menor ou igual a 2) das instâncias com número de cortes “ > 2 ” (maior que 2). Essa divisão está coerente com os níveis de produtividade apresentados nos *boxplots* da Figura 9, na Seção 4.2.1. Ao longo da árvore são realizados testes referentes aos números de cortes 1, 2 3, 4 e 5. Isso significa que o modelo trata os cortes “ >5 ” (maior que 5) da mesma forma, o que está de acordo com o período em que os canaviais passam a ser menos produtivos, conforme apresentado na Seção 2.1.

Para as instâncias com número de cortes “ ≤ 2 ” (menor ou igual a 2) são testados os valores de precipitação nas quatro fases fenológicas e os valores de temperatura nas fases de crescimento e brotação. Também são testados os valores para os atributos InsumoK, InsumoN, Solo e Variedade.

Para as instâncias com número de cortes “ > 2 ” (maior que 2) são verificados todos os atributos para a obtenção da classe (nó folha), exceto o atributo Fertilidade. Salienta-se que os valores referentes à Fertilidade são sempre iguais ao do atributo Ambiente. Assim, o algoritmo J48 consegue identificar essa correlação e não utiliza um dos atributos. Todos os valores de Ambiente (1 a 5) são testados, indicando a importância da fertilidade do solo para a discriminação do nível de produtividade. Em relação à Textura, foi verificado somente se o valor é igual a 1 ou diferente de 1, implicando que solos com textura 2 ou 3 estão sendo tratados da mesma forma pelo modelo. A textura 1 diz respeito aos solos argilosos, que são os mais férteis e de manejo mais complexo. Foram testados 4 tipos de solos e 8 tipos de variedades. Como se trata de uma árvore binária, várias generalizações são feitas com os atributos categóricos, o que não ocorre com as árvores não binárias, que geram um nó para cada valor dos atributos categóricos. Vale ressaltar que a partir dos atributos Ambiente e Textura é possível identificar grupos de solos, uma vez que esses valores são pré-fixados

para cada tipo de solo. Os atributos Condição de Corte e Tipo de Corte também são testados apenas uma vez, o que ocorre em decorrência da baixa contribuição desses atributos para a obtenção da classe, conforme identificado na Seção 4.4.2.

Os valores do InsumoK são testados em oito nós ao longo de toda a árvore, enquanto os valores do InsumoN são testados em cinco nós e do InsumoP apenas uma vez. Como mencionado na seção anterior, apenas 2% das formulações continham InsumoP, justificando assim sua pequena participação na árvore de decisão.

Em uma análise geral da árvore, pode-se afirmar que o número de cortes foi o fator preponderante para discriminar a classe de uma determinada instância. Além disso, a importância dos atributos climáticos para a produtividade da cana foi expressa no modelo uma vez que condições sobre a precipitação e temperatura estão presentes em muitos dos nós da árvore.

Para avaliação do desempenho do classificador, a ferramenta Weka fornece várias métricas, tanto para o modelo como um todo, como para cada classe. Na Tabela 8, são apresentados os indicadores de desempenho geral, na Tabela 9, os valores referentes a cada classe e, na Tabela 10, apresenta-se a Matriz de confusão para a Modelagem 10.

TABELA 8 - DESEMPENHO GERAL DO CLASSIFICADOR

Métrica	Resultado
Acurácia geral	72,158 %
Estatística <i>Kappa</i>	0,582
Erro médio absoluto	0,261

FONTE: A AUTORA

O valor da acurácia geral do modelo está de acordo com resultados obtidos em outros trabalhos que realizaram mineração de dados por meio da tarefa de classificação, conforme discutido na Seção 4.4.1. O erro médio absoluto também

pode ser considerado adequado, uma vez que valores mais próximos de zero são melhores para essa medida.

Em relação à estatística *Kappa*, embora seu valor esteja na faixa denominada moderada (0,41 a 0,60), pode-se dizer que está bem próximo da faixa de valores classificadas como substancial (0,61 a 0,80). Ressalta-se, entretanto, que em análises mais conservadoras sugerem-se índices superiores a 70%, como aqueles obtidos em Nonato e Oliveira (2013) - coeficientes *Kappa* variando de 0,93 a 0,96 - e em Vieira *et al.* (2012) - coeficiente *kappa* igual a 0,87.

TABELA 9 - DESEMPENHO DO CLASSIFICADOR POR CLASSE

TVP	TFP	Precisão	Classe
0,758	0,121	0,757	'(-inf-76.915]'
0,637	0,192	0,624	'(76.915-98.224361]'
0,770	0,104	0,787	'(98.224361-inf)'

FONTE: A AUTORA

Observa-se na Tabela 9 que o melhor valor de sensitividade ou taxa de verdadeiro positivo (TVP) se deu na classe com maior valor de produtividade, acima de 98,224 toneladas de cana por hectare (TCH), consequentemente essa classe foi a que obteve menor taxa de falsos positivos. A classe com produtividade média (entre 76, 915 e 98,224 TCH) obteve o pior valor de sensitividade, 0,637 (63,7%), ou seja, um pouco menor que o valor de acurácia geral estipulado para este trabalho. Verifica-se também que a taxa de precisão, que indica a qualidade da predição, obteve valores semelhantes aos da sensitividade.

TABELA 10 - MATRIZ DE CONFUSÃO DA MODELAGEM 10

Predita a	Predita b	Predita c	← Classificado como
6902	1781	418	a = '(-inf-76.915]'
1833	5796	1475	b = '(76.915-98.224361]'
379	1717	7007	c = '(98.224361-inf)'

FONTE: A AUTORA

As medidas de desempenho dos classificadores são obtidas a partir dos acertos e erros do modelo, explicitados na Matriz de Confusão (Tabela 10). Ressalta-se que os valores dispostos na diagonal representam os acertos dos classificadores, enquanto os outros valores representam seus erros.

4.6 IMPLEMENTAÇÃO

A ferramenta Weka gera a árvore de decisão e, a partir dessa, as respectivas regras de decisão. Essas informações, apesar de interpretáveis, fazem parte de uma estrutura, em geral, muito grande, por causa do número de atributos presente nos modelos. Desta forma, foi proposto neste trabalho, um roteiro sistematizado, isto é, um processo para dar suporte à criação de uma ferramenta que visa a propiciar aos gestores agrícolas a visualização de cada um dos diversos caminhos de uma árvore de decisão. Assim, essa ferramenta pode ser utilizada no apoio às atividades de planejamento e tomada de decisão do setor sucroenergético. Trata-se de uma atividade de pós-processamento que, apesar de relevante, não tem recebido a devida atenção nos processos de KDD apresentados na literatura.

Na Figura 19, apresenta-se como ilustração, uma árvore de decisão e, na Figura 20, as regras correspondentes a essa árvore, ambas geradas pela ferramenta Weka. Na árvore de decisão as elipses representam as condições, os ramos representam os valores associados às condições e os retângulos representam a classe atribuída aos dados que satisfizeram as condições do nó raiz até o nó folha (classe). Por exemplo, no nó raiz testa-se se a Fertilidade é igual a 1 (=1) ou diferente de 1 ($\neq 1$), para as instâncias com Fertilidade igual a 1 a classe

atribuída é *cluster0*. Nas regras de decisão cada linha diz respeito a uma condição, as regras são encadeadas e terminam na linha que tem a especificação de uma classe.

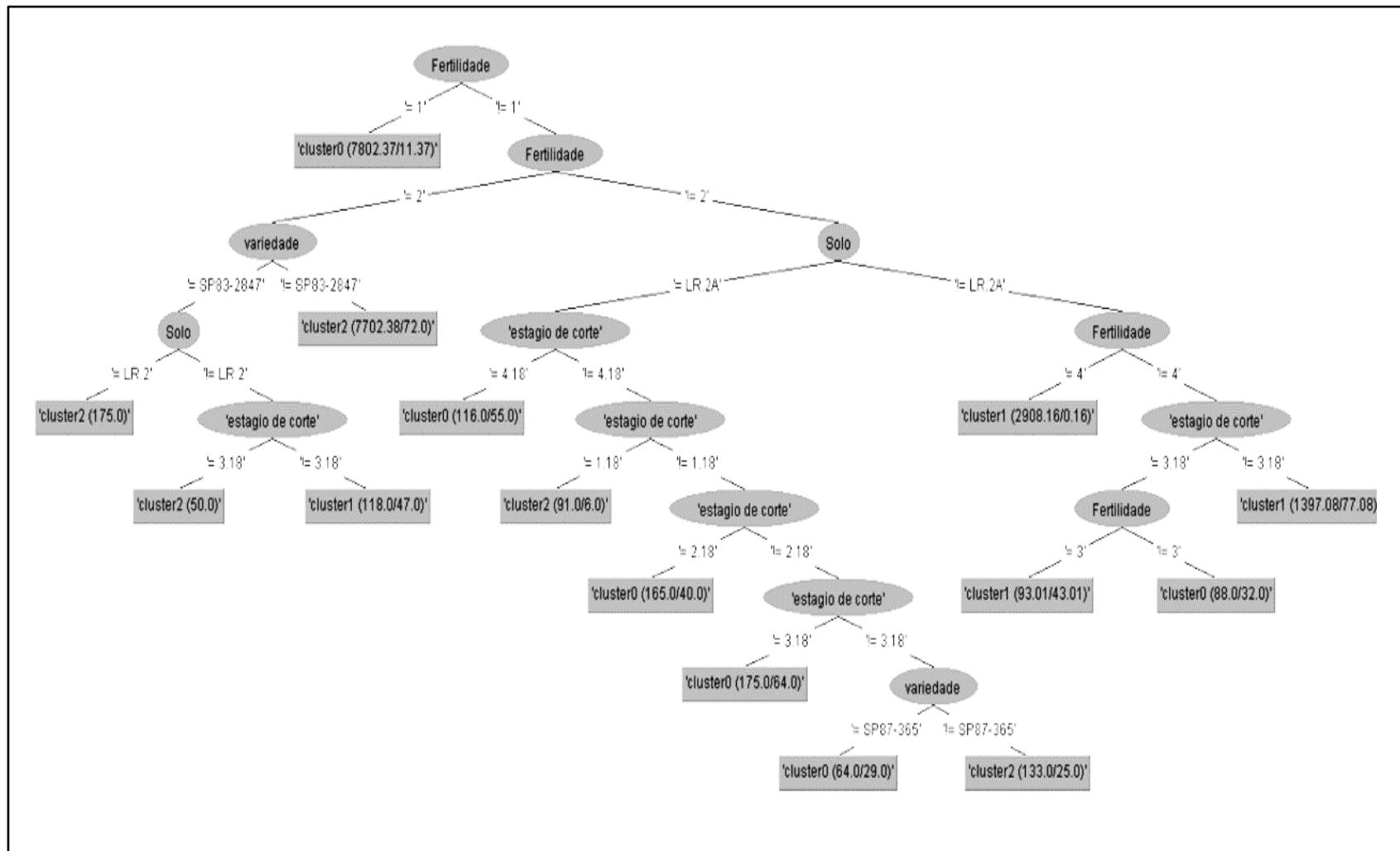


FIGURA 19 - ÁRVORE DE DECISÃO

FONTE: GERADA PELA FERRAMENTA WEKA

```

Fertilidade = 1: cluster0 (7802.37/11.37)
    Fertilidade != 1
        | Fertilidade = 2
            | variedade = SP83-2847
                | Solo = LR.2: cluster2 (175.0)
                    | Solo != LR.2
                | estagio de corte = 3.18: cluster2 (50.0)
                | estagio de corte != 3.18: cluster1 (118.0/47.0)
            | variedade != SP83-2847: cluster2 (7702.38/72.0)
                | Fertilidade != 2
                    | Solo = LR.2A
                | estagio de corte = 4.18: cluster0 (116.0/55.0)
                    | estagio de corte != 4.18
                | estagio de corte = 1.18: cluster2 (91.0/6.0)
                    | estagio de corte != 1.18
                | estagio de corte = 2.18: cluster0 (165.0/40.0)
                    | estagio de corte != 2.18
                | estagio de corte = 3.18: cluster0 (175.0/64.0)
                    | estagio de corte != 3.18
                | variedade = SP87-365: cluster0 (64.0/29.0)
                | variedade != SP87-365: cluster2 (133.0/25.0)
                    | Solo != LR.2A
    | Fertilidade = 4: cluster1 (2908.16/0.16)
        | Fertilidade != 4
            | estagio de corte = 3.18
            | Fertilidade = 3: cluster1 (93.01/43.01)
            | Fertilidade != 3: cluster0 (88.0/32.0)
        | estagio de corte != 3.18: cluster1 (1397.08/77.08)

```

FIGURA 20 - REGRAS DE DECISÃO

FONTE: GERADA PELA FERRAMENTA WEKA

No próximo capítulo, apresenta-se a proposta de um roteiro sistematizado e a criação de uma ferramenta *Web* para a visualização e exploração de cenários de produção da cana, criados a partir de árvore/regras de decisão.

4.7 SÍNTESE DOS PROCEDIMENTOS REALIZADOS NO PROCESSO DE KDD

O processo de KDD realizado neste trabalho seguiu as etapas do modelo CRISP-DM, relatadas em detalhes nas seções anteriores desse capítulo. O Quadro 17 tem o propósito de apresentar, de forma sumarizada, todos os procedimentos realizados.

QUADRO 17 - SÍNTSE DOS PROCEDIMENTOS REALIZADOS NO MODELO CRISP-DM

Etapa CRISP-DM	Atividade	Ferramenta/Método
Entendimento do negócio	Entrevista semi-estruturada	5W1H
Entendimento dos dados Escolha dos atributos	Elaboração lista inicial de atributos	Pesquisas/especialistas
	Seleção dos atributos	Método Delphi
Entendimento dos dados Atividades de exploração	Inspeção visual das planilhas	Filtros da Planilha Eletrônica (MS-Excel)
	Gráficos dos dados climáticos	Planilha Eletrônica (MS-Excel)
	Boxplots da produtividade	Matlab
Preparação de dados: Limpeza	Eliminação/correção de dados inconsistentes	Filtros - Planilha Eletrônica (MS-Excel)
Preparação de dados: Transformação	Cálculo dos valores para os insumos N, P e K	Planilha Eletrônica (MS-Excel)
	Uniformização de Siglas (manual e mecanizado)	Planilha Eletrônica (MS-Excel)
	Cálculo da precipitação acumulada por fase fenológica	Planilha Eletrônica (MS-Excel)
	Cálculo da média da temperatura por fase fenológica	Planilha Eletrônica (MS-Excel)
	Discretização do atributo produtividade	Ferramenta Weka
	Transformação de numérico para categórico - atributos Ambiente, Fertilidade e Textura	Ferramenta Weka
	Retirada dos atributos de identificação: Fazenda, Gleba e Talhão	Ferramenta Weka
Preparação de dados: Integração	Integração das planilhas dos 4 períodos	Planilha Eletrônica (MS-Excel)
	Inserção dos dados climáticos	Planilha Eletrônica (MS-Excel)
Modelagem	Parametrização de 10 experimentos: árvore binária/não binária e número de instâncias por nó	Ferramenta Weka
	Eliminação de outliers e realização dos 10 experimentos de modelagem	Planilha Eletrônica (MS-Excel) Ferramenta Weka
	Aplicação de métodos de seleção de atributos	Weka
	Realização dos 10 experimentos de modelagem com seleção de atributos	Weka
Avaliação	Análise de métricas de desempenho do classificador	Weka
	Análise da estrutura da árvore	-
Implementação	Roteiro sistematizado	-
	Desenvolvimento de ferramenta de visualização	Ruby Ruby on rails SQL Server

FONTE: A AUTORA

5 PROPOSTA DO ROTEIRO SISTEMATIZADO

Neste capítulo, é apresentado o Roteiro Sistematizado, objetivo principal deste trabalho, cuja necessidade foi identificada na RSL realizada. Na Seção 5.1, o roteiro é descrito em cinco passos, para cada um apresenta-se: *i*) uma breve descrição do passo; *ii*) as entradas esperadas para a realização desse passo; *iii*) o processamento que deve ser efetuado e; *iv*) as saídas que devem ser geradas após a execução do passo.

Na Seção 5.2, relata-se o desenvolvimento da ferramenta de apoio à decisão que implementa o Roteiro Sistematizado a partir da realização de um processo de KDD. O desenvolvimento dessa ferramenta concretiza um dos objetivos específicos propostos neste trabalho.

5.1 DESCRIÇÃO DO ROTEIRO SISTEMATIZADO

Para que um sistema informatizado possa manipular as informações presentes em uma árvore de decisão gerada a partir de um processo de KDD, propiciando não só a interação, mas principalmente a verificação dos impactos causados por mudanças em determinados atributos, essa árvore deve ser transformada e armazenada em estruturas que possibilitem a visualização de cada um de seus caminhos de maneira isolada, de maneira que possa apoiar os gestores agrícolas em suas atividades de tomada de decisão.

Essa transformação implica processos de interpretação da árvore gerada pela ferramenta de mineração de dados. Após análise, detectou-se a possibilidade de efetuar o armazenamento do arquivo texto gerado contendo a árvore de decisão, representada na forma de regras de decisão, em uma estrutura de tabela em que os registros representam os nós da árvore de decisão e seus respectivos antecessores.

Com isso, após a implementação de um sistema automatizado de consulta a essa estrutura gerada, será possível a visualização de múltiplos cenários, pela

seleção de valores dos diversos atributos, ou pela seleção de níveis de produtividade, pelo próprio usuário. Esse processo está representado na Figura 21.

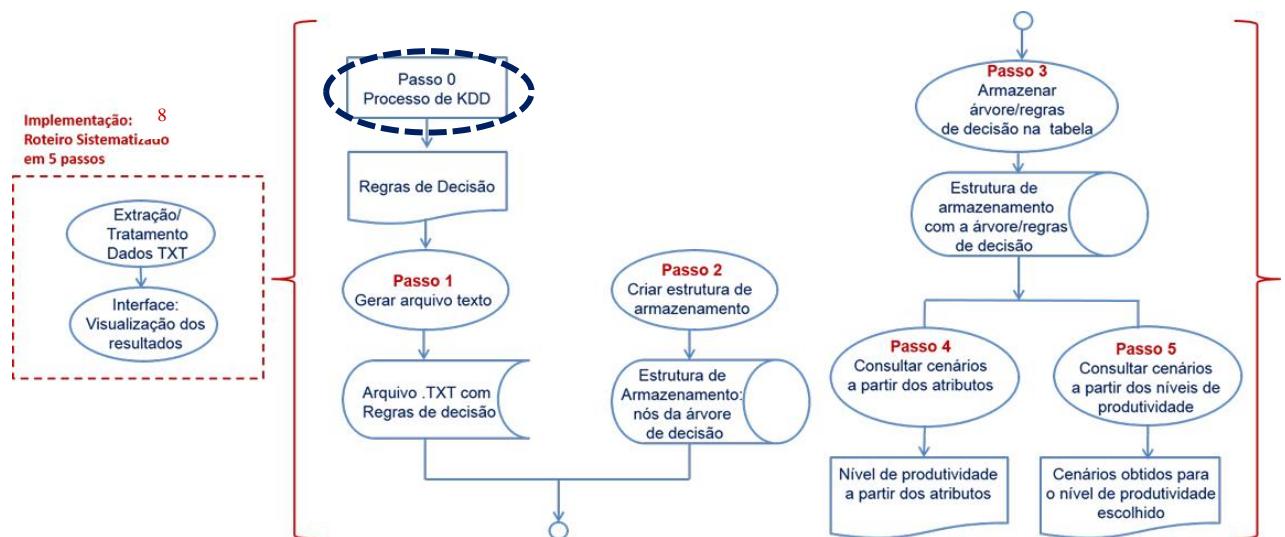


FIGURA 21 - FLUXO DE EXECUÇÃO DO ROTEIRO SISTEMATIZADO

FONTE: A AUTORA

O roteiro deve se iniciar sempre após a realização de um processo de KDD, por essa razão essa etapa é denominada de Passo 0. Destaca-se que os Passos 1 e 2 podem ser realizados de forma intercambiável, ou seja: Passo 1, Passo 2 ou Passo 2, Passo 1. O mesmo ocorre em relação aos Passos 4 e 5. Os detalhes do Roteiro Sistematizado são descritos a seguir.

Descrição do Roteiro Sistematizado:

Passo 0 – Realização do processo de KDD, segundo modelo de referência CRISP-DM. Esse passo já deve ter sido executado para iniciar o roteiro, por isso está tracejado na Figura 21.

- **Entrada:** dados para o processo de KDD.
- **Processamento:** realizar KDD – tarefa de classificação.
- **Saída:** árvore e regras de decisão.

⁸ Etapas destacadas dentro do quadro tracejado em vermelho na Figura 6 da Seção 3.2

Passo 1 – Gerar arquivo texto contendo regras de decisão, obtidas a partir de uma árvore de decisão resultante do processo de KDD. As árvores de decisão podem ser facilmente transformadas em regras de decisão; desta forma, mesmo que a ferramenta de mineração não dê suporte a essa transformação, ela poderá ser realizada sem complexidade.

- **Entrada:** árvore de decisão expressa no formato de regras de decisão.
- **Processamento:** utilizar mecanismo disponível na ferramenta de mineração de dados para gerar arquivo texto contendo regras de decisão.
- **Saída:** arquivo texto (formato TXT).

Passo 2 – Criar a estrutura de armazenamento (tabela) para armazenar os nós referentes às condições da árvore de decisão.

- **Entrada:** árvore de decisão expressa no formato de regras de decisão.
- **Processamento:** criar tabela para receber cada nó da árvore de decisão e um caminho do nó atual até o nó raiz.
- **Saída:** tabela contendo a árvore/regras de decisão.

Passo 3 – Armazenar as regras de decisão na estrutura previamente criada, de forma a garantir todos os percursos da árvore. O algoritmo responsável por essa ação deverá ser capaz de reconhecer a condição presente em cada nó da árvore, bem como seu nó antecessor e sucessor.

- **Entrada:** arquivo texto (formato TXT) com árvore/regras de decisão.
- **Processamento:** ler e interpretar cada linha do arquivo texto de forma a identificar:
 - ✓ Informações a respeito da árvore;
 - ✓ Caracteres que devem ser desconsiderados;
 - ✓ Nó raiz, nós internos e nós folha (atributo classe);
 - ✓ Nó antecessores do nó atual.
- **Saída:** estrutura de armazenamento contendo as regras de decisão.

Passo 4 – Consultar os diversos cenários a partir dos atributos.

- **Entrada:** estrutura de armazenamento (tabela) com árvore/regras de decisão e opção do usuário.
- **Processamento:** o usuário deve escolher um valor para um dos atributos que estarão disponíveis. A partir dessa escolha, será realizada uma busca na estrutura de armazenamento para selecionar todos os registros (nós da árvore) que atendam à seleção. Esse procedimento restringe os valores possíveis para os demais atributos. Sucessivas escolhas devem ser realizadas para os atributos restantes. Em cada escolha, os valores dos outros atributos, ainda não selecionados, são reduzidos de acordo com os valores dos atributos já selecionados. Esse processo é finalizado quando forem escolhidos todos os valores possíveis para cada atributo disponível para seleção. Trata-se de um procedimento de busca em uma árvore, que possui estrutura hierárquica. Assim, se o atributo escolhido não for o nó raiz, os níveis anteriores ao nó selecionado serão apenas informados, não sendo possível realizar nenhuma escolha. O processo de seleção de valores se dará para os níveis posteriores ao primeiro atributo (nó) selecionado.
- **Saída:** nível de produtividade resultante considerando os valores selecionados.

Passo 5 – Consultar os diversos cenários a partir dos níveis de produtividade.

- **Entrada:** estrutura de armazenamento com regras de decisão e opção do usuário.
- **Processamento:** o usuário deve escolher um valor para um dos níveis de produtividade que estarão disponíveis. A partir dessa escolha, será realizada uma busca na estrutura de armazenamento para selecionar todos os registros que atendam esse nível de produtividade. A escolha pelo nível de produtividade implica a seleção de todos os nós folhas que atendem aquele nível de produtividade; assim, haverá várias opções de caminhos até à consolidação do caminho completo, que se dará no nó raiz. O usuário poderá então visualizar os diversos cenários que atendam ao nível de escolhido de produtividade.

- **Saída:** Cenários completos, ou seja, atributos e seus respectivos valores, considerando nível de produtividade selecionado.

Na próxima seção, apresenta-se o processo de desenvolvimento de uma ferramenta de visualização dos cenários da cana que dá suporte às atividades propostas no roteiro sistematizado.

5.2 DESENVOLVIMENTO E IMPLEMENTAÇÃO DA FERRAMENTA DE VISUALIZAÇÃO

Ao sistema automatizado de consulta e visualização foi dado o nome de SECC – Sistema de Exploração dos Cenários da Cana. A escolha do nome deriva-se da potencialidade do sistema em apresentar diversos cenários de produção da cana-de-açúcar. A ferramenta desenvolvida implementa os passos 3, 4 e 5 do Roteiro Sistematizado (Seção 5.1), conforme esquematizado na Figura 22.

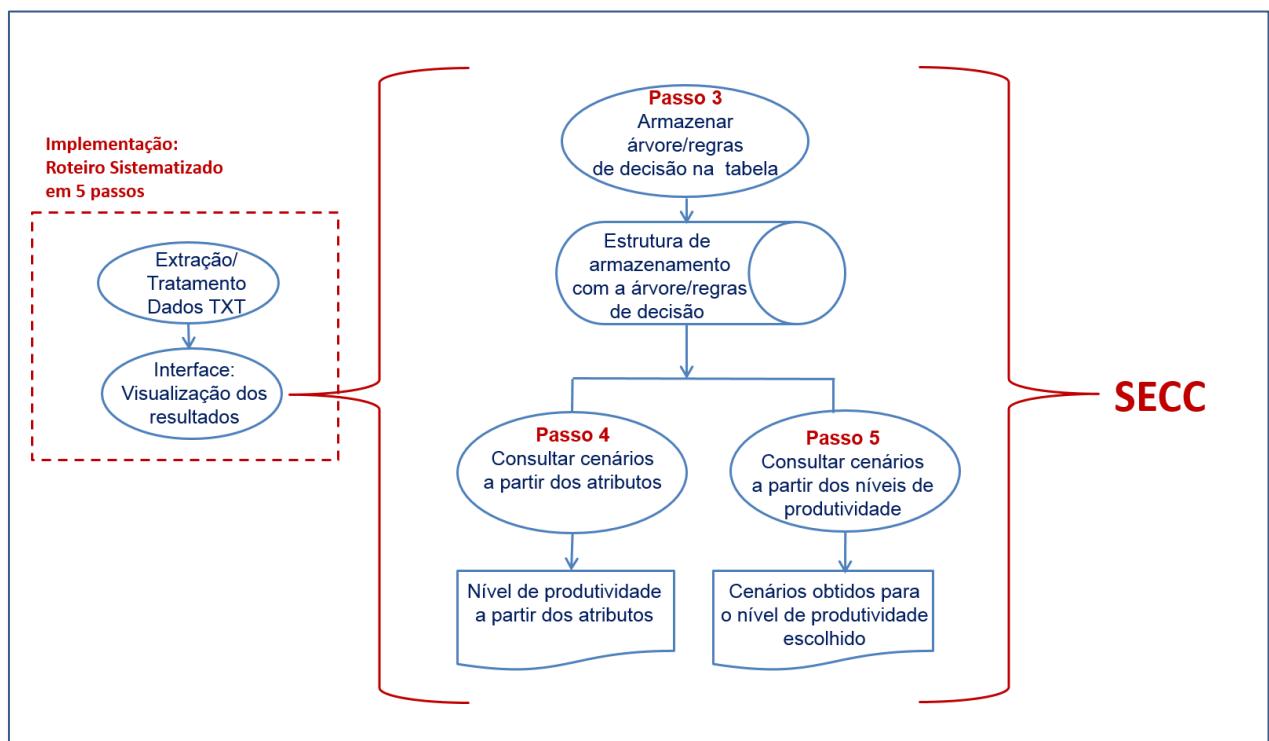


FIGURA 22 - PASSOS DO ROTEIRO IMPLEMENTADOS NA FERRAMENTA SECC

FONTE: A AUTORA

O SECC tem por objetivo facilitar a pesquisa das diversas combinações possíveis dos atributos e dos níveis de produtividade obtidos com essas

combinações. Trata-se de um sistema de apoio à decisão, que poderá ser utilizado como suporte para diversas atividades de planejamento da cadeia de valor da cana. Esse sistema poderá ser utilizado para a realização do planejamento de plantio e de tratos, identificando cenários de baixa produtividade e propondo ações que promovam seu aumento. Outro exemplo de utilização do sistema e do conhecimento gerado é para atividades de ampliação de área de plantação da cana. Pode-se escolher um cenário semelhante às características dessa nova área e, assim, identificar o nível de produtividade esperado, possibilitando, dessa forma, prospectar a quantidade de veículos para transporte da cana colhida, ou identificar as variedades que podem ser plantadas, ou a quantidade de adubo que precisará ser comprado e assim por diante. A identificação do nível de produtividade associado aos cenários de produção pode ainda contribuir para o processo de valoração na aquisição de novas áreas.

A necessidade de uma ferramenta para a visualização dos resultados de árvores de decisão, obtidos em um processo de mineração de dados, foi identificada a partir da RSL realizada, apresentada nos Capítulos 2 e 3. Foram encontrados quatro trabalhos que utilizaram ferramentas (ARCGIS, SatImagExplore, Agri-remoto, Sistema de Recomendação da Agência de Informação EMPRAPA) de visualização dos resultados da mineração de dados e nenhuma delas apresentava os resultados obtidos a partir de árvores de decisão de forma a propiciar a visualização de cada um dos caminhos dessas árvores de maneira isolada, reduzindo assim sua complexidade de interação e interpretação. Os requisitos funcionais da ferramenta computacional foram identificados por meio do desenvolvimento dos passos explicados no roteiro sistematizado (Seção 5.1).

A Figura 23 apresenta, esquematicamente o processo de desenvolvimento do SECC, que será descrito em detalhes nas próximas seções.

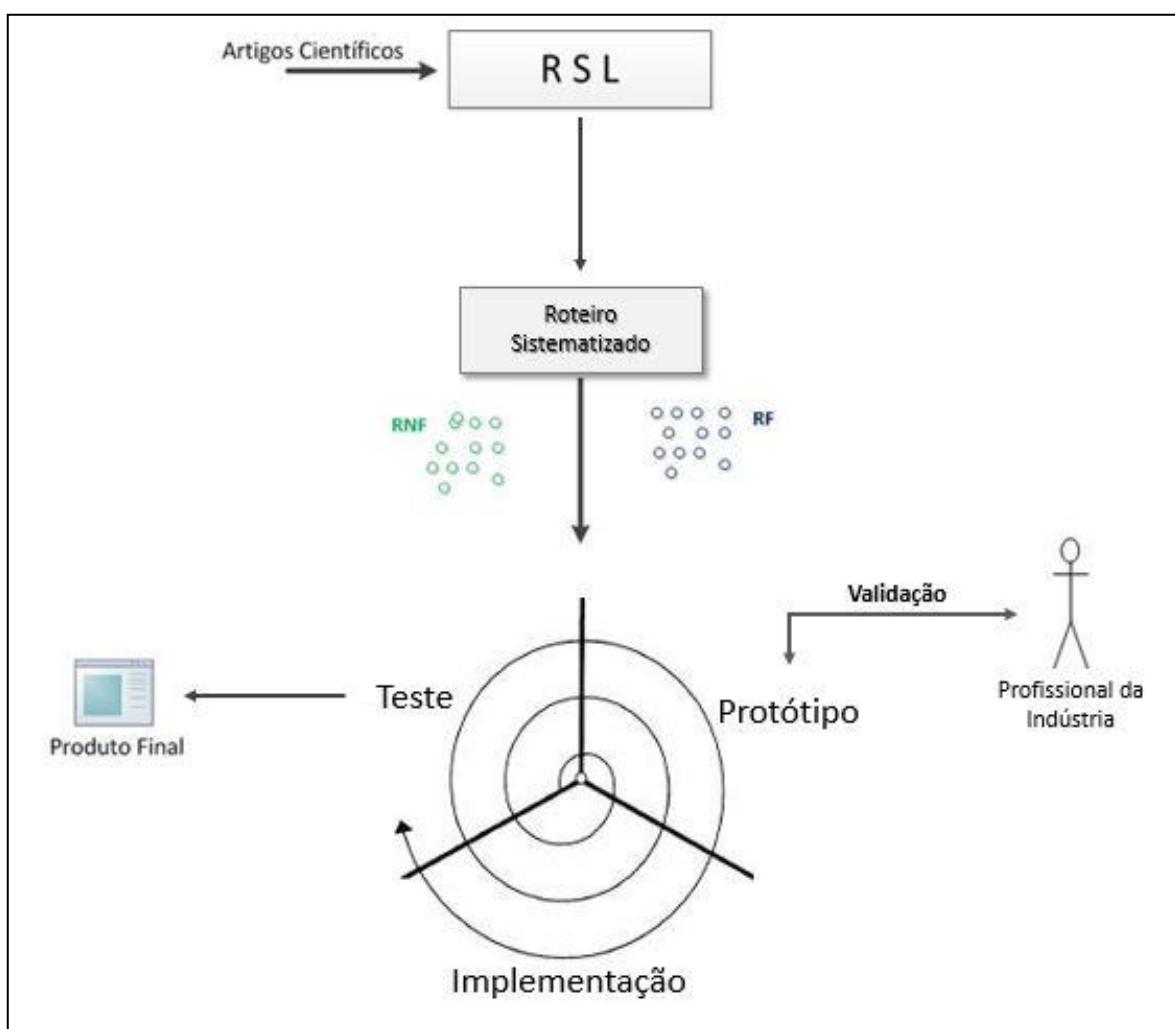


FIGURA 23 - PROCESSO DE DESENVOLVIMENTO DO SECC

FONTE: A AUTORA

O processo de desenvolvimento foi adaptado do processo espiral (PRESSMAN e MAXIM, 2016).

A definição da interface se deu por meio de protótipos de baixa fidelidade (SHARP, ROGERS e PREECE, 2013), validados pelo profissional da indústria caracterizada no Capítulo 4.

A arquitetura do software seguiu o padrão MVC (*Model, View, Controller*). Nesse padrão, a aplicação é dividida em 3 camadas: a camada de manipulação dos dados da aplicação, regras de negócios, lógica e funções (**Model**); a camada de interação do usuário que realiza a exibição dos dados (**View**), e a

camada de controle, que faz a mediação da entrada, convertendo-a em comandos para o modelo ou para a visão (**Controller**) (GAMMA, HELM e JOHNSON, 2008)

Para as atividades de teste da ferramenta SECC, foi utilizado teste funcional, ou teste de “caixa preta”, que se baseia nos requisitos funcionais do software para derivar casos de teste (MYERS, SANDLER e BADGETT, 2011).

O sistema foi desenvolvido utilizando-se ferramentas “livres” e de “código aberto”, pois, desta forma, as empresas não terão custos adicionais com sua implementação. É uma ferramenta *Web* com *layout* responsivo, ou seja, preparado para se adaptar ao formato de diversos dispositivos, como *tablets*, *smartphones* ou *notebooks*.

5.2.1 REQUISITOS FUNCIONAIS E NÃO FUNCIONAIS

Os requisitos funcionais (RF) considerados no desenvolvimento do sistema foram:

1. **RF1:** Gerenciar usuários
2. **RF2:** Realizar *login* (*login/senha*);
3. **RF3:** Carregar arquivo contendo a árvore/regras de decisão;
4. **RF5:** Interpretar regras de decisão e inserir dados na tabela;
5. **RF6:** Visualizar cenário a partir dos atributos;
6. **RF7:** Visualizar cenário a partir da produtividade;
7. **RF8:** Salvar cenários;
8. **RF9:** Gerenciar cenários salvos.

Foram considerados os seguintes requisitos não funcionais (RNF) no desenvolvimento do sistema:

- **RNF1:** Acesso à conexão *Web*;
- **RNF2:** Desempenho - para conforto do usuário e agilidade de manuseio da ferramenta é necessário que essa tenha a capacidade de processar e salvar o arquivo em um curto intervalo de tempo (menos de dois segundos);

- **RNF3:** Usabilidade - considerando que um dos focos do sistema se refere à tomada de decisão, é indispensável que o *layout* seja responsivo, para ser utilizado em *smartphones* ou *tablets*;
- **RNF4:** Segurança - como o sistema será alimentado com informações críticas de empresas, é fundamental que haja segurança durante o armazenamento de dados;
- **RNF5:** Mensagens de retorno de erros amigáveis aos usuários (gestores e técnicos da produção agrícola).

5.2.2 INTERAÇÕES DO USUÁRIO E FUNCIONALIDADES DO SISTEMA

Para melhor compreensão do escopo completo do sistema, foi desenvolvido o diagrama de casos de uso (RUMBAUGH, BOOCH e JACOBSON, 2012), apresentado na Figura 24. O diagrama de casos de uso é um artefato resultante da modelagem UML (*Unified Modeling Language*) e que representa as interações do usuário com o sistema, assim como suas funcionalidades.

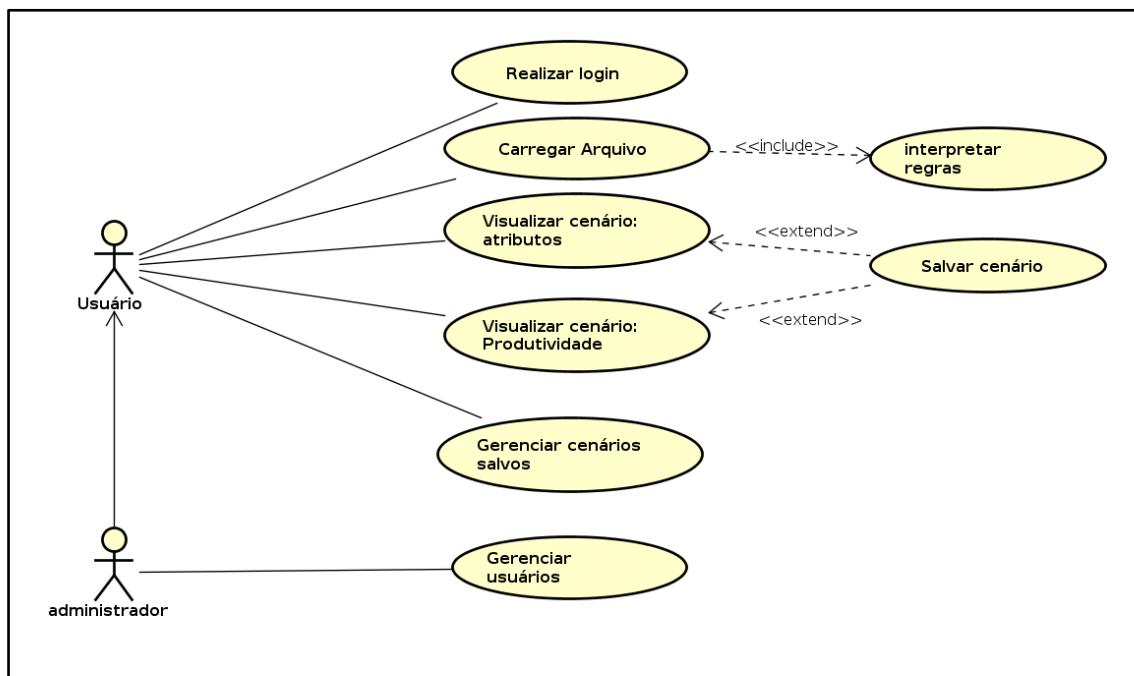


FIGURA 24 - DIAGRAMA DE CASOS DE USO

FONTE: A AUTORA

Cada uma das funções apresentadas no diagrama da Figura 24 é descrita a seguir:

Gerenciar usuários: esse módulo permite a criação, alteração e exclusão de usuários a partir de um *login* de administrador. Para a criação de um usuário devem ser informados um *e-mail* e uma senha, que serão gravados na tabela de usuários.

Realizar login (*login/senha*): os usuários poderão realizar o *login* no sistema por meio de uma senha enviada no e-mail, e como primeira ação é necessário trocar a senha. Outra funcionalidade do sistema de *login* é a opção “esqueceu a senha” que apagará a antiga e enviará uma nova por *e-mail*.

Carregar arquivo: após o *login*, o usuário poderá visualizar suas árvores já gravadas e também optar por carregar um novo arquivo. Nesse caso será disponibilizada uma tela para que o usuário insira o caminho e o nome do arquivo. Se o arquivo for válido, o sistema irá mostrar algumas informações básicas do arquivo e uma opção para visualizar os possíveis cenários de produtividade (opção “Navegar”, no protótipo da tela). Esse módulo implementa parte do **Passo 3** do Roteiro Sistematizado.

Interpretar regras: após a carga do arquivo no sistema, o módulo “Interpretar regras” será ativado. Esse módulo lê o arquivo texto, reconhece as informações gerais sobre a árvore e grava essas informações em uma tabela própria. Em seguida, cada linha do arquivo texto que representa uma regra de decisão (um nó da árvore de decisão), será armazenada na tabela de nós, que conterá também todos os nós antecessores ao nó atual, para que os caminhos da árvore possam ser recuperados pelos módulos de consulta. Esse módulo também está associado ao **Passo 3** do Roteiro Sistematizado.

Visualizar cenário a partir dos atributos: esse módulo permite a construção de cenários. Inicialmente é apresentada uma lista que contém todos os atributos presentes na árvore de decisão previamente gravada. O usuário poderá selecionar os atributos e os valores para esses atributos, porém apenas os valores que constam na árvore de decisão podem ser apresentados. Ao final da seleção dos atributos será obtido um cenário completo, um caminho do nó raiz até um nó folha. Esse módulo implementa o **Passo 4** do Roteiro Sistematizado.

Visualizar cenário a partir da produtividade: esse módulo permite a visualização de cenários completos a partir de um nível de produtividade. Uma vez escolhido o nível de produtividade, todos os cenários daquele nível estarão disponíveis para visualização do usuário, que poderá escolher quais deseja ver em detalhes. Esse módulo está associado ao **Passo 5** do Roteiro Sistematizado.

Salvar cenário: esse módulo será executado se o usuário desejar salvar o cenário gerado a partir de um dos módulos para sua visualização. Cada cenário é um dos caminhos de uma única árvore de decisão associada ao usuário corrente. O sistema grava uma linha na tabela de cenários contendo o caminho completo do cenário atual.

Gerenciar cenários salvos: esse módulo permite a visualização ou exclusão dos cenários já gravados do usuário que está *logado*.

5.2.3 CARACTERÍSTICAS DA ESTRUTURA DE ARMAZENAMENTO

A estrutura do banco de dados é representada pelo diagrama da Figura 25. A entidade **Usuário** armazena os dados de cada usuário para que seja possível realizar *login* no sistema. Cada usuário pode ter várias árvores associadas a ele, a entidade **Árvore** armazena as características gerais de cada árvore (atributos, tamanho da árvore, número de nós folha e nome atribuído pelo usuário). Uma árvore possui diversos nós, registrados na entidade **Nó**. Cada nó representa uma condição em relação a um dos atributos do modelo de produtividade. O campo **condição** armazena a condição referente a um atributo específico, armazenado no campo denominado **variável**. O campo **nível** regista o nível de profundidade do nó na árvore. O campo **ancestral** tem como objetivo salvar os ancestrais do nó atual, assim possibilitando a pesquisa recursiva até à raiz da árvore. O campo **resultado** informa se o nó atual é um nó folha ou não. A entidade **Cenário** armazena cada caminho do nó raiz a um nó folha, está ligada a uma árvore de um usuário específico. Cada instância dessa entidade representa um dos possíveis cenários de produção de uma única árvore. A criação da estrutura de armazenamento está associada ao **Passo 2** do Roteiro Sistematizado.

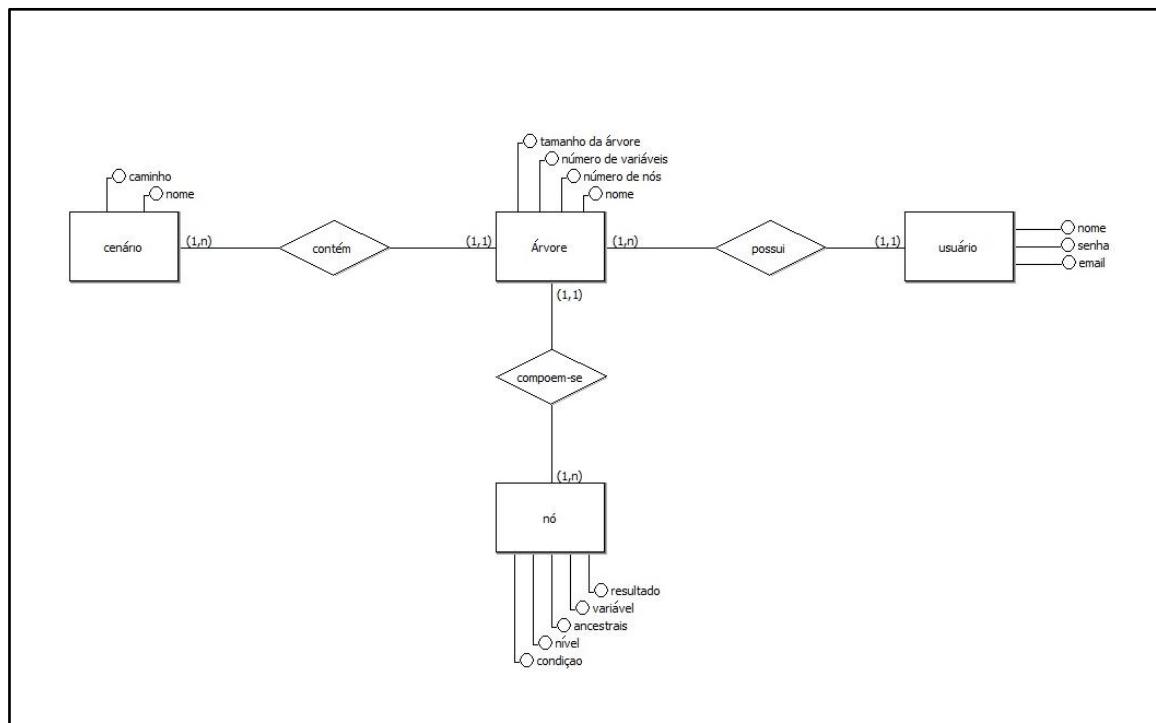


FIGURA 25 - MODELO LÓGICO DO BANCO DE DADOS

FONTE: A AUTORA

5.2.4 PROTÓTIPO DO SISTEMA

O *layout* do sistema foi projetado visando a simplificar ao máximo a visualização de resultados obtidos no processo de descoberta de conhecimento. Dessa forma, a estrutura da *interface* buscou:

- Ser de fácil compreensão para os gestores agrícolas, reduzindo ao máximo a quantidade de informações irrelevantes na tela de decisão;
- Ser minimalista;
- Utilizar o menor número de *clicks* para atingir o objetivo.

Os protótipos das telas referentes ao gerenciamento dos usuários e à tela de *Login* são apresentados no Apêndice G. Os protótipos a seguir dizem respeito às telas para utilização efetiva do sistema.

Na Figura 26, apresenta-se a tela para exibir os arquivos já carregados. Nessa tela, é apresentada a opção para carregar um novo arquivo texto contendo uma árvore de decisão de um usuário específico. O nome do usuário *logado* ficará exposto no canto superior da tela. A opção “Navegar” permite que o usuário vá para a tela de visualização de cenário, e a opção “Excluir” permite a exclusão de uma árvore previamente gravada.

SECC		Meus Cenários					Graca	Ações	▼					
Meus Arquivos >														
Meus Arquivos														
Cod	Usuario	Nome do Arquivo	Número de instâncias	Número de variáveis	Tamanho da arvore	Ações								
1	Graca	Teste de Arquivo	51	11	101	Navegar Excluir								

FIGURA 26 - PROTÓTIPO DA TELA DE GERENCIAMENTO DE ARQUIVOS

FONTE: A AUTORA

Uma vez selecionada a opção de carga de um novo arquivo, a tela ilustrada na Figura 27 é disponibilizada para que sejam informados o local e o nome do arquivo.

Conforme descrito no roteiro (Seção 5.1), existem dois tipos de análises possíveis, iniciando pelos atributos de produção e tendo como resultado um nível de produtividade, ou especificando um nível de produtividade e obtendo os diversos cenários que atendam esse nível. Na Figura 28, representa-se a consulta que se inicia pelos atributos de produção. A princípio, todos os atributos disponíveis na árvore que se está consultando são apresentados. Os valores possíveis para cada atributo estarão disponíveis para seleção. Em cada nova seleção a partir da primeira, os atributos e seus respectivos valores serão filtrados. Esse processo se repetirá até que um cenário completo (de um nó raiz até um nó folha) seja definido e o nível de produtividade para esse cenário seja

obtido. Ressalta-se que os valores disponibilizados para seleção são apenas aqueles presentes na árvore de decisão corrente.

O protótipo é uma interface web com o seguinte layout:

- Cabeçalho:** Contém o logo "SECC" e links para "Meus Arquivos" e "Meus Cenários". À direita, uma barra com o nome "Graca".
- Navegação:** Mostra o caminho "Meus Arquivos > Novo Arquivo >".
- Título:** "Novo Arquivo".
- Campos:** Um campo "Arquivo:" com uma interface "Browser" e o placeholder "sem arquivo selecionado".
- Botões:** Um botão azul "Enviar".

FIGURA 27 - PROTÓTIPO DA TELA DE IDENTIFICAÇÃO DE NOVO ARQUIVO

FONTE: A AUTORA

Por exemplo, na Figura 28, os valores disponíveis para tipo de solo seriam LR.2 e LVE.2A. Ao ser escolhido um desses dois valores, só serão disponibilizados para a próxima seleção os atributos que façam parte de um dos caminhos que levem desse atributo/valor até um nó folha. No Apêndice B, são apresentados todos os valores possíveis para os atributos utilizados neste trabalho. A opção “Salvar” presente nesta tela permite que o usuário salve o cenário obtido após suas escolhas de atributos/valores.

A interface do usuário para a consulta de atributos de produção. No topo, uma barra com o logo "SECC", links para "Meus Arquivos" e "Meus Cenários", e uma opção "Graça". Abaixo, uma barra de navegação com "Meus Arquivos > Visualização de Variáveis e Condições >". Um formulário central intitulado "Selecione Tipo de Análise:" com opções para "Ambiente de Produção" e "Produtividade". A seção "Variáveis:" contém uma lista com "Solo", "Variedade", "Tipo de corte", "Condicaes de corte", "FormulaAdubo", "Adubacao", "Ambiente", "Fertilidade" e "Textura". Um campo "Selecionar" com uma seta para baixo mostra "L.R.2" e "LVE.2A", com "LVE.2A" ressaltado em cinza. À direita, uma seção "Resultado:" com um botão "Salvar".

FIGURA 28 - PROTÓTIPO DA CONSULTA A PARTIR DOS ATRIBUTOS DE PRODUÇÃO

FONTE: A AUTORA

Na Figura 29, representa-se a consulta que se inicia pelo nível de produtividade. Uma vez definido o nível de produtividade, todos os cenários (um caminho completo do nó raiz até um nó folha), que atendam aquele nível de produtividade, são apresentados para conhecimento do usuário. Assim como na tela da Figura 28, a opção “Salvar” permite que os cenários obtidos com o nível de produtividade escolhido, sejam gravados.

A interface do usuário para a consulta de nível de produtividade. No topo, uma barra com o logo "SECC", links para "Meus Arquivos" e "Meus Cenários", e uma opção "Graça". Abaixo, uma barra de navegação com "Meus Arquivos > Visualização de Variáveis e Condições >". Um formulário central intitulado "Selecione Tipo de Análise:" com opções para "Ambiente de Produção" e "Produtividade" (selecionada). A seção "Produtividade" contém três opções: "Alta", "Média" (resaltada em cinza) e "Baixa". À direita, três caixas exibem detalhes de cenários: "Cenário 1" (1 - Ambiente = 1, 2 - Textura = 1), "Cenário 2" (1 - Ambiente = 1, 2 - Textura != 1, 3 - estagio de corte = 4, 18, 4 - variedade = SP91-1049) e "Cenário 3" (1 - Ambiente = 1, 2 - Textura != 1, 3 - estagio de corte = 4, 18, 4 - variedade != SP91-1049, 5 - Adubacao <= 0.037352). Um botão "Salvar" está no lado direito.

FIGURA 29 - PROTÓTIPO DA CONSULTA A PARTIR DO NÍVEL DE PRODUTIVIDADE

FONTE: A AUTORA

Na Figura 30, apresenta-se a tela que permite visualização ou exclusão dos cenários anteriormente gravados.

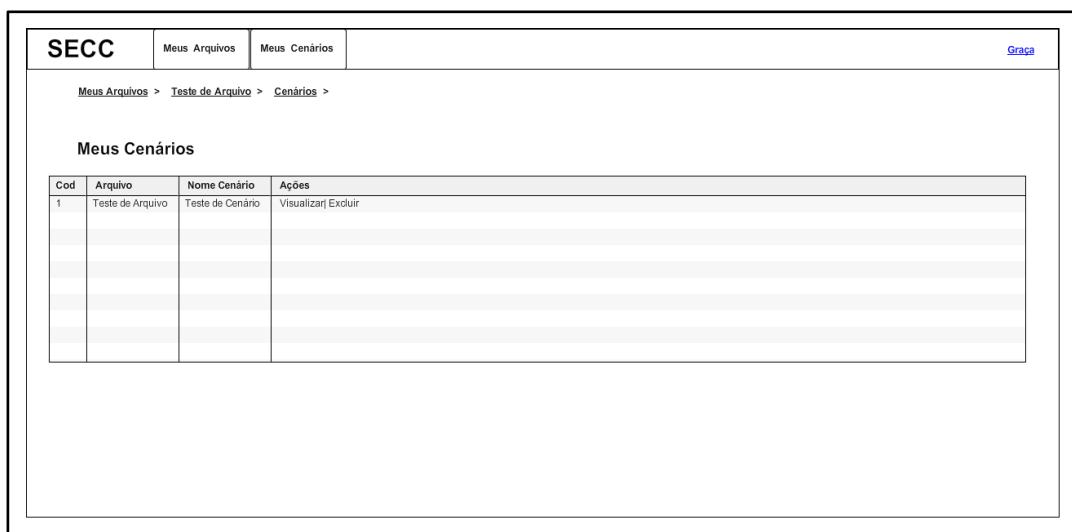


FIGURA 30 - PROTÓTIPO DA CONSULTA DOS CENÁRIOS DISPONÍVEIS

FONTE: A AUTORA

Após a escolha de um cenário específico, apresenta-se o detalhamento desse cenário, conforme pode ser visto na Figura 31.

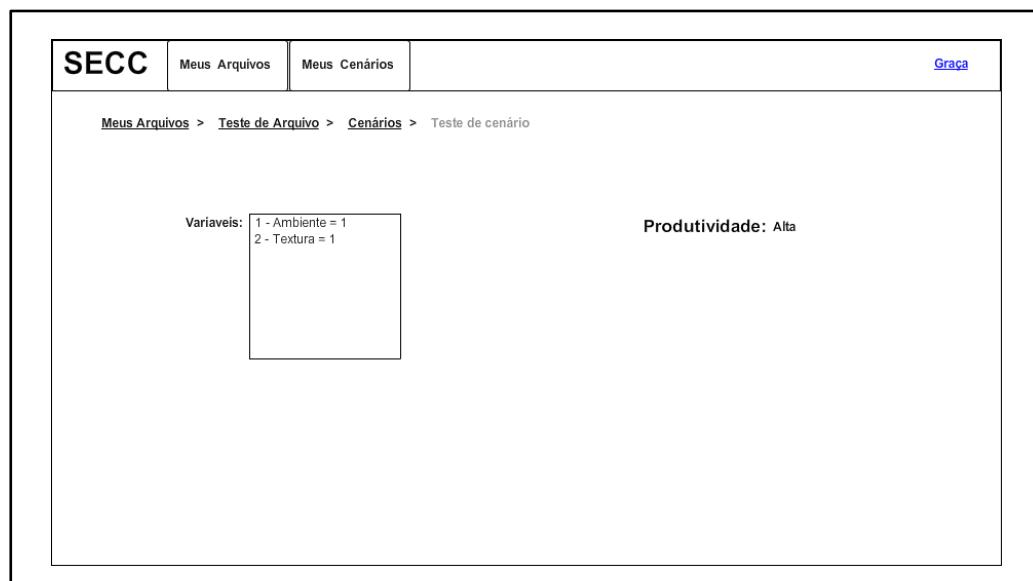


FIGURA 31 - PROTÓTIPO DO DETALHAMENTO DOS CENÁRIOS DISPONÍVEIS

FONTE: A AUTORA

5.2.5 ATIVIDADES DE TESTE

Para as atividades de teste da ferramenta SECC, foi utilizada a técnica de teste funcional. O objetivo desse tipo de teste é verificar a adequação ou não do software à sua especificação. Os principais tipos de erros que são alvos do teste funcional são: funções incorretas, erros de interface, erros na estrutura de dados ou acesso à base de dados, erros de desempenho e erros de inicialização e término (PRESSMAN e MAXIM, 2016).

Existem diversos critérios de testes funcionais. Neste trabalho, foi feita uma adaptação do critério de teste Análise de Valores Limites (MYERS, SANDLERE e BADGETT, 2011). Nesse critério, o domínio dos valores de entrada é dividido em classes, e em seguida são escolhidos valores abaixo, acima e nos limites dessas classes para serem testados. Além disso, esse critério se preocupa também em criar casos de teste baseados no espaço de saída ou resultados.

No Quadro 18, são apresentados os casos de teste utilizados para a verificação das diversas funcionalidades do sistema.

QUADRO 18 - CASOS DE TESTE

Casos de teste para o módulo Realizar Login	
Identificação	Casos de teste
1	<i>Login</i> de usuário não cadastrado
2	<i>Login</i> de usuário cadastrado digitado incorretamente
3	Senha inválida para o <i>login</i> (1 vez)
4	Senha inválida para o <i>login</i> (<i>n</i> vezes)
Caso de teste para o módulo Gerenciar Login	
Identificação	Casos de teste
5	Usuário comum tentando realizar operações de usuário administrador

(CONT.) QUADRO 18 - CASOS DE TESTES

Casos de teste para os módulos: Carregar arquivo/ Interpretar regras	
Identificação	Casos de teste
6	Arquivo texto vazio
7	Arquivo texto com conteúdo incorreto (não contém árvore/regras de decisão)
8	Arquivo texto contendo árvore inválida
9	Arquivo texto contendo mais de uma árvore

Casos de teste para os módulos: Carregar arquivo/ Interpretar regras	
Visualizar cenários a partir dos atributos/ Visualizar cenários a partir da produtividade	
Salvar cenário / Gerenciar cenários	
Identificação	Casos de teste
10	Arquivo texto contendo árvore com apenas um nó
11	Arquivo texto contendo uma árvore com o número máximo de nós permitido no contexto
12	Arquivo texto contendo uma árvore binária
13	Arquivo texto contendo uma árvore não binária

FONTE: A AUTORA

5.2.6 TECNOLOGIAS

O ambiente de desenvolvimento utilizado para o sistema proposto foi:

- Sistema operacional - *Ubuntu*⁹ desktop 15.4, por ser um sistema consolidado na área de desenvolvimento de software, ter facilidade de uso e principalmente pela sua capacidade de integração com as outras tecnologias utilizadas neste trabalho.
- Editor de texto - o *Sublime*¹⁰ Text 3 não é uma IDE (*Integrated Development Environment* ou Ambiente de Desenvolvimento Integrado), porém é uma das ferramentas mais conhecidas para editar

⁹ Disponível em: <https://www.ubuntu.com>

¹⁰ Disponível em: <https://www.sublimetext.com/3>

textos. Provê vários recursos para facilitar o desenvolvimento, como, por exemplo, o recurso “autocompletar”.

- Versionamento - para versionamento da própria linguagem, foi utilizada o RVM¹¹ (*Ruby Version Management*), uma ferramenta de versionamento do *Ruby*¹² em que é possível gerenciar todas as dependências da versão ou também criar diferentes pacotes de dependências para projetos específicos.
- Servidor *Web* – o servidor *Web* escolhido foi o WEBrick¹³, que é o padrão para aplicações *Ruby-on-rails*¹⁴.

O Sistema de **gerenciamento de banco de dados** utilizado foi o *PostgreSQL*¹⁵.

A **linguagem de programação** escolhida para concretizar a criação do SECC foi o *Ruby* 2.2. O **framework back-end** utilizado foi o *Ruby-on-Rails*, um *framework* que utiliza a linguagem *Ruby*. O *Rails*, como é amplamente conhecido, é especializado para criação de sistemas para internet.

As **bibliotecas** utilizadas que não acompanham o *framework* foram:

- *Devise*, uma solução de autenticação para *Rails*, é composta por dez módulos que podem ser utilizados para gerenciamento de *login*.
- *Ancestry*, principal biblioteca utilizada no sistema. Permite registrar um caminho completo, de um nó (o nó corrente) até o nó raiz.

Para construção do **Front-end** foram utilizados 2 *frameworks*: o primeiro, para *estilização*, foi o *Bootstrap*, que é amplamente utilizado para construção de sistemas responsivos para *Web*. O segundo *framework* utilizado foi o *Angular JS*⁴, criado pela Google, que permite a criação de páginas *Web* dinâmicas a partir de *HTML* (*Hyper Text Markup Language*).

¹¹ Disponível em: <https://rvm.io/>

¹² Disponível em: <https://www.ruby-lang.org>

¹³ Disponível em: <https://ruby-doc.org/stdlib-2.0.0/libdoc/webrick/rdoc/WEBrick.html>

¹⁴ Disponível em: <http://rubyonrails.org/>

¹⁵ Disponível em: <https://www.postgresql.org/>

5.2.7 RESULTADOS DA APLICAÇÃO DA FERRAMENTA

As telas apresentadas a seguir são um exemplo completo de utilização da ferramenta SECC. As telas referentes ao gerenciamento dos usuários e ao *Login* são apresentadas no Apêndice H.

Após o *Login* de um usuário, a tela da Figura 32 será apresentada. Nela é possível visualizar o nome das árvores que já foram carregadas para o sistema por esse usuário e carregar uma nova árvore.

Cod	Usuario	Nome da árvore	Número de variáveis	Número de folhas	Tamanho da árvore	Ações
3	admin@teste.com	Mreduzida10.txt	20	99	197	Visualizar Excluir
4	admin@teste.com	Mreduzida9.txt	20	598	663	Visualizar Excluir
6	admin@teste.com	modelagem8.txt	20	227	453	Visualizar Excluir
7	admin@teste.com	Mreduzida7.txt	20	1264	1374	Visualizar Excluir
11	admin@teste.com	modelagem6.txt	20	636	1271	Visualizar Excluir
12	admin@teste.com	modelagem10b.txt	20	97	193	Visualizar Excluir

FIGURA 32 - TELA PARA CARREGAMENTO DE NOVAS ÁRVORES

FONTE: A AUTORA

Quando o arquivo é carregado, são apresentadas informações a respeito da árvore contida no arquivo (Figura 33).

Arquivo Adicionado com sucesso.

Usuario: admin@teste.com
Nome: modelagem10b.txt
Número de variáveis: 20

Número de Folhas: 97
Tamanho da árvore: 193

[Voltar](#)

FIGURA 33 - CARACTERÍSTICAS DA ÁRVORE CARREGADA

FONTE: A AUTORA

Com a opção “Voltar” o sistema retorna à tela da Figura 32 para que o usuário possa escolher qual árvore deseja visualizar.

Depois de escolhida a árvore que se deseja visualizar, a tela da Figura 34 é apresentada ao usuário. Nela é possível selecionar o tipo de busca: a partir dos atributos ou a partir da produtividade. Na Figura 34, a opção foi pela busca a partir dos atributos. São dispostos, em ordem alfabética, todos os atributos existentes na árvore. É possível iniciar a seleção de valores a partir de qualquer um dos atributos.

Variáveis	Condições	Produtividade:
Ambiente	selecione uma opcao	
Condicao de corte	selecione uma opcao	
Corte	selecione uma opcao	
K	selecione uma opcao	
N	selecione uma opcao	
P	selecione uma opcao	
Precipitacaoob	selecione uma opcao	
Precipitacaooc	selecione uma opcao	
Precipitacaom	selecione uma opcao	
Precipitacaop	selecione uma opcao	
Solo	selecione uma opcao	

FIGURA 34 - TELA PARA GERAÇÃO DE CENÁRIOS

FONTE: A AUTORA

Na Figura 35, apresenta-se a tela obtida após as escolhas realizadas. Nos campos reservados para as condições, agora, está escrito “Nenhuma opção”, indicando que um caminho específico já foi definido. Nota-se também, no lado direito da tela, que foram selecionados valores para dois atributos. A partir dessas duas seleções, os demais valores possíveis para a concretização de um caminho foram designados aos demais atributos. A produtividade obtida para este cenário foi “<= 76.915”. Cada cenário representa um único caminho de um

nó raiz até um nó folha. A partir dessa tela, é possível também salvar cada cenário obtido a partir das escolhas realizadas.

Variáveis	Condições
Ambiente	Nenhuma opção
Condicao de corte	Nenhuma opção
Corte	Nenhuma opção
K	Nenhuma opção
N	Nenhuma opção
P	Nenhuma opção
Precipitacaob	Nenhuma opção
Precipitacaoc	Nenhuma opção
Precipitacaom	Nenhuma opção
Precipitacaop	Nenhuma opção
Solo	Nenhuma opção

Produtividade:
(76.915-98.224361)

Escolhas:

- corte: > 2.0
- solo: = AQ.2

FIGURA 35 - TELA APÓS SELEÇÃO DOS ATRIBUTOS

FONTE: A AUTORA

A tela da Figura 36 é obtida depois que um cenário é salvo. A partir dessa tela é possível selecionar o cenário que se deseja visualizar.

Nome	Classe	Ações	
cenario612	(76.95741-98.851012]	Visualizar	Excluir
cenario634	(-Inf-76.95741]	Visualizar	Excluir
testecenariopr1	(-Inf-76.915]	Visualizar	Excluir
testecenarioatr5	(98.224361-Inf)	Visualizar	Excluir
testecenario7	(76.915-98.224361]	Visualizar	Excluir
testecenarioar7	(98.224361-Inf)	Visualizar	Excluir
testecenariopr8	(-Inf-76.915]	Visualizar	Excluir
cenario6a	(76.95741-98.851012]	Visualizar	Excluir

FIGURA 36 - TELA ESCOLHER A VISUALIZAÇÃO DE CENÁRIOS JÁ GRAVADOS

FONTE: A AUTORA

Na Figura 37, é apresentado um cenário completo, após a escolha de um cenário para visualização. Os atributos são apresentados de acordo com a ordem em que estão na árvore de decisão; essa ordem implica a importância de cada atributo para obtenção da classe.



FIGURA 37 - VISUALIZAÇÃO DE UM CENÁRIO ESCOLHIDO

FONTE: A AUTORA

A tela da Figura 38 se refere à busca por produtividade. Ao escolher um dos níveis de produtividade apresentados do lado esquerdo da tela, todos os cenários associados àquele nível de produtividade são listados. O usuário pode selecionar qualquer um desses cenários para visualização.

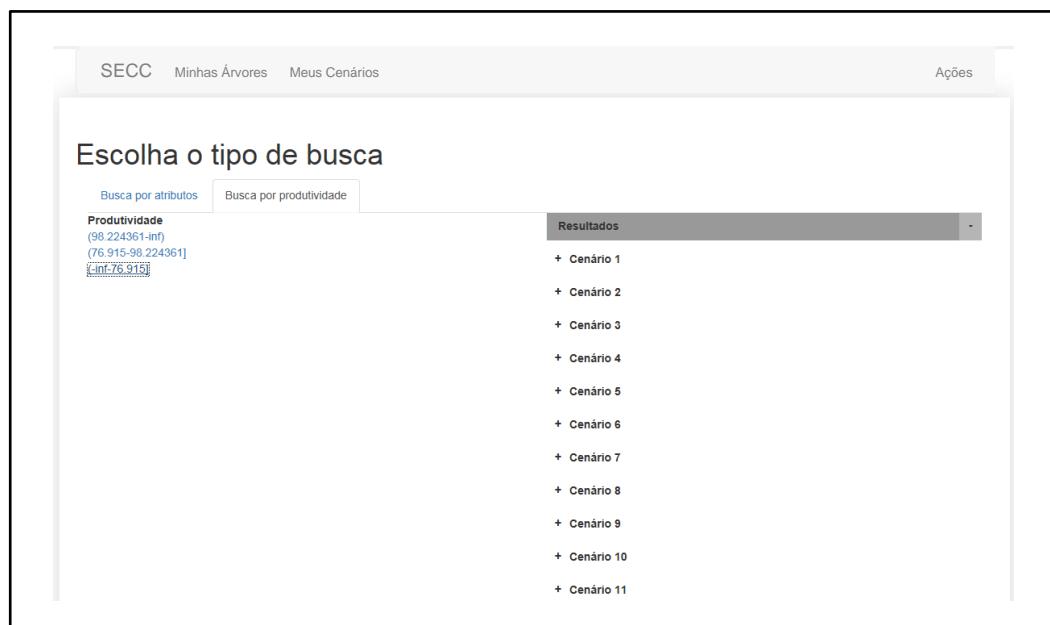


FIGURA 38 - BUSCA POR NÍVEL DE PRODUTIVIDADE

FONTE: A AUTORA

Na Figura 39, apresenta-se a mesma tela da Figura 38, mas no momento após a escolha de um cenário. No canto inferior esquerdo da tela encontra-se a opção para salvar esse cenário. Se existirem vários cenários “expandidos”, será salvo o último cenário que o usuário “expandiu”.

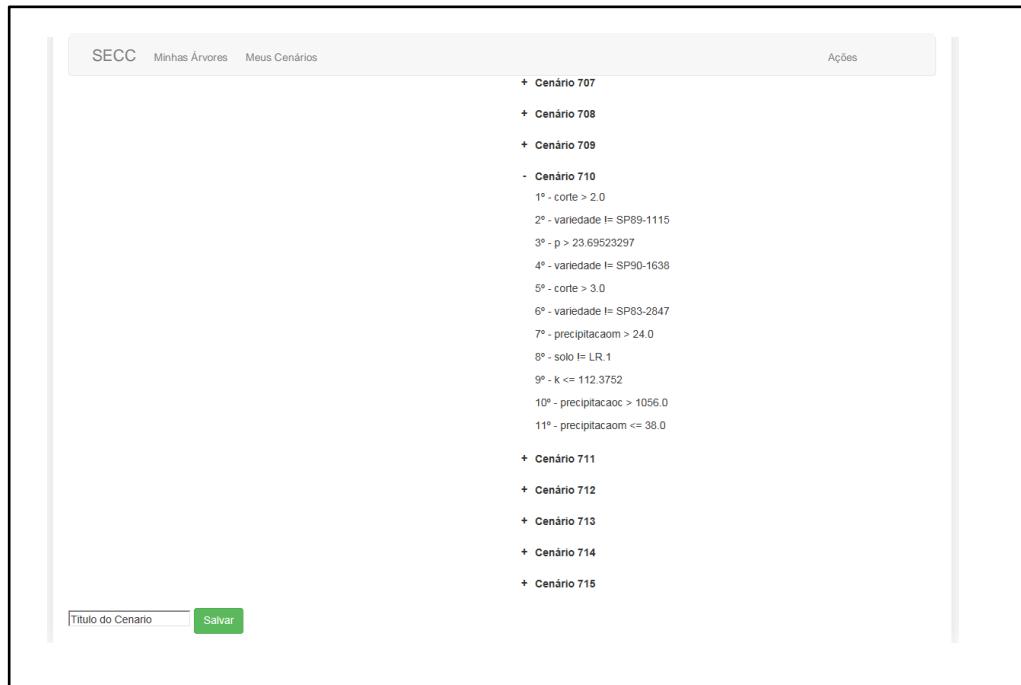


FIGURA 39 - OPÇÃO SALVAR CENÁRIO NA BUSCA POR PRODUTIVIDADE

FONTE: A AUTORA

5.2.8 CONSIDERAÇÕES SOBRE A FERRAMENTA SECC

A ferramenta SECC implementa o conhecimento obtido a partir de um processo de KDD, com base em um roteiro sistematizado. Essa ferramenta faz a leitura de um arquivo texto, isso implica que o processo de KDD pode ser realizado por qualquer ferramenta, desde que seja possível a geração de um arquivo texto contendo as regras de decisão referentes à árvore gerada. Também pode ser aplicada a qualquer conjunto de atributos, uma vez que a estrutura de armazenamento foi organizada de forma a registrar a condição do nó corrente da árvore e o caminho desse nó até o nó raiz.

A ferramenta permite a inclusão de várias árvores para cada usuário, dessa maneira podem ser realizados inúmeros experimentos de modelagem, combinando diferentes atributos, com várias parametrizações. Para cada árvore,

os usuários podem visualizar os diversos caminhos que levam de um nó raiz até um nó folha. Esses caminhos representam cenários de produção, que podem ser salvos para futuras análises. Salienta-se que a análise de uma árvore de decisão inteira, geralmente, é muito complexa; dessa forma, ao permitir que cada caminho seja visualizado e gravado de forma individual, essa ferramenta reduz a complexidade do modelo e, consequentemente, o tempo de entendimento e análise dos cenários de produção.

Além disso, a ferramenta propicia também flexibilidade nas formas de consulta. O usuário pode escolher os atributos (em qualquer ordem) e seus valores, até chegar a um nível de produtividade, ou pode escolher um nível de produtividade e obter todos os cenários que levam àquela produtividade.

As características dessa ferramenta de visualização possibilitam seu uso nas diversas atividades de planejamento e suporte à decisão da cultura da cana-de-açúcar.

6 CONCLUSÃO

Este trabalho teve por objetivo propor um roteiro sistematizado, a partir da aplicação de técnicas de mineração de dados, para dar suporte aos processos de tomada de decisão na gestão da produção de cana-de-açúcar.

Verificou-se que os fatores climáticos são os mais utilizados nos modelos de produtividade propostos. Também se observou que é bastante comum o uso de imagens de satélite, especialmente o índice NDVI e sua associação com atributos climáticos. Dados do solo também estão presentes em grande parte dos estudos sobre a cana. Destaca-se, entretanto, a grande variedade de atributos utilizados, bem como as diversas combinações desses atributos, em diferentes pesquisas analisadas. Ao todo foram identificados na literatura 154 atributos e 32 trabalhos que utilizaram esses atributos combinados.

Foi aplicado o método Delphi para a identificação de forma sistematizada dos atributos relevantes para a criação de um modelo de produtividade. Partiu-se de um conjunto de 34 atributos e, após as rodadas de análise do método Delphi, chegou-se a um conjunto de 19 atributos.

O conceito de produtividade utilizado neste trabalho contempla não apenas a produção de cana-de-açúcar por hectare, mas também qualquer redução de custo ou melhoria nos processos gerenciais que foram classificados em cinco categorias: 1) programação da colheita; 2) previsão de produtividade; 3) caracterização de áreas; 4) modelos de apoio às práticas de gestão; e 5) outros.

Foram encontrados apenas quatro trabalhos que desenvolveram ou utilizaram ferramentas que tivessem uma interface com foco em abreviar as tarefas dos gestores agrícolas durante a tomada de decisão. Especificamente, nenhuma ferramenta foi encontrada que possibilitasse a visualização de cada um dos caminhos de uma árvore de decisão, de forma a reduzir a complexidade inerente dessas estruturas.

Assim, desenvolveu-se um roteiro sistematizado, implementado por meio de uma ferramenta de visualização que apresenta os resultados obtidos com base na utilização da técnica de indução por árvore de decisão. A execução desse roteiro tem como entrada o resultado de um processo de KDD.

Esse processo de KDD foi realizado de acordo com a metodologia CRISP-DM. Foi utilizada a técnica 5W1H na condução de uma entrevista semiestruturada na etapa de entendimento do negócio. Com isso, foi possível identificar como a ferramenta proposta poderia subsidiar a tomada de decisão. Embora os resultados da mineração de dados, visualizados por meio dessa ferramenta, possam ser utilizados de diversas maneiras, destaca-se sua utilização no planejamento do plantio, uma vez que é possível associar os diversos cenários de produção a sua respectiva produtividade e, também, no planejamento de tratos, identificando cenários que levem à produtividade baixa e realizando ações corretivas.

Embora a avaliação do modelo tenha sido feita para uma modelagem específica, a ferramenta desenvolvida neste trabalho, denominada SECC, pode ser utilizada para diversos experimentos de modelagem com qualquer conjunto de atributos. É possível inserir diversas árvores e para cada árvore podem ser visualizados diversos caminhos que representam cenários de produção. A possibilidade de visualizar cada caminho separadamente reduz de maneira significativa a complexidade da análise, uma vez que a árvore de decisão inteira é uma estrutura demasiadamente grande e, por essa razão, de difícil entendimento.

Assim, decorrente desse conjunto de processos desenvolvidos neste trabalho de pesquisa, efetivam-se as seguintes contribuições:

➤ Acadêmicas

- Categorização dos modelos de produtividade de cana-de-açúcar;
- Sistematização do processo de definição de atributos para construção de modelos de produtividade de cultura de cana-de-açúcar;
- Proposta de um roteiro sistematizado para visualização dos resultados de um processo de KDD utilizando o modelo de referência CRISP-DM;

- Implementação de uma ferramenta de visualização (SECC) que apresenta cada caminho de uma árvore de decisão.

- Empresariais
 - Pela redução sistematizada dos dados de atributos, criou-se a possibilidade de ***favorecer a tomada de decisão*** sobre esse conjunto menor de uma forma mais assertiva;
 - A estrutura do modelo de referência do CRISP-DM junto com a proposta do roteiro sistematizado ***possibilita a visão geral do processo*** de preparação, modelagem e visualização dos dados referentes à cultura da cana-de-açúcar, servindo de diretriz para a execução desse processo;
 - ***O Conjunto de funcionalidades da ferramenta de visualização*** (SECC): *i)* possibilita a ***inclusão de diversos modelos***, com conjuntos distintos de atributos; *ii)* possibilita análise individual de cada cenário de produção pertencente a um modelo específico, ***reduzindo*** assim a ***complexidade e o tempo*** dessa análise.

Com base nessas contribuições, depreende-se que este trabalho atendeu à pergunta de pesquisa e ao objetivo de apresentar os meios para facilitar o processo de tomada de decisão visando a gestão da produção de cana-de-açúcar a partir da aplicação de um roteiro sistematizado.

6.1 PROPOSTA PARA TRABALHOS FUTUROS

Apresentam-se, a seguir, outras possibilidades de pesquisa a partir do proposto neste trabalho:

- Adaptar o sistema de exploração de cenários para aceitar dados gerados por diferentes ferramentas de mineração – o módulo que faz a leitura e interpretação do arquivo texto para armazenar a árvore de decisão em uma tabela trata detalhes específicos da ferramenta Weka. Assim a

utilização de outras ferramentas requer algumas modificações no processo de interpretação dos dados presentes no arquivo texto.

- Realizar uma pesquisa *survey* com gerentes das usinas – o roteiro sistematizado pode ser analisado e avaliado por gerentes de diversas usinas com vistas a validar sua utilização e propor melhorias para o sistema.
- Realizar uma pesquisa no formato de estudo multicasos – envolvendo gerentes de usinas que utilizariam o roteiro sistematizado proposto e o SECC ao longo de algumas safras (pelo menos quatro).
- Adaptar o sistema de exploração de cenários para outras culturas – como o sistema armazena uma árvore de decisão em uma tabela, independentemente de quais atributos sejam utilizados, o sistema de exploração de cenários pode ser adaptado para outras culturas.

REFERÊNCIAS

- ACOMPANHAMENTO DA SAFRA BRASILEIRA: cana-de-açúcar: monitoramento agrícola. Brasília: CONAB, v. 2, n. 1, p. 1-28, 2015. Disponível em: <http://www.conab.gov.br/OlalaCMS/uploads/arquivos/15_04_13_08_49_33_boletim_cana_portugues_-_1o_lev_-_15-16.pdf>. Acesso em: 10 fev. 2016.
- AHUMADA, O.; VILLALOBOS, J. R. Application of planning models in the agri-food supply chain: a review. **European Journal of Operational Research**, v.196, n. 1, p. 1–20, 2009.
- AMO, S. **Técnicas de mineração de dados**. Uberlândia, MG: Minas: Universidade Federal de Uberlândia, 2004. Disponível em: <<http://www.deamo.prof.ufu.br/arquivos/JAI-cap5.pdf>>. Acesso em: 03 fev. 2011.
- ANANTHARA, M.; ARUNKUMAR, T.; HEMAVATHY, R. CRY: an improved crop yield prediction model using bee hive clustering approach for agricultural data sets. In: INTERNATIONAL CONFERENCE ON PATTERN RECOGNITION, INFORMATICS AND MOBILE ENGINEERING, 2013, Salem. **Proceedings...** Salem: IEEE, 2013. p. 473-478.
- ANDRADE, M. M. **Como preparar trabalhos para cursos de pós-graduação: noções práticas**. 5. ed. São Paulo: Atlas, 2002.
- BARROS, F. M. M.; OLIVEIRA, S. R. M.; OLIVEIRA, L. H. M. Desenvolvimento e validação de um sistema de recomendação de informações tecnológicas sobre cana-de-açúcar. **Bragantia**, Campinas, v. 72, n. 4, p. 387–395, 2013.
- BERTO, R.M.V.S.; NAKANO, D. N. Métodos de pesquisa na engenharia de produção. In: ENCONTRO NACIONAL DE ENGENHARIA DA PRODUÇÃO, 18, 1998, Niterói. **Anais...** Niterói, RJ: ENEGEP, 1998.
- BIOLOCHINI, J.; MIAN, P. G.; NATALI, A. C. C. Systematic review in software engineering. **Relatório Técnico RT-679/05**: COPPE/UFRJ, Rio de Janeiro, maio 2005.
- BOCCA, F. F. **Produtividade de cana-de-açúcar**: caracterização dos contextos de decisão e utilização de técnicas de mineração de dados para modelagem. 2014. 86 f. Dissertação (Mestrado em Engenharia Agrícola) – Universidade Estadual de Campinas, Campinas, SP, 2014.
- BRASIL. Ministério da Agricultura, Pecuária e Abastecimento. **Sapcana**: Sistema de Acompanhamento de Produção Canavieira. 2014. Disponível em:<<http://www.agricultura.gov.br/comunicacao/noticias/2014/09/mapa-publica-projcoes-do-agronegocio-para-a-safra-20232024>>. Acesso em:20 jul. 2015.

BRASIL. Ministério da Agricultura, Pecuária e Abastecimento. **Sapcana**: Sistema de Acompanhamento de Produção Canavieira. 2016. Disponível em:<<http://www.agricultura.gov.br/vegetal/culturas/cana-de-acucar>>. Acesso em 10 fev. 2016.

BRUNINI,O. Ambientes climáticos e exploração agrícola da cana-de-açúcar. In: DINARDO-MIRANDA, L. L; VASCONCELOS, A. C. M.; LANDELL, M. G. A. (Ed.). **Cana-de-açúcar**. Campinas: Instituto Agronômico, 2008. p. 179-204.

CARBONELL, J.; OSORIO, C. A. Characterization of different areas with maximum potential productivity planted with sugarcane in the Cauca River Valley (Colombia). In: INTERNATIONAL SYMPOSIUM ON VORONOI DIAGRAMS IN SCIENCE AND ENGINEERING, 2010, Quebec. **Anais...** Quebec: IEEE, 2010. p.266–272.

CERRI, D.; MAGALHÃES, P. Correlation of physical and chemical attributes of soil with sugarcane yield. **Pesquisa Agropecuária Brasileira**, n. 1, p. 613–620, 2012.

CHAPMAN, P *et al.* **CRISP-DM 1.0**: step-by-step data mining guide. [S.I]: SPSS Inc., 2000.

CHEN, Y.; HU, D.; ZHANG, G. Data mining and critical success factors in data mining projects. **IFIP: Advances in Information and Communication Technology**, v. 207, n. 05, p. 281–287, 2006.

COCK, J.*et al.* Crop management based on field observations: case studies in sugarcane and coffee. **Agricultural Systems**, v. 104, n. 9, p. 755–769, 2011.

COLAK, I.; SAGIROGLU, S.; YESILBUDAK, M. Data mining and wind power prediction: a literature review. **Renewable Energy**, v. 46, p. 241–247, 2012.

EVERINGHAM, Y. L. *et al.* Advanced satellite imagery to classify sugarcane crop characteristics. **Agronomy Sustain. Dev.**, v. 27, p. 111–117, 2007.

EVERINGHAM, Y. L. *et al.* Enhanced risk management and decision-making capability across the sugarcane industry value chain based on seasonal climate forecasts. **Agricultural Systems**, v. 74, n. 3, p. 459–477, 2002.

EVERINGHAM, Y. L.; SMYTH, C. W.; INMAN-BAMBER, N. G. Ensemble data mining approaches to forecast regional sugarcane crop production. **Agricultural and Forest Meteorology**, v. 149, n. 3-4, p. 689–696, 2009.

FAYYAD, U.; PIATETSKY-SHAPIRO, G.; SMYTH, P. From data mining to knowledge discovery in databases. **AI magazine**, p. 37–54, 1996.

FERNANDES, J. L.; ROCHA, J. V.; LAMPARELLI, R. A. C. Sugarcane yield estimates using time series analysis of spot vegetation images temporais de imagens spot vegetation. **Sci. Agric.**, Piracicaba, v.68, n.2, p. 139–146, abr. 2011.

- FERRARO, D. O.; GHERSA, C. M.; RIVERO, D. E. Weed vegetation of sugarcane cropping systems of northern Argentina: data-mining methods for assessing the environmental and management effects on species composition. **Weed Science**, v. 60, n. 1, p. 27–33, 2012.
- FERRARO, D. O.; RIVERO, D. E.; GHERSA, C. M. An analysis of the factors that influence sugarcane yield in Northern Argentina using classification and regression trees. **Field Crops Research**, v. 112, n. 2-3, p. 149–157, 2009.
- GAMMA, Erich; HELM, Richard; JOHNSON, Ralph. **Padrões de projeto: soluções reutilizáveis de software orientado a objetos**. Porto Alegre: Bookman, 2008.
- GIL, A. C. **Métodos e técnicas de pesquisa social**. 5. ed. São Paulo: Atlas, 2007.
- GOLDSCHMIDT, R.; PASSOS, E. **Data mining**: um guia prático. Rio de Janeiro: Elsevier, 2005.
- GONÇALVES, R. R. V. *et al.* Analysis of NOAA/AVHRR multitemporal images, climate conditions and cultivated land of sugarcane fields applied to agricultural monitoring. In: INTERNATIONAL WORKSHOP ON THE ANALYSIS OF MULTI TEMPORAL REMOTE SENSING IMAGES, 6., 2011, Trento. **Anais...** Trento: IEEE, p.229–232, 2011.
- GONÇALVES, R. R. V. *et al.* Analysis of NDVI time series using cross-correlation and forecasting methods for monitoring sugarcane fields in Brazil. **International Journal of Remote Sensing**, v. 33, n.15, p. 4653–4672, 2012.
- GREENLAND, D. Climate variability and sugarcane yield in Louisiana. **Journal of Applied Meteorology**, v. 44, n. 11, p. 1655–1666, 2005.
- GUESSI, M.; OLIVEIRA, L. B. R.; NAKAGAWA, E. Y. Current state on representation of reference architectures. **Technical report**, Universidade de São Paulo, 2011.
- GUYON, I.; ELISSEEFF, A. An introduction to variable and feature selection. **Journal of machine learning research**, v. 3, p. 1157-1182, 2003.
- HALL, M. A.; HOLMES, G. Benchmarking attribute selection techniques for discrete class data mining. **IEEE Transactions on Knowledge and Data Engineering**, v. 15, n. 6, p. 1437-1447, 2003.
- HAJJ, M. E. L. *et al.* Integrating SPOT-5 time series, crop growth modeling and expert knowledge for monitoring agricultural practices:the case of sugarcane harvest on Reunion Island. **Remote Sensing of Environment**, v. 113, n. 10, p. 2052–2061, 2009.
- HAN, J.; KAMBER, M. **Data mining**: concepts and techniques. 2. ed. São Francisco: Morgan Kaufmann, 2006. 770 p.

HIGGINS, A. et al. Opportunities for value chain research in sugar industries. **Agricultural Systems**, v.94, n. 3, p. 611–621, 2007.

JENA, S. D.; POGGI, M. Harvest planning in the Brazilian sugar cane industry via mixed integer programming. **European Journal of Operational Research**, v. 230, n. 2, p. 374–384, 2013.

JIAO, Z.; HIGGINS, A. J.; PRESTWIDGE, D. B. An integrated statistical and optimisation approach to increasing sugar production within a mill region. **Computers and Electronics in Agriculture**, v. 48, n. 2, p. 170–181, 2005.

KAYO, E. K.; SECURATO, J. R. Método Delphi: fundamentos, críticas e vieses. **Caderno de Pesquisas em Administração**, v.1, n. 4, p. 51–61, 1997.

KITCHENHAM, B.; CHARTERS, S. Guidelines for performing Systematic Literature Reviews in Software Engineering: version 2.3. **EBSE Technical Report**, United Kingdom, 2007.

.LANDELL, M. G. A.; BRESSIANI J. A. Melhoramento genético, caracterização e manejo varietal. In: DINARDO-MIRANDA, L. L; VASCONCELOS, A. C. M.; LANDELL, M. G. A. (Ed.). **Cana-de-açúcar**. Campinas: Instituto Agronômico, 2008. p. 101-156.

LOYOLA, O.; MEDINA, M. A.; GARCÍA, M. Inducing decision trees based on a cluster quality index. **IEEE Latin America Transactions**, v. 13, n. 4, p. 1141–1147, 2015.

MARIN, F.; CARVALHO, G. Spatio-temporal variability of sugarcane yield efficiency in the state of São Paulo, Brazil. **Pesquisa Agropecuária Brasileira**, n. 1, p. 149–156, 2012.

MEINKE, H.; STONE, R. C. New tool for increasing preparedness to climate variability and change in agricultural planning and operations. **Climate Change**, v. 70, p. 221–253, 2005.

MELLO, M. P.; ATZBERGER, C.; FORMAGGIO, A. R. Near real time yield estimation for sugarcane in Brazil combining remote sensing and official statistical data. In: INTERNATIONAL GEOSCIENCE AND REMOTE SENSING SYMPOSIUM (IGARSS), 2014. **Anais...** [S.I]: IEEE, 2014. p.5064–5067.

MINAYO, M. C. S. Quantitativo-qualitativo: oposição ou complementaridade? **Cadernos de Saúde Pública**, Rio de Janeiro, jul./set. 1993.

MONARD, M. C.; BARANAUSKAS, J. A. Conceitos sobre aprendizado de máquina. In: REZENDE, S. O. (Org.). **Sistemas inteligentes**: fundamentos e aplicações. Barueri: Manole, 2003. p. 89-114.

MOREIRA, D.A. **Administração da produção e operações**. 2. ed. São Paulo: Cengage Learning, 2011. 624 p.

MYERS, G. J.; SANDLER, C.; BADGETT, T. **The art of software testing**. 3rd ed. New York: Wiley, 2011.

NASCIMENTO, C. R. *et al.* Identification of sugar cane fields in the state of São Paulo using a time series of avhrr/noaa satellite images. In: INTERNACIONAL WORKSHOP ON THE ANALYSIS OF MULTI-TEMPORAL REMOTE SENSING IMAGES, 5., 2009, Connecticut. **Anais...** Connecticut: Embrapa, 2009. p. 104–111.

NAWI, N. M. *et al.* Prediction and classification of sugar content of sugarcane based on skin scanning using visible and shortwave near infrared. **Biosystems Engineering**, v. 115, n. 2, p. 154–161, 2013.

NAWI, N. M. *et al.* Prediction of sugarcane quality parameters using visible-shortwave near infrared spectroradiometer. **Agriculture and Agricultural Science Procedia**, v. 2, p. 136–143, 2014.

NONATO, R. T.; OLIVEIRA, S. R. D. E. M. Data mining techniques for identification of sugarcane crop areas in images Landsat 5. **Engenharia Agrícola**, v. 33, n. 6, p. 1268–1280, 2013.

PACHECO, D. F.; LUCAS, T. D. P.; LIMA NETO, F. B. How preferences affect productivity in the sugarcane harvest problem a comparative study of a two-steps MOEA. In: INTERNATIONAL CONFERENCE ON HYBRID INTELLIGENT SYSTEMS, 8., 2008, Barcelona. **Anais...** Barcelona: IEEE, 2008, p.296–301.

PIEWTHONNGAM, K.; SUKSAWAT, J.; TENGLOLAI, A. Identifying efficient cane growers and exploiting their expertise in improving inefficient ones? In: INTERNATIONAL CONFERENCE ON INDUSTRIAL ENGINEERING MANAGEMENT, 2007, Singapura. **Proceedings...** Singapura: IEEE, 2007, p.1654–1658.

PRADO, H. *et al.* Solos e ambientes de produção. In: DINARDO-MIRANDA, L. L; VASCONCELOS, A. C. M.; LANDELL, M. G. A. (Ed.). **Cana-de-açúcar**. Campinas: Instituto Agronômico, 2008. p. 179-204.

PRATI, R. C.; BATISTA, G.; MONARD, M. C. Curvas ROC para avaliação e classificadores. **Revista IEEE América Latina**, v. 6, n. 2, p. 215-222, 2008.

PRESSMAN, R. S.; MAXIM, B. R. **Engenharia de software**: uma abordagem profissional. 8. ed. Porto Alegre: Bookman, 2016.

QUINLAN, J. R. **C4.5**: programs for machine learning. San Francisco: Morgan Kaufmann, 1993.

RUMBAUGH, James; BOOCHE, Grady; JACOBSON, Ivar. **UML**: guia do usuário. 2. ed. Rio de Janeiro: Campus, 2012.

ROMANI, L. A. S. et al. Aplicação de técnicas de mineração em dados climáticos e de satélite para auxiliar no acompanhamento das safras de cana-de-açúcar. In: WORKSHOP EM ALGORITMOS E APLICAÇÕES DE MINERAÇÃO DE DADOS, 4., 2008, Campinas, SP. **Anais...** Campinas, SP: SBC, 2008, p. 87–92.

ROMANI, L. et al. A new time series mining approach applied to multitemporal remote sensing imagery. **IEEE Transactions on Geoscience and Remote Sensing**, v. 51, n. 1, p. 140–150, 2013.

ROMANI, L. A. S. et al. Clustering analysis applied to NDVI/NOAA multitemporal images to improve the monitoring process of sugarcane crops. In: INTERNATIONAL WORKSHOP ON THE ANALYSIS OF MULTI-TEMPORAL REMOTE SENSING IMAGES (MULTI-TEMP), 6., 2011, Trento. **Anais...** Trento: IEEE, 2011. p. 33–36.

ROSSETTO, R.; DIAS, F.L.F; VITTI, A.C. Fertilidade do solo, nutrição e adubação. In: DINARDO-MIRANDA, L. L; VASCONCELOS, A. C. M.; LANDELL, M. G. A. (Ed.). **Cana-de-açúcar**. Campinas, SP: Instituto Agronômico, 2008, p. 221-238.

ROSSETTO, R. et al. Fósforo. In: DINARDO-MIRANDA, L. L; VASCONCELOS, A. C. M.; LANDELL, M. G. A. (Ed.). **Cana-de-açúcar**. Campinas, SP: Instituto Agronômico, 2008a, p. 271-287.

ROSSETTO, R. et al. Potássio. In: DINARDO-MIRANDA, L. L; VASCONCELOS, A. C. M.; LANDELL, M. G. A. (Ed.). **Cana-de-açúcar**. Campinas, SP: Instituto Agronômico, 2008b, p. 289-312.

SANTCHURN, D. et al. From sugar industry to cane industry: investigations on multivariate data analysis techniques in the identification of different high biomass sugarcane varieties. **Euphytica**, v. 185, n. 3, p. 543–558, 2012.

SÃO PAULO (Estado). Investe São Paulo: Agência paulista de promoção de investimentos e competitividade, 2014. Disponível em: <<http://www.investe.sp.gov.br/setores-de-negocios/agronegocios/cana-deacucar/>>. Acesso em: 17 jan. 2014.

SHARP, H.; ROGERS, Y.; PREECE, J. **Design de interação:** além da interação homem-computador. 3. ed. Porto Alegre: Bookman, 2013.

SILVA, A. F.; MARINS, F. A. S.; DIAS, E. X. Addressing uncertainty in sugarcane harvest planning through a revised multi-choice goal programming model. **Applied Mathematical Modelling**, v. 39, n. 18, p. 5540–5558, 2015.

SOUZA, Z. et al. Análise dos atributos do solo e da produtividade da cultura de cana-de-açúcar com o uso da geo estatística e árvore de decisão. **Ciência Rural**, v. 40, n. 4, p. 840–847, 2010.

STAPLES, M.; NIAZI, M. Systematic review of organizational motivations for adopting CMM-based SPI. **Information and Software Technology**, v. 50, n. 7–8, p. 605–620, jun. 2008.

THUANKAEWSING, S.; PATHUMNAKUL, S.; PIEWTHONGNGAM, K. Using an artificial neural network and a mathematical model for sugarcane harvesting scheduling. In: INTERNATIONAL CONFERENCE ON INDUSTRIAL ENGINEERING MANAGEMENT, 2011, Singapura. **Anais...** Singapura: IEEE, 2011, p.308–312.

TRIVIÑOS, A. N. S. **Introdução à pesquisa em ciências sociais:** a pesquisa qualitativa em educação. São Paulo: Atlas, 1987.

TSAI, H. H. Global data mining: an empirical study of current trends, future forecasts and technology diffusions. **Expert Systems with Applications**, v. 39, n. 9, p. 8172–8181, 2012.

TSAI, H. H. Knowledge management vs. data mining: research trend, forecast and citation approach. **Expert Systems with Applications**, v. 40, n. 8, p. 3160–3173, 2013.

UNICA (União da indústria de cana-de-açúcar). 2016. Disponível em: <<http://www.unica.com.br/empresa/5573206/sao-martinho>>. Acesso em: 11 jan. 2016.

UNIVASO, P.; ALE, J. M.; GURLEKIAN, J. A. Data mining applied to forensic speaker identification. **IEEE**, v. 13, n. 4, p. 1098–1111, 2015.

VICENTE, A. A.; DELAMARO, M. E.; MALDONADO, J. C. Uma revisão sistemática sobre a atividade de teste de software em métodos ágeis. In: CONFERÊNCIA LATINOAMERICANA DE INFORMÁTICA, 35., 2009, São Carlos, SP. **Anais...** São Carlos, SP: CLEI, 2009.

VIEIRA, M. A. *et al.* Object based image analysis and data mining applied to a remotely sensed landsat time-series to map sugarcane over large areas. **Remote Sensing of Environment**, v. 123, p. 553–562, 2012.

VINTROU, E. *et al.* Data mining, a promising tool for large-area cropland mapping. **IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing**, v. 6, n. 5, p. 2132–2138, 2013.

VITTI, A. C. *et al.* Nitrogênio. In: DINARDO-MIRANDA, L. L; VASCONCELOS, A. C. M.; LANDELL, M. G. A. (Ed.). **Cana-de-açúcar**. Campinas: Instituto Agronômico, 2008. p. 239-269.

WITTEN, I. H.; FRANK, E. **Data mining:** practical machine learning tools and techniques. 2nd ed. San Francisco: Morgan Kaufmann, 2005.

ZHOU, M. M.*et al.* Artificial neural network models as a decision support tool for selection in sugarcane: a case study using seedling populations. **Crop Science**, v. 51, n. 1, p.21, 2011.

APÊNDICE A – DESCRIÇÃO DA APLICAÇÃO DO MÉTODO DELPHI

Para a seleção dos atributos que mais interferem na produtividade da cana-de-açúcar e que, portanto, deveriam compor o modelo de produtividade proposto neste trabalho, foi utilizado o método Delphi, que envolve a aplicação sucessiva de questionários sobre um determinado tema a um grupo de especialistas. Cada etapa do método denomina-se rodada. No intervalo de cada rodada são feitas análises estatísticas das respostas e o resultado é compilado em novos questionários que são novamente distribuídos ao grupo. O objetivo principal é tentar obter consenso, ou quase consenso, entre os especialistas (DALKEY; HELMER, 1963 apud KAYO; SECURATO, 1997). A aplicação do método é descrita a seguir.

No total, quatro especialistas participaram do processo de seleção de atributos, conforme caracterizado no Quadro 19. Dois dos especialistas são pesquisadores de universidades públicas com mais de 30 anos na área agrícola. Os outros dois trabalham em empresas da iniciativa privada, um com 25 anos de atuação na área agrícola e outro com 6 anos. O especialista E2 é suporte técnico na Qualidade Agrícola da usina de cana-de-açúcar, que cedeu os dados da produção da cana para a realização deste trabalho, ou seja, apesar de menor tempo de atuação, sua experiência é focada exatamente na área de interesse deste trabalho.

QUADRO 19 - CARACTERIZAÇÃO DOS ESPECIALISTAS

Especialista	Empresa/ Instituição	Formação	Área de atuação	Área de pesquisa	Tempo de atuação	Cargo
E1	Tecnologia agroindustrial	Mestrado	Máquinas e implementos agrícolas	-	25	Diretor Vendas e marketing
E2	Usina de cana-de- açúcar	Graduação	Qualidade agrícola	-	6	Suporte técnico
E3	Universidade pública	Doutorado	Administração	Mecaniza- ção/ Cons. solo	30	Diretor
E4	Universidade pública	Doutorado	Mecanização/ gestão	Mecaniza- ção/ Gestão	35	Professor

FONTE: A AUTORA

Com base nos atributos utilizados nos modelos de produtividade da cana em Ferraro, Rivero e Ghersa (2009) e Bocca (2014), dois trabalhos que se assemelham a esta pesquisa, foi elaborada uma lista inicial de atributos. Essa lista foi apresentada a dois dos especialistas participantes do processo de seleção de atributos, para que fosse realizada uma validação desse conjunto inicial, bem como um ajuste fino desses atributos.

Uma vez consolidado o conjunto de atributos para realização do método Delphi, foi elaborada uma tabela contendo 27 atributos, agrupados por tipos, a saber: *i*) atributos de solo; *ii*) atributos de manejo e *iii*) atributos climáticos. A primeira coluna da tabela especificava o atributo e as demais colunas continham números, de 0 a 10, que representavam o grau de importância daquele atributo para a produtividade da cana-de-açúcar, sendo 0 – para o de menor importância e 10 – para o de maior importância. Essa tabela e um documento solicitando a participação na pesquisa e explicando o funcionamento do método Delphi foram enviados, por e-mail, para os quatro especialistas, previamente contatados.

Os especialistas atribuíram notas de 0 a 10 para cada um dos atributos. A frequência dessas notas foi então contabilizada, conforme apresentado no Quadro 20, que também contém os comentários emitidos por esses especialistas.

Para iniciar a segunda rodada do método Delphi foram selecionados apenas os atributos com frequência maior que dois em notas superiores a sete. Também foram incluídos os atributos sugeridos pelos especialistas E1 e E2. Assim, foi enviado novo e-mail aos especialistas contendo: documento explicando os procedimentos para a realização dessa segunda rodada, os resultados da primeira rodada (Quadro 20) e uma nova tabela com os atributos selecionados na primeira rodada. Foi solicitado aos especialistas que atribuissem nota maior ou igual a oito para os atributos mais importantes para a produtividade da cana, critério que já havia sido utilizado pelo especialista E3 na primeira rodada.

QUADRO 20 - RESULTADO DA PRIMEIRA RODADA DO MÉTODO DELPHI

Notas	0	1	2	3	4	5	6	7	8	9	10
Atributos de solo											
Tipo do solo										3	1
Fertilidade					1				1	1	1
Textura						1	1			2	
Atributos de manejo											
Data do plantio						1			1	2	
Número de cortes	1						1	2			
Insumos											
Nitrogênio									2		2
Potássio					1			1		2	
Fósforo					1			1		2	
PH	1						2			1	
Torta de Filtro					1		1	1			1
Tipo da cana (18 meses, 12 meses, inverno)							2	2			
Variedade							1	2	1		
Ambiente de manejo								2	1	1	
Atributos Climáticos											
Temperatura média mensal						1	1	1			1
Média das Temperaturas relacionadas a cada fase* do período de desenvolvimento *(brotação, perfilhamento, crescimento e maturação)								1	1	1	2
Precipitação acumulada mensal						1		1			2
Precipitação Acumulada por fase*								1	1	2	
Estiagem (número de dias consecutivos sem chuva) – por fase*						1	1				2
Comentários da primeira rodada											
E1 (especialista 1)											
Incluir também aspectos relacionados ao sistema de produção, do tipo:											
- Tipo de preparo do Solo											
- Plantio Mecanizado											
- Tratos Culturais											
- Colheita Mecanizada											
- Manejo de Biomassa											
E2 (especialista 2)											
Incluir o atributo Época de colheita (em relação ao ambiente)											
E3 (especialista 3)											
Atributos com valores abaixo de 8 podem ser controlados ou são atípicos, interferem menos no processo produtivo, mais especificamente na produtividade.											

FONTE: A AUTORA

Nessa segunda rodada não houve retorno da avaliação do especialista E2. Dessa forma, optou-se por considerar as notas atribuídas na primeira rodada,

por esse especialista, na contagem final. No Quadro 21 são apresentadas as frequências obtidas, em cada nota, na segunda rodada do método Delphi.

QUADRO 21- RESULTADO DA SEGUNDA RODADA DO MÉTODO DELPHI

Notas	0	1	2	3	4	5	6	7	8	9	10
Atributos de solo											
Tipo do solo									4		
Fertilidade					1			1	2		
Textura								3	1		
Atributos de manejo											
Data do plantio						1	1	1	1		
Número de cortes							1	1	1		
Insumos											
Nitrogênio								1	1	2	
Potássio							1	1		2	
Fósforo								2		2	
PH	1				1			1	1		
Torta de Filtro						1	2			1	
Tipo da cana(18 meses, 12 meses, inverno)						1	2	1			
Variedade							1	2	1		
Ambiente de manejo								3	1		
Atributos Climáticos											
Temperatura média mensal								2	2		
Média das Temperaturas relacionadas a cada fase* do período de desenvolvimento (brotação, perfilhamento, crescimento e maturação)										3	1
Precipitação acumulada mensal							2		1	1	
Precipitação Acumulada por fase*							1	1		2	
Estiagem (número de dias consecutivos sem chuva) – por fase*						1	1			2	
Sistema de produção											
Tipo de preparo do Solo							1	2			
Plantio Mecanizado						1			2		
Tratos Culturais							1	1	1		
Colheita Mecanizada								1	2		
Manejo de Biomassa								3			

FONTE: A AUTORA

Segundo Kayo e Securato (1997), a grande maioria das pesquisas que utilizam o método Delphi produzem, no máximo, quatro rodadas. Os autores salientam ainda que nada impede que se faça um número menor, desde que os objetivos tenham sido atingidos. Assim, por restrições de tempo e por entender que os objetivos da aplicação do método foram alcançados, o procedimento foi encerrado na segunda rodada, com os seguintes atributos: Tipo do Solo, fertilidade, textura, número de cortes, insumos (N,P,K), variedade da cana, ambiente de manejo, média das temperaturas em cada fase do desenvolvimento

fenológico, precipitação acumulada por fase do desenvolvimento fenológico, tipo de preparo do solo, plantio mecanizado, tratos culturais, colheita mecanizada e manejo da biomassa.

APÊNDICE B – CONJUNTO DE VALORES POSSÍVEIS PARA OS ATRIBUTOS DA ÁRVORE DE DECISÃO

QUADRO 22 - FERTILIDADE

Código	Desc. Abreviada	Descrição
1	ALTA	Alta
2	MALT	Média alta
3	MED	Média
4	MBX	Média Baixa
5	BAIX	Baixa
99	ADEF	A Definir

QUADRO 23 - TEXTURA

Código	Desc. Abreviada	Descrição
1	ARG	Argiloso
2	ARE	Arenoso
3	ARGA	Argiloso/Arenoso
4	INDE	Indefinido
5	DESC	Desconhecido
99	ADEF	A Definir

QUADRO 24 - AMBIENTE DE PRODUÇÃO

Código	Desc. Abreviada	Descrição	Descrição
1	A	Ambiente A	Alta
2	B	Ambiente B	Média alta
3	C	Ambiente C	Média
4	D	Ambiente D	Média Baixa
5	E	Ambiente E	Baixa
99	ADEF	A Definir	A Definir

QUADRO 25 - FÓRMULA DO ADUBO

N.	Descrição Abreviada
1	Adubo 27-00-24
2	Adubo 32-00-02 Nit.Amonia
3	Adubo GR 20-05-20
4	Adubo GR 21-00-21
5	Adubo GR 31-00-02 Nitrato
6	Adubo Gran. Plantio 00.20
7	Oxido Magnesio 86/90 MgO-
8	Ureia Agricola (46-00-00)

QUADRO 26 - VARIEDADE

N.	Variedade	N.	Variedade	N.	Variedade	N.	Variedade
1	CT92-1634	22	IACSP93-3046	43	RB885007	64	SP83-5073
2	CTC1	23	IACSP93-6006	44	RB886022	65	SP84-5560
3	CTC10	24	IACSP95-3028	45	RB906022	66	SP85-3877
4	CTC12	25	IACSP95-5000	46	RB925211	67	SP86-155
5	CTC13	26	Q-138	47	RB925268	68	SP86-42
6	CTC14	27	RB72454	48	RB925345	69	SP87-365
7	CTC15	28	RB825336	49	RB92579	70	SP88-725
8	CTC17	29	RB835054	50	RB928064	71	SP89-1115
9	CTC2	30	RB835089	51	RB935744	72	SP90-1107
10	CTC3	31	RB835486	52	RB935907	73	SP90-1638
11	CTC4	32	RB845210	53	RB935925	74	SP90-1644
12	CTC5	33	RB845257	54	RB946022	75	SP90-3414
13	CTC6	34	RB855035	55	SP79-1011	76	SP90-3723
14	CTC7	35	RB855113	56	SP79-2233	77	SP91-1049
15	CTC8	36	RB855156	57	SP80-1816	78	SP91-1285
16	CTC9	37	RB855453	58	SP80-1842	79	SP91-1397
17	DIVERSAS	38	RB855536	59	SP80-185	80	SP91-3011
18	IAC86-2210	39	RB855595	60	SP80-3280	81	SP92-1634
19	IAC87-3184	40	RB855598	61	SP80-3480	82	VIV.EXP.
20	IAC87-3396	41	RB865230	62	SP81-3250		
21	IAC91-2195	42	RB867515	63	SP83-2847		

QUADRO 27 - TIPO DE SOLO

Código	Desc. Abrev.	Descrição
1	Gera	Geral
16	LR.1	LR1 Text. Finas,Eutrof ou Endoeutroficos
19	LR.4	LR4 Text. Finas,Conc. Eutrof. Distr. Epie
21	LR.5	LR5 Text. Finas,Conc. Distrof. ou Epieutr
22	LR.2	LR2 Text. Finas,Distrof. ou Epieutroficos
23	LR.3	LR3 Latossolo Roxo
24	LR.2 ^a	LR2A Text. Finas,Atritos ou Endoatricos
32	LVE.3	LVE3 Latossolo Vermelho Escuro
34	LVE.2 ^a	LVE2A T.Fin.,Atricos ou Endoatricos
35	LVE.2	LVE2 T.Fin. Distroficos ou Epieutroficos
36	LVE.1	LVE1 T.Fin. Eutroficos ou Endoeutroficos
45	LVE.4	LVE4 T.Med. Eutroficos ou Endoeutroficos
47	LVE.6	LVE6 T.Med. Alicos ou Endoalicos
48	LVE.8	LVE8 T.Gros. Distroficos ou Epieutroficos
49	LVE.7	LVE7 T.Gros. Eutroficos ou Endoeutroficos
51	LVE.5	LVE5 T.Med. Distroficos ou Epieutroficos
52	LVE.9	LVE9 T.Gros. Alicas ou Epial. ou Endoalic
53	LVE.11	LVE11 T.Gros. Alicas ou Epial. ou Endoalic
74	LVA.4	LVA4 T.Média, Eutroficos ou Endoeutroficos
75	LVA.1	LVA1 T.Finas, Eutroficos ou Endoeutroficos
76	LVA.2	LVA2 T.Finas, Distroficos ou Epieutoficos
77	LVA.3	LVA3 T.Fin,Alicos,Epialicos ou Endoalicos
78	LVA.5	LVA5 T.Med, Distroficos ou Epieutroficos
79	LVA.8	LVA8 T.Gros, Distroficos ou Epieutroficos
80	LVA.12	LVA12 T.Mui.Gr,Alicos ou Epial ou Endoal
81	LVA.6	LVA6 T.Med, Alicos ou Epialicos ou Endoal
82	LVA.7	LVA7 T.Gros, Eutroficos ou Endoeutrofico
83	LVA.11	LVA11 T.Mui.Gr,Alicos ou Epial ou Endoal
84	LVA.9	LVA9 T.Mui.Gr,Alicos ou Epial ou Endoal
85	CB.29	CB29 T.Fin.Substr.Basalto,Eutr. ou Endoeutrof
86	LVA.2 ^a	LVA2A T.Finas, Distroficos ou Epieutoficos
100	LVU.1	LVU1 Una Text. Finas, Eutrof. ou Endoeutro
101	LVU.2	LVU2 Una Text. Finas, Eutrof. ou Epieutrof
257	PVA.2	PVA2 Podzolico Vermelho Amarelo
258	PVA.25	PVA25 Podzolico Vermelho Amarelo
259	PVA.22	PVA22 Podzolico Vermelho Amarelo
300	TRE.4	TRE4 T.F.,Rasa-Fundas, Pedr,Eutr,Endoeutr
321	TRE.1	TRE1 T. Fin. Eutrof. ou Endoeutroficos
322	TRE.3	TRE3 Terra Roxa Estruturada

(CONT.) QUADRO 27 - TIPO DE SOLO

Código	Desc. Abrev.	Descrição
326	TRE.7	TRE7 Latos.,T.Fin Eutr.ou Endoeutroficos
405	AQ.1	AQ1 Eutroficas ou Endoeutroficas
406	AQ.2	AQ2 Distroficas ou Epieutroficas
407	AQ.3	AQ3 Areia Quartzosas
408	AQ.4	AQ4 Areia Quartzosas
409	AQ.5	AQ5 Areia Quartzosas
442	LI	LI Litolicos, T. Fin. Subst, Basalt. Eutrof
443	LI.1	LI1 Litolicos, T. Fin. Subst, Basalt. Eutrof
500	HI	HI Hidromorficos Indiscriminados
600	GX	GX Gleissolo Haplico
999	ADEF	A Definir

APÊNDICE C – MAPA DAS ESTAÇÕES PLUVIOMÉTRICAS

QUADRO 28 - FAZENDAS

Cod. Faz.	Dist.(Km)	Município	Setor	Estação Pluviométrica
0001	11,5	Barrinha	1	Barrinha
0002	15,0	Barrinha	1	Barrinha
0003	16,0	Barrinha	1	Barrinha
0005	14,0	Barrinha	1	Fundão(Alem)
0012	5,5	Pradópolis	1	Fundão(Alem)
0019	5,5	Pradópolis	1	Fundão(Alem)
0102	15	Barrinha	1	Agua Branca
0109	9,5	Pradópolis	2	Carabolante
0201	27,0	Guatapará	3	Barreiro
0210	24,0	Guatapará	3	Figueira
210 ^a	26,0	Guatapará	3	Figueira
210B	26,0	Guatapará	3	Figueira
210C	24,0	Guatapará	3	Figueira
210C	24,0	Guatapará	4	Figueira
0212	28,0	Guatapará	3	Barreiro
212 ^a	26,0	Guatapará	3	Barreiro
212B	25,0	Guatapará	3	Barreiro
212C	24,0	Guatapará	3	Barreiro
212E	31,0	Guatapará	3	Barreiro
212F	29,0	Guatapará	3	Barreiro
0228	33,0	Guatapará	3	Barreiro
0242	25,5	Guatapará	3	Santa Margarida
0265	36,0	Guatapará	3	Barreiro
0279	40,0	Guatapará	3	Barreiro
0280	42,0	Guatapará	3	Barreiro
0282	41,0	Guatapará	3	Barreiro
0285	39,0	Guatapará	3	Barreiro
0287	40,0	Guatapará	3	Barreiro
0291	39,0	Guatapará	3	Barreiro
0295	40,0	Guatapará	3	Barreiro
0296	39,5	Guatapará	3	Barreiro
0297	41,0	Guatapará	3	Barreiro
0298	41,0	Guatapará	3	Barreiro
0299	39,5	Guatapará	3	Barreiro
0305	36,0	Guatapará	3	Barreiro
0306	37,0	Guatapará	3	Barreiro

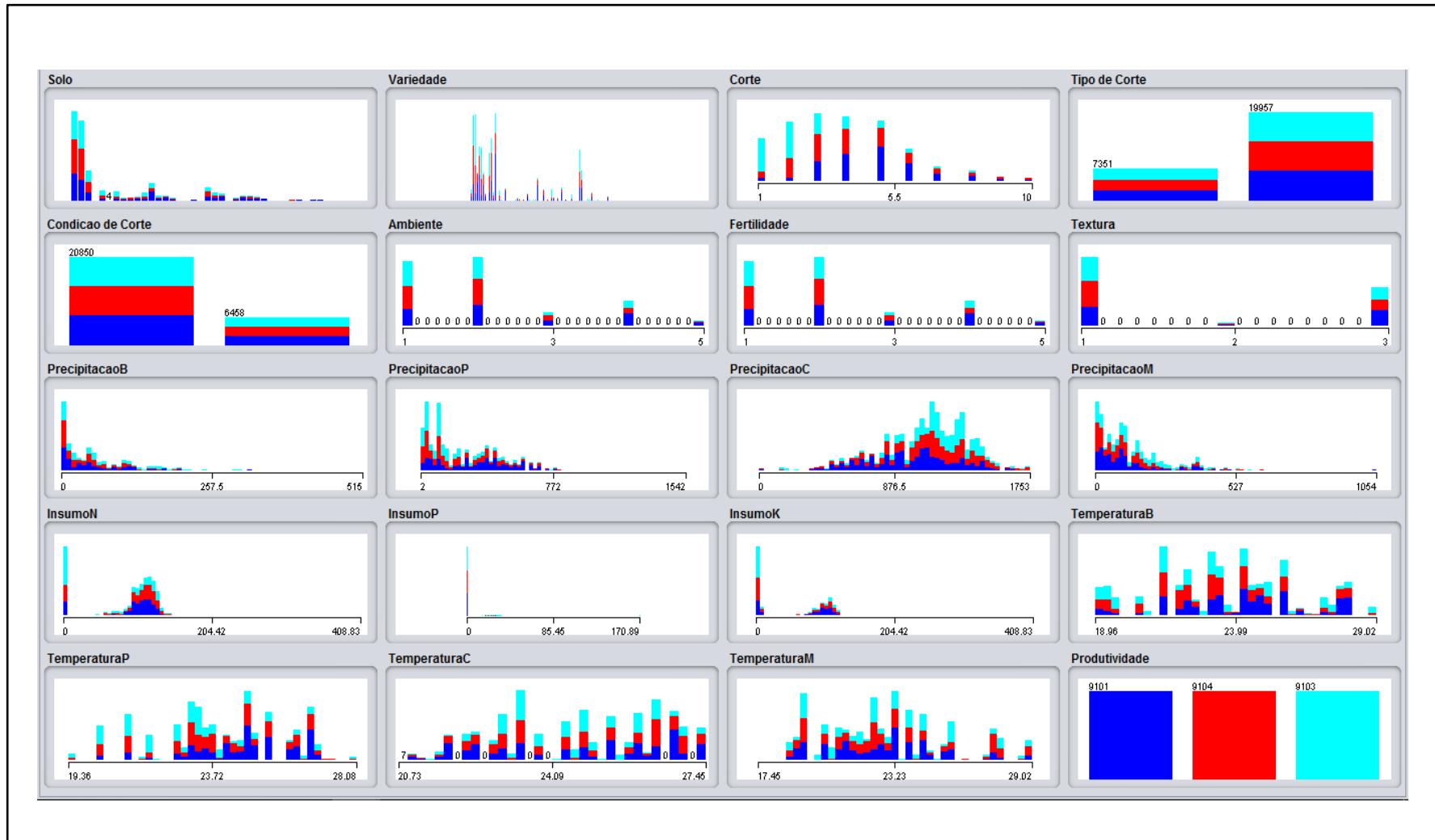
(CONT) QUADRO 28 - FAZENDAS

Cod. Faz.	Dist.(Km)	Município	Setor	Estação Pluviométrica
0307	37,0	Guatapará	3	Barreiro
0308	36,0	Guatapará	3	Barreiro
0309	39,0	Guatapará	3	Barreiro
0310	37,0	Guatapará	3	Barreiro
0353	32,0	Guatapará	3	Barreiro
0382	33,0	Guatapará	3	Barreiro
0420	19,0	Guatapará	4	Aparecida
420 ^a	23,0	Guatapará	4	Aparecida
0431	27,5	Ribeirão Preto	4	Estreito
0481	34,0	Ribeirão Preto	4	Cruz do Pedro
0495	36,0	Ribeirão Preto	4	Aparecida
0547	9,0	Guariba	5	São Bento
0550	16,0	Guariba	5	São Bento
0551	13,0	Guariba	5	São Bento
0552	11,5	Guariba	5	São Bento
0553	16,5	Guariba	5	São Bento
0554	16,0	Guariba	5	São Bento
0556	11,0	Jaboticabal	5	São Bento
0557	16,5	Guariba	5	São Bento
0559	16,0	Guariba	5	São Bento
0560	16,0	Guariba	5	São Bento
0640	25,0	Motuca	6	Limeira
640 ^a	24,0	Motuca	6	Limeira
0642	27,5	Motuca	6	Limeira
0644	19,5	Motuca	6	Limeira
644 ^a	20,0	Motuca	6	Limeira
644B	23,0	Motuca	6	Limeira
0648	20,0	Motuca	6	Limeira
0679	28,0	Motuca	6	Limeira
0690	25,0	Motuca	6	Limeira
0700	43,0	Rincão	8	Modelo
0701	44,0	Rincão	8	Modelo
0702	51,0	Rincão	8	Modelo
0707	43,0	Rincão	6	Modelo
0708	53,0	Rincão	8	Modelo
0711	54,0	Araraquara	8	Modelo
0761	48,0	Luis Antonio	7	Fortaleza
0789	51,0	Luiz Antonio	7	América
789 ^a	45,0	Luiz Antonio	7	América
789B	52,0	Luiz Antonio	7	América
789C	52,0	Luiz Antonio	7	América

(CONT) QUADRO 28 - FAZENDAS

Cod. Faz.	Dist.(Km)	Município	Setor	Estação Pluviométrica
0795	60,0	Luiz Antonio	7	Limoeiro
795 ^a	61,0	Luiz Antonio	7	Limoeiro
795B	59,0	Luiz Antonio	7	Limoeiro
795C	57,0	Luiz Antonio	7	Limoeiro
0901	24,5	Motuca	6	Santa Luiza
0902	31,5	Motuca	6	Santa Luiza
0903	24,5	Motuca	6	Santa Luiza
0904	26,0	Motuca	6	Santa Luiza
0905	26,0	Motuca	8	Santa Luiza
0906	24,5	Motuca	6	Santa Luiza
0907	25,0	Motuca	6	Santa Luiza
0908	26,5	Motuca	8	Santa Luiza
908A	27,5	Motuca	8	Santa Luiza
908B	27,5	Motuca	8	Santa Luiza
0909	27,0	Motuca	6	Santa Luiza
0909	27,0	Motuca	8	Santa Luiza
0910	23,5	Motuca	6	Santa Luiza
0913	27,5	Motuca	8	Santa Luiza
0914	30,5	Motuca	8	Santa Luiza
0915	30,5	Motuca	8	Santa Luiza
0916	30,5	Motuca	8	Santa Luiza
0917	39,0	Rincão	8	Modelo
0918	32,5	Rincão	8	Modelo
0919	39,5	Rincão	8	Modelo
0920	36,5	Araraquara	8	Modelo
0921	40,0	Araraquara	8	Modelo
0922	39,5	Araraquara	8	Modelo
0923	44,5	Araraquara	8	Modelo
923A	43,5	Araraquara	8	Modelo
0924	48,5	Matão	8	Modelo
0925	48,5	Matão	8	Modelo
0927	21,5	Motuca	6	Limeira

APÊNDICE D – HISTOGRAMAS DOS ATRIBUTOS UTILIZADOS NOS MODELOS DE PRODUTIVIDADE



APÊNDICE E – DESCRIÇÃO DE MÉTODOS DE SELEÇÃO DE ATRIBUTOS DA FERRAMENTA WEKA

Método	Característica
<i>Correlation</i>	Avalia o valor de um atributo, mensurando a correlação (Pearson's) entre este e sua classe.
<i>Gain Ratio</i>	Avalia o valor de um atributo, mensurando a relação/proporção de ganho, relacionado a classe
Information gain	Avalia o valor de um atributo, mensurando o ganho de informação, relacionado a classe
<i>OneR</i>	Avalia o valor de um atributo usando o classificador OneR
<i>ReliefF</i>	Avalia o valor de um atributo, amostrando repetidamente uma instância e considerando o valor do dado atributo para a instância mais próxima da mesma classe e de uma classe diferente.
<i>SymmetricalUncert</i>	Avalia o valor de um atributo mensurando a incerteza simétrica relacionada a classe.

FONTE: APRESENTADO NA FERRAMENTA WEKA (LIVRE TRADUÇÃO)

APÊNDICE F – REGRAS DE DECISÃO REFERENTE À MODELAGEM 10

==== Run information ====

SCHEME: WEKA.CLASSIFIERS.TREES.J48 -C 0.25 -B -M 100

RELATION: CENSOVARIETAL SOMENTE VALORES_SUCESSO_CONJUNTOREDUZIDO-WEKA.FILTERS.UNSUPERVISED.ATTRIBUTE.REMOVE-R1-3-WEKA.FILTERS.UNSUPERVISED.ATTRIBUTE.DISCRETIZE-F-B3-M-1.0-R4-WEKA.FILTERS.UNSUPERVISED.ATTRIBUTE.NUMERICToNOMINAL-R7-9

INSTANCES: 27308

ATTRIBUTES: 20

SOLO

VARIÉDADE

CORTE

PRODUTIVIDADE

TIPO DE CORTE

CONDICAO DE CORTE

AMBIENTE

FERTILIDADE

TEXTURA

PRECIPITACAOB

PRECIPITACAOP

PRECIPITACAOC

PRECIPITACAO M

INSUMON

INSUMOP

INSUMOK

TEMPERATURAB

TEMPERATURAP

TEMPERATURAC

TEMPERATURAM

TEST MODE: 10-FOLD CROSS-VALIDATION

==== CLASSIFIER MODEL (FULL TRAINING SET) ====

J48 PRUNED TREE

CORTE <= 2.0

| PRECIPITACAOB <= 318.5

| | TEMPERATURAC <= 22.11899631

| | | TEMPERATURAB <= 23.29: '(98.224361-INF)' (121.0/31.0)

```

| | | TEMPERATURAB > 23.29: '(76.915-98.224361]' (324.0/99.0)
| | | TEMPERATURAC > 22.11899631
| | | INSUMOK <= 87.91818
| | | | PRECIPITACAO P <= 363.5: '(98.224361-INF)' (3730.0/421.0)
| | | | PRECIPITACAO P > 363.5
| | | | | CORTE <= 1.0: '(98.224361-INF)' (462.0/103.0)
| | | | | CORTE > 1.0: '(76.915-98.224361]' (207.0/92.0)
| | | | INSUMOK > 87.91818
| | | | SOLO = AQ.2: '(76.915-98.224361]' (139.0/57.0)
| | | | SOLO != AQ.2
| | | | | VARIEDADE = RB855453: '(98.224361-INF)' (204.0/24.0)
| | | | | VARIEDADE != RB855453
| | | | | | TEMPERATURAB <= 21.51333333
| | | | | | TEMPERATURAC <= 26.12329032: '(98.224361-INF)' (589.0/93.0)
| | | | | | TEMPERATURAC > 26.12329032: '(76.915-98.224361]' (138.0/69.0)
| | | | | | TEMPERATURAB > 21.51333333
| | | | | | INSUMON <= 112.176: '(76.915-98.224361]' (325.0/88.0)
| | | | | | INSUMON > 112.176
| | | | | | | PRECIPITACAO C <= 1476.0
| | | | | | | PRECIPITACAO M <= 180.0
| | | | | | | | INSUMOK <= 109.5507: '(76.915-98.224361]' (242.0/98.0)
| | | | | | | | INSUMOK > 109.5507
| | | | | | | | | VARIEDADE = SP91-1049: '(76.915-98.224361]' (138.0/63.0)
| | | | | | | | | VARIEDADE != SP91-1049: '(98.224361-INF)' (342.0/132.0)
| | | | | | | | | PRECIPITACAO M > 180.0: '(98.224361-INF)' (325.0/77.0)
| | | | | | | | | PRECIPITACAO C > 1476.0: '(76.915-98.224361]' (145.0/42.0)
| | | | | | PRECIPITACAO B > 318.5
| | | | | TEMPERATURAC <= 20.9787276: '(-INF-76.915]' (131.0/49.0)
| | | | | TEMPERATURAC > 20.9787276: '(76.915-98.224361]' (148.0/55.0)
CORTE > 2.0
| | | VARIEDADE = SP89-1115
| | | TEMPERATURAC <= 26.83344086
| | | | TEXTURA = 1
| | | | AMBIENTE = 1
| | | | | TEMPERATURAB <= 21.51333333: '(98.224361-INF)' (169.0/39.0)
| | | | | TEMPERATURAB > 21.51333333
| | | | | | PRECIPITACAO B <= 1.0: '(76.915-98.224361]' (123.0/16.0)
| | | | | | PRECIPITACAO B > 1.0
| | | | | | | TEMPERATURAP <= 24.68338185
| | | | | | | INSUMOK <= 105.6672: '(76.915-98.224361]' (216.0/110.0)
| | | | | | | INSUMOK > 105.6672: '(98.224361-INF)' (102.0/60.0)
| | | | | | | TEMPERATURAP > 24.68338185: '(98.224361-INF)' (190.0/64.0)
| | | | | | AMBIENTE != 1: '(76.915-98.224361]' (319.0/107.0)
| | | | | | TEXTURA != 1: '(98.224361-INF)' (278.0/69.0)
| | | | | | TEMPERATURAC > 26.83344086: '(76.915-98.224361]' (160.0/33.0)

```

```

| VARIEDADE != SP89-1115
| | TEMPERATURAP <= 23.69523297
| | | AMBIENTE = 5: '(-INF-76.915]' (204.0/17.0)
| | | AMBIENTE != 5
| | | | AMBIENTE = 4
| | | | | PRECIPITACAO C <= 1243.0
| | | | | INSUMON <= 101.52538
| | | | | CORTE <= 4.0
| | | | | | VARIEDADE = SP83-2847: '(-INF-76.915]' (176.0/44.0)
| | | | | | VARIEDADE != SP83-2847: '(76.915-98.224361]' (100.0/38.0)
| | | | | | CORTE > 4.0: '(-INF-76.915]' (307.0/12.0)
| | | | | | N > 101.52538: '(-INF-76.915]' (106.0/54.0)
| | | | | | PRECIPITACAO C > 1243.0: '(76.915-98.224361]' (132.0/72.0)
| | | | | AMBIENTE != 4
| | | | | SOLO = LVE.3: '(-INF-76.915]' (319.0/124.0)
| | | | | SOLO != LVE.3
| | | | | VARIEDADE = RB835486
| | | | | | PRECIPITACAO B <= 22.0: '(-INF-76.915]' (142.0/27.0)
| | | | | | PRECIPITACAO B > 22.0: '(76.915-98.224361]' (149.0/49.0)
| | | | | | VARIEDADE != RB835486
| | | | | | TEMPERATURAP <= 23.578198
| | | | | | INSUMOP <= 26.0
| | | | | | | VARIEDADE = SP80-1842: '(76.915-98.224361]' (346.0/114.0)
| | | | | | | VARIEDADE != SP80-1842
| | | | | | | | TIPO DE CORTE = MA: '(76.915-98.224361]' (1243.0/624.0)
| | | | | | | | TIPO DE CORTE != MA
| | | | | | | | | PRECIPITACAO C <= 1495.0
| | | | | | | | | CORTE <= 3.0
| | | | | | | | | | TEMPERATURA C <= 23.466066: '(98.224361-INF)' (517.0/128.0)
| | | | | | | | | | TEMPERATURA C > 23.466066
| | | | | | | | | | | VARIEDADE = RB855453: '(98.224361-INF)' (171.0/56.0)
| | | | | | | | | | | VARIEDADE != RB855453: '(76.915-98.224361]' (307.0/122.0)
| | | | | | | | | | | CORTE > 3.0
| | | | | | | | | | | INSUMOK <= 30.01572
| | | | | | | | | | | AMBIENTE = 3: '(76.915-98.224361]' (192.0/48.0)
| | | | | | | | | | | AMBIENTE != 3
| | | | | | | | | | | INSUMOK <= 6.4028
| | | | | | | | | | | | PRECIPITACAO B <= 111.25
| | | | | | | | | | | | PRECIPITACAO B <= 77.75
| | | | | | | | | | | | | PRECIPITACAO B <= 50.25
| | | | | | | | | | | | | PRECIPITACAO M <= 385.0
| | | | | | | | | | | | | CORTE <= 5.0
| | | | | | | | | | | | | SOLO = LVE.2: '(76.915-98.224361]' (100.0/26.0)
| | | | | | | | | | | | | SOLO != LVE.2
| | | | | | | | | | | | | | VARIEDADE = SP91-1049: '(76.915-98.224361]' (116.0/51.0)

```

||||| VARIEDADE != SP91-1049: '(98.224361-INF)' (276.0/78.0)
||||| CORTE > 5.0
||||| PRECIPITACAO C <= 1223.25: '(76.915-98.224361]' (194.0/43.0)
||||| PRECIPITACAO C > 1223.25
||||| INSUMOK <= 4.39994: '(-INF-76.915]' (113.0/63.0)
||||| INSUMOK > 4.39994: '(76.915-98.224361]' (107.0/53.0)
||||| PRECIPITACAO M > 385.0: '(98.224361-INF)' (107.0/27.0)
||||| PRECIPITACAO B > 50.25
||||| CORTE <= 4.0: '(98.224361-INF)' (130.0/8.0)
||||| CORTE > 4.0
||||| AMBIENTE = 1: '(98.224361-INF)' (160.0/32.0)
||||| AMBIENTE != 1: '(76.915-98.224361]' (115.0/37.0)
||||| PRECIPITACAO B > 77.75: '(76.915-98.224361]' (170.0/40.0)
||||| PRECIPITACAO B > 111.25: '(98.224361-INF)' (167.0/79.0)
||||| INSUMOK > 6.4028: '(98.224361-INF)' (166.0/43.0)
||||| INSUMOK > 30.01572
||||| TEMPERATURAP <= 21.08831064: '(76.915-98.224361]' (149.0/30.0)
||||| TEMPERATURAP > 21.08831064
||||| PRECIPITACAO B <= 12.0: '(-INF-76.915]' (114.0/60.0)
||||| PRECIPITACAO B > 12.0
||||| TEMPERATURAM <= 22.45666667: '(76.915-98.224361]' (223.0/66.0)
||||| TEMPERATURAM > 22.45666667: '(-INF-76.915]' (229.0/102.0)
||||| PRECIPITACAO C > 1495.0: '(-INF-76.915]' (48.0/16.0)
||||| UNSUMOP > 26.0
||||| CORTE <= 4.0: '(76.915-98.224361]' (107.0/56.0)
||||| CORTE > 4.0: '(-INF-76.915]' (119.0/37.0)
||||| TEMPERATURAP > 23.578198: '(76.915-98.224361]' (557.0/196.0)
|| TEMPERATURAP > 23.69523297
|| VARIEDADE = SP90-1638: '(98.224361-INF)' (197.0/88.0)
|| VARIEDADE != SP90-1638
|| CORTE <= 3.0
|| PRECIPITACAO C <= 665.0: '(-INF-76.915]' (491.0/100.0)
|| PRECIPITACAO C > 665.0
|| TEXTURA = 1
|| PRECIPITACAO C <= 906.25
|| PRECIPITACAO P <= 580.0: '(76.915-98.224361]' (169.0/40.0)
|| PRECIPITACAO P > 580.0: '(-INF-76.915]' (147.0/30.0)
|| PRECIPITACAO C > 906.25
|| PRECIPITACAO C <= 1267.0
|| TEMPERATURAB <= 22.45666667: '(76.915-98.224361]' (159.0/43.0)
|| TEMPERATURAB > 22.45666667
|| VARIEDADE = SP91-1049: '(76.915-98.224361]' (135.0/36.0)
|| VARIEDADE != SP91-1049
|| PRECIPITACAO C <= 1010.0: '(76.915-98.224361]' (152.0/55.0)
|| PRECIPITACAO C > 1010.0: '(98.224361-INF)' (290.0/82.0)

```

||||| PRECIPITACAO C > 1267.0: '(76.915-98.224361]' (157.0/81.0)
||||| TEXTURA != 1
||||| AMBIENTE = 2: '(76.915-98.224361]' (215.0/56.0)
||||| AMBIENTE != 2
||||| TEMPERATURA B <= 25.18016129: '(-INF-76.915]' (402.0/170.0)
||||| TEMPERATURA B > 25.18016129: '(76.915-98.224361]' (160.0/45.0)
|||| CORTE > 3.0
|||| VARIEDADE = SP83-2847: '(-INF-76.915]' (757.0/55.0)
|||| VARIEDADE != SP83-2847
|||| PRECIPITACAO M <= 24.0
|||| VARIEDADE = RB835486: '(-INF-76.915]' (175.0/37.0)
|||| VARIEDADE != RB835486
|||| PRECIPITACAO C <= 1048.0
|||| PRECIPITACAO P <= 370.0
|||| INSUMO K <= 100.1256
|||| INSUMO N <= 123.418: '(76.915-98.224361]' (132.0/62.0)
|||| INSUMO N > 123.418: '(-INF-76.915]' (119.0/25.0)
|||| INSUMO K > 100.1256: '(-INF-76.915]' (125.0/5.0)
|||| PRECIPITACAO P > 370.0
|||| SOLO = LR.2: '(-INF-76.915]' (102.0/43.0)
|||| SOLO != LR.2: '(76.915-98.224361]' (110.0/35.0)
|||| PRECIPITACAO C > 1048.0
|||| VARIEDADE = SP81-3250: '(-INF-76.915]' (116.0/33.0)
|||| VARIEDADE != SP81-3250
|||| N <= 112.176
|||| CONDICAO DE CORTE = C
|||| PRECIPITACAO C <= 1114.0: '(76.915-98.224361]' (269.0/84.0)
|||| PRECIPITACAO C > 1114.0: '(-INF-76.915]' (109.0/30.0)
|||| CONDICAO DE CORTE != C: '(76.915-98.224361]' (103.0/56.0)
|||| N > 112.176: '(76.915-98.224361]' (311.0/89.0)
|||| PRECIPITACAO M > 24.0
|||| SOLO = LR.1
|||| CORTE <= 4.0
|||| PRECIPITACAO M <= 62.0: '(-INF-76.915]' (108.0/24.0)
|||| PRECIPITACAO M > 62.0
|||| INSUMO K <= 46.9968: '(76.915-98.224361]' (110.0/15.0)
|||| INSUMO K > 46.9968
|||| TEMPERATURA C <= 26.77819355: '(76.915-98.224361]' (153.0/70.0)
|||| TEMPERATURA C > 26.77819355: '(-INF-76.915]' (103.0/47.0)
|||| CORTE > 4.0: '(-INF-76.915]' (830.0/248.0)
|||| SOLO != LR.1
|||| K <= 112.3752
|||| PRECIPITACAO C <= 1056.0: '(-INF-76.915]' (2018.0/245.0)
|||| PRECIPITACAO C > 1056.0
|||| PRECIPITACAO M <= 38.0: '(-INF-76.915]' (104.0/6.0)

```

```

||||| PRECIPITACAO M > 38.0
||||| AMBIENTE = 4: '(-INF-76.915]' (281.0/41.0)
||||| AMBIENTE != 4
||||| VARIEDADE = SP80-1816: '(-INF-76.915]' (255.0/40.0)
||||| VARIEDADE != SP80-1816
||||| PRECIPITACAO P <= 403.0
||||| INSUMO N <= 115.0184
||||| PRECIPITACAO C <= 1185.5: '(76.915-98.224361]' (160.0/76.0)
||||| PRECIPITACAO C > 1185.5: '(-INF-76.915]' (203.0/52.0)
||||| N > 115.0184: '(76.915-98.224361]' (149.0/42.0)
||||| PRECIPITACAO P > 403.0: '(-INF-76.915]' (354.0/92.0)
||||| INSUMO K > 112.3752: '(-INF-76.915]' (233.0/97.0)

```

NUMBER OF LEAVES : 99

SIZE OF THE TREE : 197

TIME TAKEN TO BUILD MODEL: 1.28 SECONDS

==== STRATIFIED CROSS-VALIDATION ===

==== SUMMARY ===

CORRECTLY CLASSIFIED INSTANCES	19705	72.1583 %
INCORRECTLY CLASSIFIED INSTANCES	7603	27.8417 %
KAPPA STATISTIC	0.5824	
MEAN ABSOLUTE ERROR	0.2618	
ROOT MEAN SQUARED ERROR	0.3648	
RELATIVE ABSOLUTE ERROR	58.9002 %	
ROOT RELATIVE SQUARED ERROR	77.3849 %	
TOTAL NUMBER OF INSTANCES	27308	

==== DETAILED ACCURACY BY CLASS ===

TP RATE	FP RATE	PRECISION	RECALL	F-MEASURE	MCC	ROC AREA	PRC AREA	CLASS
0,758	0,121	0,757	0,758	0,758	0,637	0,897	0,803	'(-INF-76.915]'
0,637	0,192	0,624	0,637	0,630	0,442	0,782	0,627	'(76.915-98.224361]'
0,770	0,104	0,787	0,770	0,778	0,670	0,908	0,812	'(98.224361-INF)'
WEIGHTED AVG.	0,722	0,139	0,723	0,722	0,722	0,583	0,862	0,747

==== CONFUSION MATRIX ===

A	B	C	<- CLASSIFIED AS
6902	1781	418	A = '(-INF-76.915]'
1833	5796	1475	B = '(76.915-98.224361]'
379	1717	7007	C = '(98.224361-INF)'

APÊNDICE G – PROTÓTIPO DAS TELAS PARA GERENCIAR USUÁRIOS DO SISTEMA

A Tela da Figura 40 tem por objetivo o gerenciamento dos usuários. A partir dela é possível visualizar, editar ou excluir usuários ou dar início ao cadastro de um novo usuário. As ações das telas das Figuras 40 e 41 podem ser realizadas apenas por usuários administradores do sistema.

FIGURA 40 - PROTÓTIPO DA TELA PARA GERENCIAR USUÁRIOS

FONTE: A AUTORA

O cadastro de novos usuários será realizado a partir da tela proposta na Figura 41.

O protótipo da tela para cadastro de usuário SECC é apresentado em uma interface web. No topo, há uma barra com o logo "SECC" e três links: "Meus Arquivos", "Meus Cenários" e "Graça". Abaixo da barra, uma barra de navegação mostra "Úsuarios > Incluir >". O formulário principal, intitulado "Novo Usuário", contém campos para "Nome" (campo com placeholder), "E-mail" (campo com placeholder), "CPF" (campo com placeholder) e "Senha" (campo com placeholder). Abaixo dos campos, há um botão "salvar".

FIGURA 41 - PROTÓTIPO DA TELA PARA CADASTRO DE USUÁRIO

FONTE: A AUTORA

A tela inicial da ferramenta permite o *login* de um usuário previamente cadastrado. A Figura 42 é uma proposta para essa interface.

SECC - Sistema de Exploração dos Cenários da Cana

O protótipo mostra uma interface de usuário com um formulário para login. No topo, o título "SECC - Sistema de Exploração dos Cenários da Cana" é exibido em negrito. Abaixo, há dois campos de texto rotulados "E-mail" e "Senha". À esquerda, uma link rotulado "Esqueceu a senha?". À direita, um botão rotulado "Login".

FIGURA 42 - PROTÓTIPO DA TELA DE LOGIN E SENHA

FONTE: A AUTORA

APÊNDICE H – TELAS PARA GERENCIAR USUÁRIOS DO SISTEMA

As telas apresentadas a seguir (Figuras 43 e 44) são um exemplo de utilização do SECC, reservado apenas aos usuários com status de administradores do sistema, para gerenciar os demais usuários (criação, alteração e exclusão). A partir da tela da Figura 43 é possível visualizar os usuários já cadastrados, bem como dar início ao cadastro de um novo usuário.

The screenshot shows a user interface for managing users. At the top, there is a navigation bar with links 'SECC', 'Minhas Árvores', 'Meus Cenários', and 'Ações'. Below the navigation, a green success message box displays the text 'Usuário deletado com sucesso.' (User deleted successfully.). To the right of the message is a close button (X). Below the message, the title 'Lista de Usuários' (User List) is centered above a table. A blue button labeled 'Novo Usuário' (New User) is located in the top right corner of the table area. The table has columns: 'Cod', 'Nome', 'Email', 'CPF', 'Cadastrado-em', and 'Ações'. It contains two rows of data:

Cod	Nome	Email	CPF	Cadastrado-em	Ações
2	Graça Tomazela	gtomazela@fatecindaiatuba.edu.br	08510705828	2017-03-13 16:53:04 -0300	Editar Deletar
1	admin	admin@teste.com	32221315942	2016-11-28 14:56:17 -0200	Editar Deletar

FIGURA 43 - TELA PARA INSERÇÃO DE NOVO USUÁRIO

FONTE: A AUTORA

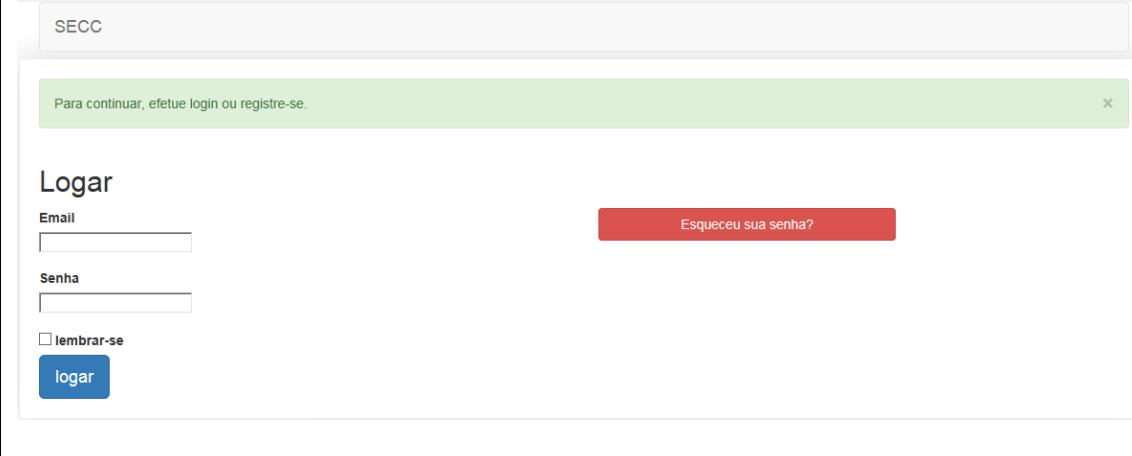
Após selecionar a opção “Novo Usuário”, a tela da Figura 44 será disponibilizada para cadastro de um novo usuário.

The screenshot shows a user interface for creating a new user. At the top, there is a navigation bar with links 'SECC', 'Minhas Árvores', 'Meus Cenários', and 'Ações'. Below the navigation, the title 'Novo Usuário' (New User) is displayed. The form consists of four input fields: 'Name', 'Email', 'Cpf', and 'Password'. Each field has a corresponding label and an input box. Below the input boxes are two buttons: a blue 'Enviar' (Send) button on the left and an orange 'Voltar' (Back) button on the right.

FIGURA 44 - TELA PARA CADASTRO DE USUÁRIOS

FONTE: A AUTORA

A tela da Figura 45 tem o propósito de efetuar o *login*, no sistema, para um usuário já cadastrado.



The image shows a login interface for a system named 'SECC'. At the top, there is a green header bar with the text 'Para continuar, efetue login ou registre-se.' (To continue, log in or register). Below this, the word 'Logar' (Log in) is centered. There are two input fields: 'Email' and 'Senha' (Password), each with a corresponding input box. To the right of the 'Senha' field is a red button labeled 'Esqueceu sua senha?' (Forgot password?). Below the input fields is a checkbox labeled 'Lembrar-se' (Remember me). At the bottom is a blue button labeled 'logar' (Log in).

FIGURA 45 - TELA DE LOGIN

FONTE: A AUTORA