Kelly Marques de Oliveira Lopes

Modelos Baseados em Data Mining Para Classificação Multitemporal de Culturas No Mato Grosso Utilizando Dados de NDVI/MODIS

CAMPINAS
2013

UNIVERSIDADE ESTADUAL DE CAMPINAS

Instituto de Matemática, Estatística
e Computação Científica

KELLY MARQUES DE OLIVEIRA LOPES

MODELOS BASEADOS EM DATA MINING PARA CLASSIFICAÇÃO MULTITEMPORAL DE CULTURAS NO MATO GROSSO UTILIZANDO DADOS DE NDVI/MODIS

Dissertação apresentada ao Instituto de Matemática, Estatística e Computação Científica da Universidade Estadual de Campinas como parte dos requisitos exigidos para a obtenção do título de Mestra em matemática aplicada.

Orientador: Laércio Luis Vendite

Coorientador: Stanley Robson de Medeiros Oliveira

ESTE EXEMPLAR CORRESPONDE À VERSÃO FINAL DA DISSERTAÇÃO DEFENDIDA PELA ALUNA KELLY MARQUES DE OLIVEIRA LOPES, E ORIENTADA PELO PROF. DR. LAÉRCIO LUIS VENDITE.

Assinatura do Orientador

Assinatura do Coorientador

CAMPINAS
2013

Ficha catalográfica
Universidade Estadual de Campinas
Biblioteca do Instituto de Matemática, Estatística e Computação Científica
Maria Fabiana Bezerra Muller - CRB 8/6162

L881m

Lopes, Kelly Marques de Oliveira, 1982-
Modelos baseados em data mining para classificação multitemporal de culturas em Mato Grosso utilizando dados de NDVI/MODIS / Kelly Marques de Oliveira Lopes. – Campinas, SP : [s.n.], 2013.

Orientador: Laércio Luis Vendite.
Coorientador: Stanley Robson de Medeiros Oliveira.
Dissertação (mestrado) – Universidade Estadual de Campinas, Instituto de Matemática, Estatística e Computação Científica.

1. Mineração de dados. 2. Sensoriamento remoto. 3. Mapeamento da cobertura do solo. 4. Reconhecimento de padrões. 5. Modelagem de dados. I. Vendite, Laércio Luis,1954-. II. Oliveira, Stanley Robson de Medeiros. III. Universidade Estadual de Campinas. Instituto de Matemática, Estatística e Computação Científica. IV. Título.

Informações para Biblioteca Digital

Título em outro idioma: Models based on data mining for classification multitemporal crop in Mato Grosso data using NDVI/MODIS
Palavras-chave em inglês:
Data mining
Remote sensing
Mappind land cover
Pattern recognition
Data modeling
Área de concentração: Matemática Aplicada e Computacional
Titulação: Mestra em Matemática Aplicada e Computacional
Banca examinadora:
Laércio Luis Vendite [Orientador]
Laécio Carvalho de Barros
Julio César Dalla Mora Esquerdo
Data de defesa: 09-08-2013
Programa de Pós-Graduação: Matemática Aplicada e Computacional

Dissertação de Mestrado defendida em 09 de agosto de 2013 e aprovada

Pela Banca Examinadora composta pelos Profs. Drs.

Prof.(a). Dr(a). LAÉRCIO LUIS VENDITE

Prof.(a). Dr(a). LAÉCIO CARVALHO DE BARROS

Prof.(a). Dr(a). JULIO CÉSAR DALLA MORA ESQUERDO

Abstract

The development of studies in the field of geotechnology and increased ability to store data have improved the exploration and study of satellite images obtained by satellite sensors. The mapping of land cover, estimates of crop productivity and crop forecasting is important information for the farmer and for the government, because this information is essential to support decisions related to production, estimates of purchase and sale, import and calculations and export. An alternative use for data analysis and coverage will be obtained by means of sensors, is the use of data mining techniques since these techniques can be used to transform data and information on the knowledge that will support decisions on agricultural planning. In this work, we used data on the multitemporal vegetation index NDVI derived from MODIS images for monitoring crops of cotton, soybean and corn in the state of Mato Grosso, in the period of the crop year 2008/2009. The dataset supplied by Embrapa Agricultural Informatics, comprised 24 columns and 728 rows, where the 23 first columns refer to the values of NVDI, and the last, the soil cover. The methodology used was based on the model CRISP-DM (Cross Industry Standard Process for Data Mining). Predictive models to classify data on these cultures were prepared and analyzed by machine learning algorithms such as decision trees (J48 and PART), Random Forests (Random Forest). The feature selection improved the *Kappa* index values and accuracy of the models. Classification rules were generated to map the cultures studied (soy, corn and cotton). The results show that the machine learning algorithms are promising for the problem of classification of land cover. In particular, the J48 algorithm, used in conjunction with feature selection done by principal component analysis, stood out against the other by the simplicity and the values presented. The results also revealed the presence of regions of cotton cultivation in other areas of the state, out of those studied.

Keywords: Data mining, Remote sensing, Mapping of land cover, Recognition Patterns, Data modeling.

Resumo

O desenvolvimento de estudos na área de geotecnologia e o aumento na capacidade de armazenar dados têm melhorado a exploração e os estudos de imagens de satélites obtidas através de sensores orbitais. O mapeamento da cobertura da terra, estimativas de produtividade de culturas e a previsão de safras são informações importantes para o agricultor e para o governo, pois essas

informações são essenciais para subsidiar decisões relacionadas à produção, estimativas de compra e venda, e cálculos de importação e exportação. Uma das alternativas para analisar dados de uso e cobertura da terra, obtidos por meio de sensores, é o uso de técnicas de mineração de dados, uma vez que essas técnicas podem ser utilizadas para transformar dados e informações em conhecimentos que irão subsidiar decisões relativas ao planejamento agrícola. Neste trabalho, foram utilizados dados multitemporais sobre o índice de vegetação NDVI, derivados de imagens do sensor MODIS, para o monitoramento das culturas de algodão, soja e milho no estado do Mato Grosso, no período do ano-safra de 2008/2009. O conjunto de dados, fornecido pela Embrapa Informática Agropecuária, foi composto por 24 colunas e 728 linhas, onde as 23 primeiras colunas referem-se aos valores do NVDI, e a última, à cobertura do solo. A metodologia utilizada teve como base o modelo CRISP-DM (Cross Industry Standard Process for Data Mining). Modelos preditivos para classificar dados sobre essas culturas foram elaborados e avaliados por algoritmos de aprendizado de máquina, tais como árvores de decisão (J48 e PART), florestas aleatórias (Random Forest). A seleção de atributos melhorou os valores do índice *Kappa* e a acurácia dos modelos. Foram geradas regras de classificação para mapear as culturas estudadas (soja, milho e algodão). Os resultados revelaram que os algoritmos de aprendizado de máquina são promissores para o problema de classificação de cobertura do solo. Em particular o algoritmo J48, utilizado em conjunto com a seleção de atributos feito por meio de análise de componentes principais, destacou-se em relação ao demais pela simplicidade e pelos valores apresentados. Os resultados também evidenciaram a presençe regiões de cultivo do algodão em outras áreas do estado, fora daquelas estudadas.

Palavras chaves: Mineração de dados, Sensoriamento remoto, Mapeamento da cobertura do solo, Reconhecimento de padrões, Modelagem de dados.

Sumário

Dedicatória **xi**

Agradecimentos **xiii**

Lista de Figuras **xv**

Lista de Tabelas **xviii**

1 Introdução **1**
1.1 O problema de Pesquisa . . . 2
1.2 Hipótese . . . 3
1.3 Objetivos . . . 3

2 Revisão Bibliográfica **5**
2.1 Previsão de Safras . . . 5
2.2 Sensoriamento Remoto . . . 6
2.2.1 Aquisição de Informação . . . 6
2.2.2 Sistemas Sensores . . . 10
2.3 Sensor MODIS . . . 11
2.4 Índice Vegetativo . . . 14
2.5 Mineração de Dados . . . 16
2.5.1 Conceitos de Mineração de Dados . . . 16
2.5.2 Classificação . . . 20
2.6 Métodos para Seleção de Atributos . . . 24
2.6.1 PCA . . . 25
2.6.2 X^2 . . . 27
2.6.3 Gain Ratio e Info Gain . . . 28
2.6.4 CFS . . . 29
2.6.5 Wrapper . . . 30
2.7 Valores Faltantes . . . 30
2.7.1 K-NN . . . 31
2.8 Aprendizado com Classes Desbalanceadas . . . 31
2.9 Validação de Modelos Preditivos . . . 32
2.10 Medidas de Avaliação de Regras . . . 33

2.10.1 Matriz de Confusão . . . 33
2.10.2 Matriz de Contingência . . . 34
2.11 Modelo do processo de descoberta de conhecimento em bases de dados . . . 36

3 Materiais e Métodos **39**
3.1 Considerações Iniciais . . . 39
3.2 Compreensão do domínio . . . 39
3.3 Entendimento dos dados . . . 40
3.3.1 Coleção inicial dos dados e descrição . . . 40
3.3.2 Exploração dos dados . . . 42
3.4 Preparação dos dados . . . 44
3.4.1 Especificação do atributo meta . . . 45
3.4.2 Especificação dos atributos preditivos espectrais . . . 45
3.5 Modelagem . . . 45
3.6 Softwares e parâmetros . . . 46

4 Resultados e Discussões **49**
4.1 Análise exploratória . . . 49
4.1.1 Modelo de predição utilizando regras de classificação . . . 52
4.1.2 Testes com o PCA . . . 52
4.1.3 Testes com o CFS . . . 55
4.1.4 Testes com o Info Gain . . . 58
4.1.5 Testes com o Wrapper . . . 61
4.2 Base de Conhecimentos . . . 82
4.3 Mapas para interpretação das regras . . . 97
4.4 Modelo para interpretação das regras de classificação . . . 99

5 Conclusões **105**

Referência Bibliográfica **108**

Dedico aos meus pais, esposo e filha
e a todos aqueles que se alegram com minha conquista.

Agradecimentos

Agradeço primeiramente a Deus pela vida, por ter colocado em mim a vontade de vencer.

Em segundo lugar minha família que tanto me apoiou durante todo esse tempo. Meus pais José Rodrigues e Rita Marques que sempre me ensinaram que é possível vencer e conquistar basta lutar. Meu querido esposo Mauricelio, esse que ficou quatro anos distantes só para que eu pudesse ver o meu sonho realizado. Minha filha Kethy, que nasceu durante esse período. Aos meus irmãos: Karen, Nathaly e Deivid . Agradeço a Deus por ter vocês em minha vida.

Não poderia deixar de fora meus queridos tios e tias, que direta ou indiretamente me influenciaram, obrigada.

Aos meus amigos do IMECC, pessoas extraordinárias que sempre me ajudaram nas dificuldades. Aos velhos amigos de Rio Claro: Fernanda; Paulo; Sergio; Rodrigo; Claudia; Aline; Moara e minha grande amiga Vanessa, obrigada por estarem presentes quando mais precisei.

Agradeço ao meu orientador Dr. Laércio Luis Vendite, por ter confiado, acreditado e investido em mim. Também ao meu coorientador Dr. Stanley Robson por sua paciência e seus ensinamentos.

À Embrapa Informática Agropecuária, que me forneceu os dados.

Quero expressar também minha gratidão ao Júlio Esquerdo, que muito me ajudou. Ao Luíz Manoel, Vanderson, Flávio, Porfírio, Raniere e ao Estavão pelos momentos de discussão, meu muito obrigada.

Agradeço aos professores Aurélio, Laécio e Joni, que sempre me trataram muito bem, e que me ofereceram uma estrutura um crescimento profissional e pessoal.

Agradeço à Coordenação de Aperfeiçamento de Pessoal de Nível Superior (Capes), pelo auxílio no período de mestrado.

Obrigada IMECC, obrigada Unicamp!

Lista de Figuras

2.1 Interação da radiação eletromagnética fonte-alvo-sensor(ANTUNES, 2005) 7
2.2 Interação da radiação eletromagnética solar com a folha (Adaptado de MOREIRA, 2003) 8
2.3 Curva padrão de refletância de uma vegetação verde (Adaptado de HOFFER, 1978) 9
2.4 Imagem dividida em pixels (Adaptado de UFG, 2003) 13
2.5 Variação do NDVI em uma planta saudável e outra em estado senescente. 15
2.6 Processo KDD (Adaptado FAYYAD et al., 1996) 17
2.7 Etapas da Mineração de Dados(Data Mining, 2009, p.4) 18
2.8 Modelo de classificação (Data Mining, 2009, p.175) 21
2.9 Modelo de uma árvore de decisão (Data Mining: Concepts and Techniques, 2011, p.18) 22
2.10 Análise com o PCA 26
2.11 Modelo de um método de seleção de subconjuntos de atributo utilizando o Wrapper (JONH et al.,1994) 30
2.12 Árvore de decisão induzida dos dados da Tabela 2.5 32
2.13 O processo CRISP-DM(CHAPMAN, 2000) 37

3.1 Composição de 16 dias do NDVI no Estado de Grosso 40
3.2 Série temporal da vegetação 43

4.1 Mapa Geral de Mato Grosso 98
4.2 Áreas de plantio de algodão em Mato Grosso 99
4.3 Árvore de decisão gerada com o J48 e o PCA 99

Lista de Tabelas

2.1 Produtos gerados pelo sensor MODIS(Silva, 2004) 12
2.2 Exemplo de transação de cestas de compras(Data Mining, 2009, p.390) 19
2.3 Exemplo de descarte de variável segundo critério de Jolliffe (Adaptado do Conjunto de dados Iris) 27
2.4 Exemplo de descarte de variável segundo critério de Jolliffe (Adaptado do Conjunto de dados Iris) 27
2.5 Exemplo de conjunto de treinamento para classificar mamíferos. 31
2.6 Matriz de Confusão 33
2.7 Matriz de Contingência 35
2.8 Tabela da classificação *Kappa* segundo LANDIS e KOCH (1977) 36

3.1 Disposição inicial dos dados. 41
3.2 Disposição final dos dados. 41
3.3 Amostra do conjunto de dados no Excel 42
3.4 Tabela com os dados redundantes 44

4.1 Teste Sem Seleção de Atributo e Sem Balanceamento com o J48 49
4.2 Teste Sem Seleção de Atributo e Sem Balanceamento com o Random Forest 50
4.3 Teste Sem Seleção de Atributo e Sem Balanceamento com o PART 50
4.4 Frequência da classe meta de 2003 a 2009 51
4.5 Frequência das classes: soja, milho e algodão do ano-safra 2008/2009 52
4.6 Testes utilizando J48 com o PCA e o Resample 53
4.7 Testes utilizando o PCA e o Cross Validation 53
4.8 Resultado do teste utilizando o J48 com o PCA e o Resample 54
4.9 Conjunto de regras gerado com o J48 com o PCA e o Resample 54
4.10 Análise das regras com o J48, o PCA e o Resample 55
4.11 Testes utilizando o CFS e o Resample 56
4.12 Número de Folds utilizado no Cross Validation 56
4.13 Testes utilizando o CFS com Cross Validation 57
4.14 Resultado do teste utilizando CFS com Cross Validation 57
4.15 Conjunto de regras gerado pelo CFS com o PART 58
4.16 Análise das regras geradas pelo CFS com o PART 58
4.17 Testes utilizando o Info Gain e o Resample 59
4.18 Testes utilizando o Info Gain e o Cross Validation 59

4.19 Resultado do teste utilizando o Info Gain e o Resample 60
4.20 Conjunto de regras gerado pelo Info Gain com o Resample 60
4.21 Análise das regras geradas pelo Info Gain e o Resample 61
4.22 Testes utilizando o Wrapper com o Resample 61
4.23 Testes utilizando o Wrapper com o Cross Validation 62
4.24 Poda das árvores 62
4.25 Teste utilizando o Wrapper e o Random Forest 63
4.26 Conjunto de regras geradas pela 1^0 árvore do modelo 64
4.27 Análise das regras geradas pela 1^o árvore do modelo 65
4.28 Conjunto de regras geradas pela 2^o árvore do modelo 66
4.29 Análise das regras geradas pela 2^o árvore do modelo 66
4.30 Conjunto de regras geradas pela 3^o árvore do modelo 67
4.31 Análise das regras geradas pela 3^o árvore do modelo 68
4.32 Conjunto de regras geradas pela 4^o árvore do modelo 69
4.33 Análise das regras geradas pela 4^o árvore do modelo 70
4.34 Conjunto de regras geradas pela 5^0 árvore do modelo 71
4.35 Análise das regras geradas pela 5^o árvore do modelo 72
4.36 Conjunto de regras geradas pela 6^0 árvore do modelo 73
4.37 Análise das regras geradas pela 6^o árvore do modelo 74
4.38 Conjunto de regras geradas pela 7^0 árvore do modelo 75
4.39 Análise das regras geradas pela 7^o árvore do modelo 76
4.40 Conjunto de regras geradas pela 8^o árvore do modelo 77
4.41 Análise das regras geradas pela 8^o árvore do modelo 78
4.42 Conjunto de regras geradas pela 9^o árvore do modelo 79
4.43 Análise das regras geradas pela 9^o árvore do modelo 80
4.44 Conjunto de regras geradas pela 10^o árvore do modelo 81
4.45 Análise das regras geradas pela 10^o árvore do modelo 81
4.46 Tabela com todas as regras geradas pelos classificadores 88
4.47 Tabela com todas as regras ranqueadas 94
4.48 Melhores regras geradas pelos classificadores 97
4.49 Análise das regras por data 101
4.50 Verificação da safra por período 103

Capítulo 1

Introdução

O Brasil é um país essencialmente agrícola, possui um grande estoque de terras agricultáveis e um clima favorável à produção. A atividade agrícola sempre teve um papel importante no cenário nacional, desde o monocultivo da cana-de-açúcar até o desenvolvimento de uma policultura com o cultivo de grãos, frutas, cereais entre outros.

Esse desenvolvimento e o crescimento do setor agrícola têm requerido cada vez mais agilidade e precisão na antecedência da época da safra. Por exemplo, o transporte, a importação e exportação, a cotação, tudo isso depende da época da safra e do seu cultivo. Nesse sentido, estimativas confiáveis de produção agrícola são ferramentas poderosas para orientar o produtor nas questões ligadas ao plantio, auxiliar as indústrias no setor operacional e de comercialização, indicando ao governo números confiáveis que permitam sua intervenção para reduzir impactos negativos na economia ou aproveitar antecipadamente os benefícios de uma situação favorável [5].

A previsão de safra é feita hoje no Brasil de modo indireto por meio de questionários aplicados ao agricultor e a entidades agrícolas. Esse processo de coleta de dados requer um alto valor em capital e um tempo elevado para ser desenvolvido [16]. A falta de confiabilidade pode ser minimizada por meio de técnicas alternativas como: monitoramento de imagens através de sensores, modelagem dos dados, análise de séries temporais, entre outras.

Para auxiliar a previsão de safras por meio de geotecnologias existem índices que são utilizados para essas análises, sendo o Índice de Vegetação Por Diferença Normalizada (NDVI) um dos mais utilizados.

Os perfis de série temporais de NDVI são gerados a partir de observações sequenciais desse índice, que mostra o desenvolvimento do estádio fenológico da planta. Problemas envolvendo perfis de série temporais já vêm sendo estudado por vários pesquisadores tais como [17], [10], [11]. Esse tipo de estudo, permite obter uma análise quantitativa e qualitativa do que está sendo cultivado, mesmo antes da sua colheita [16].

Em geral, o NDVI é obtido através de imagens geradas, a partir de sensores orbitais. Em particular, o sensor MODIS (Moderate Resolution Imaging Spectroradiometer) fornece imagens com coberturas de áreas com grande extensão e alta periodicidade, características fundamentais para o monitoramento de culturas agrícolas. No entanto, sua baixa resolução espacial pode dificultar a distinção entre uma cultura e outra.

Em particular, a Embrapa vem executando alguns projetos para mapeamento da cobertura da

terra em algumas áreas de Mato Grosso, utilizando dados do sensor MODIS. Algumas culturas agrícolas vêm recebendo atenção especial, como por exemplo, soja, algodão e milho, devido à sua importância para a região.

Em particular, a cultura da soja recebe um destaque especial no estado de Mato Grosso. Segundo a Companhia Nacional de Abastecimento (CONAB), o Brasil é o segundo maior produtor de soja do mundo, perdendo apenas para os Estados Unidos. A soja é considerada hoje o principal produto de exportação do país. Ainda segundo a CONAB, o estado do Mato Grosso lidera a produção nacional com um volume estimado entre 23,15 e 24,24 milhões de toneladas produzidas no ano-safra de 2012/2013, com uma área estimada de 26,42 e 27,32 milhões de hectares, com uma produção média média de 81,44 milhões de toneladas[2].

Vários estudos envolvendo as imagens MODIS têm sido propostos para o desenvolvimento de técnicas alternativas de estimativa agrícola. No entanto, a baixa resolução espacial de dados extraídos do sensor MODIS pode dificultar a distinção entre uma cultura e outra, por parte de especialistas. O uso de técnicas de mineração de dados (MD) é uma alternativa promissora para melhorar a análise de dados de sensoriamento remoto.

As técnicas de mineração de dados auxiliam na interpretação desses dados, uma vez que estas têm por objetivo encontrar padrões novos, válidos e potencialmente úteis em grandes bancos de dados e transformá-los em conhecimento que irão subsidiar decisões estratégicas concernentes à previsão de safras [18].

Dessa forma, por meio de técnicas de mineração de dados para geração de regras de classificação, com dados fornecidos pela Embrapa, tem-se como objetivo desenvolver modelos preditivos, baseados em regras de classificação para a variável safra, em função de séries temporais de NDVI, utilizando os algoritmos **J48**, **Random Forest** e **PART** [13]. Em seguida, obter o mapeamento das áreas de plantio em Mato Grosso e conseguir fazer a diferenciação das culturas de soja, algodão e milho.

Este trabalho está organizado na forma de capítulos. No Capítulo 2, apresenta-se a revisão da literatura, que abrange os assuntos de Sensoriamento Remoto, Índice Vegetativo, Análise Estatística e Mineração de Dados. No Capítulo 3, tem-se as ferramentas necessárias para o desenvolvimento desse trabalho, a aplicação dos modelos aos dados e várias discussões envolvendo o índice *Kappa*, a *acurácia* do modelo e a *precisão* de cada classe (soja, milho e algodão) referente às coberturas de solos. Os resultados da validação dos modelos desenvolvidos, com as devidas comparações e considerações, são apresentados no Capítulo 4. Finalmente, no Capítulo 5, as conclusões são expostas a partir dos resultados dos testes e da avaliação de um especialista. Apresentam-se também sugestões para a realização de trabalhos futuros.

1.1 O problema de Pesquisa

O sensor MODIS fornece imagens com coberturas de áreas com grande extensão a alta periodicidade, características fundamentais para o monitoramento de culturas agrícolas. No entanto, sua a baixa resolução espacial dificulta a distinção sobre o que está sendo cultivado em uma região, por parte de um especialista. Assim, dado um conjunto de séries temporais com valores do NDVI, onde cada séries possui a indicação da cobertura do solo, deseja-se obter intervalos de NDVI que

consiga descrever qual é o tipo de cultura que está sendo cultivada em áreas específicas do estado de Mato Grosso.

1.2 Hipótese

É possível melhorar a classificação automática de culturas agrícolas no estado de Mato Grosso, tais como soja, algodão e milho, por meio da aplicação de técnicas de mineração de dados em séries temporais de NDVI obtidas a partir do sensor MODIS.

1.3 Objetivos

O objetivo geral do trabalho é desenvolver modelos baseados em técnicas de mineração de dados para classificar as culturas de soja, milho e algodão, no estado de Mato Grosso, considerando séries temporais de NDVI para o ano safra de 2008/2009. Dentre os objetivos específicos, tem-se:

- Identificar intervalos de NDVI que melhorem o mapeamento das culturas de soja, milho e algodão, considerando métodos para seleção de atributos e para discretizar variáveis.

- Gerar uma base de conhecimento para cada cultura estudada, considerando as regras de classificação geradas por meio dos algortimos J48, Random Forest e PART.

Capítulo 2

Revisão Bibliográfica

Neste capítulo, será apresentada uma breve discussão sobre a utilidade do Sensoriamento Remoto na previsão e no mapeamento de Safras, bem como a importância do Sensor MODIS na aquisição dos dados para o estudo de imagens através do NDVI. Posteriormente, serão apresentados conceitos relativos à mineração de dados com destaque para a geração de regras de classificação. No final do capítulo, as Medidas para Avaliação das regras e dos modelos serão apresentadas.

2.1 Previsão de Safras

A previsão de safras é hoje uma ferramenta muito útil e necessária no setor das políticas agrícolas de um país. No setor agrícola, para que o agricultor saiba como está o desenvolvimeto da agricultura no seu estado, ou no seu país, podendo assim entender o mercado que ele investe. Já na política, o governo consegue saber quanto terá para vender; podendo estabelecer valores; quanto terá para exportar; investimentos futuros; fazer o mapeamento das regiões produtoras e assim por diante.

No Brasil, o Instituo de Brasileiro de Geografia e Estatística (IBGE) realiza o levantamento de pesquisa utilizado o método conhecido por Levantamento Sistemático da Produção Agropecuária (LSPA), este que tem por objetivo fazer o censo das regiões de plantio e cultivo. Essa pesquisa é realizada junto à entidades ligadas ao setor agrícola por meio de questionários, abrangendo todo o país[3].

Existe também a CONAB, orgão público que tem por finalidade fazer o acompanhamento da trajetória agrícola do plantio até à colheita. Ela realiza pesquisas de intenção de plantio e colheita seis vezes durante o ano-safra, por meio de enquetes com as entidades ligadas ao agronegócio. Essa pesquisa é realizada por meio de uma amostra pré-estabelecida nos municípios mais representativos de cada cultura[2].

Segundo Junges et al.(2010), esse tipo de levantamento de dados é muito oneroso, difícil de se obter dados confiáveis, requer um grande investimento na questão do tempo, além de ser subjetivo, o que leva a imprecisões no final do estudo.

Para auxiliar esse tipo de abordagem, vem crescendo o uso de modelos que têm por finalidade descrever o corportamento das culturas no campo, utilizando variavéis espectrais, derivadas do

sensoriamento remoto.

2.2 Sensoriamento Remoto

O sensoriamento remoto começou no Brasil na década de 60, com o desenvolvimento do projeto **RADAMBRASIL**, que tinha por objetivo fazer um levantamento integrado dos recursos naturais do país. Isso permitiu o desenvolvimento de várias técnicas na área que até então só conheciam a fotografia área [22]. Na década de 80, com o avanço da tecnologia, os sistemas especializados em processamento de imagens de sensoriamento remoto tornaram-se mais acessíveis atingindo uma maior quantidade de usuários, fazendo com que o recurso fosse utilizado em diversas áreas. Essa evolução trouxe sensores cada vez mais potentes e com imagens com melhores qualidades, fazendo com que o processo fosse aplicado a diversas áreas do conhecimento como: recursos ambientais, geologia, agricultura, reflorestamento, estudos urbanos entre outros [23].

2.2.1 Aquisição de Informação

Campbell(1987) define o sensoriamento remoto como sendo a ciência de obter informações da superfície terrestre a partir de imagens tomadas a uma certa distância. A aquisição das informações depende da fonte de energia eletromagnética que interage com a matéria. O Sol é a principal fonte de radiação. Essa radiação que é capaz de atravessar o vácuo, incide sobre a superfície terrestre, que por sua vez interage com os objetos que estão sobre a superfície. Parte dessa radiação retorna do objeto para o sensor. Esse processo, gera um ciclo necessário para a coleta dos dados.

Os sensores orbitais foram desenvolvidos para entrarem em contato com a energia eletromagnética refletida pelos alvos terrestres, como mostra a Figura 2.1:

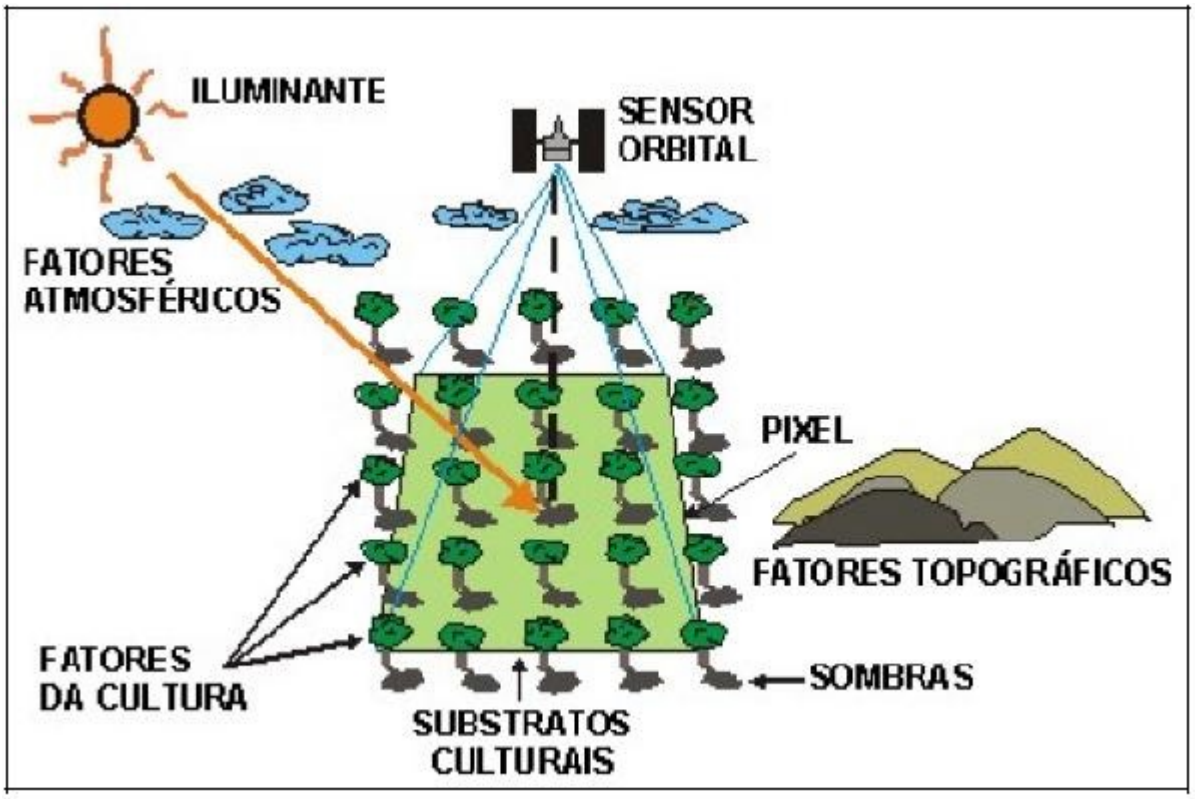

Figura 2.1: Interação da radiação eletromagnética fonte-alvo-sensor(ANTUNES, 2005)

Um fluxo de energia eletromagnética ao se propagar no espaço interage com os objetos da superfície da terra. Pode-se ver na Figura 2.2 que essa energia pode ser absorvida (A), refletida (R) ou ainda transmitida (T), ou seja, a soma das energias absorvida, refletida ou transmitida é igual ao fluxo de energia incidente (I). As variações de energia absorvida, refletida e transmitida podem variar de acordo com a composição físico-química de cada objeto [25].

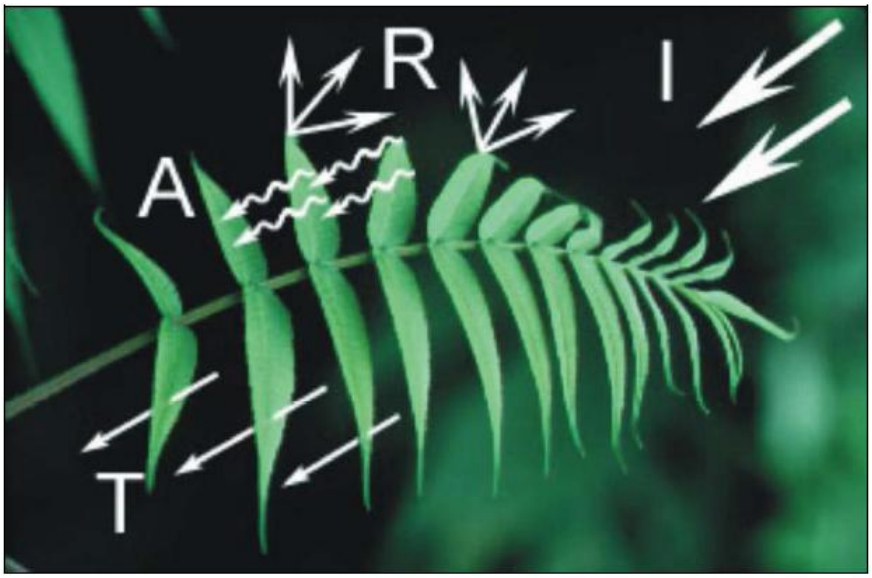

Figura 2.2: Interação da radiação eletromagnética solar com a folha (Adaptado de MOREIRA, 2003)

É possível acompanhar o desenvolvimento de uma cultura agrícola por meio da radiação eletromagnética por ela refletida, sendo as folhas estruturas de maior influência no comportamento espectral das plantas. As características espectrais dadas pelo sensor tendo como fonte as folhas ocorrem pela quantidade de pigmento, estrutura e água contida nelas. Isso é devido ao fato do processo conhecido como fotossíntese. A folha é a parte da planta que está preparada para o melhor aproveitamento da água, do ar e dos raios solares para a realização da fotossíntese [19].

A radiação solar quando atinge a atmosfera é espalhada, refletida, ou consegue atravessar as partículas existentes nela. Quando a radiação que atravessa a atmosfera consegue atingir o alvo, podem acontecer três fenômenos na interação da radiação com o objeto que são: refletância, transmitância e a absortância [5].

- A refletância

É dada como sendo o fluxo de radiação refletido pelo alvo, dividido pelo fluxo de radiação eletromagnética incidente.

- A transmitância

É dada como sendo o fluxo de radiação transmitido pelo alvo, dividido pelo fluxo de radiação incidente.

- A absortância

Já a absortância é dada como sendo o fluxo de radiação absorvido pelo alvo, dividido pelo fluxo de radiação incidente [24].

A característica multiespectral desses sensores permite medir e registrar a energia eletromagnética em determinadas faixas de comprimentos de onda (Figura 2.3), chamadas de bandas espectrais, que podem ser transformadas em uma imagem digital, a qual pode ser interpretada e analisada [5].

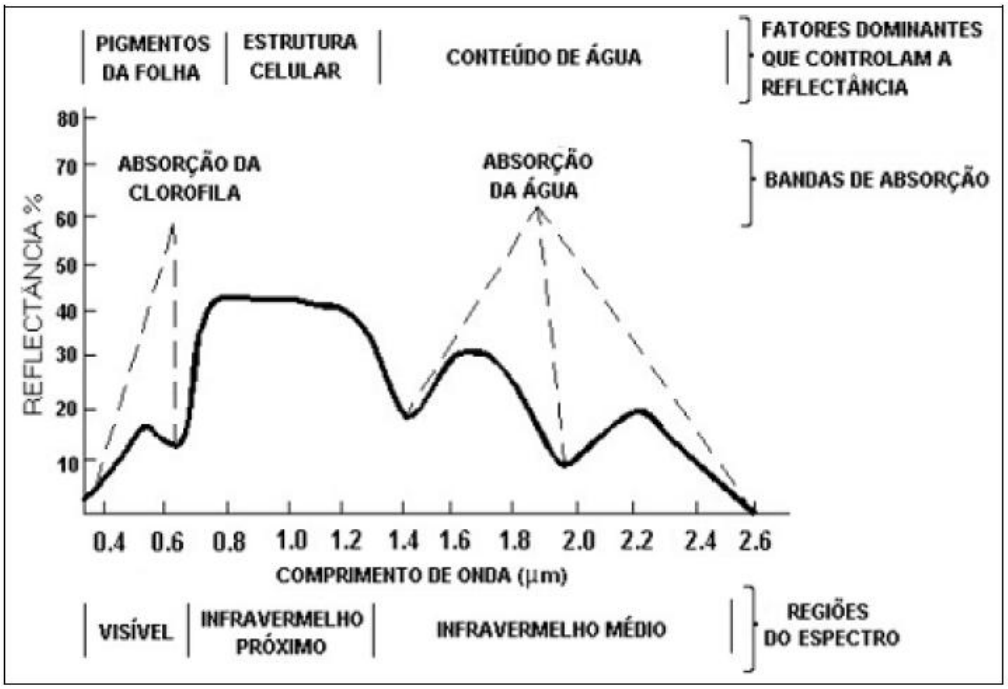

Figura 2.3: Curva padrão de refletância de uma vegetação verde (Adaptado de HOFFER, 1978)

A refletância de uma vegetação verde sadia é distinta e completamente variável com o comprimento de onda nas regiões do espectro eletromagnético do visível, do infravermelho próximo e do infravermelho médio [5].

Na região do visível (0,4 μm a 0,72 μm), a presença de pigmentos clorofilados da planta é responsável pela maior absorção da radiação nas porções azul e vermelho, sendo menos absorvida na porção do verde. O pico de refletância máximo no visível ocorre no comprimento de onda do verde (0,54 μm) e não ultrapassa os 20%. No infravermelho próximo (0,72 μm a 1,3 μm), a estrutura celular é responsável por altos valores de reflectância e transmitância (quase 50%) e baixos valores de absortância (menor que 5%). No infravermelho médio (1,3 μm a 2,6 μm), as propriedades ópticas internas das plantas e a presença de água provocam alta absorção da radiação nos comprimentos de onda de 1,4 μm, 1,9 μm e 2,6 μm [14].

A região entre $0,38\mu$m e $3,00$ μm é chamada de região de energia refletida do espectro. Essa energia é basicamente a energia refletida pelos objetos da superfície terrestre.

2.2.2 Sistemas Sensores

Os sensores são os responsáveis pela transformação de energia em informação na forma de um gráfico ou de uma imagem, que possam transmitir algum tipo de informação para o analista, ou especialista da área.

Segundo Novo (2008), resolução espacial é a capacidade que um sensor tem em representar a menor feição passível de detecção pelo instrumento em questão. Então, uma imagem com alta resolução nos permite detectar objetos de pequena dimensão como, por exemplo, casas, veículos, e a distinção entre plantações. Ao contrário de uma imagem que possui uma baixa resolução espacial, onde podemos observar grandes objetos como grandes construções ou grandes áreas de plantação.

> A resolução radiométrica representa a menor diferença de brilho que um sistema sensor é capaz de perceber, determinando se o alvo pode ser visto na imagem em função de seu contraste com alvos vizinhos. Essa resolução está relacionada ao número de níveis de cinza usados para expressar os dados coletados pelo sensor, considerando que, quanto maior o número de níveis de cinza, maior a resolução radiométrica [6].

Os sistemas sensores podem ser divididos em dois grupos: os passivos e os ativos [21].

- Sensores passivos:

São aqueles que detectam a radiação emitida por objetos da superfície. Dependem, portanto, de uma fonte de radiação externa para que possam gerar algum tipo de informação. Os sensores passivos que possuem espelhos ou prisma lentes são classificados como sensores ópticos.

- Sensores ativos:

São aqueles que produzem sua própria radiação, como por exemplo, os radares e lasers.
Ainda é possível dividir os sensores em Imageadores e Não-imageadores:

- Sensores Imageadores

Têm a capacidade de produzir uma imagem bidimensional radiância, emitância ou retroespalhamento do terreno, gerando como resultado uma imagem da área estudada.

- Sensores Não-imageadores

Permitem medir a intensidade da energia proveniente do objeto de estudo sem necessariamente produz uma imagem.

A representação continua da radiação eletromagnética em termos de comprimento de onda é denominado de espectro eletromagnético. Esse por sua vez é dividido em faixas com características próprias, sendo elas:

- Raios cósmicos e raios gama: $< 0{,}003$ a $0{,}4$ μm.

- Raios X: $<$ 0,03 a 3,0 μm.
- Ultravioleta: $<$ 0,003 a 0,4 μm.
- Visível: $<$ 0,4 a 0,72 μm, sendo azul (0,45 a 0,5 μm), verde (0,5 a 0,54 μm), vermelho (0,56 a 0,72 μm)
- Infravermelho próximo: 0,72 a 1,3 μm
- Infravermelho médio: 1,3 a 4,0 μm
- Infravermelho distante: 4,0 a 300 μm
- Microondas: 1,0 a 100 cm
- Ondas de radio: $>$ 100 cm

O desenvolvimento desse trabalho está restrito ao visível e ao infravermelho próximo.

2.3 Sensor MODIS

O primeiro sensor MODIS foi lançando em Dezembro de 1999, pela *National Aerospace and Space Administration* (NASA), a bordo do satélite TERRA. Ele é um dentre outros quatro sensores que o satélite possui. O segundo foi lançado em Maio de 2002 a bordo do satélite AQUA, também sendo um dentre quatro satélites. Ambos fazem parte de um programa denominado *Earth Observing System* (EOS), que tem por objetivo fazer o monitoramento das mudanças climáticas do planeta Terra. Os satélites foram projetados de forma a conceberem uma visão global da Terra. Pois possuem sensores para a captação das informações necessárias para isso, e alguns desses sensores são comuns nos satélites, para a troca e a comparação de informações. O MODIS é o principal sensor a bordo dos satélites TERRA e AQUA, pois ele tem a função de mensurar os processos biológicos e físicos de toda a superfície terrestre a cada um ou dois dias, adquirindo dados de alta sensibilidade radiométrica em 36 bandas espectrais que se situam entre $0,4$ μm e $14,4$ μm, e são distribuídos em diferentes grupos de resolução [6].

O sensor está localizado a 705 km de altitude e consegue fazer uma varredura de 2.330 km de largura na faixa de cobertura espacial. Os dados brutos desse sensor são coletados e enviados para um centro de pesquisa onde são analisados, processados e divididos em um dos cinco grupos (0 a 4), onde recebem o nome de produto [6].

A Tabela 2.1, mostra uma lista dos produtos gerados pelo sensor MODIS.

Aplicação	Produto
Calibração	MOD 01-Nível IA Radiância MOD 02-Nível IB Radiância calibrada e georretificada MOD 03-Dados de georreferenciamento
Atmosfera	MOD 04-Aerossol MOD 05-Vapor d'água MOD 06-Nuvem MOD 07-Perfil atmosférico MOD 08-Produto atmosférico em formato matricial MOD 35-Máscara de nuvem
Terra	MOD 09-Reflectância de superfície MOD 11-Temperatura superficial do terreno e emissividade MOD 12-Cobertura do solo/mudanças na cobertura MOD 13-Índices de vegetação em formato matricial (NDVI máximo e MVI integrado) MOD 14-Anomalias termais, fogo e temperatura MOD 15-Índice de área foliar e fPAR MOD 16-Evapotranspiração MOD 17-Fotossíntese e produtividade primária bruta MOD 43-Reflectância de superfície MOD 44-Conversão de cobertura vegetação
Crisofera	MOD 10-Cobertura de neve MOD 29-Cobertura do gelo do mar
Oceano	MOD 18-Radiância normalizada da água de saída MOD 19-Concentração de pigmentos MOD 20-Fluorecência de clorofila MOD 21-Concentração de pigmentos da clorofila_a MOD 22-Radiação fotossinteticamente disponível (PAR) MOD 23-Concentração de sólidos em suspensão MOD 24-Concentração de matéria orgônica MOD 25-Concentração de cocólitos MOD 26-Coeficiente de atenuação de água oceânica MOD 27-Produtividade primária dos oceanos MOD 28-Temperatura da superfície do mar MOD 31-Concentração de ficoeritrina MOD 36-Coefifiente de absorção total MOD 37-Propriedades dos aerossóis oceânicos MOD 39-ϵ (água límpida)

Tabela 2.1: Produtos gerados pelo sensor MODIS(Silva, 2004)

As imagens geradas pelo sensor MODIS estão organizadas em uma estrutura de grade, como pode ser observado na Figura 2.4. As imagens são quadriculadas e possuem uma referência

horizontal e vertical para facilitar a sua localização. Cada imagem representa uma matriz numérica, onde cada elemento é denominado pixel e representa a unidade de imagem.

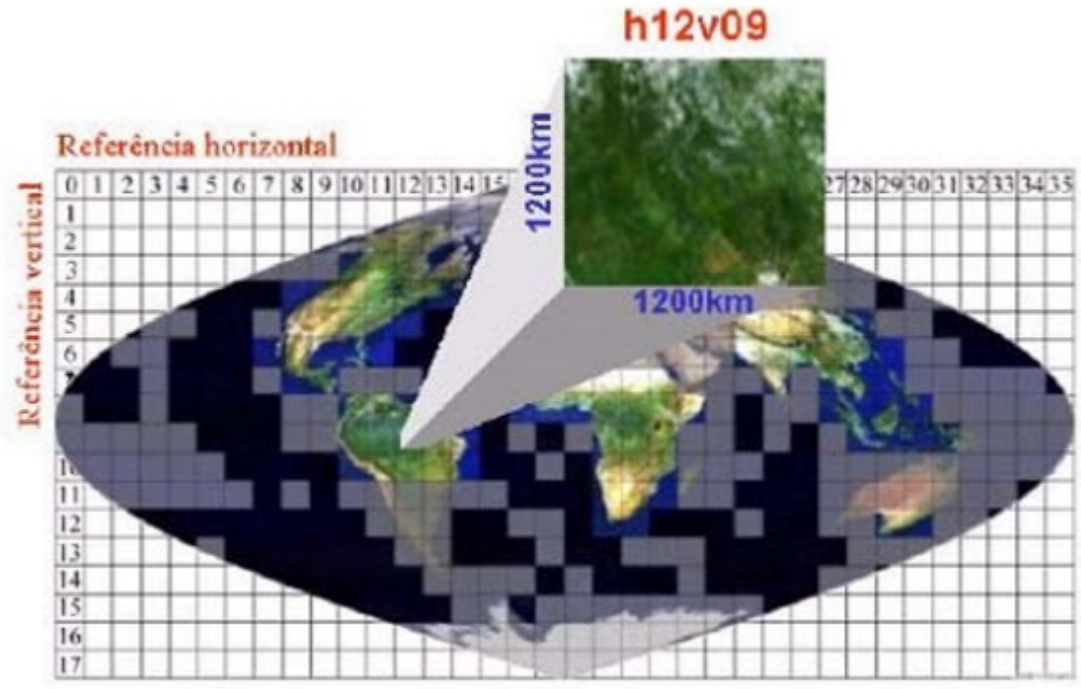

Figura 2.4: Imagem dividida em pixels (Adaptado de UFG, 2003)

O sensor MODIS foi desenvolvido para estudar: atmosfera, oceano e terra, com bandas de resolução espacial e espectral desenvolvidos para estes objetivos, fazendo uma varredura diária das localidades [24].

> Dada as características dos dados MODIS (largura da faixa de imageamento, alta resolução temporal e baixa resolução espacial), Anderson(2003) relata que as imagens dos índices de vegetação, obtidas dos dados MODIS (MOD13) têm como objetivo fornecer dados consistentes para comparações temporais e espaciais das condições da vegetação, em nível global, ou seja, monitorar a atividade fotossintética da vegetação para detectar mudanças no vigor vegetativo e associar as condições biofísicas e fenológicas das mesmas [6].

Dependendo da faixa espectral, a resolução espacial do sensor MODIS varia de 250 a 1000 m, sendo considerada baixa e indicada apenas para mapeamento em escala regional e nacional, que são consideradas imagens de baixa resolução espacial.

Para fazer a retirada de ruídos como as nuvens por exemplo, se utiliza um processo chamado de composição do valor máximo ou *Maximum Value Composition* (MVC). Esse método busca o

valor máximo do NDVI em um dado pixel, em uma série temporal de imagens, formando uma nova imagem a partir dos maiores valores dos pixels em um período de 16 dias.

Pelo fato das imagens terem um fácil acesso e serem distribuídas gratuitamente, isso aumenta a sua utilização e seu estudo no mundo inteiro.

2.4 Índice Vegetativo

O Índice de Vegetação pela Diferença Normaliza (NDVI) tem auxiliado cientistas do mundo todo nos estudos e reconhecimentos de imagens. Este índice tem por finalidade estimar a quantidade de biomassa verde da vegetação. Sua observação frequente a partir de sensores de elevada resolução temporal permite o mapeamento de áreas agrícolas, bem como a detecção de áreas de supressão vegetal, causadas pelos desmatamentos ou por processos intensivos de desertificação [1].

É atribuída a Jordan (1969) a origem dos índices de vegetação, que criou uma expressão simples utilizando a razão entre as refletâncias no infravermelho próximo e no vermelho. Mas os pioneiros nos estudos foram Pearson e Miller (1972) que realizaram pesquisas para medir o fator de reflectância de gramíneas e descobriram que os valores obtidos entre $0,68$ μm e $0,78$ μm nos comprimento de ondas, foram os melhores para distinguir a vegetação do solo [5].

O valor do NDVI pode variar de -1 a 1, onde os valores mais altos implicam em vegetação mais sadia. Esse fato pode ser observado na Figura 2.5, que exemplifica o comportamento do NDVI em função da vegetação.

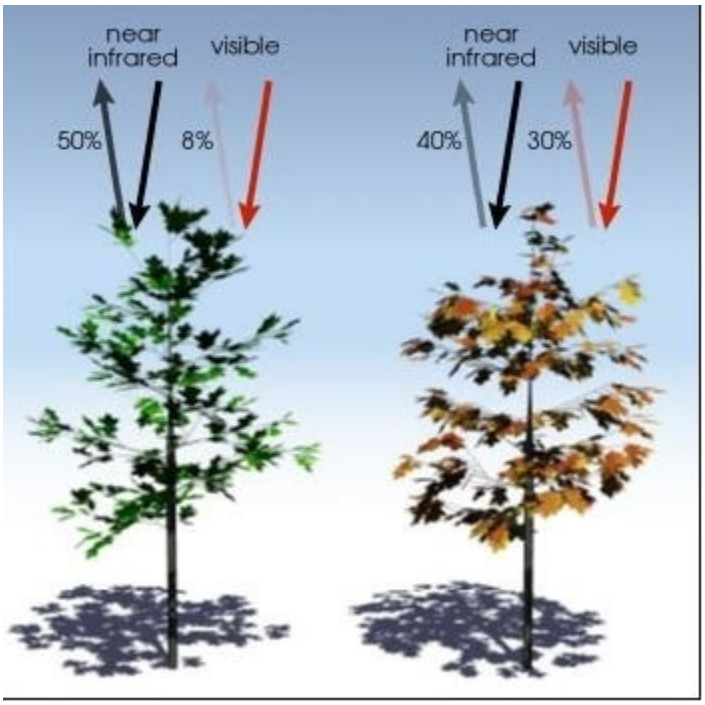

Figura 2.5: Variação do NDVI em uma planta saudável e outra em estado senescente.

Fonte:
http://www.earthobservatory.nasa.gov/Features/MeasuringVegetation/measuring_vegetation_2.php

Assim, o NDVI é dado pela expressão a seguir:

$$NVDI = (IVP - VER)/(IVP + VER) \tag{2.4.1}$$

onde,
IVP= Banda infravermelho próximo
VER= Banda vermelho

O NDVI é calculado a partir da luz visível e do infravermelho próximo, que são refletidas pela vegetação. A vegetação saudável absorve a maior parte da luz visível, e reflete uma grande parte da luz do infravermelho próximo, enquanto que uma planta senescente ou doente tem seu comportamento espectral alterado, reduzindo a refletância no infravermelho próximo e aumentando no vermelho visível.

Vegetação verde, saudável possui uma variação no espectro eletromagnético nas faixas do visível e do infravermelho próximo. A pigmentação da planta provoca uma absorção da energia solar para a realização da fotossíntese. Essa absorção atinge picos do vermelho e do azul no espectro visível, produzindo a aparência verde da planta. Já no infravermelho próximo, a planta reflete ou transmite toda a energia incidente, fazendo pouco uso dela. Essa quantidade de energia refletida no vermelho e no infravermelho próximo tem recebido atenção especial de pesquisadores e estudos recentes [24].

A energia eletrogenética vinda do Sol e refletida pela superfície retorna ao espaço e é captada pelos sensores remotos, denominados radiômetros, a bordo dos satélites orbitais, que trabalham em determinadas faixas espectrais. Os sensores remotos são desenvolvidos para medirem essa variação eletromagnética refletida por um alvo terrestre que, no escopo deste trabalho referem-se às culturas de soja, milho e algodão.

Pode-se observar que quase todos os Índices de Vegetação utilizados para estudos são obtidos de medidas da refletância nas faixas das bandas vermelho e infravermelho próximo. Isso se dá porque nessas faixas, presentes na maioria dos satélites de recursos terrestres, a refletância da vegetação sadia apresenta comportamento contrastante, sendo alta no IVP e baixa no VER. [7].

2.5 Mineração de Dados

2.5.1 Conceitos de Mineração de Dados

Com o avanço da tecnologia, a quantidade de dados armazenados vem aumentando significativamente. Bancos, lojas, financeiras, centro de pesquisas, e as mais diversas áreas do conhecimento têm feito o armazenamento dessas informações. Técnicas tradicionais de análise de dados já vinham sendo utilizadas nos estudos desses dados, mas elas normalmente eram caras e ineficazes para grandes bancos de dados. Muitas vezes a natureza dos dados acabava dificultando a análise, ou seja, as abordagens tradicionais não eram suficientes para uma conclusão satisfatória. Com isso tem-se o nascimento de novos métodos e estudos voltados para a Descoberta de Conhecimento em Bases de Dados (Knowledge Discovery in Databases-KDD), com o intuito de transforma−los em conhecimento.

O KDD nada mais é do que a extração automática de regras, padrões desconhecidos em grandes bancos de dados, data warehouses, web, ou qualquer outro grande repositório de informações [13].

Apesar de muitos pesquisadores utilizarem o termo mineração de dados como sinônimo do KDD, o Knowledge Discovery in Databases é o processo completo de extração de informações, enquanto que o mineração de dados é apenas uma etapa desse processo.

A Figura 2.6 representa uma visão geral do processo KDD

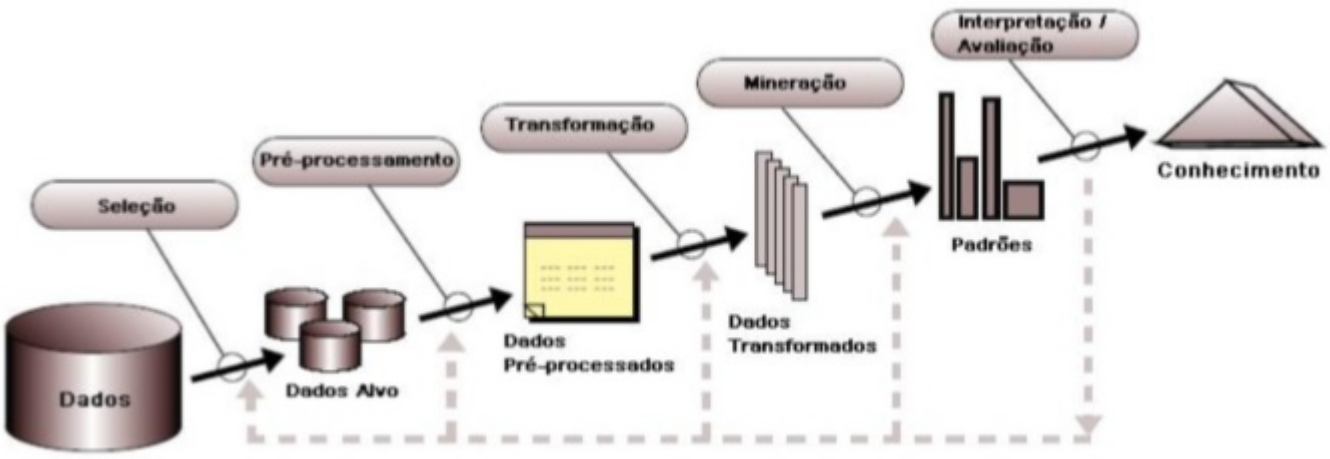

Figura 2.6: Processo KDD (Adaptado FAYYAD et al., 1996)

Mineração de Dados (ou Data Mining) nada mais é do que o processo de descoberta de novas informações a partir de grandes massas de dados, que poderiam permanecer ignoradas. Tendo também a capacidade de fazer previsões sobre alguma observação [27].

Essa ferramenta tem a vantagem de analisar novos tipos de dados e de explorar dados antigos obtendo novos resultados [27].

Entendo a necessidade de métodos para explorar grandes quantidades de dados, pesquisadores de diferentes disciplinas começaram a enfocar o desenvolvimento de ferramentas mais eficientes e escaláveis que pudessem lidar com diferentes tipos de dados. Este trabalho culminou na área de Mineração de Dados, que agrega técnicas e abordagens como a amostragem, estimativas e testes de hipóteses a partir de estatísticas e algoritmos de busca, técnicas de modelagem, teorias de aprendizagem de inteligência artificial, reconhecimento de padrões e aprendizagem de máquina. Mineração de Dados foi rápida em adotar idéias de outras áreas, incluindo otimização, computação evolutiva, teoria da informação, processamento de sinal, visualização e recuperação de informações [27].

O processo de Mineração de Dados pode ser dividido em três grandes áreas: Pré-processamento, Mineração e Pós-processamento. Como pode ser visto na Figura 2.7:

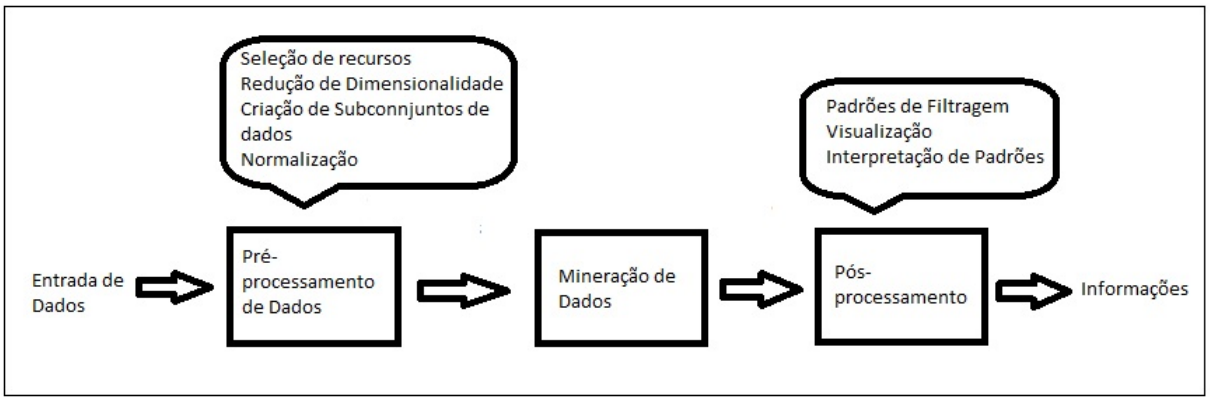

Figura 2.7: Etapas da Mineração de Dados(Data Mining, 2009, p.4)

Pré-processamento

Nessa etapa o objetivo é transformar os dados de entrada em um formato apropriado para o software que será utilizado no processo de mineração de dados.

A integração dos dados é feita de tal forma que os dados de diversas fontes são ajuntados. Em seguida, alguns procedimentos são aplicados para obtenção da qualidade dos dados, como por exemplo, remoção de ruídos e inconsistência, preenchimento de valores faltantes, identificação de informações duplicadas e tudo aquilo que seja rrelevante para a tarefa de mineração. Tudo isso para garantir a confiabilidade dos dados utilizados.

Pelo fato da coleta de dados poder ser feita por várias pessoas e de diferentes maneiras, o pré-processamento é a fase mais trabalhosa e demorada, e o mais importante na Mineração [27].

Ainda dentro do pré-processamento pode se utilizar testes de hipótese, medições estatísticas ou seleção de amostras, para eliminar resultados não legítimos da mineração [27].

Mineração

De posse dos dados pré-processados, utiliza-se técnicas específicas, sendo elas supervisionadas ou não supervisionadas, com a finalidade de descobrir padrões, regras de associação ou agrupamento de dados (*clusterização*).

As tarefas de mineração podem ser divididas em dois grupos: Tarefas Preditivas e Tarefas Descritivas.

Tarefas Preditivas:

Tem por objetivo prever o valor de um atributo em função dos outros. O atributo a ser previsto é conhecido como variável dependente ou atributo meta, e os atributos utilizados para a previsão são as variáveis independentes. Dentro das tarefas de previsão temos, por exemplo: a classificação (usada quando o atributo meta é uma variável discreta ou nominal) e a regressão (usada quando o atributo meta é uma variável contínua).

Tarefas Descritivas:

O objetivo dessa tarefa é encontrar padrões. Essa tarefa é muitas vezes exploratória, ou seja, requer um pós-processamento para validar e explicar os resultados. Dentro das tarefas descritivas temos: clusterização e associação.

- Clusterização

As técnicas de clusterização servem para dividir instâncias em grupos que possuem alguma característica em comum [28].

A clusterização está diretamente relacionada com técnicas que são utilizadas para dividir objetos em grupos. Essa divisão é baseada nos rótulos existentes nos objetos. O objetivo é fazer agrupamentos (ou divisões) entre objetos que sejam semelhantes entre si. Quanto maior a semelhança entre os objetos do grupo e maior a diferença entre os grupos, melhor será o agrupamento [27].

- Associação

Uma regra de associação é algo do tipo: $X \rightarrow Y$, onde X e Y são conjuntos disjuntos de itens [27].

A intenção das técnicas de associação é descobrir padrões em grandes bancos de dados e avaliá-los de tal forma que esses padrões sejam consistentes, evitando assim a geração de resultados falsos [27].

O atrativo das regras de associação é que o especialaista pode estabelecer um limite mínimo de confiança e de suporte para as regras. Onde o suporte determina a frequência na qual a regra aparece, enquanto a confiança determina a frequência na qual os itens em Y aparecem nas transações que contenham X . Isso garante a qualidade das regras [13].

TIPO	ITENS
1	Pão, Leite
2	Pão, Fraldas, Cerveja, Ovos
3	Fraldas, Cerveja, Cola, Leite
4	Pão, Leite, Fraldas, Cerveja
5	Pão, Leite, Fraldas, Cola

Tabela 2.2: Exemplo de transação de cestas de compras(Data Mining, 2009, p.390)

Por exemplo, analisando a regra $\{Leite, Fralda\} \rightarrow \{Cerveja\}$ extraída da Tabela 2.2, calcula-se:

$$\text{suporte} = \frac{\text{frequência}\,\{Leite, Fralda, Cerveja\}}{\text{total}} = \frac{2}{5} \tag{2.5.1}$$

$$\text{confiança} = \frac{\text{frequência}\,\{Leite, Fralda, Cerveja\}}{\{Leite, Fralda\}} = \frac{2}{3} \tag{2.5.2}$$

Ou seja, apesar da regra $\{Leite, Fralda\} \rightarrow \{Cerveja\}$ ter um suporte baixo (talvez não seja tão interessante) ela possui uma alta confinaça (mas é confiável).

Pós-processamento

A etapa do Pós-processamento integra e interpreta os resultados da mineração com os sistemas de apoio a decisões. Essa etapa é necessária para que um especialista da área estudada, analise os resultados obtidos na mineração com a finalidade de se trabalhar apenas com os resultados válidos e úteis para o caso em questão.

Nessa etapa pode-se utilizar medições estatísticas ou métodos de testes de hipóteses para eliminar resultados inválidos.

2.5.2 Classificação

A classificação consiste em examinar uma certa característica nos dados e atribuir uma classe previamente definida.

> Classificação é a tarefa de aprender uma função alvo f que mapeia cada conjunto de atributos x para um dos rótulos de classes y pré-determinados [27].

As técnicas de classificação são utilizadas para a construção de modelos a partir de um conjunto de dados de entrada. Elas são apropriadas para prever ou descrever conjuntos de dados com categorias nominais ou binárias.

A classificação organiza os objetos dentre várias categorias pré-definidas.

É um problema que engloba diversas aplicações, como por exemplo: detecção de fraudes, classificação de tumores malignos ou benignos, classificação de cobertura de solos, concessão de empréstimos, entre outros.

Um algoritmo de classificação gera regras do tipo $X \rightarrow Y$ (lê-se: Se X então Y), onde Y representa o atributo objetivo e X representa um conjunto de valores tomado por atributos [26].

Como por exemplo:

Se x= 1 e y= 1, então classe A
Se z= 1 e w= 1, então classe B

Para a geração das regras de classificação, a base de dados é dividida em dois conjuntos: o conjunto de treinamento e o conjunto teste. A divisão desses dois conjuntos pode ser aleatória, mas também determinística.

O algoritmo de classificação como pode ser visto em Figura 2.8, recebe o conjunto de treinamento, aplica técnicas estatísticas ou de aprendizado de máquina e gera as regras de classificação, com base nos valores encontrados nos atributos de cada um dos exemplos. Em um segundo momento, o algoritmo aplica as regras de classificação obtidas no conjunto teste, então mede o quão confiáveis são as regras geradas no processo anterior. Para tal medição utilizou se a acurácia, quanto maior o valor da acurácia mais precisas são as regras obtidas [26].

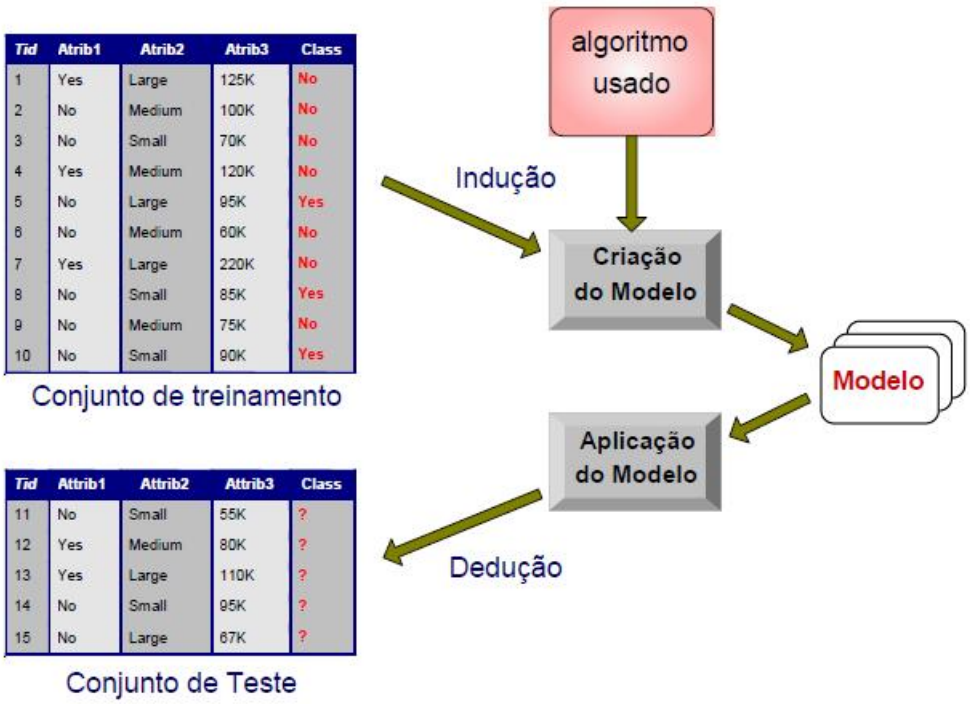

Figura 2.8: Modelo de classificação (Data Mining, 2009, p.175)

Árvore de Decisão (J48)

Uma Árvore de Decisão é um fluxograma com sua estrutura semelhante à de uma árvore, onde cada nodo (ou nó) interno denota um teste realizado em um atributo, cada ramo representa o resultado e cada folha representa a distribuição dos registros. É recomendado o treinamento do método utilizando-se várias amostras nos dados [9].

A árvore de decisão é muito utilizada devida sua facilidade de interpretação. Os resultados são visivelmente dados pela estrutura simbólica sem grandes dificuldades. Um exemplo de uma árvore de decisão é dada pela Figura 2.9, na qual tem-se um modelo que avalia qual a classe do indivíduo de acordo com sua idade e renda financeira. Nesse caso, o atributo meta está dividido em três categorias: classe A, classe B e classe C.

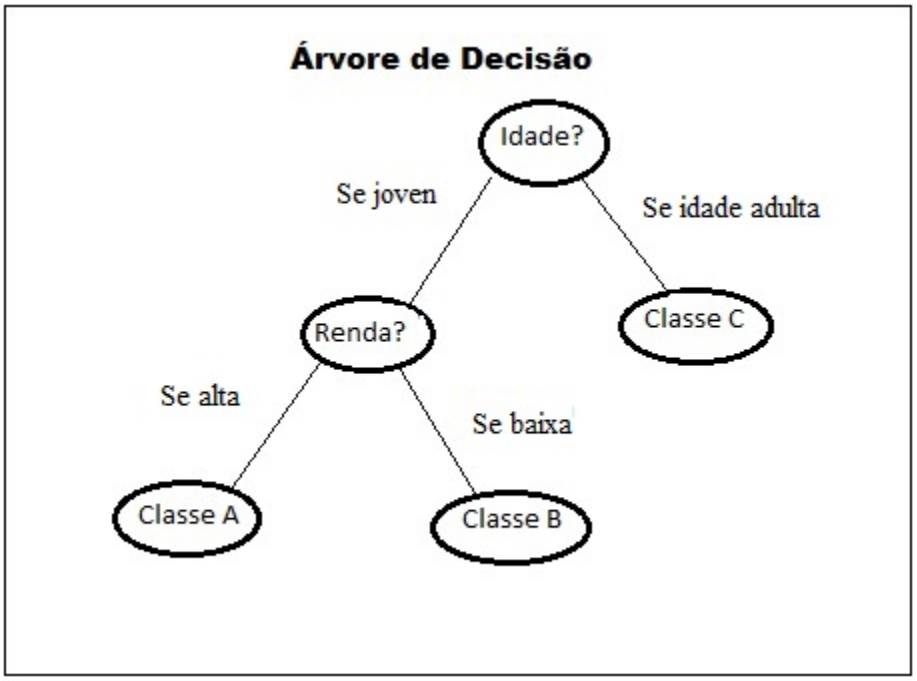

Figura 2.9: Modelo de uma árvore de decisão (Data Mining: Concepts and Techniques, 2011, p.18)

A interpretação da árvore de decisão pode ser dificultada se a árvore for muito grande. Mas esse problema pode ser contornado se extraída suas regras. As regras de classificação são de fácil interpretação para o ser humano. Uma regra é criada por cada caminho da árvore, que se dá da raiz até o nó da folha [13].

No caso da Figura 2.9, as seguintes regras podem ser extraídas:

Se idade = adulta então classe C.

Se idade = jovem e renda = baixa então classe B.

Se idade = jovem e renda = alta então classe A.

A construção de uma árvore consiste em duas fases: a primeira que constitui a construção propriamente dita da Árvore com os particionamentos dos atributos, já a segunda fase constitui-se da poda. Essa que identifica e remove ramos que representem ruídos e ou *outliers* do banco de dados.

Indução de Árvore de Decisão

A indução da árvore de decisão é o processo que analisa qual ou quais atributos descreverão melhor a estrutura da árvore. Qual o atributo que será selecionado para a raiz, quais os atributos que serão selecionados para o segundo nível da árvore gerando as folhas e assim por diante.

Esse ranqueamento das variáveis se dá em um processo interno do próprio algoritmo conhecido como ganho de informação. Os atributos da árvore são ranqueados pelo ganho de informação (menor entropia), que representa a variação de impurezas. O maior valor do ganho de informação determinará o atributo que por sua vez apresentará na raiz da árvore. Calculado novamente o

ganho de informação, esse começará a determinar os nós internos até chegar nas folhas, onde o ganho de informação será o menor de todos ([13],[27]).

Classificar um conjunto de dados é rápido e direto, tendo a árvore já construída. Começando da raiz, aplica-se um conjunto de teste ao conjunto de banco de dados e segui-se as ramificações determinadas, que levará a um nodo no qual um novo teste será aplicado e assim sucessivamente até se chegar nas folhas. O rótulo da classe associado a folha é então atribuído ao registro.

Tendo a árvore já construída, podem existir regras que dão falsas informações devido à influência de ruídos ou *outliers*. Esse tipo de problema pode ser evitado se utilizado um processo chamado de poda da árvore. Existem métodos para a pré-poda e a pós-poda. A pré-poda busca interromper o algoritmo mais cedo evitando assim o crescimento indevido da árvore, já a pós-poda realiza cortes na árvore eliminando ramos indesejados.

Random Forest

Random Forest ou Floresta Aleatória é uma classe de métodos de conjuntos desenvolvidos para classificadores de árvore de decisão. Ele combina previsões feitas por múltiplas árvores de decisão, onde cada árvore é gerada a partir dos valores de um vetor aleatório. Estes vetores são gerados a partir de uma distribuição de probabilidade fixa [27].

Esse vetor pode selecionar aleatóriamente F características de entrada dos dados, e baseado nessas F características fazer a divisão dos nós da árvore. Depois das construções das árvores, as previsões são combinadas utilizando a votação da maioria. Essa abordagem é conhecida como Floresta-RI, onde RI faz referência à seleção aleatória de entrada. O processo conhecido como *bagging*[1] também pode ser utilizado para gerar amostras de *bootstrap* para a Floresta-RI [27].

Esse F escolhido determina a correlação entre as árvores, se F for suficientemente pequeno, então as árvores tendem a se tornar menos correlacionadas. Por outro lado, se for grande a força do classificador tende a melhorar.

Tem-se uma outra abordagem conhecida por Floresta-RC, onde o número de entradas das características é um número pequeno, dificultando a escolha do F. Dessa forma, é criado um subconjunto através de combinações lineares das características de entrada[13].

Existe uma terceira abordagem, em que o algoritmo escolhe as F melhores divisões de cada nó da árvore. Esse fato pode levar em árvores mais correlacionadas, mas pelo fato de ter que fazer essa procura pelos melhores F, essa abordagem acaba gastando um tempo maior de processamento que as demais abordagens.

Todas as árvores são geradas em sua integridade, ou seja, não há poda em nenhuma delas [13].

PART

O algoritmo PART (parte da árvore de decisão) foi desenvolvido por Frank e Witten em 1998, sendo também um algoritmo de classificação. Ele é uma variação do algoritmo J48, pois constrói um conjunto de regras do tipo SE $\rightarrow$ ENTÃO atráves de uma árvore de decisão construída pelo J48. A cada iteração ele transforma a melhor folha da árvore em uma regra. Para cada regra

[1] *bagging* também conhecido como agregação de *bootstrap*, é uma técnica que cria amostras repetidamente a partir de um conjunto de dados de acordo com uma distribuição uniforme de probabilidade.

criada, é estimada sua cobertura das instâncias da base. Após a escolha da melhor regra, ele retira todas as instâncias que se encaixam na regra gerada para a construção de uma nova árvore. Esse processo é repedito até que todas as instâncias sejam cobertas. As regras com coberturas mais altas são apresentadas e as demais descartadas. Dessa forma o número de regras apresentadas pelo PART é menor que a árvore que seria construida pelo J48, logo esse conjunto de regras é mais precisa([12],[20]).

2.6 Métodos para Seleção de Atributos

Quando a dimensionalidade aumenta, têm-se a dispersão dos dados no espaço, isso significa que não há objetos suficientes no espaço para a criação de um modelo. Por isso é necessário fazer uma redução na dimensionalidade, ou seja, uma Seleção de Atributos.

Muitos algoritmos funcionam melhor se a dimensionalidade, o número de atributos dos dados for menor. Isso se deve ao fato da redução de dimensionalidade eliminar atributos altamente correlacionados, irrelevantes e os ruídos [27].

Dependendo dos tipos de dados que se tem, é necessário fazer uma seleção de atributos, ou seja, selecionar aqueles atributos que são mais importantes para o problema, e retirar aqueles que não são. Com isso têm-se uma melhor performance do algoritmo, pois:

- Os atributos altamente correlacionados são retirados:

Fazendo a retirada dos atributos irrelevantes que não agregam informações significativas para o problema, é possível reduzir significativamente o tamanho do conjunto de dados que está sendo trabalhado, sem perda de qualidade, pois as informações mais relevantes são filtradas.

- Redução do tempo de execução:

Mesmo se a redução na dimensionalidade não conseguir reduzir os dados de maneira significativa, consegue-se algumas combinações dos atributos antigos, gerando um novo subconjunto de novos atributos. Com isso reduz se o tempo de execução dos algoritmos.

- Melhor entendimento sobre os resultados obtidos:

A presença de atributos irrelevantes pode de alguma maneira levar o algoritmo a uma classificação equivocada dos dados, levando o processo a produzir resultados pouco significativo ou pouco confiável. Trabalhando apenas com atributos relevantes, os resultados sairão apenas em função deles, o que implica em resultados e interpretações mais simples.

- Remoção de ruídos:

Ruído é considerado o componente aleatório de um erro de medição. Ele pode envolver a distorção de um valor ou a adição de objetos ilegítimos [27].

Quando se realiza uma redução na dimensionalidade, consegue-se reduzir esses ruídos.

Esse trabalho utilizou alguns algoritmos para reduzirmos a dimensionalidade, como: PCA, X^2, Info Gain, Gain Ratio, CFS e o Wrapper.

2.6.1 PCA

Análise de Componentes Principais (PCA) é uma técnica da álgebra linear para atributos contínuos. Seu objetivo é encontrar um novo subconjunto de dados que capture melhor a variabilidade dos dados. Inicialmente o algoritmo escolhe um vetor que descreve a maior variabilidade possível, o segundo vetor será ortogonal ao primeiro, e assim sucessivamente [27].

O PCA tem o poder de identificar padrões forte nos dados. Um segundo fato interessante é que muitas vezes uma pequena quantidade dos dados pode descrever todo o conjunto. Logo, o PCA é excelente para a redução de dimensionalidade. Em terceiro lugar, como os ruídos em geral são mais fracos que os demais dados, quando se utiliza o PCA, indiretamente elimina-se os ruídos.

Sejam D uma matriz m por n (onde as n colunas são os atributos) dos dados, S a matriz de covariância (matriz semidefinida). Se a matriz D for pré-processada, de modo que a média de cada atributo seja 0, então: $S = D^t D$.

Suponhamos que $\lambda_1 \geq \lambda_2 \geq \ldots \lambda_m$ são os autovalores da matriz S, e U a matriz dos autovetores de S, então:

- $D^{'} = DU$ é o conjunto de dados transformados que satisfaz a condição acima.
- Cada novo atributo é uma combinação linear dos atributos originais.
- A variância do novo atributo de índice i é λ_i.
- A soma das variâncias dos atributos originais é igual a soma dos novos.
- Os novos atributos são chamados de componentes principais.
- Os autovetores de S definem um novo conjunto de eixos. Assim, o PCA pode ser visto como uma rotação do eixo das coordenadas originais para a nova, preservando a variabilidade dos dados.

A Figura 2.10 ilustra que, o autovetor (v_1) associado ao maior autovalor indica a direção que os dados possuem maior variabilidade, o autovetor (v_2) associado ao segundo maior autovalor dá a segunda maior direção de variabilidade, que por sua vez é ortogonal ao primeiro, e assim por diante [27]

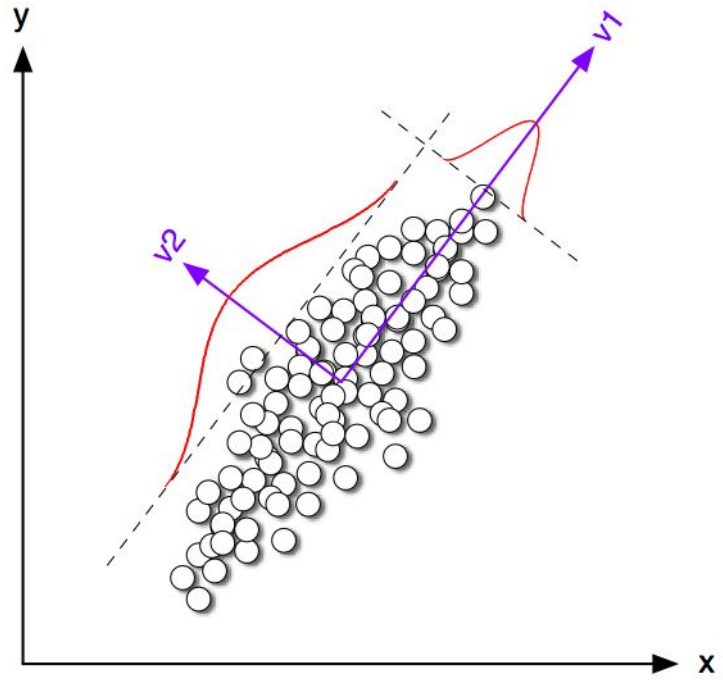

Figura 2.10: Análise com o PCA

Fonte: http://web.media.mit.edu/ tristan/phd/dissertation/chapter5.html

Critério de Jolliffe

O critério de Jolliffe é utilizado no algoritmo do PCA para o descarte de variáveis produzindo uma redução de dimensionalidade. Segundo Jolliffe, quando se possui um alto número de variáveis, é possível que muitas delas sejam redundantes, podendo portanto serem eliminadas. Na seção acima foi visto que os maiores autovetores associados aos maiores autovalores descrevem a direção de maior variabilidade do conjunto de dados. Sabendo que os autovalores são dados em ordem decrescente, os últimos, ou os menores autovalores associados aos menores autovetores, descrevem menos os dados, logo são menos importantes ou até redundantes, podendo ser eliminados [15].

Em [15], Jolliffe, estabelece que o corte para o descarte das variáveis seja feita para autovalores menores que $0,7$. É verificado todos os autovalores menores que $0,7$, então no autovetor associado a ele busca-se a variável com menor valor em módulo, essa variável é então descarta do conjunto. Segue um exemplo(Tabela 2.3):

Número de componentes		1	2	3	4
	1	0,52	-0,37	0,72	0,26
	2	-0,26	-0,92	-0,24	-0,12
Número de variáveis	3	0,58	-0,02	-0,14	-0,80
	4	0,56	-0,65	-0,63	0,52
Autovalores		2,91	0,92	0,14	0,02

Tabela 2.3: Exemplo de descarte de variável segundo critério de Jolliffe (Adaptado do Conjunto de dados Iris)

De acordo com os valores da tabela acima, tem-se que o terceiro e o quarto autovalor são menores que $0,70$. Então, de acordo com o critério de Jolliffe, deve-se buscar nos respectivos autovetores (terceira e quarta coluna) seus maiores valores em módulo. Observa-se que as variáveis de número 1 e 3 são as que possuem os maiores valores em módulo. Logo, a primeira e a terceira variável serão descartas do conjunto de dados, obtendo a seguinte Tabela 2.4:

Número de componentes		1	2	3	4
	2	-0,26	-0,92	-0,24	-0,12
Número de variáveis	4	0,56	-0,65	-0,63	0,52

Tabela 2.4: Exemplo de descarte de variável segundo critério de Jolliffe (Adaptado do Conjunto de dados Iris)

2.6.2 X^2

X^2 é um teste de hipótese que se destina a calcular a relação entre duas ou mais variáveis qualitativas. Seu princípio básico é comparar proporções, as possíveis divergências entre as frequências Observadas e as frequências Esperadas [4].

Esse método é não paramétrico, ou seja, não depende das médias com relação à classe meta ou ao atributo meta. As condições necessárias para se aplicar o teste são:

- Os grupos devem ser independentes;
- Os itens de cada grupo devem ser selecionados aleatoriamente;
- As observações devem ser frequência ou contagem;
- Cada observação pertence apenas a uma categoria;
- A amostra deve ser relativamente grande.

O valor do X^2 é dado por:

$$X^2 = \frac{\sum (\text{Observada} - \text{Esperada})^2}{\text{Esperada}} \tag{2.6.1}$$

onde a frequência Observada é retirada da amostra e a Esperada é calculada a partir destas. Quanto maior valor do X^2 maior é a relação entre as variáveis.

Os testes de hipótese são divididos em duas análises. Se aceitamos ou rejeitamos o H_0:

H_0 : os atributos são independentes.

H_1 : os atributos são dependentes, ou seja, existe correlação entre eles.

Para se verificar a aceitação ou não, se calcula o X^2 observado com base na amostra e se compara com o X^2 esperado de uma tabela, que por sua vez depende do grau de liberdade (GL) e o nível de significância adotado. Em geral, para duas variáveis utiliza-se o nível de significância de 5%. Assim

- Se X^2 calculado é $\geq X^2$ esperado, então se rejeita o H_0.
- Se X^2 calculado é $< X^2$ esperado, então se aceita o H_0.

O nível de significância representa a máxima probabilidade de erro que se tem ao rejeitar uma hipótese. Já o grau de liberdade é definido com

$$\text{GL=número de classes} - 1 \tag{2.6.2}$$

2.6.3 Gain Ratio e Info Gain

O método Info Gain mede a proporção do ganho de informação de cada atributo com relação a classe. Atributos com maior ganho de informação possuem menor entropia. Em uma árvore de decisão, o atributo com maior ganho de informação fica localizado na raiz, os demais atributos ficam localizados nos níveis seguintes da árvore, ranqueados pelo valor da entropia. Logo, os atributos com maior ganho de entropia descrevem melhor os dados (o atributo meta) [13].

Assim, a informação esperada ou entropia, para classificar um atributo em D, onde D é o conjunto de dados, será dado por:

$$Info(D) = -\sum_{i=1}^{m} p_i log_2(p_i) \tag{2.6.3}$$

Onde p_i é a probabilidade de uma instância em D pertencer à classe C_i sendo estimada por $|Ci, D|/|D|$.

E,

$$Info_A(D) = \sum_{j=1}^{v} \frac{|D_j|}{|D|} * Info(D_j) \tag{2.6.4}$$

Onde $Info_A(D)$ é a informação necessária para a classificação de uma informação em D. A razão $|D_j|/|D|$ atua como um peso para a partição j.

Assim, o ganho de informação será dado por:

$$Gain(A) = Info(D) - Info_A(D) \tag{2.6.5}$$

Portanto, o atributo com maior ganho de informação será escolhido para a divisão das ramificações da árvore.

Por outro lado, o método Gain Ratio (Taxa ou razão ganho de informação), utiliza uma normalização chamada de informação de divisão, que serve como um peso linearizador para os atributos, dado por:

$$SplitInfo_A(D) = -\sum_{j=1}^{v} \frac{|D_j|}{|D|} * log_2(\frac{|D_j|}{|D|}) \tag{2.6.6}$$

este valor representa a informação da divisão dos atributos.

Assim, o ganho de informação será definido por;

$$GainRatio(A) = \frac{Gain(A)}{SplitInfo_A(D)} \tag{2.6.7}$$

O atributo com maior ganho de informação será selecionado para a divisão dos atributos. Atributos com o *Ranked* = 0 são descartados, pois não acrescentam informações ao modelo. É dessa forma que o ganho de informação é considerado um selecionador de atributos, pois aqueles que não acrescentam informações são descartados.

2.6.4 CFS

O algorimo *Correlation Based Feature Selection* (CFS) é baseado nas relações entre os atributos. Ele calcula uma matriz de correlações entre atributo-atributo e entre atributo-classe, um peso S é utilizado, dando origem a seguinte fórmula [28]:

$$\text{Mérito(S)} = \frac{k * r_{ac}}{\sqrt{k + k(k-1)r_{aa}}} \tag{2.6.8}$$

onde:

- Mérito(S) é o mérito de um subconjunto de atributos S contendo k atributos.
- r_{ac} é a medida de correlação entre atributo-classe.
- r_{aa} é a medida de correlação atributo-atributo.

O numerador pode ser visto como um indicador de poder preditivo do conjunto de atributos, e o denominador indica o grau de redundância, que existe entre os atributos.

O algoritmo começa com um subconjunto vazio de atributos e utiliza a heurística da primeira procura com um critério de parada de 5 consecutivos subconjuntos que não melhoram o mérito. O subconjunto com o maior mérito encontrado pela heurística será o escolhido.

2.6.5 Wrapper

Esse algoritmo pode ser dividido em três grandes etapas, como mostrado na Figura 2.11. A primeira etapa que é a entrada dos dados, a segunda etapa que é chamada de filtragem, onde o algoritmo avalia o conjunto de dados utilizando um algoritmo de aprendizado de máquina, de acordo com as características gerais do conjunto de dados. Na segunda etapa, que chamamos de *wrapper*, ele encontra um subconjunto do banco de dados utilizando um algoritmo de classificação. É por essa razão que o método *wrapper* é chamado de caixa preta, pois seu funcionamento interno não é conhecido [28].

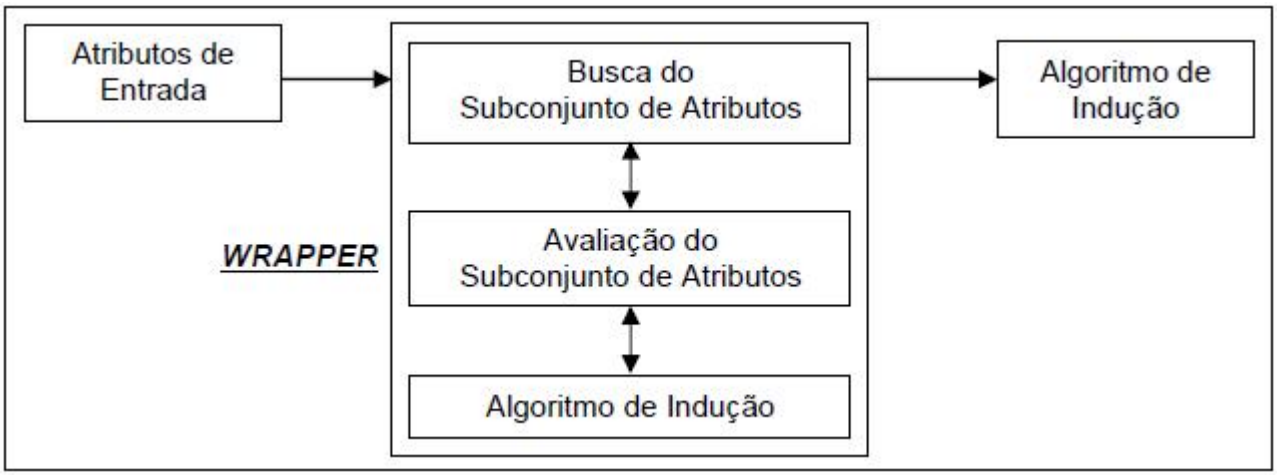

Figura 2.11: Modelo de um método de seleção de subconjuntos de atributo utilizando o Wrapper (JONH et al.,1994)

No modelo Wrapper, a validação da precisão é de extrema importância, e segundo John(1994), o mais indicado para isso é o Cross Validation.

2.7 Valores Faltantes

É comum encontrarmos conjuntos de dados faltando informações, porque o dado não foi coletado, ou se perdeu, havia dados ilegítimos e foi necessário retirá-los, isto é, são inúmeras as razões para não se ter um dado. Independente do que aconteceram durante a fase de coleta, os valores faltantes devem ser considerados no modelo. Existem inúmeras maneiras de se tratar esse tipo de problema.

Esse trabalho utilizou o algoritmo dos K vizinhos mais próximos (K-NN), que classifica por similaridade os K vizinhos mais próximos do valor faltante. Nesse trabalho será tratado apenas dos valores faltantes referentes ao atributo meta, ou seja, a safra para as culturas da soja, milho e algodão, pois o restante do conjunto de dados não possuía valor faltante.

2.7.1 K-NN

No algoritmo K-NN, os exemplos dos conjuntos de dados são armazenados. Quando um novo exemplo precisa ser classificado é verificado sua similariedade, então esse novo exemplo é classificado segundo sua proximidade com os exemplos de treinamento.

O parâmetro K é o número de vizinhos mais próximos considerados. Geralmente, K é um número pequeno e ímpar para evitar empates.

Quando utilizamos $K = 1$, estamos considerando apenas o exemplo mais próximo. Isso pode nos levar a uma incorreta classificação caso o conjunto contenha ruídos. O algoritmo K-NN assume que todos os exemplos são pontos no espaço e a distância entre esses pontos normalmente é medida pela distância Euclidiana. Para isso ele considera que os atributos são normalizados e possuem todos os mesmos pesos, ou seja, mesma importância. Os K vizinhos mais próximos podem ser representados tanto por um atributo classe Y contínuo quanto discreto.

2.8 Aprendizado com Classes Desbalanceadas

Trabalhar com conjuntos de dados com classes desbalanceadas pode causar um problema conhecido como *overfitting*.

Esse desbalanceamento ocorre quando existe uma grande diferença entre as classes. Quando uma ou mais classes estão com mais frequência que as demais, o modelo classifica ou acerta a classe que está com maior frequência e classifica incorretamente as classes com menor frequência (minoritárias). O problema com classes desbalanceadas se torna complicado, pois o modelo pode fazer uma classificação errada dos dados.

Chama-se a classe menos frequente de classe positiva, que é a classe onde se encontram as regras mais interessantes. Um classificador que erra muito a classe positiva é inútil.

Toma-se como exemplo de classes desbalanceadas o conjunto de dados dado por [27].

NOME	TEMPERATURA DO CORPO	ORIGINA	QUATRO PATAS	HIBERNA	RÓTULO DA CLASSE
salamandra	sangue frio	não	sim	sim	não
peixe	sangue frio	sim	não	não	não
águia	sangue quente	não	não	não	não
poorwill	sangue quente	não	não	sim	não
platypus	sangue quente	não	sim	sim	sim

Tabela 2.5: Exemplo de conjunto de treinamento para classificar mamíferos.

De acordo com a árvore de decisão (Figura 2.12) gerada pelo conjunto de dados acima (Tabela 2.5), pode se verificar que os humanos, elefantes e golfinhos são classificados de forma errada, pois a árvore de decisão classifica todos os invertebrados de sangue quente que não hibernam como mamíferos. A árvore faz essa classificação, pois existe apenas um registro para o treinamento, que é a águia.

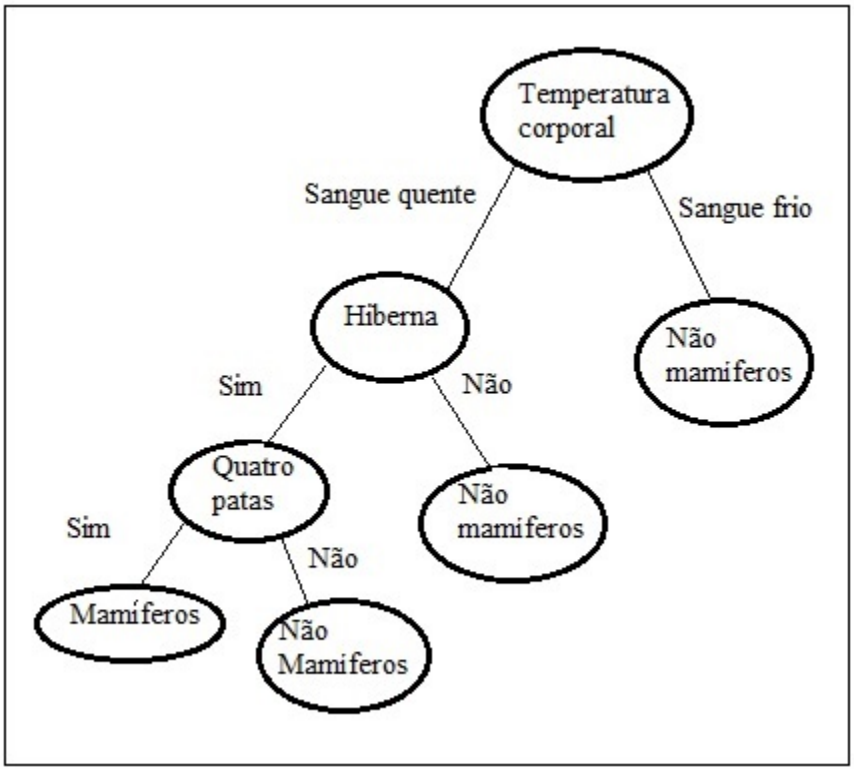

Figura 2.12: Árvore de decisão induzida dos dados da Tabela 2.5

Para contornar esse tipo de problema, nesse trabalho foi utilizado um filtro chamado de *Resample*, onde se pode variar o seu coeficiente de balanceamento, ou seja, aplicando o filtro e fazendo o seu coeficiente de balanceamento igual a 0.0, ele balanceia os dados de tal forma a não alterar a sua distribuição; e com o coeficiente de balanceamento igual a 1.0, faz-se uma distribuição uniforme dos dados.

O processo de balanceamento de dados é aplicado somente nos dados do conjunto de treinamento, enquanto o conjunto teste não pode ser alterado para não inviezar os dados.

2.9 Validação de Modelos Preditivos

A validação dos modelos preditivos é feita para testar o desempenho dos modelos gerados pelos algoritmos de classificação. Normalmente o conjunto de dados é dividido em duas partes, sendo uma para o conjunto de treinamento e a outra para o conjunto de teste.

Dentre os métodos mais utilizados tem-se o *Cross Validation* (Validação Cruzada) e o *Holdout*. O método de validação cruzada particiona o conjunto de dados em um número k de mesmo tamanho, e utilizam essas k divisões para gerar o modelo e as mesmas k divisões para o treinamento.

Durante cada execução, uma das partições é utilizada para o teste e as demais para o treinamento. Esse processo é repetido k vezes, de tal forma que cada partição seja utilizada uma única vez como teste e como treinamento. O erro total desse algoritmo é dado pela média dos k erros.

Já no método conhecido como *Holdout*, os exemplos rotulados são normalmente divididos em 2/3 para o conjunto de treinamento e 1/3 para o conjunto de teste. Dessa forma o modelo é induzido pelo conjunto de treinamento e seu desempenho é avaliado a partir do conjunto teste. Esse método possui algumas limitações como, uma pequena quantidade de exemplos rotulados pode fazer parte do conjunto de treinamento, fazendo com que o modelo não aprenda o que realmente deveria aprender, não tendo os exemplos na memória para se aplicar no conjunto teste. Outra limitação que o método possui é que o modelo pode ser altamente dependente do conjunto teste e do treinamento. Quanto menor o conjunto de treinamento maior a variância do modelo. Mas se o conjunto de treinamento for grande demais, então a precisão estimada calculada a partir do conjunto teste será menor, logo menos confiável [27].

No caso da validação cruzada, existem estudos mostrando que o número adequado para esse divisão é em dez partes, ou seja, número de Folds=10, pois ele minimiza o erro [18].

A utilização desses métodos garante que atributos com poucas representações entrem no conjunto de treinamento, evitando assimo o *overfitting*.

2.10 Medidas de Avaliação de Regras

A avaliação de regras é necessária para que o analista consiga saber quais são as regras mais interessantes, quais são aquelas que realmente são sustentadas pelos dados, ou ainda, quais são as regras que dão informações surpreendentes.

Mesmo as regras que possuem uma boa avaliação, podem fornecer informações irrelevantes para o analista. Por isso a importância da análise quantitativa, que conseguirá distinguir entre uma regra relevante ou não.

Nesta seção serão apresentadas a Matriz de Confusão e a Matriz de Contingência, que serão utilizadas nas análises dos modelos preditivos e de regras de classificação.

2.10.1 Matriz de Confusão

Seja uma Matriz de Confusão com duas classes, sendo uma Positiva (C_+) e outra negativa (C_-). Para casos com dimensões maiores, ela pode ser estendida sem perda de generalidade.

Uma matriz de confusão é dada da seguinte forma (Tabela 2.6):

	C_+	C_-	total
C_+	VP	FN	P
C_-	FP	VN	N
total	P'	N'	$P+N$

Tabela 2.6: Matriz de Confusão

Onde:
P= Positivos
$P^{'}$= Complementar dos positivos
N= Negativos
$N^{'}$= Complementar dos negativos
$P + N$= Soma dos positivos com os negativos
VP= Verdadeiro Positivos
FP= Falsos positivos
VN= Verdadeiros negativos
FN= Falsos negativos

Um bom modelo apresenta valores altos em sua diagonal principal e valores baixo na diagonal secundária.

Para a avaliação dos modelos, serão utilizadas as seguintes métricas:

1. Taxa de acerto ou *acurácia*:

$$\text{acurácia} = \frac{VP + VN}{P + N} \tag{2.10.1}$$

2. *Kappa*:

$$\text{Kappa} = \frac{P_0 - P_e}{1 - P_e}; \tag{2.10.2}$$

onde: $P_0 = \frac{VP+VN}{P+N}$ e $P_e = \frac{P^{'}P+N^{'}N}{(P+N)^2}$

3. *Precisão* por classe

$$\text{Precisão} = \frac{VP}{VP + FP} \tag{2.10.3}$$

2.10.2 Matriz de Contingência

Uma matriz de contingência é construída a partir da veracidade e falsidade dos elementos de uma regra.

Uma regra é definida como sendo: ANTECEDENTE (B) $\rightarrow$ CONSEQUENTE (H).

Sejam,

B: conjunto de exemplos para os quais o corpo ANTECEDENTE da regra é verdadeiro.

$\bar{B}$:Conjunto de exemplos para os quais a regra é falsa (Complemento de B).

H:conjunto de exemplos para os quais a cabeça CONSEQUENTE da regra é verdadeiro.

$\bar{H}$:Conjunto de exemplos para os quais a regra é falsa (Complemento de H).

Uma matriz de contingência é dada da seguinte forma (Tabela 2.7):

	H	$\bar{H}$	
B	hb	$\bar{h}b$	b
$\bar{B}$	$h\bar{b}$	$\bar{h}\bar{b}$	$\bar{b}$
	h	$\bar{h}$	n

Tabela 2.7: Matriz de Contingência

hb = Número de exemplos para os quais H é verdade e B é verdade.
$\bar{h}b$ = Número de exemplos para os quais H é falso e B é verdade.
$h\bar{b}$ = Número de exemplos para os quais H é verdade e B é falso.
$\bar{h}\bar{b}$ = Número de exemplos para os quais H é falso e B é falso.
b = Número de exemplos para os quais B é verdade.
$\bar{b}$ = Número de exemplos para os quais B é falso.
h = Número de exemplos para os quais H é verdade.
$\bar{h}$ = Número de exemplos para os quais B é falso.
n = Número total de exemplos.
Para a avaliação das regras, serão utilizadas as seguintes métricas:
1. Acurácia:
A acuráracia mede quanto uma regra é específica para o problema

$$Acc(R) = \frac{hb}{b} \tag{2.10.4}$$

2. Erro:
Quanto maior o erro, menos a regra cobre corretamente os exemplos

$$\text{Err(R)= 1- Acc(R)} \tag{2.10.5}$$

3. Sensitividade:
Sensitividade ou completeza mede a fração de verdadeiros positivos que são corretamente classificados

$$Sens(R) = \frac{hb}{h} \tag{2.10.6}$$

4. Especificidade:
É equivalente a Completeza, mas para os casos que não são cobertos pela regra

$$Sens(R) = \frac{\bar{h}\bar{b}}{\bar{h}} \tag{2.10.7}$$

5. Novidade:
Mostra o quanto uma regra é interessante

$$Nov(R) = \frac{1}{n}(hb - \frac{(h * b)}{n}) \tag{2.10.8}$$

6. Cobertura:
Mede o número de exemplos cobertos pela regra. Mede a generalidade da regra

$$Cov(R) = \frac{b}{n} \tag{2.10.9}$$

7. Suporte:
Mede o número de exemplos cobertos corretamente pela regra

$$Sens(R) = \frac{hb}{n} \tag{2.10.10}$$

8. *Kappa*:
A estatística *Kappa* é uma medida utilizada para verificar a concordância entre dois valores preditos e observados num processo de classificação.
O valor do *Kappa* é calculado da seguinte maneira:

$$\text{Kappa} = \frac{P_0 - P_e}{1 - P_e} \tag{2.10.11}$$

onde: $P_0 = \frac{VP+VN}{P+N}$ e $P_e = \frac{(P'P)+(N'N)}{(P+N)^2}$
Seu valor varia no intervalo $[0, 1]$, como pode ser observado na Tabela 2.8.

Estatística Kappa	**Qualidade**
$< 0{,}00$	concordância pobre
0,00-0,20	ligeira concordância
0,21-0,40	concordância considerável
0,41-0,60	concordância moderada
0,61-0,80	concordância substancial
0,81-1,00	concordância excelente

Tabela 2.8: Tabela da classificação *Kappa* segundo LANDIS e KOCH (1977)

2.11 Modelo do processo de descoberta de conhecimento em bases de dados

Tendo por objetivo padronizar o processo de descoberta de conhecimento, em 1996 foi criado o modelo CRISP-DM (*Cross-Industry Standart Process for Data Mining*), onde o ciclo de vida de um projeto de mineração é dividido em seis etapas (Figura 2.11), sendo elas: compreensão do domínio, entendimento dos dados, preparação dos dados, modelagem, avaliação e distribuição [8].
Esse trabalho seguiu esse modelo que será brevemente discutido a seguir:

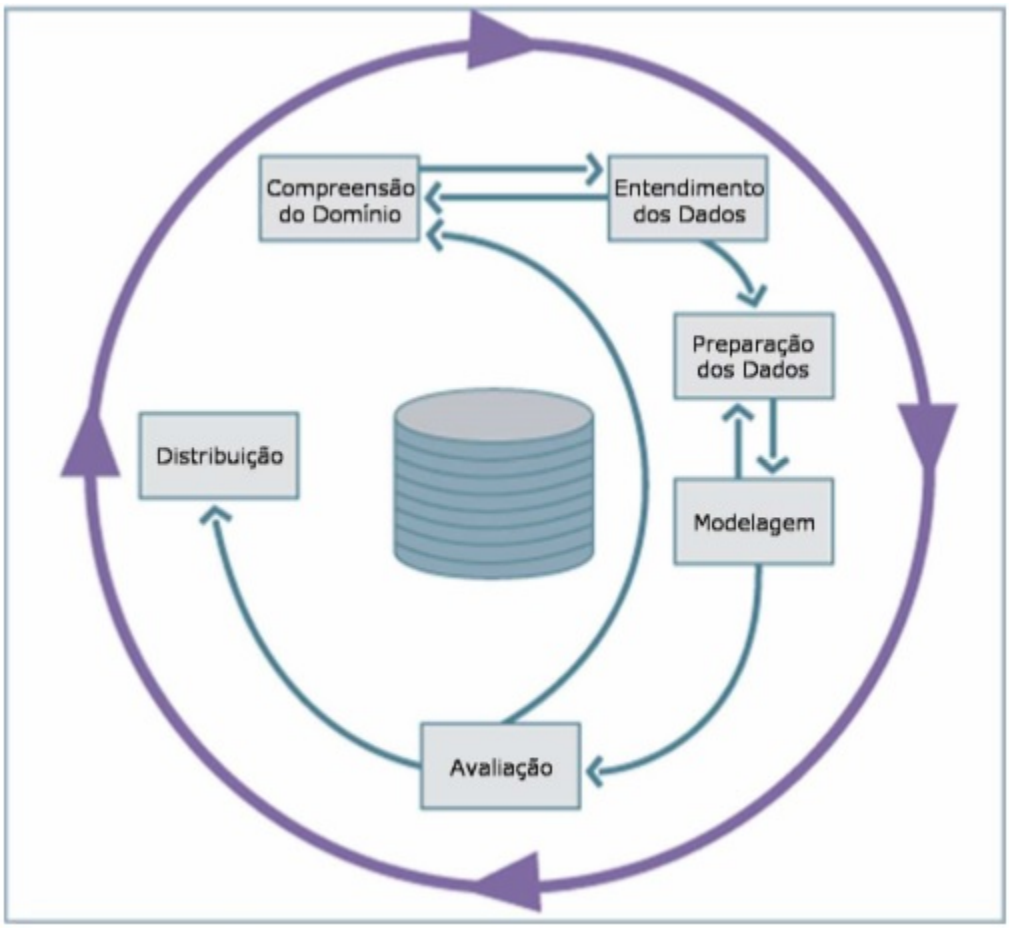

Figura 2.13: O processo CRISP-DM(CHAPMAN, 2000)

- Compreensão do domínio:

Compreender os objetivos e requisitos do projeto e transformar esses conhecimentos em um problema de mineração de dados; Definir um plano preliminar para atingir os objetivos.

- Entendimento dos dados:

Esse processo se inicia com a coleção de dados, são realizadas atividades de exploração de dados, identificação de problemas de qualidade, identificação das primeiras hipóteses e possíveis subconjuntos que possam fornecer informações ocultas sobre os dados.

- Preparação dos dados:

A fase de preparação dos dados contempla todas as atividades necessárias para a construção do conjunto de dados final, no qual serão aplicadas as técnicas de modelagem. As atividades incluem, por exemplo: limpeza de dados, seleção e transformação de atributos.

- Modelagem:

Nessa fase são escolhidas e aplicadas técnicas de mineração de dados, e seus parâmetros são calibrados de acordo com cada situação. Diversas técnicas podem ser aplicadas ao mesmo problema, embora, cada técnica necessite de formatos específicos e necessite voltar para a fase de preparação de dados.

- Avaliação:

Nesse estágio, tem-se o modelo (ou modelos) com boa qualidade. Os resultados são comparados e interpretados conforme a área de aplicação. É importante reavaliar todas as etapas do processo, para ter a certeza de que o modelo atende às necessidades e aos objetivos do projeto.

- Distribuição:

A criação de modelos geralmente não finaliza um projeto; O conhecimento obtido deve ser documentado, organizado e apresentado aos usuários, para que os mesmos saibam quais ações devem ser tomadas para melhor aproveitamento dos modelos criados.

Capítulo 3

Materiais e Métodos

3.1 Considerações Iniciais

Para o desenvolvimento deste trabalho foram utilizadas imagens do sensor MODIS a bordo do satélite TERRA. Este, por sua vez, fornece imagens com resolução espacial de 250 m. Essas imagens são disponibilizadas pelo LP-DAAC/NASA (Land Processes Distributed Active Archive Center) que é um centro de processamento e distribuição ligado à NASA, que por sua vez, foram repassadas pela Embrapa Informática Agropecuária.

Foram realizadas tarefas de exploração, limpeza dos dados e o preenchimento de valores faltantes dos dados, com o objetivo de minimizar os ruídos, aumentando assim o potencial de descoberta de conhecimento dos dados.

Para conseguir informações sobre a classificação das culturas, foram utilizadas técnicas de modelagem com os classificadores J48, Random Forest e PART, todos pertencentes ao software Weka [1], ambiente que contém um conjunto de algoritmos de aprendizado de máquina, para solucionar problemas de mineração de dados [13].

Para dar suporte aos procedimentos realizados neste trabalho, optou-se por seguir um modelo de processo descrito por [8] para projetos de descoberta de conhecimento em bases de dados, conhecido como CRISP-DM (CRoss-Industry Standart Process for Data Mining). A aplicação do modelo CRISP-DM neste trabalho, com as atividades desenvolvidas nas respectivas fases, podem ser acompanhadas nas seções seguintes.

3.2 Compreensão do domínio

Esta fase teve como resultado o Capítulo 2, com a revisão da literatura, reuniões com especialistas das áreas de geotecnologia e mineração de dados, para melhor conhecimento dos problemas, propriedades e as restrições que cercam o assunto da soja, do milho e do algodão.

[1]Software de domínio público Weka (Waikato Environment for Knowledge Analysis), da Universidade de Waikato, Nova Zelândia.

3.3 Entendimento dos dados

3.3.1 Coleção inicial dos dados e descrição

O sensor MODIS faz a varredura diária de regiões agrícolas tradicionais do Mato Grosso; nos municípios de Campo Novo do Parecis, Campos de Júlio, Nova Mutum, Sapezal, Lucas do Rio Verde, Sinop e Sorriso (região central e oeste do estado), tirando e armazenando imagens de onde foram obtidos os dados deste trabalho. A Figura 3.1, mostra as regiões de Mato Grosso que serão examinadas neste trabalho.

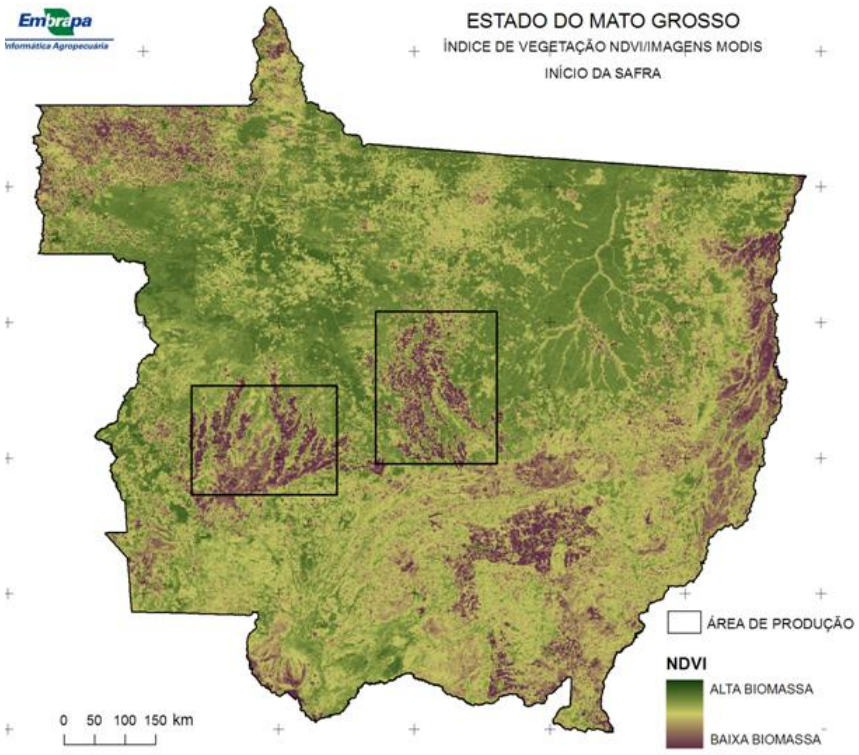

Figura 3.1: Composição de 16 dias do NDVI no Estado de Grosso

Fonte: Embrapa Informática Agropecuária

Para o desenvolvimento do trabalho, foram fornecidos quatorze conjuntos de dados, variando nos anos safra de 2003/2004 a 2008/2009, onde sete deles são os dados brutos dos sensores e os outros sete, são dados com filtros, ou seja, os dados possuem alguma variação devido à retirada de ruídos por um especialista da Embrapa. Depois de vários testes e experimentos, verificou-se que os dados do ano-safra de 2008/2009 do estado de Mato Grosso-Brasil, sendo eles filtrados, eram os melhores conjuntos de dados que poderiam descrever melhor o problema, e assim obter melhores resutados na classificação.

Os registros dos dados utilizados neste trabalho, estavam organizados em uma planilha do Excel, então eles foram organizados em um único arquivo no formato de tabela tipo .csv (comma separated values - valores separados por vírgulas), no qual cada linha representa um registro de ocorrência e cada coluna representa um atributo relacionado à ocorrência.

Dados espectrais

As planilhas foram agrupadas de tal forma que se preservou o rótulo do ano safra da planilha com as datas mais recentes, e as planilhas com as datas mais velhas eram colocadas logo abaixo, conseguindo assim um grande conjunto de dados (matriz de pé) com um único rótulo.

As duas primeiras colunas do conjunto de dados fazem referência à localização geográfica de cada talhão: a latitude e a longitude. Já a terceira é referente ao código do talhão (um código que identifica cada gleba analisada). Da quarta até a vigésima quinta coluna tem-se os valores do NDVI. A penúltima coluna é a Safra e na última coluna é a Safrinha.

Safrinha é a cultura cultivada naquele solo após o período de Safra. Se faz necessário o cultivo da safrinha para a proteção do solo, para evitar perdas de nutrientes na exposição ao sol ou à chuva, para a preparação daquele solo para o próximo plantio da Safra e para geração de renda no período fora da safra.

No período da safrinha, a maior parte dos registros referem-se ao milho, milheto, algodão, sorgo, feijão, girassol, pasto, crotalária e arroz. Eventualmente, outros usos não agrícolas, como mata e reflorestamento, também foram encontrados. Como o objetivo principal era conseguir distinguir a soja, o milho e o algodão no período da Safra, com a ajuda do especialista e após uma série de testes, optou-se por não trabalhar com a Safrinha, excluindo-se portanto a última e as três primeiras colunas do conjunto de dados. Sendo assim, foram utilizados apenas os 24 atributos, sendo 23 referentes ao NDVI e 1 atributo referente à Safra.

As Tabelas 3.1 e 3.2 a seguir, dão uma idéia de como era e como ficou a disposição dos dados desse trabalho:

LATITUDE col 1	LONGITUDE col 2	C. TALHÃO col 3	NDVI col 4-25	SAFRA col 26	SAFRINHA col 27
valores númericos	valores númericos	valores númericos	valores NDVI	culturas	culturas

Tabela 3.1: Disposição inicial dos dados.

NDVI col 4-25	SAFRA col 26
valores NDVI	culturas

Tabela 3.2: Disposição final dos dados.

3.3.2 Exploração dos dados

Na fase de exploração dos dados buscou-se entender a sua distribuição, para isso foi utilizado o Excel, para a visualização de tabelas e gráficos.

Nesta fase, foram feitos vários testes e análises combinando vários conjuntos de dados, com a finalidade de se encontrar aquele ou aqueles que melhor descreveríam o problema estudado.

Cada conjunto de dados possuia 413 observações (instâncias ou linhas) e 27 atributos (colunas). Os atributos são dados em datas, sempre iniciando em 13/09 e são somados 16 dias para a obtenção do próximo atributo, como pode ser observado na Tabela 3.3:

	A	B	C	D	E	F
1	Latitude	Longitude	Codigo Talhao	20080913	20080929	20081015
2	-14,045	-55,51385	194	0,344626	0,282709	0,28197
3	-14,0259	-55,54369	192	0,327164	0,32333	0,305982
4	-14,017	-55,56328	193	0,353502	0,319844	0,305423
5	-14,015	-55,54163	191	0,363057	0,332345	0,319672
6	-13,9885	-58,92519	307	0,164856	0,165362	0,282343
7	-13,9888	-58,96782	308	0,529457	0,589184	0,59363
8	-13,9848	-58,95	3071	0,249922	0,27662	0,345411

Tabela 3.3: Amostra do conjunto de dados no Excel

Elaborado pelo autor

Cada imagem gerada pelo sensor, pode ser vista como uma matriz, onde cada elemento da matriz é chamado de pixel. A partir dos pixels localizados dentro das áreas de referência, visitadas em campo, os valores de NDVI são extraídos em cada uma das datas, obtendo-se uma série temporal, a qual pode ser representadas a partir de gráficos de perfis (Figura 3.2)

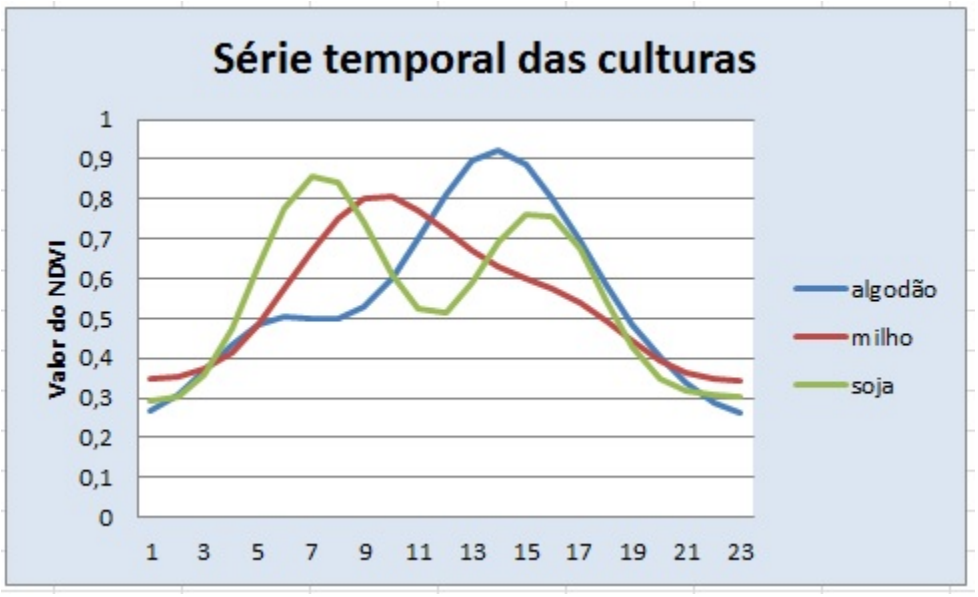

Figura 3.2: Série temporal da vegetação

Então, para cada imagem gerada pelo sensor é analisado cada um dos seus pixels (cada elemento da matriz). Aqueles que apresentam algum tipo de interferência é substituido por outro pixel, entre as imagens dos 16 dias, conseguindo assim montar uma imagem com os melhores pixels, ou os mais limpos (sem interferências ou ruídos). Todo esse trabalho foi realizado por um especialista da Embrapa.

Nesse conjunto de dados, foram encontrados dados redundantes, mostrados na Tabela 3.4. Dados com baixa frequência foram retirados, pois não fariam nenhuma diferença na análise pela sua baixa frequência.

cultura	frequencia
abertura	1
pipoca	1
crotalia	1
enleiramento	2
milheto/soja	2
milho branco	1
nim	2
no data	4
pasto(b)	2
pasto/boi	8
pasto/eucalipto	4

Tabela 3.4: Tabela com os dados redundantes

Como os algoritmos de aprendizagem do Weka trabalham apenas com uma variável-meta, cujos valores são únicos e não compostos, retiraram-se também as colunas que continham duas culturas, como por exemplo: **pasto/boi** e **pasto/eucalipto**.

No caso do **no data**, entendeu-se que era um sinônimo de um valor faltante, então foi deixado em branco para ser preenchido com os demais valores faltantes.

3.4 Preparação dos dados

Após as fases de compreensão do domínio e entendimento dos dados, proseguiu-se com a fase de preparação dos dados que foram:

- Verificação e preenchimento de valores faltantes;
- Integração os dados em um formato específico, necessário para aplicar as técnicas de classificação.

Para definir o período em que as séries temporais seriam estudadas, foi considerado apenas o período Safra, levando em consideração o período de cada cultura (detalhado mais adiante), no caso; soja, milho e algodão.

3.4.1 Especificação do atributo meta

O atributo meta, também conhecido por classe ou variável dependente, é um atributo especial para problemas de classificação, e deve ser categórico, ou seja, não numérico. Neste caso, ele assume valores de três tipos de cultura: soja, milho e algodão.

3.4.2 Especificação dos atributos preditivos espectrais

Os atributos preditivos (as variáveis independentes) são utilizados para explicar o atributo meta (variável dependente). Como o objetivo geral deste trabalho é desenvolver modelos para classificar culturas de soja, milho e algodão com a utilização do NDVI. Foram descartados os atributos que possuíam qualquer outro tipo de cobertura de solo que não fossem soja, milho ou o algodão.

Os dados deste trabalho foram coletados pessoalmente por pesquisadores da Embrapa em um trabalho de campo. O gerente ou o funcionário responsável pela fazenda era abordado, e o pesquisador perguntava a ele qual era a cultura que havia sido cultivada no ano anterior, no seu antecessor, e assim por diante. Se não houvesse um livro registro ou algo parecido, as informações se tornariam não confiáveis e inconsistêntes com o passar dos anos. Com isso os dois últimos anos (2008 e 2009) foram bem preenchidos e conforme os anos se passavam as informações se perderiam. Isso justifica o porquê do trabalho ser realizado apenas com os dois últimos anos, para evitar inconsistências de dados e valores faltantes (missing).

Outras atividades de extrema importância para a preparação dos dados estão descritas no próximo capítulo.

3.5 Modelagem

As técnicas de modelagem escolhidas foram a indução de árvores de decisão, para classificação por meio do classificador J48 (uma versão adaptada do algoritmo clássico C4.5), Random Forest e o PART, todos estão disponíveis no ambiente Weka.

Os primeiros testes foram realizados para a verificação de valores faltantes. Para isso, ultilizou-se o algoritmo K-NN também disponível no Weka, para o preenchimento dos mesmos.

De posse do conjunto de dados já preparado (também chamado de conjunto de treinamento) foram aplicados diversos experimentos a fim de se avaliar e caracterizar o desempenho dos classificadores escolhidos. Esses experimentos possuem suas diferenças nos subconjuntos de treinamento utilizados, tanto em relação aos registros, quanto nas diferenças dos parâmetros utilizados para a geração dos modelos.

No caso dos experimentos realizados com o J48 e o PART, é realizada uma pós-poda, que tem como finalidade deixar o modelo mais genérico; Uma vez que ele altera a quantidade de objetos que vão passar por aquela folha, ele está alterando a divisão dos atributos, reduzindo a quantidade de ramos da árvore a partir do MinNumObj. Para os testes realizados com esses algoritmos, os parâmetros que representam o número de objetos por folha utilizado foi o *default* do software [2].

[2]Software de domínio público Weka (Waikato Environment for Knowledge Analysis), da Universidade de Waikato, Nova Zelândia.

No caso dos testes realizados com o Random Forest, para o número de árvores da floresta representado pelo parâmetro NumTrees, fez-se uso do *default* que é igual a 10; E para o número de folhas da árvore, utilizou-se MaxDepht = 5. A conclusão de como variar esses parâmetros foi obtida através de vários testes, sempre buscando altos valores para a estatística *Kappa*, *acurácia* e a *precisão* de cada classes.

A grande dificuldade nos testes foi o desbalanceamento das classes, a classe soja era predominante com relação às demais classes. Por esse motivo, foi necessário a utilização do filtro *Resample*, que tenta contornar esse tipo de problema fazendo um balanceamento, ou seja diminuindo a classe majoritária e aumentando as classes minoritárias.

Para a validação dos modelos foi utilizado um filtro para instâncias supervisionadas chamado StratifiedRemoveFolds, onde ele divide o conjunto de dados em duas partes, sendo uma para o conjunto de treinamento e o outro para o conjunto teste. Para os mesmos testes foram também utilizados o CrossValidation, onde ele particiona o conjunto de dados em K partes, treina o conjunto em uma das K partes e testa com as K-1 partes, o processo é repetido até que todas as partes sejam utilizadas no treinamento.

Depois dos modelos gerados e escolhidos, foi analisada cada uma das regras para verificar possíveis inconsistências, como por exemplo: dois valores distintos de NDVI para uma mesma data. Este tipo de regra foi retirada do conjunto.

E para encerrar os estudos, foram gerados mapas das regiões agrícolas de Mato Grosso, para verificar se, o que os modelos estavam descrevendo, condizia com a realidade.

3.6 Softwares e parâmetros

Os principais softwares utilizados para a realização deste trabalho foram o Weka [13] versão 3.7.9 e o Excel. A plataforma e o sistema operacional foi o Windows. A principal característica do software Weka utilizado é que, ele é livre e gratuito, de fácil instalação e sem a necessidade de licenças especiais para realizar este trabalho.

O software Weka (Waikato Environment for Knowledge Analysis) foi desenvolvido na Universidade de Waikato, Nova Zelândia, ele possui uma coleção de algoritmos de aprendizado de máquina e outras ferramentas de análise, que o permite fazer uma série de testes e experimentos no processo de mineração de dados [28]. Ele pode ser adquirido gratuitamente no site (http://www.cs.waikato.ac.nz/ml/weka/) e distribuído sob a licença de uso GNU GPL. Ele foi escolhido por sua facilidade de uso e ter sido apontado como um dos principais softwares livres utilizados para mineração de dados.

O Weka possui vários tipos de ferramenta para o mineração de dados, sendo eles: Explorer, Experimenter, KnowledgeFlow e Simple CLI. Para o desenvolvimento desse trabalho foi utilizado o ambiente gráfico interativo Explorer. Neste ambiente, assim como nos outros, cada algoritmo possui parâmetros e características próprias, sendo possível alterá-los de acordo com as necessidades de cada problema.

Para cada algoritmo utilizado nesse trabalho, fez-se uma adaptação nos parâmetros dos filtros e dos métodos de seleção (quando houve necessidade) com a finalidade de se obter melhores resultados. Ao utilizar o classificador J48, foi mantido os valores dos seus parâmetros originais, ou

seja, foi utilizado o *default* do software. No Random Forest, foram ajustados: o número de árvores da floresta representado por Max Deph, na Tabela 4.24 e o número mínimo de objetos por folha, representado por MinNumObj utilizou-se o *defauld* do software . Já no caso da validação cruzada, o número utlizado foi 10, que por sua vez foi justificado com uma série de testes apresentados na Tabela 4.12. Para o filtro Resample foi variado o número de balanceamento representado por BiasToUniformClass (que no texto foi chamado de R=0,0; R=0,5 e R=1,0), de tal forma que quando se utiliza R=0,0 não ocorre alteração na distribuição dos dados, já quando se utiliza R=1,0, têm-se uma distribuição uniforme dos dados.

Capítulo 4

Resultados e Discussões

4.1 Análise exploratória

Exploração dos dados

Neste capítulo serão apresentados os principais resultados obtidos nos testes. Buscou-se modelos em que os valores do *Kappa* fossem acima de 60%, onde a concordância é considerada substancial; a *acurácia* o mais próximo possível do 100%; e modelos que fizessem a classificação das três classes (precisão por classe) estudas.

Os primeiros testes realizados foram feitos para a verificação de valores faltantes. Para isso, utilizou-se inicialmente os anos safra de 2003/2004 a 2008/2009, depois apenas o ano-safra de 2008/2009. Nos testes pode-se verificar a existência de 64 valores faltantes, para os anos safra de 2003/2004 a 2008/2009, o que representa 4% dos dados. Enquanto que no ano-safra de 2008/2009, tem-se apenas 1 valor faltante, representando 0% do conjunto de dados. Mesmo assim, foram realizados os testes com os anos safras de 2003/2004 a 2008/2009 e 2008/2009, utilizando os classificadores J48, Random Forest e o PART sem seleção de Atributos e sem Balanceamento. Mas os resultados não foram satisfatórios, pois o valor do *Kappa* estava abaixo do desejado, como pode ser visto nas Tabelas 4.1, 4.2 e 4.3:

Teste com o J48	
Acurácia	85,68 %
Kappa	45,12 %

Tabela 4.1: Teste Sem Seleção de Atributo e Sem Balanceamento com o J48

Teste com o Random Forest	
Acurácia	87,16 %
Kappa	44,98 %

Tabela 4.2: Teste Sem Seleção de Atributo e Sem Balanceamento com o Random Forest

Teste com o PART	
Acurácia	84,63 %
Kappa	44,34 %

Tabela 4.3: Teste Sem Seleção de Atributo e Sem Balanceamento com o PART

Como pode ser observado, o valor do *Kappa* ficou em torno de 45% nos três testes. Mas como o objetivo desse trabalho é encontrar modelos com valores do *Kappa* acima do 60%, esses testes foram desprezados e novos testes foram realizados.

Como o ano-safra de 2008/2009 continha apenas 1 valor faltante, optou-se por trabalhar com ele restringindo as classes em: soja, milho e algodão. Mas mesmo assim o conjunto de dados estava muito mal balanceado, ou seja, a frequência da cultura soja era muito grande com relação as demais classes, como pode ser observado na Figura 4.4 a seguir:

Classes	Frequência
soja	1297
milho	12
algodão	61
pasto	77
pasto/eucalipto	0
pasto+boi	0
pousio	15
eucalipto	3
milheto/soja	0
milho branco	0
juquira	7
amendoin	5
jeijão	0
crotalária	0
pasto degenerado	10
arroz	32
no data	0
reflorestamento	4
nim	0
pasto(b)	4
mata	14
milheto	4
pato(brizantão)	0
abertura	0
enleiramento	0
soja precoce	4
pipoca	0

Tabela 4.4: Frequência da classe meta de 2003 a 2009

Como já discutido anteriormente, esse desbalanceamento nos dados prejudica muito os resulatdos dos teste. Pois o modelo pode ser treinado apenas com a classe mais representativa, no caso a soja, deixando assim de classificar as classes minoritárias.

Assim, optou-se por trabalhar apenas com o ano-safra de 2008/2009, contendo apenas as classes: soja, milho e algodão, fazendo o preenchimento dos valores faltantes e retirando as redundâncias. Dessa forma, a classe meta passou para 3 classes distintas, sendo elas: soja, milho e algodão, com as seguintes frequências.

Classes	Frequência
soja	678
milho	8
algodão	39

Tabela 4.5: Frequência das classes: soja, milho e algodão do ano-safra 2008/2009

Pode-se observar na Tabela 4.5 o desbalanceamento nas classes, ou seja, uma diferença muito grande com relação ao número de frequência das culturas. Por conta desse desbalanceamento, foi necessário a utilização de um filtro específico para tentar contornar o problema e evitar assim o *overfitting*, ou a classificação errônea das classes. Para tanto, utilizamos o filtro supervisionado *Resample*.

Como já pode ser observado nas Tabelas 4.1, 4.2, 4.3 o valor do *Kappa* não estava satisfatório (estava em 45%, enquanto que o objetivo era algo acima de 60%), por isso foi realizada uma seleção de atributos para tentar aumentar o valor do mesmo. Para tanto, foi utilizado os seguintes algoritmos: X^2, PCA, CFS, Gain Ratio, Info Gain e Wrapper.

4.1.1 Modelo de predição utilizando regras de classificação

O conjunto de dados utilizado para gerar os resultados desse trabalho foi o conjunto de dados do ano-safra de 2008/2009. Como já foi discutido, para a realização dos testes foram utilizados os seguintes selecionadores de atributos: X^2, PCA, CFS, Gain Ratio, Info Gain e Wrapper. Para contornar a questão do desbalanceamento foi utilizado o filtro *Resample* com o StratifiedRemoveFolds e posteriormente com o Cross Validation.

A seguir serão apresentados os testes realizados por cada selecionador de atributo alternando com os classificadores. O primeiro teste será mais detalhado com a finalidade de familiarizar-se com as Tabelas, os próximos testes serão realizados de maneira análoga.

4.1.2 Testes com o PCA

Neste primeiro teste será realizada uma selação nos atributos utilizando o selecionador de atributos PCA. De posse do conjunto selecionado será realizado o balanceamento utilizando o filtro *Resample*, variando seu coeficiente de balanceamento. Em seguida será aplicado os seguintes classificadores: J48, Random Forest e o PART.

Assim para cada coeficiente de balanceamento (*Resample*=0,0; *Resample*=0,5 e *Resample*=1,0) será gerado um teste para cada classificador (grupo de três linhas da primeira Tabela) utilizando o StratifiedRemoveFolds. Nestes testes serão verificados os valores da *acurácia*, do *Kappa* e a *precisão* de cada cultura, onde os resultados serão imprensos na Tabela 4.6

PCA utilizando Conjunto teste de 10%					
classificador	acurácia	kappa	precisão da soja	precisão do milho	precisão do algodão
Resample=0.0					
J48	98,630	0,852	0,972	1,000	1,000
Random Forest	97,260	0,737	0,971	0,000	0,667
PART	95,890	0,649	0,985	0,000	0,750
Resample=0.5					
J48	90,411	0,423	0,984	0,000	0,333
Random Forest	95,890	0,648	0,985	0,000	0,600
PART	95,890	0,649	0,985	0,000	0,750
Resample=1.0					
J48	91,781	0,468	0,985	0,500	0,333
Random Forest	95,890	0,648	0,985	0,000	0,600
PART	89,041	0,388	0,984	0,000	0,333

Tabela 4.6: Testes utilizando J48 com o PCA e o Resample

De maneira análoga, com o mesmo conjunto de dados serão realizados os mesmos testes (um selecionador de atributo com os três classificadores), mas utilizando agora o Cross Validation, obtendo a segunda Tabela 4.7:

PCA utilizando o Cross Validation					
classificador	acurácia	kappa	precisão da soja	precisão do milho	precisão do algodão
J48	95,604	0,599	0,970	0,000	0,788
Random Forest	96,429	0,650	0,970	0,000	0,667
PART	95,742	0,616	0,972	0,000	0,788

Tabela 4.7: Testes utilizando o PCA e o Cross Validation

Será comparado os valores das duas Tabelas, buscando sempre o teste em que houve a classificação das três classes e os maiores possíveis valores para o *kappa* e para a *acurácia*.

De acordo com a Tabela 4.7, pode-se verificar que não houve a classificação do milho (precisão do milho $= 0,000$) no entanto, na Tabela 4.6 houve a classificação das três classes com uma *precisão* de quase 100% para cada uma delas, têm-se também o valor do *kappa* igual a $85,16\%$, que é maior que o mínimo desejado e uma *acurácia* com valor igual a $98,603\%$. Logo, os valores obtidos nesse teste são os melhores para a geração do modelo.

A seguir, será apresentada uma Tabela com o melhor resultado dos testes (Tabela 4.8):

Analise do Modelo PCA utilizando Resample com R=0 e conjunto teste de 10%					
algoritmo	acurácia	kappa	precisão da soja	precisão do milho	precisão do algodão
J48	98,6301	0,8516	0,986	1	1

Tabela 4.8: Resultado do teste utilizando o J48 com o PCA e o Resample

Em seguida, será apresentada uma quarta Tabela (Tabela 4.9) com as regras de classificação obtidas no respectivo teste.

Regras	PCA utilizando Resample com R=0 e conjunto teste de 10%
R1	Se 20090306 <= 0,8648 e 20090101 <= 0,7134 e 20090202 <= 0,7093 e 20090218 <= 0,288509 então algodão (2)
R2	Se 20090306 <= 0,8648 e 20090101 <= 0,7134 e 20090202 <= 0,7093 e 20090218 > 0,288509 e 20081015 <= 0,2626 então algodão (3/1)
R3	Se 20090306 <= 0,8648 e 20090101 <= 0,7134 e 20090202 <= 0,7093 e 20090218 > 0,288509 e 20081015 > 0,2626 então soja (32)
R4	Se 20090306 <= 0,8648 e 20090101 > 0,7134 então soja (572/7)
R5	Se 20090306 <= 0,8648 e 20090101 <= 0,7134 e 20090202 > 0,7093 e 20090101 <= 0,6761 e 20081015 <= 0,7411 então soja (7)
R6	Se 20090306 <= 0,8648 e 20090101 <= 0,7134 e 20090202 > 0,7093 e 20090101 <= 0,6761 e 20081015 > 0,7411 então milho (3)
R7	Se 20090306 <= 0,8648 e 20090101 <= 0,7134 e 20090202 > 0,7093 e 20090101 > 0,6761 então milho (6)
R8	Se 20090306 > 0,8648 e 20081202 <= 0,6228 então algodão (19/1)
R9	Se 20090306 > 0,8648 e 20081202 > 0,6228 então soja (11)

Tabela 4.9: Conjunto de regras gerado com o J48 com o PCA e o Resample

E para encerrar a abordagem, a retirada se necessário de alguma regra devido algum tipo de incompatibilidade discutido juntamente com o especialista da área. E a análise de todas as regras utilizando os valores obtidos nas matrizes de contingência, obtendo assim uma quinta Tabela com esses valores. As regras com os maiores valores, serão classificadas como as melhores regras do teste.

Nesse primeiro teste, a regra R4 foi retirada do conjunto de regras. Segundo o especialista da Embrapa, o valor do NDVI dado por 0,8648 em Março não pode ser maior que seu valor dado por 0,7134 em Janeiro, onde se encontra o pico da soja.

Desse conjunto de regras, foi analisada cada uma delas utilizando as matrizes de contingência. De posse dessas matrizes pode-se calcular a *acurácia*, o *erro*, a *sensitividade*, a *especificidade*, a *cobertura*, o *suporte* e a *novidade* de cada regra, explicitados na Tabela 4.10:

Analise das regras							
regras	acurácia	erro	sensitividade	especificidade	cobertura	suporte	novidade
R1	1,000	0,000	0,830	1,000	0,003	0,003	0,003
R2	0,750	0,250	0,125	0,998	0,006	0,005	0,004
R3	1,000	0,000	0,051	1,000	0,049	0,049	0,002
R5	1,000	0,000	0,011	1,000	0,011	0,011	0,001
R6	1,000	0,000	0,330	1,000	0,005	0,005	0,005
R7	1,000	0,000	0,660	1,000	0,009	0,009	0,009
R8	0,950	0,050	0,971	0,998	0,031	0,029	0,028
R9	1,000	0,000	0,170	1,000	0,017	0,017	0,001

Tabela 4.10: Análise das regras com o J48, o PCA e o Resample

De acordo com os valores dados pela tabela acima, pode-se verificar que as regras, R7 e R8, são as que possuem os maiores valores para a *acurácia*, a *sensitividade*, a *especificidade* e o *suporte*. Logos essas são as melhores regras geradas por esse modelo.

4.1.3 Testes com o CFS

Para a realização dos testes com o CFS, foram utilizados o J48, o PART e o Random Forest juntamente com o Resample, obtendo como resultado a Tabela 4.11:

CFS utilizando Conjunto teste de 10%					
classificador	acurácia	kappa	precisão da soja	precisão do milho	precisão do algodão
Resample=0,0					
J48	97,260	0,737	0,986	0,000	0,750
Random Forest	97,260	0,737	0,986	0,000	0,750
PART	97,260	0,737	0,986	0,000	0,750
Resample=0,5					
J48	93,151	0,516	0,985	0,000	0,429
Random Forest	97,260	0,737	0,986	0,000	0,750
PART	93,151	0,516	0,985	0,000	0,429
Resample=1,0					
J48	86,301	0,328	0,984	0,000	0,273
Random Forest	97,260	0,788	1,000	0,000	0,600
PART	86,301	0,235	0,968	0,000	0,222

Tabela 4.11: Testes utilizando o CFS e o Resample

Foram realizados alguns testes para verificar qual seria o melhor valor para o número de Folds, na realização dos testes com o Cross Validation, explicitado na Tabela 4.12.

Número de folds utilizado no Cross Validation					
folds	acurácia	kappa	precisão da soja	precisão do milho	precisão do algodão
2	94,643	0,532	0,968	0,000	0,686
3	95,742	0,640	0,978	0,200	0,692
4	94,231	0,512	0,968	0,000	0,632
5	94,506	0,499	0,964	0,000	0,688
6	95,192	0,601	0,974	0,000	0,707
7	94,918	0,547	0,968	0,000	0,750
8	95,330	0,592	0,971	0,000	0,692
9	95,604	0,617	0,974	0,222	0,781
10	95,742	0,607	0,970	0,500	0,735
11	95,330	0,609	0,975	0,143	0,711
12	95,192	0,580	0,971	0,000	0,703

Tabela 4.12: Número de Folds utilizado no Cross Validation

Com esses valores é possível justificar o porquê de se utilizar o número de Folds = 10 com o Cross Validaton, que por sua vez é o *defauld* do software.

Agora, é testado o CFS com o J48, o PART e o Random Forest com o Cross Validation e o número de Folds=10, obtendo o seguinte resultado (Tabela 4.13):

CFS utilizando o Cross Validation					
classificador	acurácia	kappa	precisão da soja	precisão do milho	precisão do algodão
J48	95,154	0,649	0,977	0,000	0,848
Random Forest	96,154	0,656	0,975	0,000	0,757
PART	95,742	0,607	0,970	0,500	0,735

Tabela 4.13: Testes utilizando o CFS com Cross Validation

Observando os valores da tabela acima, pode-se concluir que os melhores valores será dado por (Tabela 4.14):

Analise do Modelo CFS com Cross Validation					
algoritmo	acurácia	kappa	precisão da soja	precisão do milho	precisão do algodão
PART	95,742	0,607	0,970	0,500	0,735

Tabela 4.14: Resultado do teste utilizando CFS com Cross Validation

Onde o valor do *Kappa* é igual a $60,70\%$, a *acurácia* com valor igual a $95,741\%$, *precisão* da soja 97%, *precisão* do milho 50%, e a *precisão* do algodão $73,5\%$.

Nesse modelo foi utilizado o classificador PART juntamente o Cross Validation.

A Tabela 4.15 mostra o conjunto de regras geradas por esse modelo.

Regras	CFS com Cross Validation
R1	Se 20090306 <= 0,865337 e 20081218 > 0,7067 então soja (612/7)
R2	Se 200903022 <= 0,8141 e 20090306 <= 0,7173 então soja (46/1)
R3	Se 200900101 <= 0,5635 e 20081202 <= 0,6475 então algodão (36/5)
R4	Se 20081202 > 0,63206 então soja (19/1)
R5	Se 20081202 <= 0,4775 então soja (7)
R6	Se 20090306 > 0,5921 então milho (5)
R7	Se 20090306 > 0,5921 então soja (3/1)

Tabela 4.15: Conjunto de regras gerado pelo CFS com o PART

Da mesma forma, com as matrizes de contingência, foi calculado a *acurácia*, o *erro*, a *sensitividade*, a *especificidade*, a *cobertura*, o *suporte* e a *novidade* de cada regra, com seus valores dados na Tabela 4.16:

Analise das regras							
regras	acurácia	erro	sensitividade	especificidade	cobertura	suporte	novidade
R1	0,989	0,011	0,891	0,829	0,850	0,841	0,038
R2	0,979	0,021	0,067	0,976	0,065	0,063	0,002
R3	0,878	0,122	1,000	0,993	0,056	0,049	0,047
R4	0,950	0,050	0,028	0,976	0,027	0,026	0,000
R5	1,000	0,000	0,010	1,000	0,010	0,010	0,001
R6	1,000	0,000	1,000	1,000	0,007	0,007	0,007
R7	0,750	0,250	0,004	0,976	0,005	0,004	-0,001

Tabela 4.16: Análise das regras geradas pelo CFS com o PART

De acordo com os resultaados dados pela Tabela acima, pode-se verificar que a regra R1 possui um alto valor de *acurácia*, a *sensitividade*, e a *especificidade* e de *suporte*, já as regras R3 e R6 possuem altos valores de *acurácia*, a *sensitividade*, e a *especificidade*. Logos essas três regras são as melhores regras geradas por esse modelo.

4.1.4 Testes com o Info Gain

Nessa altura do trabalho, pode-se verificar que os testes realizados com o Gain Ratio, Info Gain e o X^2 deram os mesmos resultados. Assim, será apresentado apenas os procedimentos com o Info Gain, levando em consideração que para os demais selecionadores de atributos os resultados serão os mesmos.

Para a realização dos testes com o Info Gain utilizando o J48, o PART e o Random Forest, fazendo um balanceamento nos dados com o Resample, obteve-se os seguintes resultados (Tabela 4.17):

Info Gain utilizando Conjunto teste de 10%					
classificador	acurácia	kappa	precisão da soja	precisão do milho	precisão do algodão
Resample=0,0					
J48	97,260	0,737	0,986	0,000	0,750
Random Forest	95,890	0,553	0,971	0,000	0,667
PART	94,521	0,477	0,971	0,000	0,667
Resample=0,5					
J48	91,781	0,467	0,985	0,000	0,429
Random Forest	97,260	0,737	0,986	0,000	0,750
PART	97,260	0,788	1,000	1,000	0,600
Resample=1,0					
J48	86,301	0,328	0,870	0,000	0,273
Random Forest	95,890	0,648	0,985	0,000	0,600
PART	93,161	0,471	0,985	0,000	0,400

Tabela 4.17: Testes utilizando o Info Gain e o Resample

Analogamente, testando o Info Gain com o J48, o PART e o Random Forest com o Cross Validation, obtêve-se os seguintes resultados (Tabela 4.18):

Info Gain utilizando o Cross Validation					
classificador	acurácia	kappa	precisão da soja	precisão do milho	precisão do algodão
J48	95,604	0,616	0,972	0,000	0,778
Random Forest	95,604	0,580	0,967	0,000	0,727
PART	96,017	0,633	0,973	0,000	0,813

Tabela 4.18: Testes utilizando o Info Gain e o Cross Validation

Observando as Tabelas 4.17 e 4.18, conclui-se que os melhores valores para a geração do modelo será dado por Tabela 4.19:

Analise do Modelo Info Gain utilizando Resample R=0,5 e conjunto teste de 10%					
algoritmo	acurácia	kappa	precisão da soja	precisão do milho	precisão do algodão
PART	97,260	0,788	1,000	1,000	0,600

Tabela 4.19: Resultado do teste utilizando o Info Gain e o Resample

Pois valor do *Kappa* é igual a $78,80\%$, a *acurácia* com valor igual a $97,26\%$, *precisão* da soja e do milho em 100%, e a *precisão* do algodão em 60%.

Nesse modelo foi aplicado o classificador PART em conjunto com o filtro Resample com um balanceamento R=$0,5$.

A Tabela 4.20 ilustra o conjunto de regras geradas por esse modelo:

Regras	Info Gain utilizando Resample R=0,5 e conjunto teste de 10%
R1	Se 20090101 > 0,8495 e 20081202 <= 0,986 e 20090218 > 0,3052 e 20090509 > 0,6953 e 20090101 <= 0,9542 então soja (255)
R2	Se 200900424 > 0,8039 e 20081202 <= 0,7814 e 20090525 <= 0,853 e 20090525 > 0,652 então algodão (88)
R3	Se 20081218 > 0,8086 e 20090423 <= 0,8039 então soja (67)
R4	Se 20090407 > 0,7757 e 20090509 > 0,82 então soja (28)
R5	Se 20090407 > 0,7757 e 20081202 <= 0,5351 então algodão (17)
R6	Se 20090101 <= 0,8599 e 20090306 > 0,6535 e 20081031 <= 0,6705 e 20090218 <= 0,8646 e 20090306 <= 0,8013 e 20081116 <= 0,571 então milho (104/1)
R7	Se 20090306 <= 0,7888 e 20090423 <= 0,8039 então soja (35)
R8	Se 20081202 <= 0,5526 e 20081031 > 0,502 então milho (12)
R9	Se 20090322 <= 0,7906 e 20090423 <= 0,8538 e 20090509 > 0,6142 e 20090101 <= 0,9584 então algodão (24)
R10	Se 20090322 <= 0,7906 e 20090423 <= 0,8538 e 20090509 > 0,6142 e 20090101 <= 0,9584 então soja (25)

Tabela 4.20: Conjunto de regras gerado pelo Info Gain com o Resample

A regra R4 também foi retirada do conjunto de regras, pois o valor de NDVI igual a $0,82$ em Maio não pode ser maior que seu valor em Abril que era de $0,775$, pois o pico da soja é dado em Janeiro, logo nesse período o valor de NDVI tem que estar necessariamente decrescendo.

Com as matrizes de contingência calculou-se a *acurácia*, o *erro*, a *sensitividade*, a *especificidade*, a *cobertura*, o *suporte* e a *novidade* de cada regra (Tabela 4.21):

Analise das regras							
regras	acurácia	erro	sensitividade	especificidade	cobertura	suporte	novidade
R1	1,000	0,000	0,622	1,000	0,389	0,389	0,146
R2	1,000	0,000	0,682	1,000	0,134	0,134	0,108
R3	1,000	0,000	0,163	1,000	0,102	0,102	0,038
R5	1,000	0,000	0,132	1,000	0,026	0,026	0,021
R6	0,990	0,010	0,897	0,998	0,160	0,159	0,130
R7	1,000	0,000	0,085	1,000	0,053	0,053	0,020
R8	1,000	0,000	0,103	1,000	0,018	0,018	0,015
R9	1,000	0,000	0,186	1,000	0,037	0,037	0,029
R10	1,000	0,000	0,061	1,000	0,038	0,038	0,014

Tabela 4.21: Análise das regras geradas pelo Info Gain e o Resample

De acordo com os valores da Tabela acima, pode-se verificar que R1, R2 e R6 possuem os maiores valores para: a *acurácia*, a *sensitividade* e a *especificidade*. Logos as melhores regras geradas por esse modelo são: R1, R2 e R6 .

4.1.5 Testes com o Wrapper

Para encerrar essa abordagem, foram realizados testes utilizando o selecionador de atributos Wrapper. Os resultados dos testes gerados com o J48, o PART e o Random Forest, com o Resample foram (Tabela 4.22):

Wrapper utilizando o Resample					
classificador	acurácia	kappa	precisão da soja	precisão do milho	precisão do algodão
J48					
Resample=0,0	97,260	0,656	0,972	0,000	1,000
Resample=0,5	95,890	0,656	0,971	0,000	0,667
Resample=1,0	89,041	0,291	0,969	0,000	0,333
Random Forest					
Resample=0,0	98,630	0,883	1,000	1,000	0,750
Resample=0,5	95,890	0,553	0,971	0,000	0,667
Resample=1,0	94,521	0,475	0,971	0,000	0,500
PART					
Resample=0,0	95,890	0,555	0,971	0,000	1,000
Resample=0,5	93,151	0,413	0,971	0,000	0,400
Resample=1,0	91,781	0,365	0,970	0,000	0,400

Tabela 4.22: Testes utilizando o Wrapper com o Resample

Da mesma forma, testando o J48, PART e o Random Forest com o o Wrapper e o Cross Validation, tem-se (Tabela 4.23):

Wrapper utilizando o Cross Validation					
classificador	acurácia	kappa	precisão da soja	precisão do milho	precisão do algodão
J48	96,153	0,624	0,968	0,000	0,862
Random Forest	96,841	0,698	0,974	0,000	0,875
PART	96,566	0,664	0,971	0,000	0,839

Tabela 4.23: Testes utilizando o Wrapper com o Cross Validation

De acordo com as Tabelas: 4.22 e 4.23, o classificador Random Forest apresentou os melhores resultados. Será feita mais algumas investigações com ele, para verificar se realmente estes são os melhores valores para o modelo.

O Random Forest classifica as regras gerando uma floresta de árvores com as respectivas regras. Esse número de árvores pode ser aleatório ou determinístico. Pode-se também olhar para o número de observações que passam por cada folha da árvore, ou seja, realizar a poda nas árvores. Será dado um enfoque a essas duas características importantes do Random Forest para se conseguir bons valores para o *Kappa*, a *acurácia*, e *as precisões* por classe.

Variando o valor do *MaxDepth* no Random Forest, está sendo realizada uma variação no número de observações que passam por cada folha da árvore, com isso obtêm-se a redução do tamanho das árvores (poda nas árvores), consequentemente há uma redução na quantidade de regras geradas pelas mesmas, e com isso é gerado um aumento nos valores do *Kappa* e da *acurácia* do modelo.

A Tabela 4.24 a seguir, mostra a variação do *MaxDepth* com a precisão de cada classe:

Poda das árvores					
max depth	acurácia	kappa	precisão da soja	precisão do milho	precisão do algodão
2	97,260	0,737	0,986	0,000	0,750
3	97,260	0,737	0,986	0,000	0,750
5	98,630	0,883	1,000	1,000	0,750
6	98,630	0,737	0,986	0,000	0,750

Tabela 4.24: Poda das árvores

De acordo com resultados mostrados na Tabela 4.24, pode-se verificar bons valores para as classes, utilizando o *MaxDepth*=5. Por esse motivo, foi feita a poda nas árvores, porém não houve

a necessidade de alterar o número de árvores geradas pelo algoritmo, fazendo uso do *default* do software, que é quantidade de árvores = 10, pois os valores das classes soja, milho e algodão já estavam dentro dos valores desejado.

Então, de acordo com os resultados obtidos nas Tabelas: 4.22, 4.23 e 4.24, pode-se concluir que os melhores valores para a gerção do modelo é (Tabela 4.25):

Analise do Modelo Wrapper utilizando Resample com R=0 e o Random Forest					
algoritmo	acurácia	kappa	precisão da soja	precisão do milho	precisão do algodão
Random Forest	98,630	0,883	1,000	1,000	0,750

Tabela 4.25: Teste utilizando o Wrapper e o Random Forest

Onde se encontra um ótimo valor do *Kappa* igual a $88,3\%$, uma excelente *acurácia* com valor igual a $98,630\%$, precisão da soja e do milho em 100%, e a precisão do algodão em 75%.

Nesse modelo foi feita uma seleção nos atributos com o Wrappper, fazendo o balanceamento nos dados utilizando o filtro *Resample* com o coeficiente de balanceamento R=$0,0$. Esse modelo gerou um conjunto de 10 árvores.

Agora, será explorado o conjunto de regras de cada árvore da floresta gerada pelo Random Forest .

Segue o conjunto de regras (Tabela 4.26) gerado pela 1^o árvore:

Regras da árvore 1	Wrapper utilizando Resample com R=0 e o Random Forest
R1	Se 20090101 < 0,71 e 20081202 < 0,8 e 20090322 < 0,81 e 20090101 < 0,69 e 20081202 < 0,32 então soja (4/1)
R2	Se 20090101 < 0,71 e 20081202 < 0,8 e 20090322 < 0,81 e 20090101 < 0,69 e 20081202 >= 0,32 então soja (11)
R3	Se 20090101 < 0,71 e 20081202 < 0,8 e 20090322 < 0,81 e 20090101 >= 0,69 e 20090306 < 0,62 então soja (3)
R4	Se 20090101 < 0,71 e 20081202 < 0,8 e 20090322 < 0,81 e 20090101 >= 0,69 e 20090306 >= 0,62 então milho (6)
R5	Se 20090101 < 0,71 e 20081202 < 0,8 e 20090322 >= 0,81 e 20080913 < 0,62 e 20090101 < 0,21 então soja (1)
R6	Se 20090101 < 0,71 e 20081202 < 0,8 e 20090322 >= 0,81 e 20080913 < 0,62 e 20090101 >= 0,21 então algodão (26/1)
R7	Se 20090101 < 0,71 e 20081202 < 0,8 e 20090322 >= 0,81 e 20080913 >= 0,62 então soja (1)
R8	Se 20090101 < 0,71 e 2001202 >= 0,8 então soja (27)
R9	Se 20090101 >= 0,71 e 20080913 < 0,2 e 20090322 < 0,78 então soja (6)
R10	Se 20090101 >= 0,71 e 20080913 < 0,2 e 20090322 >= 0,78 e 20090306 < 0,54 então soja (1)
R11	Se 20090101 >= 0,71 e 20080913 < 0,2 e 20090322 >= 0,78 e 20090306 >= 0,54 então algodão (5)
R12	Se 20090101 >= 0,71 e 20080913 < 0,2 e 20090101 >= 0,93 então soja (17)
R13	Se 20090101 >= 0,71 e 20080913 >= 0,2 e 20081202 < 0,99 e 20090306 < 0,8 e 20080913 < 0,41 então soja (497)
R14	Se 20090101 >= 0,71 e 20080913 >= 0,2 e 20081202 < 0,99 e 20090306 < 0,8 e 20080913 >= 0,41 então soja (25/1)
R15	Se 20090101 >= 0,71 e 20080913 >= 0,2 e 20081202 < 0,99 e 20090306 >= 0,8 e 20090306 < 0,81 então milho (1)
R16	Se 20090101 >= 0,71 e 20080913 >= 0,2 e 20081202 < 0,99 e 20090306 >= 0,8 e 20090306 >= 0,81 então soja(22)
R17	Se 20090101 >= 0,71 e 20080913 >= 0,2 e 20081202 >= 0,99 então algodão (2)

Tabela 4.26: Conjunto de regras geradas pela 1^0 árvore do modelo

Calculando a *acurácia*, o *erro*, a *sensitividade*, a *especificidade*, a *cobertura*, o *suporte* e a *novidade* da 1^o árvore:

Analise das regras da árvore 1							
regras	acurácia	erro	sensitividade	especificidade	cobertura	suporte	novidade
R1	0,800	0,200	0,007	0,975	0,008	0,006	-0,001
R2	1,000	0,000	0,018	1,000	0,017	0,017	0,001
R3	1,000	0,000	0,005	1,000	0,005	0,005	0,000
R4	1,000	0,000	0,857	1,000	0,009	0,009	0,009
R5	1,000	0,000	0,002	1,000	0,002	0,002	0,000
R6	0,963	0,037	0,788	0,998	0,041	0,040	0,038
R7	1,000	0,000	0,002	1,000	0,002	0,002	0,000
R8	1,000	0,000	0,044	1,000	0,041	0,041	0,003
R9	1,000	0,000	0,010	1,000	0,009	0,009	0,001
R10	1,000	0,000	0,002	1,000	0,002	0,002	0,000
R11	1,000	0,000	0,152	1,000	0,008	0,008	0,007
R12	1,000	0,000	0,028	1,000	0,026	0,026	0,002
R13	1,000	0,000	0,808	1,000	0,759	0,759	0,046
R14	0,962	0,038	0,041	0,975	0,040	0,038	0,001
R15	1,000	0,000	0,143	1,000	0,002	0,002	0,002
R16	1,000	0,000	0,036	1,000	0,034	0,034	0,002
R17	1,000	0,000	0,061	1,000	0,003	0,003	0,003

Tabela 4.27: Análise das regras geradas pela 1^o árvore do modelo

De acordo com os resultados obtidos na Tabela 4.27 acima, pode-se verificar, que R4, R6 e R13 possuem altos valores da *acurácia*, *sensitividade*, e *especificidade*. Logo, essas são as melhores regras geradas pela 1^o árvore.

Segue o segundo conjunto de regras (Tabela 4.28) geradas pela 2^o árvore:

Regras da árvore 2	Wrapper utilizando Resample com R=0 e o Random Forest
R1	Se 20090101 < 0,71 e 20090101 < 0,7 e 20081202 < 0,41 e 20080913 < 0,29 então algodão (16)
R2	Se 20090101 < 0,71 e 20090101 < 0,7 e 20081202 < 0,41 e 20080913 >= 0,29 então milho (1)
R3	Se 20090101 < 0,71 e 20090101 < 0,7 e 20081202 >= 0,41 e 20090306 < 0,72 então soja (21)
R4	Se 20090101 < 0,71 e 20090101 < 0,7 e 20081202 >= 0,41 e 20090306 >= 0,72 e 20081202 < 0,81 então algodão (18/9)
R5	Se 20090101 < 0,71 e 20090101 < 0,7 e 20081202 >= 0,41 e 20090306 >= 0,72 e 20081202 >= 0,81 então soja (13)
R6	Se 20090101 < 0,71 e 20090101 >= 0,7 então milho (7)
R7	Se 20090101 >= 0,71 e 20080913 < 0,2 e 20080913 < 0,19 então soja (26)
R8	Se 20090101 >= 0,71 e 20080913 < 0,2 e 20080913 >= 0,19 então algodão (6)
R9	Se 20090101 >= 0,71 e 20080913 >= 0,2 e 20081202 < 0,99 então soja (546)
R10	Se 20090101 >= 0,71 e 20080913 >= 0,2 e 20081202 >= 0,99 então algodão (1)

)

Tabela 4.28: Conjunto de regras geradas pela 2^o árvore do modelo

Calculando a *acurácia*, o *erro*, a *sensitividade*, a *especificidade*, a *cobertura*, o *suporte* e a *novidade* da 2^o árvore:

Analise das regras da árvore 2							
regras	acurácia	erro	sensitividade	especificidade	cobertura	suporte	novidade
R1	1,000	0,000	0,390	1,000	0,024	0,024	0,023
R2	1,000	0,000	0,125	1,000	0,002	0,002	0,002
R3	1,000	0,000	0,035	1,000	0,032	0,032	0,002
R4	0,667	0,333	0,439	0,985	0,041	0,027	0,025
R5	1,000	0,000	0,021	1,000	0,020	0,020	0,001
R6	1,000	0,000	0,875	1,000	0,011	0,011	0,011
R7	1,000	0,000	0,043	1,000	0,040	0,040	0,003
R8	1,000	0,000	0,146	1,000	0,009	0,009	0,009
R9	1,000	0,000	0,901	1,000	0,834	0,834	0,062
R10	1,000	0,000	0,024	1,000	0,002	0,002	0,001

Tabela 4.29: Análise das regras geradas pela 2^o árvore do modelo

De acordo com os resultados da Tabela 4.29, pode-se verificar, que as regras R6 e R9 são as

regras que possuem os mais altos valores da *acurácia*, a *sensitividade*, e a *especificidade*. Logo, essas são as melhores regras geradas pela 2^o árvore.

Segue o terceiro conjunto de regras (Tabela 4.30) geradas pela 3^o árvore:

Regras da árvore 3	Wrapper utilizando Resample com R=0,0 e o Random Forest
R1	Se 20090101 < 0,71 e 20081202 < 0,62 e 20090322 < 0,8 e 20081202 < 0,48 e 20081202 < 0,36 então milho (3/1)
R2	Se 20090101 < 0,71 e 20081202 < 0,62 e 20090322 < 0,8 e 20081202 < 0,48 e 20081202 >= 0,36 então soja (5)
R3	Se 20090101 < 0,71 e 20081202 < 0,62 e 20090322 < 0,8 e 20081202 >= 0,48 então milho (11)
R4	Se 20090101 < 0,71 e 20081202 < 0,62 e 20090322 >= 0,8 então algodão (25)
R5	Se 20090101 < 0,71 e 20081202 >= 0,62 e 20081202 < 0,81 e 20080913 < 0,44 então soja (8)
R6	Se 20090101 < 0,71 e 20081202 >= 0,62 e 20081202 < 0,81 e 20080913 >= 0,44 então algodão (3)
R7	Se 20090101 < 0,71 e 20081202 >= 0,62 e 20081202 >= 0,81então soja (26)
R8	Se 20090101 >= 0,71 e 20090322 < 0,7 então soja (311)
R9	Se 20090101 >= 0,71 e 20090322 >= 0,7 e 20090101 < 0,9 e 20080913 < 0,2 então algodão (3)
R10	Se 20090101 >= 0,71 e 20090322 >= 0,7 e 20090101 < 0,9 e 20080913 >= 0,2 e 20081202 < 0,62 então milho (1)
R11	Se 20090101 >= 0,71 e 20090322 >= 0,7 e 20090101< 0,9 e 20080913 >= 0,2 e 20081202 >= 0,62 então soja (64/2)
R12	Se 20090101 >= 0,71 e 20090322 >= 0,7 20090101 >= 0,9 e 20090322 < 0,71 então algodão (1)
R13	Se 20090101 >= 0,71 e 20090322 >= 0,7 20090101 >= 0,9 e 20090323 >= 0,71 então soja (194)

Tabela 4.30: Conjunto de regras geradas pela 3^o árvore do modelo

Calculando a *acurácia*, o *erro*, a *sensitividade*, a *especificidade*, a *cobertura*, o *suporte* e a *novidade* da 3^o árvore:

Analise das regras da árvore 3							
regras	acurácia	erro	sensitividade	especificidade	cobertura	suporte	novidade
R1	0,750	0,250	0,200	0,998	0,006	0,005	0,004
R2	1,000	0,000	0,008	1,000	0,008	0,008	0,001
R3	1,000	0,000	0,733	1,000	0,017	0,017	0,016
R4	1,000	0,000	0,781	1,000	0,038	0,038	0,036
R5	1,000	0,000	0,013	1,000	0,012	0,012	0,001
R6	1,000	0,000	0,094	1,000	0,005	0,005	0,004
R7	1,000	0,000	0,043	1,000	0,040	0,040	0,003
R8	1,000	0,000	0,512	1,000	0,475	0,475	0,034
R9	1,000	0,000	0,094	1,000	0,005	0,005	0,004
R10	1,000	0,000	0,067	1,000	0,002	0,002	0,001
R11	0,970	0,030	0,105	0,957	0,101	0,098	0,004
R12	1,000	0,000	0,031	1,000	0,002	0,002	0,001
R13	1,000	0,000	0,319	1,000	0,296	0,296	0,021

Tabela 4.31: Análise das regras geradas pela 3^o árvore do modelo

De acordo com os resultados da Tabela 4.31, pode-se verificar que as regras R3, R4, R8 e R13 são as regras que possuem os mais altos valores da *acurácia*, a *sensitividade*, e a *especificidade*. Logo, essas são as melhores regras geradas pela 3^o árvore do modelo.

Segue o quarto conjunto de regras (Tabela 4.32) geradas pela 4^o árvore:

Regras da árvore 4	Wrapper utilizando Resample com R=0,0 e o Random Forest
R1	Se 20090101 < 0,54 e 20081202 < 0,63 e 20080913 < 0,29 e 20090306 < 0,57 então soja (1)
R2	Se 20090101 < 0,54 e 20081202 < 0,63 e 20080913 < 0,29 e 20090306 >= 0,57 então algodão (19)
R3	Se 20090101 < 0,54 e 20081202 < 0,63 e 20080913 >= 0,29 e 20090101 < 0,51 e 20081202 < 0,36 então milho (3)
R4	Se 20090101 < 0,54 e 20081202 < 0,63 e 20080913 >= 0,29 e 20090101 < 0,51 e 20081202 >= 0,36 então soja (2)
R5	Se 20090101 < 0,54 e 20081202 < 0,63 e 20080913 >= 0,29 e 20090101 >= 0,51 então soja (3)
R6	Se 20090101 < 0,54 e 20081202 >= 0,63 então soja (13)
R7	Se 20090101 >= 0,54 e 20090306 < 0,74 e 20090322 < 0,7 então soja (314)
R8	Se 20090101 >= 0,54 e 20090306 < 0,74 e 20090322 >= 0,7 e 20080913 < 0,2 e 20090306 < 0,54 então soja (13)
R9	Se 20090101 >= 0,54 e 20090306 < 0,74 e 20090322 >= 0,7 e 20080913 < 0,2 e 20090306 >= 0,54 então algodão (9/3)
R10	Se 20090101 >= 0,54 e 20090306 < 0,74 e 20090322 >= 0,7 e 20080913 >= 0,2 e 20090322 < 0,71 então algodão (2)
R11	Se 20090101 >= 0,54 e 20090306 < 0,74 e 20090322 >= 0,7 e 20080912 >= 0,2 e 20090322 >= 0,2 e 20090322 < 0,71 então soja (200/3)
R12	Se 20090101 >= 0,54 e 20090306 >= 0,74 e 20081202 < 0,67 e 20080913 < 0,25 então soja (10)
R13	Se 20090101 >= 0,54 e 20090306 >= 0,74 e 20081202 < 0,67 e 20080913 >= 0,25 e 20090101 < 0,81 então milho (6)
R14	Se 20090101 >= 0,54 e 20090306 >= 0,74 e 20081202 < 0,67 e 20080913 >= 0,25 e 20090101 >= 0,81 então soja (2)
R15	Se 20090101 >= 0,54 e 20090306 >= 0,74 e 20081202 >= 0,67 e 20081202 < 0,77 e 20081202 < 0,75 então soja (7)
R16	Se 20090101 >= 0,54 e 20090306 >= 0,74 e 20081202 >= 0,67 e 20081202 < 0,77 e 20081202 >= 0,75 então algodão (2)
R17	Se 20090101 >= 0,54 e 20090306 >= 0,74 e 20081202 >= 0,67 e 20081202 >= 0,77 então soja (49)

Tabela 4.32: Conjunto de regras geradas pela 4º árvore do modelo

Calculando a *acurácia*, o *erro*, a *sensitividade*, a *especificidade*, a *cobertura*, o *suporte* e a *novidade* da 4º árvore:

Analise das regras da árvore 4							
regras	acurácia	erro	sensitividade	especificidade	cobertura	suporte	novidade
R1	1,000	0,000	0,002	1,000	0,002	0,002	0,000
R2	1,000	0,000	0,594	1,000	0,029	0,029	0,028
R3	1,000	0,000	0,333	1,000	0,005	0,005	0,005
R4	1,000	0,000	0,003	1,000	0,003	0,003	0,000
R5	1,000	0,000	0,005	1,000	0,005	0,005	0,000
R6	1,000	0,000	0,021	1,000	0,020	0,020	0,001
R7	1,000	0,000	0,511	1,000	0,479	0,479	0,030
R8	1,000	0,000	0,021	1,000	0,020	0,020	0,001
R9	0,750	0,250	0,281	0,995	0,018	0,014	0,013
R10	1,000	0,000	0,063	1,000	0,003	0,003	0,003
R11	0,985	0,015	0,326	0,927	0,310	0,305	0,015
R12	1,000	0,000	0,016	1,000	0,015	0,015	0,001
R13	1,000	0,000	0,667	1,000	0,009	0,009	0,009
R14	1,000	0,000	0,003	1,000	0,003	0,003	0,000
R15	1,000	0,000	0,011	1,000	0,011	0,011	0,001
R16	1,000	0,000	0,063	1,000	0,003	0,003	0,003
R17	1,000	0,000	0,080	1,000	0,075	0,075	0,005

Tabela 4.33: Análise das regras geradas pela 4^o árvore do modelo

De acordo com os resultados da Tabela 4.33, pode-se verificar que as regras R2, R7, R11 e R13 são as regras que possuem os mais altos valores da *acurácia*, a *sensitividade*, e a *especificidade*. Logo, essas são as melhores regras geradas pela 4^o árvore do modelo.

Segue o quinto conjunto de regras (Tabela 4.34) geradas pela 5^o árvore:

Regras da árvore 5	Wrapper utilizando Resample com R=0,0 e o Random Forest
R1	Se 20090101 < 0,71 e 20081202 < 0,63 e 20090322 < 0,59 então soja (6)
R2	Se 20090101 < 0,71 e 20081202 < 0,63 e 20090322 >= 0,59 e 20080913 < 0,25 então soja (3)
R3	Se 20090101 < 0,71 e 20081202 < 0,63 e 20090322 >= 0,59 e 20080913 >= 0,25 então milho (13)
R4	Se 20090101 < 0,71 e 20081202 < 0,63 e 20090322 >= 0,80 e 20080913 <0,51 e 20080913 < 0,22 então algodão (12/1)
R5	Se 20090101 < 0,71 e 20081202 < 0,63 e 20090322 >= 0,80 e 20080913 <0,51 e 20080913 >= 0,22 então algodão (16)
R6	Se 20090101 < 0,71 e 20081202 < 0,63 e 20090322 >= 0,80 e 20080913 >=0,51 então soja (1)
R7	Se 20090101 < 0,71 e 20081202 >= 0,63 e 20080913 < 0,46 então soja (29)
R8	Se 20090101 < 0,71 e 20081202 >= 0,63 e 20080913 >= 0,46 e 20090101 < 0,59 então soja (3)
R9	Se 20090101 < 0,71 e 20081202 >= 0,63 e 20080913 >= 0,46 e 20090101 >= 0,59 então algodão (1)
R10	Se 20090101 >= 0,71 e 20080913 < 0,67 e 20080913 < 0,2 e 20090322 < 0,76 então soja (17)
R11	Se 20090101 >= 0,71 e 20080913 < 0,67 e 20080913 < 0,2 e 20090322 >= 0,76 e 20090101 < 0,93 então algodão (5/2)
R12	Se 20090101 >= 0,71 e 20080913 < 0,67 e 20080913 < 0,2 e 20090322 >= 0,76 e 20090101 >= 0,93 então soja (9)
R13	Se 20090101 >= 0,71 e 20080913 < 0,67 e 20080913 >= 0,2 e 20081202 < 0,99 e 20090101 < 0,88 então soja (127/2)
R14	Se 20090101 >= 0,71 e 20080913 < 0,67 e 20080913 >= 0,2 e 20081202 < 0,99 e 20090101 >= 0,88 então soja (409)
R15	Se 20090101 >= 0,71 e 20080913 < 0,67 e 20080913 >= 0,2 e 20081202 >= 0,99 então algodão (1)
R16	Se 20090101 >= 0,71 e 20080913 >= 0,67 e 20080913 < 0,77 então milho (2)
R17	Se 20090101 >= 0,71 e 20080913 >= 0,67 e 20080913 >= 0,77 então soja (1)

Tabela 4.34: Conjunto de regras geradas pela 5^0 árvore do modelo

Calculando a *acurácia*, o *erro*, a *sensitividade*, a *especificidade*, a *cobertura*, o *suporte* e a *novidade* da 5^o árvore:

Analise das regras da árvore 5							
regras	acurácia	erro	sensitividade	especificidade	cobertura	suporte	novidade
R1	1,000	0,000	0,010	1,000	0,009	0,009	0,001
R2	1,000	0,000	0,005	1,000	0,005	0,005	0,000
R3	1,000	0,000	0,867	1,000	0,020	0,020	0,019
R4	0,923	0,077	0,343	0,998	0,020	0,018	0,017
R5	1,000	0,000	0,457	1,000	0,024	0,024	0,023
R6	1,000	0,000	0,002	1,000	0,002	0,002	0,000
R7	1,000	0,000	0,048	1,000	0,044	0,044	0,003
R8	1,000	0,000	0,005	1,000	0,005	0,005	0,000
R9	1,000	0,000	0,029	1,000	0,002	0,002	0,001
R10	1,000	0,000	0,028	1,000	0,026	0,026	0,002
R11	0,714	0,286	0,143	0,997	0,011	0,008	0,007
R12	1,000	0,000	0,015	1,000	0,014	0,014	0,001
R13	0,984	0,016	0,210	0,960	0,197	0,194	0,012
R14	1,000	0,000	0,676	1,000	0,624	0,624	0,048
R15	1,000	0,000	0,029	1,000	0,002	0,002	0,001
R16	1,000	0,000	0,133	1,000	0,003	0,003	0,003
R17	1,000	0,000	0,002	1,000	0,002	0,002	0,000

Tabela 4.35: Análise das regras geradas pela 5^o árvore do modelo

De acordo com os resultados da Tabela 4.35, pode-se verificar que as regras R3, R5, e R14 são as regras que possuem os mais altos valores da *acurácia*, *sensitividade*, e a *especificidade*. Logos essas são as melhores regras geradas pela 5^o árvore do modelo.

Segue o sexto conjunto de regras (Tabela 4.36) gerada pela 6^o árvore:

Regras da árvore 6	Wrapper utilizando Resample com R=0,0 e o Random Forest
R1	Se 20090309 < 0,73 e 20090322 < 0,75 então soja (361)
R2	Se 20090309 < 0,73 e 20090322 >= 0,75 e 20090322 < 0,78 e 20090322 < 0,78 e 20080913 < 0,41 então soja (18)
R3	Se 20090309 < 0,73 e 20090322 >= 0,75 e 20090322 < 0,78 e 20090322 < 0,78 e 20080913 >= 0,41 então soja (3/1)
R4	Se 20090309 < 0,73 e 20090322 >= 0,75 e 20090322 < 0,78 e 20090322 >= 0,78 então algodão (3)
R5	Se 20090309 < 0,73 e 20090322 >= 0,75 e 20090322 >= 0,78 então soja (158)
R6	Se 20090309 >= 0,73 e 20081202 < 0,77 e 20090322 < 0,81 e 20081202 < 0,48 então soja (7)
R7	Se 20090101 >= 0,73 e 20081202 < 0,77 e 20090322 < 0,81 e 20081202 >= 0,48 e 20081202 < 0,62 então milho (11)
R8	Se 20090101 >= 0,73 e 20081202 < 0,77 e 20090322 < 0,81 e 20081202 >= 0,48 e 20081202 >= 0,62 então soja (7)
R9	Se 20090101 >= 0,73 e 20081202 < 0,77 e 20090322 >= 0,81 e 20080913 < 0,26 e 20081202 < 0,44 então algodão (8)
R10	Se 20090101 >= 0,73 e 20081202 < 0,77 e 20090322 >= 0,81 e 20080913 < 0,26 e 20081202 >= 0,44 então soja (4/2)
R11	Se 20090306 >= 0,73 e 20081202 < 0,77 e 20090322 >= 0,81 e 20080913 >= 0,26 então soja (13)
R12	Se 20090306 >= 0,73 e 20081202 >= 0,77 então soja (62)

Tabela 4.36: Conjunto de regras geradas pela 6^0 árvore do modelo

De acordo com a Tabela 4.36, pode-se verificar que a regra R4 é redundante, pois tem-se 20090322< 0,78 e 20090322>=0,78. Logo essa regra será excluída do conjunto de regras.

Calculando a *acurácia*, o *erro*, a *sensitividade*, a *especificidade*, a *cobertura*, o *suporte* e a *novidade* da 6^o árvore:

Analise das regras da árvore 6							
regras	acurácia	erro	sensitividade	especificidade	cobertura	suporte	novidade
R1	1,000	0,000	0,582	1,000	0,551	0,551	0,029
R2	1,000	0,000	0,029	1,000	0,027	0,027	0,001
R3	0,750	0,250	0,005	0,971	0,006	0,005	-0,001
R4	1,000	0,000	0,125	1,000	0,005	0,005	0,004
R5	1,000	0,000	0,255	1,000	0,241	0,241	0,013
R6	1,000	0,000	0,011	1,000	0,011	0,011	0,001
R7	1,000	0,000	1,000	1,000	0,017	0,017	0,017
R8	1,000	0,000	0,011	1,000	0,011	0,011	0,001
R9	1,000	0,000	0,333	1,000	0,012	0,012	0,012
R10	0,667	0,333	0,006	0,943	0,009	0,006	-0,003
R11	1,000	0,000	0,542	1,000	0,020	0,020	0,019
R12	1,000	0,000	0,100	1,000	0,095	0,095	0,005

Tabela 4.37: Análise das regras geradas pela 6^o árvore do modelo

De acordo com os resultados da Tabela 4.37, pode-se verificar que as regras R1, R7, R9 e R11 são as regras que possuem os mais altos valores da *acurácia*, *sensitividade*, e *especificidade*. Logo essas são as melhores regras geradas pela 6^o árvore do modelo.

Segue o sétimo conjunto de regras (Tabela 4.38) gerada pela 7^o árvore:

Regras da árvore 7	Wrapper utilizando Resample com R=0,0 e o Random Forest
R1	Se 20090101 < 0,71 e 20081202 < 0,63 e 20090322 < 0,8 e 20080913 < 0,26 então soja (7)
R2	Se 20090101 < 0,71 e 20081202 < 0,63 e 20090322 < 0,8 e 20080913 >= 0,26 e 20080322 < 0,59 então soja (1)
R3	Se 20090101 < 0,71 e 20081202 < 0,63 e 20090322 < 0,8 e 20080913 >= 0,26 e 20080322 >= 0,59 então milho (7)
R4	Se 20090101 < 0,71 e 20081202 < 0,63 e 20090322 >= 0,8 então algodão (17)
R5	Se 20090101 < 0,71 e 20081202 > 0,63 e 20080913 <0.44 então soja (37)
R6	Se 20090101 < 0,71 e 20081202 > 0,63 e 20080913 >=0.44 e 20080913 < 0,5 então algodão (1)
R7	Se 20090101 < 0,71 e 20081202 > 0,63 e 20080913 >=0.44 e 20080913 >= 0,5 então soja (1)
R8	Se 20090101 >= 0,71 e 20081202 <0,5 e 20080913 <0,58 então soja (23)
R9	Se 20090101 >= 0,71 e 20081202 <0,5 e 20080913 >=0,58 então milho (1)
R10	Se 20090101 >= 0,71 e 20081202 >= 0,51 e 20090306 < 0,68 então soja (443)
R11	Se 20090101 >= 0,71 e 20081202 >= 0,51 e 20090306 >= 0,68 e 20080913 < 0,19 e 20090101 <0,93 então algodão (1)
R12	Se 20090101 >= 0,71 e 20081202 >= 0,51 e 20090306 >= 0,68 e 20080913 < 0,19 e 20090101 >= 0,93 então soja (1)
R13	Se 20090101 >= 0,71 e 20081202 >= 0,51 e 20090306 >= 0,68 e 20080913 >= 0,19 então soja (115)

Tabela 4.38: Conjunto de regras geradas pela 7^0 árvore do modelo

Calculando a *acurácia*, o *erro*, a *sensitividade*, a *especificidade*, a *cobertura*, o *suporte* e a *novidade* da 7^o árvore:

Analise das regras da árvore 7							
regras	acurácia	erro	sensitividade	especificidade	cobertura	suporte	novidade
R1	1,000	0,000	0,011	1,000	0,011	0,011	0,000
R2	1,000	0,000	0,002	1,000	0,002	0,002	0,000
R3	1,000	0,000	0,875	1,000	0,011	0,011	0,011
R4	1,000	0,000	0,895	1,000	0,026	0,026	0,025
R5	1,000	0,000	0,059	1,000	0,056	0,056	0,002
R6	1,000	0,000	0,053	1,000	0,002	0,002	0,001
R7	1,000	0,000	0,002	1,000	0,002	0,002	0,000
R8	1,000	0,000	0,037	1,000	0,035	0,035	0,001
R9	1,000	0,000	0,125	1,000	0,002	0,002	0,002
R10	1,000	0,000	0,705	1,000	0,676	0,676	0,028
R11	1,000	0,000	0,053	1,000	0,002	0,002	0,001
R12	1,000	0,000	0,002	1,000	0,002	0,002	0,000
R13	1,000	0,000	0,183	1,000	0,176	0,176	0,007

Tabela 4.39: Análise das regras geradas pela 7^o árvore do modelo

De acordo com os resultados da Tabela 4.39, pode-se verificar que as regras R3, R4, e R10 são as regras que possuem os mais altos valores da *acurácia*, *sensitividade*, e *especificidade*. Logo essas são as melhores regras geradas pela 7^o árvore do modelo.

Segue o oitavo conjunto de regras (Tabela 4.40) gerada pela 8^o árvore:

Regras da árvore 8	Wrapper utilizando Resample com R=0,0 e o Random Forest
R1	Se 20090101 < 0,71 e 20081202 < 0,63 e 20090306 < 0,85 e 20090101 < 0,68 e 20090322 < 0,81 então soja (9/2)
R2	Se 20090101 < 0,71 e 20081202 < 0,63 e 20090306 < 0,85 e 20090101 < 0,68 e 20090322 >= 0,81 então soja (3/1)
R3	Se 20090101 < 0,71 e 20081202 < 0,63 e 20090306 < 0,85 e 20090101 >= 0,68 então milho (9)
R4	Se 20090101 < 0,71 e 20081202 < 0,63 e 20090306 >= 0,85 e 20090101 < 0,21 então soja (2)
R5	Se 20090101 < 0,71 e 20081202 < 0,63 e 20090306 >= 0,85 e 20090101 >= 0,21 então algodão (22)
R6	Se 20090101 < 0,71 e 20081202 >= 0,63 e 20081202 < 0,8 e 20080913 < 0,44 então soja (7)
R7	Se 20090101 < 0,71 e 20081202 >= 0,63 e 20081202 < 0,8 e 20080913 >= 0,44 então algodão (1)
R8	Se 20090101 < 0,71 e 20081202 >= 0,63 e 20081202 >= 0,85 então soja (28)
R9	Se 20090101 >= 0,71 e 20080913 < 0,71 e 20090322 < 0,75 então soja (352)
R10	Se 20090101 >= 0,71 e 20080913 <0,71 e 20090322 >= 0,75 e 20090322 < 0,79 e 20080913 < 0,19 então algodão (1)
R11	Se 20090101 >= 0,71 e 20080913 <0,71 e 20090322 >= 0,75 e 20090322 < 0,79 e 20080913 >= 0,19 então soja (49/1)
R12	Se 20090101 >= 0,71 e 20080913 < 0,71 e 20090322 >= 0,75 e 20090322 >= 0,79 então soja (171)
R13	Se 20090101 >= 0,71 e 20080913 >= 0,71 então milho (1)

Tabela 4.40: Conjunto de regras geradas pela 8º árvore do modelo

Calculando a *acurácia*, o *erro*, a *sensitividade*, a *especificidade*, a *cobertura*, o *suporte* e a *novidade* da 8º árvore:

Analise das regras da árvore 8							
regras	acurácia	erro	sensitividade	especificidade	cobertura	suporte	novidade
R1	0,818	0,182	0,015	0,946	0,017	0,014	-0,002
R2	0,750	0,250	0,111	0,998	0,006	0,005	0,004
R3	1,000	0,000	0,900	1,000	0,014	0,014	0,014
R4	1,000	0,000	0,003	1,000	0,003	0,003	0,000
R5	1,000	0,000	0,815	1,000	0,034	0,034	0,032
R6	1,000	0,000	0,011	1,000	0,011	0,011	0,001
R7	1,000	0,000	0,037	1,000	0,002	0,002	0,001
R8	1,000	0,000	0,045	1,000	0,043	0,043	0,002
R9	1,000	0,000	0,570	1,000	0,537	0,537	0,030
R10	1,000	0,000	0,037	1,000	0,002	0,002	0,001
R11	0,980	0,020	0,079	0,973	0,076	0,075	0,003
R12	1,000	0,000	0,277	1,000	0,261	0,261	0,015
R13	1,000	0,000	0,100	1,000	0,002	0,002	0,002

Tabela 4.41: Análise das regras geradas pela 8^o árvore do modelo

De acordo com os resultados da Tabela 4.41, pode-se verificar que as regras R3, R5, R9 e R12 são as regras que possuem os mais altos valores da *acurácia*, *sensitividade*, e *especificidade*. Logo essas são as melhores regras geradas pela 8^o árvore do modelo.

Segue o nono conjunto de regras (Tabela 4.42) gerada pela 9^o árvore:

Regras da árvore 9	Wrapper utilizando Resample com R=0,0 e o Random Forest
R1	Se 20090101 < 0,48 e 20090322 < 0,88 e 20090322 < 0,83 então soja (11)
R2	Se 20090101 < 0,48 e 20090322 < 0,88 e 20090322 >= 0,83 e 20091202 < 0,65 então algodão (1)
R3	Se 20090101 < 0,48 e 20090322 < 0,88 e 20090322 >= 0,83 e 20091202 >= 0,65 então algodão (1)
R4	Se 20090101 < 0,48 e 20090322 >= 0,88 então algodão (14)
R5	Se 20090101 >= 0,48 e 20090101 < 0,81 e 20090306 < 0,74 então soja (49)
R6	Se 20090101 >= 0,48 e 20090101 < 0,81 e 20080306 >= 0,74 e 20081202 < 0,63 e 20090101 < 0,69 então soja (12/2)
R7	Se 20090101 >= 0,48 e 20090101 < 0,81 e 20080306 >= 0,74 e 20081202 < 0,63 e 20090101 >= 0,69 então milho (7)
R8	Se 20090101 >= 0,48 e 20090101 < 0,81 e 20080306 >= 0,74 e e 20081202 >= 0,63 então soja (17)
R9	Se 20090101 >= 0,48 e 20090101 >= 0,81 e 20090322 ,< 0,7 então soja (275)
R10	Se 20090101 >= 0,48 e 20090101 >= 0,81 e 20090322 >= 0,7 e 20090101 < 0,88 e 20080913 < 0,38 então soja (22)
R11	Se 20090101 >= 0,48 e 20090101 >= 0,81 e 20090322 >= 0,7 e 20090101 < 0,88 e 20080913 >= 0,38 então algodão (11/5)
R12	Se 20090101 >= 0,48 e 20090101 >= 0,81 e 20090322 >= 0,7 e 20090101 >= 0,88 e 20081202 < 0,99 então soja (234/2)
R13	Se 20090101 >= 0,48 e 20090101 >= 0,81 e 20090322 >= 0,7 e 20090101 >= 0,88 e 20081202 >= 0,99 então algodão (1)

Tabela 4.42: Conjunto de regras geradas pela 9º árvore do modelo

Calculando a *acurácia*, o *erro*, a *sensitividade*, a *especificidade*, a *cobertura*, o *suporte* e a *novidade* da 9º árvore:

Analise das regras da árvore 9							
regras	acurácia	erro	sensitividade	especificidade	cobertura	suporte	novidade
R1	1,000	0,000	0,018	1,000	0,017	0,017	0,001
R2	1,000	0,000	0,037	1,000	0,002	0,002	0,001
R3	1,000	0,000	0,002	1,000	0,002	0,002	0,000
R4	1,000	0,000	0,519	1,000	0,021	0,021	0,020
R5	1,000	0,000	0,079	1,000	0,075	0,075	0,004
R6	0,857	0,143	0,019	0,941	0,021	0,018	-0,002
R7	1,000	0,000	1,000	1,000	0,011	0,011	0,011
R8	1,000	0,000	0,027	1,000	0,026	0,026	0,001
R9	1,000	0,000	0,443	1,000	0,420	0,420	0,022
R10	1,000	0,000	0,035	1,000	0,034	0,034	0,002
R11	0,688	0,313	0,407	0,992	0,024	0,017	0,016
R12	0,992	0,008	0,377	0,941	0,360	0,357	0,016
R13	1,000	0,000	0,037	1,000	0,002	0,002	0,001

Tabela 4.43: Análise das regras geradas pela 9^o árvore do modelo

De acordo com os resultados da Tabela 4.43, pode-se verificar que as regras R4, R7, R9 e R12 são as regras que possuem os mais altos valores da *acurácia*, *sensitividade*, e *especificidade*. Logo essas são as melhores regras geradas pela 9^o árvore do modelo.

Segue o décimo e último conjunto de regras (Tabela 4.44) geradas pela 10^o árvore:

Regras da árvore 10	Wrapper utilizando Resample com R=0,0 e o Random Forest
R1	Se 20090101 < 0,71 e 20090322 < 0,81 e 20090322 < 0,73 e 20090306 < 0,71 então soja (9)
R2	Se 20090101 < 0,71 e 200900233 < 0,81 e 20090322 < 0,73 e 20090306 >= 0,71 então milho (14)
R3	Se 20090101 < 0,71 e 20090322 < 0,81 e 20090322 >= 0,73 então soja (17)
R4	Se 20090101 < 0,71 e 20090322 >= 0,81 e 20081202 < 0,82 e 20090101 < 0,21 então soja (1)
R5	Se 20090101 < 0,71 e 20090322 >= 0,81 e 20081202 < 0,82 e 20090101 >= 0,21 e 20081202 < 0,64 então algodão (27)
R6	Se 20090101 < 0,71 e 20090322 >= 0,81 e 20081202 < 0,82 e 20090101 >= 0,21 e 20081202 >= 0,64 então soja (4/2)
R7	Se 20090101 < 0,71 e 20090322 >= 0,81 e 20081202 >= 0,82 então soja (12)
R8	Se 20090101 >= 0,71 e 20090322 < 0,7 então soja (307)
R9	Se 20090101 >= 0,71 e 20090322 < 0,7 e 20090322 < 0,79 e 20080913 < 0,71 e 20080913 < 0,2 então soja (7/3)
R10	Se 20090101 >= 0,71 e 20080913 <0,7 e 20090322 >= 0,75 e 20090322 < 0,71 e 20080913 >= 0,2 então soja (97/3)
R11	Se 20090101 >= 0,71 e 20080913 <0,7 e 20090322 >= 0,75 e 20080913 >= 0,71 então milho (1)
R12	Se 20090101 >= 0,71 e 20080922 >= 0,7 e 20090322 >= 0,79 então soja (159)

Tabela 4.44: Conjunto de regras geradas pela 10^o árvore do modelo

Calculando a *acurácia*, o *erro*, a *sensitividade*, a *especificidade*, a *cobertura*, o *suporte* e a *novidade* da 10^0 árvore:

Analise das regras da árvore 10							
regras	acurácia	erro	sensitividade	especificidade	cobertura	suporte	novidade
R1	1,000	0,000	0,015	1,000	0,014	0,014	0,001
R2	1,000	0,000	0,933	1,000	0,021	0,021	0,021
R3	1,000	0,000	0,028	1,000	0,026	0,026	0,002
R4	1,000	0,000	0,002	1,000	0,002	0,002	0,000
R5	1,000	0,000	1,000	1,000	0,041	0,041	0,040
R6	0,667	0,333	0,007	0,952	0,009	0,006	-0,002
R7	1,000	0,000	0,020	1,000	0,018	0,018	0,001
R8	1,000	0,000	0,501	1,000	0,469	0,469	0,030
R9	0,700	0,300	0,011	0,929	0,015	0,011	-0,004
R10	0,970	0,030	0,158	0,929	0,153	0,148	0,005
R11	1,000	0,000	0,067	1,000	0,002	0,002	0,001
R12	1,000	0,000	0,259	1,000	0,243	0,243	0,016

Tabela 4.45: Análise das regras geradas pela 10^o árvore do modelo

De acordo com os resultados da Tabela 4.45, pode-se verificar que as regras R2, R5, R8, e R12

são as regras que possuem os mais altos valores da *acurácia*, *sensitividade*, e *especificidade*. Logo essas são as melhores regras geradas pela 10^o árvore do modelo.

4.2 Base de Conhecimentos

De acordo com os testes realizados até o momento, apresenta-se uma base de regras geradas pelos respectivos testes. Essas regras são apresenadas na próxima Tabela 4.2

Banco de dados com todas as regras		
Quantidade	**Regras**	**PCA utilizando Resample com R=0 e conjunto teste de 10%**
1	R1	Se 20090306 <= 0,8648 e 20090101 <= 0,7134 e 20090202 <= 0,7093 e 20090218 <= 0,288509 então algodão (2)
2	R2	Se 20090306 <= 0,8648 e 20090101 <= 0,7134 e 20090202 <= 0,7093 e 20090218 > 0,288509 e 20081015 <= 0,2626 então algodão (3/1)
3	R3	Se 20090306 <= 0,8648 e 20090101 <= 0,7134 e 20090202 <= 0,7093 e 20090218 > 0,288509 e 20081015 > 0,2626 então soja (32)
4	R5	Se 20090306 <= 0,8648 e 20090101 <= 0,7134 e 20090202 > 0,7093 e 20090101 <= 0,6761 e 20081015 <= 0,7411 então soja (7)
5	R6	Se 20090306 <= 0,8648 e 20090101 <= 0,7134 e 20090202 > 0,7093 e 20090101 <= 0,6761 e 20081015 > 0,7411 então milho (3)
6	R7	Se 20090306 <= 0,8648 e 20090101 <= 0,7134 e 20090202 > 0,7093 e 20090101 > 0,6761 então milho (6)
7	R8	Se 20090306 > 0,8648 e 20081202 <= 0,6228 então algodão (19/1)
8	R9	Se 20090306 > 0,8648 e 20081202 > 0,6228 então soja (11)
	Regras	**CFS com Cross Validation**
9	R1	Se 20090306 <= 0,865337 e 20081218 > 0,7067 então soja (612/7)
10	R2	Se 200903022 <= 0,8141 e 20090306 <= 0,7173 então soja (46/1)
11	R3	Se 200900101 <= 0,5635 e 20081202 <= 0,6475 então algodão (36/5)
12	R4	Se 20081202 > 0,63206 então soja (19/1)
13	R5	Se 20081202 <= 0,4775 então soja (7)
14	R6	Se 20090306 > 0,5921 então milho (5)
15	R7	Se 20090306 > 0,5921 então soja (3/1)
	Regras	**Info Gain utilizando Resample R=0,5 e conjunto teste de 10%**
16	R1	Se 20090101 > 0,8495 e 20081202 <= 0,986 e 20090218 > 0,3052 e 20090509 > 0,6953 e 20090101 <= 0,9542 então soja (255)
17	R2	Se 200900424 > 0,8039 e 20081202 <= 0,7814 e 20090525 <= 0,853 e 20090525 > 0,652 então algodão (88)
18	R3	Se 20081218 > 0,8086 e 20090423 <= 0,8039 então soja (67)
19	R5	Se 20090407 > 0,7757 e 20081202 <= 0,5351 então algodão (17)
20	R6	Se 20090101 <= 0,8599 e 20090306 > 0,6535 e 20081031 <= 0,6705 e 20090218 <= 0,8646 e 20090306 <= 0,8013 e 20081116 <= 0,571 então milho (104/1)
21	R7	Se 20090306 <= 0,7888 e 20090423 <= 0,8039 então soja (35)
22	R8	Se 20081202 <= 0,5526 e 20081031 > 0,502 então milho (12)
23	R9	Se 20090322 <= 0,7906 e 20090423 <= 0,8538 e 20090509 > 0,6142 e 20090101 <= 0,9584 então algodão (24)
24	R10	Se 20090322 <= 0,7906 e 20090423 <= 0,8538 e 20090509 > 0,6142 e 20090101 <= 0,9584 então soja (25)

	Regras da árvore 1	Wrapper utilizando Resample com R=0 e o Random Forest
25	**R1**	Se 20090101 < 0,71 e 20081202 < 0,8 e 20090322 < 0,81 e 20090101 < 0,69 e 20081202 < 0,32 então soja (4/1)
26	**R2**	Se 20090101 < 0,71 e 20081202 < 0,8 e 20090322 < 0,81 e 20090101 < 0,69 e 20081202 >= 0,32 então soja (11)
27	**R3**	Se 20090101 < 0,71 e 20081202 < 0,8 e 20090322 < 0,81 e 20090101 >= 0,69 e 20090306 < 0,62 então soja (3)
28	**R4**	Se 20090101 < 0,71 e 20081202 < 0,8 e 20090322 < 0,81 e 20090101 >= 0,69 e 20090306 >= 0,62 então milho (6)
29	**R5**	Se 20090101 < 0,71 e 20081202 < 0,8 e 20090322 >= 0,81 e 20080913 < 0,62 e 20090101 < 0,21 então soja (1)
30	**R6**	Se 20090101 < 0,71 e 20081202 < 0,8 e 20090322 >= 0,81 e 20080913 < 0,62 e 20090101 >= 0,21 então algodão (26/1)
31	**R7**	Se 20090101 < 0,71 e 20081202 < 0,8 e 20090322 >= 0,81 e 20080913 >= 0,62 então soja (1)
32	**R8**	Se 20090101 < 0,71 e 2001202 >= 0,8 então soja (27)
33	**R9**	Se 20090101 >= 0,71 e 20080913 < 0,2 e 20090322 < 0,78 então soja (6)
34	**R10**	Se 20090101 >= 0,71 e 20080913 < 0,2 e 20090322 >= 0,78 e 20090306 < 0,54 então soja (1)
35	**R11**	Se 20090101 >= 0,71 e 20080913 < 0,2 e 20090322 >= 0,78 e 20090306 >= 0,54 então algodão (5)
36	**R12**	Se 20090101 >= 0,71 e 20080913 < 0,2 e 20090101 >= 0,93 então soja (17)
37	**R13**	Se 20090101 >= 0,71 e 20080913 >= 0,2 e 20081202 < 0,99 e 20090306 < 0,8 e 20080913 < 0,41 então soja (497)
38	**R14**	Se 20090101 >= 0,71 e 20080913 >= 0,2 e 20081202 < 0,99 e 20090306 < 0,8 e 20080913 >= 0,41 então soja (25/1)
39	**R15**	Se 20090101 >= 0,71 e 20080913 >= 0,2 e 20081202 < 0,99 e 20090306 >= 0,8 e 20090306 < 0,81 então milho (1)
40	**R16**	Se 20090101 >= 0,71 e 20080913 >= 0,2 e 20081202 < 0,99 e 20090306 >= 0,8 e 20090306 >= 0,81 então soja(22)
41	**R17**	Se 20090101 >= 0,71 e 20080913 >= 0,2 e 20081202 >= 0,99 então algodão (2)
	Regras da árvore 2	**Wrapper utilizando Resample com R=0 e o Random Florest**
42	**R1**	Se 20090101 < 0,71 e 20090101 < 0,7 e 20081202 < 0,41 e 20080913 < 0,29 então algodão (16)
43	**R2**	Se 20090101 < 0,71 e 20090101 < 0,7 e 20081202 < 0,41 e 20080913 >= 0,29 então milho (1)
44	**R3**	Se 20090101 < 0,71 e 20090101 < 0,7 e 20081202 >= 0,41 e 20090306 < 0,72 então soja (21)
45	**R4**	Se 20090101 < 0,71 e 20090101 < 0,7 e 20081202 >= 0,41 e 20090306 >= 0,72 e 20081202 < 0,81 então algodão (18/9)
46	**R5**	Se 20090101 < 0,71 e 20090101 < 0,7 e 20081202 >= 0,41 e 20090306 >= 0,72 e 20081202 >= 0,81 então soja (13)
47	**R6**	Se 20090101 < 0,71 e 20090101 >= 0,7 então milho (7)
48	**R7**	Se 20090101 >= 0,71 e 20080913 < 0,2 e 20080913 < 0,19 então soja (26)
49	**R8**	Se 20090101 >= 0,71 e 20080913 < 0,2 e 20080913 >= 0,19 então algodão (6)
50	**R9**	Se 20090101 >= 0,71 e 20080913 >= 0,2 e 20081202 < 0,99 então soja (546)
51	**R10**	Se 20090101 >= 0,71 e 20080913 >= 0,2 e 20081202 >= 0,99 então algodão (1)

	Regras da árvore 3	Wrapper utilizando Resample com R=0 e o Random Forest
52	**R1**	Se 20090101 < 0,71 e 20081202 < 0,62 e 20090322 < 0,8 e 20081202 < 0,48 e 20081202 < 0,36 então milho (3/1)
53	**R2**	Se 20090101 < 0,71 e 20081202 < 0,62 e 20090322 < 0,8 e 20081202 < 0,48 e 20081202 >= 0,36 então soja (5)
54	**R3**	Se 20090101 < 0,71 e 20081202 < 0,62 e 20090322 < 0,8 e 20081202 >= 0,48 então milho (11)
55	**R4**	Se 20090101 < 0,71 e 20081202 < 0,62 e 20090322 >= 0,8 então algodão (25)
56	**R5**	Se 20090101 < 0,71 e 20081202 >= 0,62 e 20081202 < 0,81 e 20080913 < 0,44 então soja (8)
57	**R6**	Se 20090101 < 0,71 e 20081202 >= 0,62 e 20081202 < 0,81 e 20080913 >= 0,44 então algodão (3)
58	**R7**	Se 20090101 < 0,71 e 20081202 >= 0,62 e 20081202 >= 0,81então soja (26)
59	**R8**	Se 20090101 >= 0,71 e 20090322 < 0,7 então soja (311)
60	**R9**	Se 20090101 >= 0,71 e 20090322 >= 0,7 e 20090101 < 0,9 e 20080913 < 0,2 então algodão (3)
61	**R10**	Se 20090101 >= 0,71 e 20090322 >= 0,7 e 20090101 < 0,9 e 20080913 >= 0,2 e 20081202 < 0,62 então milho (1)
62	**R11**	Se 20090101 >= 0,71 e 20090322 >= 0,7 e 20090101< 0,9 e 20080913 >= 0,2 e 20081202 >= 0,62 então soja (64/2)
63	**R12**	Se 20090101 >= 0,71 e 20090322 >= 0,7 20090101 >= 0,9 e 20090322 < 0,71 então algodão (1)
64	**R13**	Se 20090101 >= 0,71 e 20090322 >= 0,7 20090101 >= 0,9 e 20090323 >= 0,71 então soja (194)
	Regras da árvore 4	**Wrapper utilizando Resample com R=0 e o Random Forest**
65	**R1**	Se 20090101 < 0,54 e 20081202 < 0,63 e 20080913 < 0,29 e 20090306 < 0,57 então soja (1)
66	**R2**	Se 20090101 < 0,54 e 20081202 < 0,63 e 20080913 < 0,29 e 20090306 >= 0,57 então algodão (19)
67	**R3**	Se 20090101 < 0,54 e 20081202 < 0,63 e 20080913 >= 0,29 e 20090101 < 0,51 e 20081202 < 0,36 então milho (3)
68	**R4**	Se 20090101 < 0,54 e 20081202 < 0,63 e 20080913 >= 0,29 e 20090101 < 0,51 e 20081202 >= 0,36 então soja (2)
69	**R5**	Se 20090101 < 0,54 e 20081202 < 0,63 e 20080913 >= 0,29 e 20090101 >= 0,51 então soja (3)
70	**R6**	Se 20090101 < 0,54 e 20081202 >= 0,63 então soja (13)
71	**R7**	Se 20090101 >= 0,54 e 20090306 < 0,74 e 20090322 < 0,7 então soja (314)
72	**R8**	Se 20090101 >= 0,54 e 20090306 < 0,74 e 20090322 >= 0,7 e 20080913 < 0,2 e 20090306 < 0,54 então soja (13)
73	**R9**	Se 20090101 >= 0,54 e 20090306 < 0,74 e 20090322 >= 0,7 e 20080913 < 0,2 e 20090306 >= 0,54 então algodão (9/3)
74	**R10**	Se 20090101 >= 0,54 e 20090306 < 0,74 e 20090322 >= 0,7 e 20080913 >= 0,2 e 20090322 < 0,71 então algodão (2)
75	**R11**	Se 20090101 >= 0,54 e 20090306 < 0,74 e 20090322 >= 0,7 e 20080912 >= 0,2 e 20090322 >= 0,2 e 20090322 < 0,71 então soja (200/3)
76	**R12**	Se 20090101 >= 0,54 e 20090306 >= 0,74 e 20081202 < 0,67 e 20080913 < 0,25 então soja (10)
77	**R13**	Se 20090101 >= 0,54 e 20090306 >= 0,74 e 20081202 < 0,67 e 20080913 >= 0,25 e 20090101 < 0,81 então milho (6)
78	**R14**	Se 20090101 >= 0,54 e 20090306 >= 0,74 e 20081202 < 0,67 e 20080913 >= 0,25 e 20090101 >= 0,81 então soja (2)
79	**R15**	Se 20090101 >= 0,54 e 20090306 >= 0,74 e 20081202 >= 0,67 e 20081202 < 0,77 e 20081202 < 0,75 então soja (7)
80	**R16**	Se 20090101 >= 0,54 e 20090306 >= 0,74 e 20081202 >= 0,67 e 20081202 < 0,77 e 20081202 >= 0,75 então algodão (2)
81	**R17**	Se 20090101 >= 0,54 e 20090306 >= 0,74 e 20081202 >= 0,67 e 20081202 >= 0,77 então soja (49)

	Regras da árvore 5	Wrapper utilizando Resample com R=0 e o Random Forest
82	**R1**	Se 20090101 < 0,71 e 20081202 < 0,63 e 20090322 < 0,59 então soja (6)
83	**R2**	Se 20090101 < 0,71 e 20081202 < 0,63 e 20090322 >= 0,59 e 20080913 < 0,25 então soja (3)
84	**R3**	Se 20090101 < 0,71 e 20081202 < 0,63 e 20090322 >= 0,59 e 20080913 >= 0,25 então milho (13)
85	**R4**	Se 20090101 < 0,71 e 20081202 < 0,63 e 20090322 >= 0,80 e 20080913 <0,51 e 20080913 < 0,22 então algodão (12/1)
86	**R5**	Se 20090101 < 0,71 e 20081202 < 0,63 e 20090322 >= 0,80 e 20080913 <0,51 e 20080913 >= 0,22 então algodão (16)
87	**R6**	Se 20090101 < 0,71 e 20081202 < 0,63 e 20090322 >= 0,80 e 20080913 >=0,51 então soja (1)
88	**R7**	Se 20090101 < 0,71 e 20081202 >= 0,63 e 20080913 < 0,46 então soja (29)
89	**R8**	Se 20090101 < 0,71 e 20081202 >= 0,63 e 20080913 >= 0,46 e 20090101 < 0,59 então soja (3)
90	**R9**	Se 20090101 < 0,71 e 20081202 >= 0,63 e 20080913 >= 0,46 e 20090101 >= 0,59 então algodão (1)
91	**R10**	Se 20090101 >= 0,71 e 20080913 < 0,67 e 20080913 < 0,2 e 20090322 < 0,76 então soja (17)
92	**R11**	Se 20090101 >= 0,71 e 20080913 < 0,67 e 20080913 < 0,2 e 20090322 >= 0,76 e 20090101 < 0,93 então algodão (5/2)
93	**R12**	Se 20090101 >= 0,71 e 20080913 < 0,67 e 20080913 < 0,2 e 20090322 >= 0,76 e 20090101 >= 0,93 então soja (9)
94	**R13**	Se 20090101 >= 0,71 e 20080913 < 0,67 e 20080913 >= 0,2 e 20081202 < 0,99 e 20090101 < 0,88 então soja (127/2)
95	**R14**	Se 20090101 >= 0,71 e 20080913 < 0,67 e 20080913 >= 0,2 e 20081202 < 0,99 e 20090101 >= 0,88 então soja (409)
96	**R15**	Se 20090101 >= 0,71 e 20080913 < 0,67 e 20080913 >= 0,2 e 20081202 >= 0,99 entâo algodão (1)
97	**R16**	Se 20090101 >= 0,71 e 20080913 >= 0,67 e 20080913 < 0,77 entâo milho (2)
98	**R17**	Se 20090101 >= 0,71 e 20080913 >= 0,67 e 20080913 >= 0,77 entâo soja (1)
	Regras da árvore 6	**Wrapper utilizando Resample com R=0 e o Random Forest**
99	**R1**	Se 20090309 < 0,73 e 20090322 < 0,75 então soja (361)
100	**R2**	Se 20090309 < 0,73 e 20090322 >= 0,75 e 20090322 < 0,78 e 20090322 < 0,78 e 20080913 < 0,41 então soja (18)
101	**R3**	Se 20090309 < 0,73 e 20090322 >= 0,75 e 20090322 < 0,78 e 20090322 < 0,78 e 20080913 >= 0,41 então soja (3/1)
102	**R4**	Se 20090309 < 0,73 e 20090322 >= 0,75 e 20090322 < 0,78 e 20090322 >= 0,78 então algodão (3)
103	**R5**	Se 20090309 < 0,73 e 20090322 >= 0,75 e 20090322 >= 0,78 então soja (158)
104	**R6**	Se 20090309 >= 0,73 e 20081202 < 0,77 e 20090322 < 0,81 e 20081202 < 0,48 então soja (7)
105	**R7**	Se 20090101 >= 0,73 e 20081202 < 0,77 e 20090322 < 0,81 e 20081202 >= 0,48 e 20081202 < 0,62 então milho (11)
106	**R8**	Se 20090101 >= 0,73 e 20081202 < 0,77 e 20090322 < 0,81 e 20081202 >= 0,48 e 20081202 >= 0,62 então soja (7)
107	**R9**	Se 20090101 >= 0,73 e 20081202 < 0,77 e 20090322 >= 0,81 e 20080913 < 0,26 e 20081202 < 0,44 então algodão (8)
108	**R10**	Se 20090101 >= 0,73 e 20081202 < 0,77 e 20090322 >= 0,81 e 20080913 < 0,26 e 20081202 >= 0,44 então soja (4/2)
109	**R11**	Se 20090306 >= 0,73 e 20081202 < 0,77 e 20090322 >= 0,81 e 20080913 >= 0,26 então soja (13)
110	**R12**	Se 20090306 >= 0,73 e 20081202 >= 0,77 então soja (62)

	Regras da árvore 7	Wrapper utilizando Resample com R=0 e o Random Forest
111	**R1**	Se 20090101 < 0,71 e 20081202 < 0,63 e 20090322 < 0,8 e 20080913 < 0,26 então soja (7)
112	**R2**	Se 20090101 < 0,71 e 20081202 < 0,63 e 20090322 < 0,8 e 20080913 >= 0,26 e 20080322 < 0,59 então soja (1)
113	**R3**	Se 20090101 < 0,71 e 20081202 < 0,63 e 20090322 < 0,8 e 20080913 >= 0,26 e 20080322 >= 0,59 então milho (7)
114	**R4**	Se 20090101 < 0,71 e 20081202 < 0,63 e 20090322 >= 0,8 então algodão (17)
115	**R5**	Se 20090101 < 0,71 e 20081202 > 0,63 e 20080913 <0.44 então soja (37)
116	**R6**	Se 20090101 < 0,71 e 20081202 > 0,63 e 20080913 >=0.44 e 20080913 < 0,5 então algodão (1)
117	**R7**	Se 20090101 < 0,71 e 20081202 > 0,63 e 20080913 >=0.44 e 20080913 >= 0,5 então soja (1)
118	**R8**	Se 20090101 >= 0,71 e 20081202 <0,5 e 20080913 <0,58 então soja (23)
119	**R9**	Se 20090101 >= 0,71 e 20081202 <0,5 e 20080913 >=0,58 então milho (1)
120	**R10**	Se 20090101 >= 0,71 e 20081202 >= 0,51 e 20090306 < 0,68 então soja (443)
121	**R11**	Se 20090101 >= 0,71 e 20081202 >= 0,51 e 20090306 >= 0,68 e 20080913 < 0,19 e 20090101 <0,93 então algodão (1)
122	**R12**	Se 20090101 >= 0,71 e 20081202 >= 0,51 e 20090306 >= 0,68 e 20080913 < 0,19 e 20090101 >= 0,93 então soja (1)
123	**R13**	Se 20090101 >= 0,71 e 20081202 >= 0,51 e 20090306 >= 0,68 e 20080913 >= 0,19 então soja (115)
	Regras da árvore 8	**Wrapper utilizando Resample com R=0 e o Random Forest**
124	**R1**	Se 20090101 < 0,71 e 20081202 < 0,63 e 20090306 < 0,85 e 20090101 < 0,68 e 20090322 < 0,81 então soja (9/2)
125	**R2**	Se 20090101 < 0,71 e 20081202 < 0,63 e 20090306 < 0,85 e 20090101 < 0,68 e 20090322 >= 0,81 então soja (3/1)
126	**R3**	Se 20090101 < 0,71 e 20081202 < 0,63 e 20090306 < 0,85 e 20090101 >= 0,68 então milho (9)
127	**R4**	Se 20090101 < 0,71 e 20081202 < 0,63 e 20090306 >= 0,85 e 20090101 < 0,21 então soja (2)
128	**R5**	Se 20090101 < 0,71 e 20081202 < 0,63 e 20090306 >= 0,85 e 20090101 >= 0,21 então algodão (22)
129	**R6**	Se 20090101 < 0,71 e 20081202 >= 0,63 e 20081202 < 0,8 e 20080913 < 0,44 então soja (7)
130	**R7**	Se 20090101 < 0,71 e 20081202 >= 0,63 e 20081202 < 0,8 e 20080913 >= 0,44 então algodão (1)
131	**R8**	Se 20090101 < 0,71 e 20081202 >= 0,63 e 20081202 >= 0,85 então soja (28)
132	**R9**	Se 20090101 >= 0,71 e 20080913 < 0,71 e 20090322 < 0,75 então soja (352)
133	**R10**	Se 20090101 >= 0,71 e 20080913 <0,71 e 20090322 >= 0,75 e 20090322 < 0,79 e 20080913 < 0,19 então algodão (1)
134	**R11**	Se 20090101 >= 0,71 e 20080913 <0,71 e 20090322 >= 0,75 e 20090322 < 0,79 e 20080913 >= 0,19 então soja (49/1)
135	**R12**	Se 20090101 >= 0,71 e 20080913 < 0,71 e 20090322 >= 0,75 e 20090322 >= 0,79 então soja (171)
136	**R13**	Se 20090101 >= 0,71 e 20080913 >= 0,71 então milho (1)

	Regras da árvore 9	Wrapper utilizando Resample com R=0 e o Random Forest
137	R1	Se 20090101 < 0,48 e 20090322 < 0,88 e 20090322 < 0,83 então soja (11)
138	R2	Se 20090101 < 0,48 e 20090322 < 0,88 e 20090322 >= 0,83 e 20091202 < 0,65 então algodão (1)
139	R3	Se 20090101 < 0,48 e 20090322 < 0,88 e 20090322 >= 0,83 e 20091202 >= 0,65 então algodão (1)
140	R4	Se 20090101 < 0,48 e 20090322 >= 0,88 então algodão (14)
141	R5	Se 20090101 >= 0,48 e 20090101 < 0,81 e 20090306 < 0,74 então soja (49)
142	R6	Se 20090101 >= 0,48 e 20090101 < 0,81 e 20080306 >= 0,74 e 20081202 < 0,63 e 20090101 < 0,69 então soja (12/2)
143	R7	Se 20090101 >= 0,48 e 20090101 < 0,81 e 20080306 >= 0,74 e 20081202 < 0,63 e 20090101 >= 0,69 então milho (7)
144	R8	Se 20090101 >= 0,48 e 20090101 < 0,81 e 20080306 >= 0,74 e e 20081202 >= 0,63 então soja (17)
145	R9	Se 20090101 >= 0,48 e 20090101 >= 0,81 e 20090322 ,< 0,7 então soja (275)
146	R10	Se 20090101 >= 0,48 e 20090101 >= 0,81 e 20090322 >= 0,7 e 20090101 < 0,88 e 20080913 < 0,38 então soja (22)
147	R11	Se 20090101 >= 0,48 e 20090101 >= 0,81 e 20090322 >= 0,7 e 20090101 < 0,88 e 20080913 >= 0,38 então algodão (11/5)
148	R12	Se 20090101 >= 0,48 e 20090101 >= 0,81 e 20090322 >= 0,7 e 20090101 >= 0,88 e 20081202 < 0,99 então soja (234/2)
149	R13	Se 20090101 >= 0,48 e 20090101 >= 0,81 e 20090322 >= 0,7 e 20090101 >= 0,88 e 20081202 >= 0,99 então algodão (1)
	Regras da árvore 10	Wrapper utilizando Resample com R=0 e o Random Forest
150	R1	Se 20090101 < 0,71 e 20090322 < 0,81 e 20090322 < 0,73 e 20090306 < 0,71 então soja (9)
151	R2	Se 20090101 < 0,71 e 200900233 < 0,81 e 20090322 < 0,73 e 20090306 >= 0,71 então milho (14)
152	R3	Se 20090101 < 0,71 e 20090322 < 0,81 e 20090322 >= 0,73 então soja (17)
153	R4	Se 20090101 < 0,71 e 20090322 >= 0,81 e 20081202 < 0,82 e 20090101 < 0,21 então soja (1)
154	R5	Se 20090101 < 0,71 e 20090322 >= 0,81 e 20081202 < 0,82 e 20090101 >= 0,21 e 20081202 < 0,64 então algodão (27)
155	R6	Se 20090101 < 0,71 e 20090322 >= 0,81 e 20081202 < 0,82 e 20090101 >= 0,21 e 20081202 >= 0,64 então soja (4/2)
156	R7	Se 20090101 < 0,71 e 20090322 >= 0,81 e 20081202 >= 0,82 então soja (12)
157	R8	Se 20090101 >= 0,71 e 20090322 < 0,7 então soja (307)
158	R9	Se 20090101 >= 0,71 e 20090322 < 0,7 e 20090322 < 0,79 e 20080913 < 0,71 e 20080913 < 0,2 então soja (7/3)
159	R10	Se 20090101 >= 0,71 e 20080913 <0,7 e 20090322 >= 0,75 e 20090322 < 0,71 e 20080913 >= 0,2 então soja (97/3)
160	R11	Se 20090101 >= 0,71 e 20080913 <0,7 e 20090322 >= 0,75 e 20080913 >= 0,71 então milho (1)
161	R12	Se 20090101 >= 0,71 e 20080922 >= 0,7 e 20090322 >= 0,79 então soja (159)

Tabela 4.46: Tabela com todas as regras geradas pelos classificadores

De posse dessa base de conhecimento, será feito um ranqueamento nas regras (Tabela **??**), pretende-se com isso saber qual regra é melhor que a outra olhando primeiramente para a sensitividade, depois para o suporte e então para a cobertura:

Ranqueamento das regras				
algoritmo	Regras	sensitividade	suporte	cobertura
B3	Se 200900101 <= 0,5635 e 20081202 <= 0,6475 então algodão (36/5)	1	0,049450549	0,056318681
M5	Se 20090101 < 0,71 e 20090322 >= 0,81 e 20081202 < 0,82 e 20090101 >= 0,21 e 20081202 < 0,64 então algodão (27)	1	0,041221374	0,041221374
I7	Se 20090101 >= 0,73 e 20081202 < 0,77 e 20090322 < 0,81 e 20081202 >= 0,48 e 20081202 < 0,62 então milho (11)	1	0,016793893	0,016793893
L7	Se 20090101 >= 0,48 e 20090101 < 0,81 e 20080306 >= 0,74 e 20081202 < 0,63 e 20090101 >= 0,69 então milho (7)	1	0,010687023	0,010687023
B6	Se 20090306 > 0,5921 então milho (5)	1	0,006868132	0,006868132
M2	Se 20090101 < 0,71 e 200900233 < 0,81 e 20090322 < 0,73 e 20090306 >= 0,71 então milho (14)	0,933333333	0,021374046	0,021374046
A4	Se 20090306 <= 0,8648 e 20090101 > 0,7134 então soja (572/7)	0,919614148	0,873282443	0,883969466
E9	Se 20090101 >= 0,71 e 20080913 >= 0,2 e 20081202 < 0,99 então soja (546)	0,900990099	0,833587786	0,833587786
K3	Se 20090101 < 0,71 e 20081202 < 0,63 e 20090306 < 0,85 e 20090101 >= 0,68 então milho (9)	0,9	0,013740458	0,013740458
C6	Se 20090101 <= 0,8599 e 20090306 > 0,6535 e 20081031 <= 0,6705 e 20090218 <= 0,8646 e 20090306 <= 0,8013 e 20081116 <= 0,571 então	0,896551724	0,158778626	0,160305344
J4	Se 20090101 < 0,71 e 20081202 < 0,63 e 20090322 >= 0,8 então algodão (17)	0,894736842	0,025954198	0,025954198
B1	Se 20090306 <= 0,865337 e 20081218 > 0,7067 então soja (612/7)	0,890829694	0,840659341	0,850274725
E6	Se 20090101 < 0,71 e 20090101 >= 0,7 então milho (7)	0,875	0,010687023	0,010687023
J3	Se 20090101 < 0,71 e 20081202 < 0,63 e 20090322 < 0,8 e 20080913 >= 0,26 e 20080322 >= 0,59 então milho (7)	0,875	0,010687023	0,010687023
H3	Se 20090101 < 0,71 e 20081202 < 0,63 e 20090322 >= 0,59 e 20080913 >= 0,25 então milho (13)	0,866666667	0,019847328	0,019847328
D4	Se 20090101 < 0,71 e 20081202 < 0,8 e 20090322 < 0,81 e 20090101 >= 0,69 e 20090306 >= 0,62 então milho (6)	0,857142857	0,009160305	0,009160305
K5	Se 20090101 < 0,71 e 20081202 < 0,63 e 20090306 >= 0,85 e 20090101 >= 0,21 então algodão (22)	0,814814815	0,033587786	0,033587786
D13	Se 20090101 >= 0,71 e 20080913 >= 0,2 e 20081202 < 0,99 e 20090306 < 0,8 e 20080913 < 0,41 então soja (497)	0,808130081	0,758778626	0,758778626
A8	Se 20090306 > 0,8648 e 20081202 <= 0,6228 então algodão (19/1)	0,791666667	0,029007634	0,030534351
D6	Se 20090101 < 0,71 e 20081202 < 0,8 e 20090322 >= 0,81 e 20080913 < 0,62 e 20090101 >= 0,21 então algodão (26/1)	0,787878788	0,039694656	0,041221374
F4	Se 20090101 < 0,71 e 20081202 < 0,62 e 20090322 >= 0,8 então algodão (25)	0,78125	0,038167939	0,038167939
F3	Se 20090101 < 0,71 e 20081202 < 0,62 e 20090322 < 0,8 e 20081202 >= 0,48 então milho (11)	0,733333333	0,016793893	0,016793893
J10	Se 20090101 >= 0,71 e 20081202 >= 0,51 e 20090306 < 0,68 então soja (443)	0,705414013	0,676335878	0,676335878
c2	Se 200900424 > 0,8039 e 20081202 <= 0,7814 e 20090525 <= 0,853 e 20090525 > 0,652 então algodão (88)	0,682170543	0,134351145	0,134351145
H14	Se 20090101 >= 0,71 e 20080913 < 0,67 e 20080913 >= 0,2 e 20081202 < 0,99 e 20090101 >= 0,88 então soja (409)	0,676033058	0,624427481	0,624427481
A7	Se 20090306 <= 0,8648 e 20090101 <= 0,7134 e 20090202 > 0,7093 e 20090101 > 0,6761 então milho (6)	0,666666667	0,009160305	0,009160305
G13	Se 20090101 >= 0,54 e 20090306 >= 0,74 e 20081202 < 0,67 e 20080913 >= 0,25 e 20090101 < 0,81 então milho (6)	0,666666667	0,009160305	0,009160305

C1	Se 20090101 > 0,8495 e 20081202 <= 0,986 e 20090218 > 0,3052 e 20090509 > 0,6953 e 20090101 <= 0,9542 então soja (255)	0,62195122	0,389312977	0,389312977
G2	Se 20090101 < 0,54 e 20081202 < 0,63 e 20080913 < 0,29 e 20090306 >= 0,57 então algodão (19)	0,59375	0,029007634	0,029007634
I1	Se 20090309 < 0,73 e 20090322 < 0,75 então soja (361)	0,582258065	0,551145038	0,551145038
K9	Se 20090101 >= 0,71 e 20080913 < 0,71 e 20090322 < 0,75 então soja (352)	0,569579288	0,53740458	0,53740458
I11	Se 20090306 >= 0,73 e 20081202 < 0,77 e 20090322 >= 0,81 e 20080913 >= 0,26 então soja (13)	0,541666667	0,019847328	0,019847328
L4	Se 20090101 < 0,48 e 20090322 >= 0,88 então algodão (14)	0,518518519	0,021374046	0,021374046
G7	Se 20090101 >= 0,54 e 20090306 < 0,74 e 20090322 < 0,7 então soja (314)	0,511400651	0,479389313	0,479389313
F8	Se 20090101 >= 0,71 e 20090322 < 0,7 então soja (311)	0,511513158	0,47480916	0,47480916
M8	Se 20090101 >= 0,71 e 20090322 < 0,7 então soja (307)	0,500815661	0,46870229	0,46870229
H5	Se 20090101 < 0,71 e 20081202 < 0,63 e 20090322 >= 0,80 e 20080913 <0,51 e 20080913 >= 0,22 então algodão (16)	0,457142857	0,024427481	0,024427481
L9	Se 20090101 >= 0,48 e 20090101 >= 0,81 e 20090322 ,< 0,7 então soja (275)	0,442834138	0,419847328	0,419847328
E4	Se 20090101 < 0,71 e 20090101 < 0,7 e 20081202 >= 0,41 e 20090306 >= 0,72 e 20081202 < 0,81 então algodão (18/9)	0,43902439	0,027480916	0,041221374
L11	Se 20090101 >= 0,48 e 20090101 >= 0,81 e 20090322 >= 0,7 e 20090101 < 0,88 e 20080913 >= 0,38 então algodão (11/5)	0,407407407	0,016793893	0,024427481
E1	Se 20090101 < 0,71 e 20090101 < 0,7 e 20081202 < 0,41 e 20080913 < 0,29 então algodão (16)	0,390243902	0,024427481	0,024427481
L12	Se 20090101 >= 0,48 e 20090101 >= 0,81 e 20090322 >= 0,7 e 20090101 >= 0,88 e 20081202 < 0,99 então soja (234/2)	0,376811594	0,357251908	0,360305344
H4	Se 20090101 < 0,71 e 20081202 < 0,63 e 20090322 >= 0,80 e 20080913 <0,51 e 20080913 < 0,22 então algodão (12/1)	0,342857143	0,018320611	0,019847328
A6	Se 20090306 <= 0,8648 e 20090101 <= 0,7134 e 20090202 > 0,7093 e 20090101 <= 0,6761 e 20081015 > 0,7411 então milho (3)	0,333333333	0,004580153	0,004580153
G3	Se 20090101 < 0,54 e 20081202 < 0,63 e 20080913 >= 0,29 e 20090101 < 0,51 e 20081202 < 0,36 então milho (3)	0,333333333	0,004580153	0,004580153
I9	Se 20090101 >= 0,73 e 20081202 < 0,77 e 20090322 >= 0,81 e 20080913 < 0,26 e 20081202 < 0,44 então algodão (8)	0,333333333	0,01221374	0,01221374
G11	Se 20090101 >= 0,54 e 20090306 < 0,74 e 20090322 >= 0,7 e 20080912 >= 0,2 e 20090322 >= 0,2 e 20090322 < 0,71 então soja (200/3)	0,325732899	0,305343511	0,309923664
F13	Se 20090101 >= 0,71 e 20090322 >= 0,7 20090101 >= 0,9 e 20090323 >= 0,71 então soja (194)	0,319078947	0,296183206	0,296183206
G9	Se 20090101 >= 0,54 e 20090306 < 0,74 e 20090322 >= 0,7 e 20080913 < 0,2 e 20090306 >= 0,54 então algodão (9/3)	0,28125	0,013740458	0,018320611
K12	Se 20090101 >= 0,71 e 20080913 < 0,71 e 20090322 >= 0,75 e 20090322 >= 0,79 então soja (171)	0,276699029	0,261068702	0,261068702
M12	Se 20090101 >= 0,71 e 20080922 >= 0,7 e 20090322 >= 0,79 então soja (159)	0,259380098	0,242748092	0,242748092
I5	Se 20090309 < 0,73 e 20090322 >= 0,75 e 20090322 >= 0,78 então soja (158)	0,25483871	0,241221374	0,241221374
H13	Se 20090101 >= 0,71 e 20080913 < 0,67 e 20080913 >= 0,2 e 20081202 < 0,99 e 20090101 < 0,88 então soja (127/2)	0,209917355	0,19389313	0,196946565
F1	Se 20090101 < 0,71 e 20081202 < 0,62 e 20090322 < 0,8 e 20081202 < 0,48 e 20081202 < 0,36 então milho (3/1)	0,2	0,004580153	0,00610687
C9	Se 20090322 <= 0,7906 e 20090423 <= 0,8538 e 20090509 > 0,6142 e 20090101 <= 0,9584 então algodão (24)	0,186046512	0,036641221	0,036641221

J13	Se 20090101 >= 0,71 e 20081202 >= 0,51 e 20090306 >= 0,68 e 20080913 >= 0,19 então soja (115)	0,183121019	0,175572519	0,175572519
C3	Se 20081218 > 0,8086 e 20090423 <= 0,8039 então soja (67)	0,163414634	0,102290076	0,102290076
M10	Se 20090101 >= 0,71 e 20080913 <0,7 e 20090322 >= 0,75 e 20090322 < 0,71 e 20080913 >= 0,2 então soja (97/3)	0,158238173	0,148091603	0,152671756
D11	Se 20090101 >= 0,71 e 20080913 < 0,2 e 20090322 >= 0,78 e 20090306 >= 0,54 então algodão (5)	0,151515152	0,007633588	0,007633588
E8	Se 20090101 >= 0,71 e 20080913 < 0,2 e 20080913 >= 0,19 então algodão (6)	0,146341463	0,009160305	0,009160305
H11	Se 20090101 >= 0,71 e 20080913 < 0,67 e 20080913 < 0,2 e 20090322 >= 0,76 e 20090101 < 0,93 então algodão (5/2)	0,142857143	0,007633588	0,010687023
D15	Se 20090101 >= 0,71 e 20080913 >= 0,2 e 20081202 < 0,99 e 20090306 >= 0,8 e 20090306 < 0,81 então milho (1)	0,142857143	0,001526718	0,001526718
H16	Se 20090101 >= 0,71 e 20080913 >= 0,67 e 20080913 < 0,77 entâo milho (2)	0,133333333	0,003053435	0,003053435
C5	Se 20090407 > 0,7757 e 20081202 <= 0,5351 então algodão (17)	0,131782946	0,025954198	0,025954198
A2	Se 20090306 <= 0,8648 e 20090101 <= 0,7134 e 20090202 <= 0,7093 e 20090218 > 0,288509 e 20081015 <= 0,2626 então algodão (3/1)	0,125	0,004580153	0,00610687
I34	Se 20090309 < 0,73 e 20090322 >= 0,75 e 20090322 < 0,78 e 20090322 >= 0,78 então algodão (3)	0,125	0,004580153	0,004580153
E2	Se 20090101 < 0,71 e 20090101 < 0,7 e 20081202 < 0,41 e 20080913 >= 0,29 então milho (1)	0,125	0,001526718	0,001526718
J9	Se 20090101 >= 0,71 e 20081202 <0,5 e 20080913 >=0,58 então milho (1)	0,125	0,001526718	0,001526718
K2	Se 20090101 < 0,71 e 20081202 < 0,63 e 20090306 < 0,85 e 20090101 < 0,68 e 20090322 >= 0,81 então soja (3/1)	0,111111111	0,004580153	0,00610687
F11	Se 20090101 >= 0,71 e 20090322 >= 0,7 e 20090101< 0,9 e 20080913 >= 0,2 e 20081202 >= 0,62 então soja (64/2)	0,105263158	0,097709924	0,100763359
C8	Se 20081202 <= 0,5526 e 20081031 > 0,502 então milho (12)	0,103448276	0,018320611	0,018320611
I12	Se 20090306 >= 0,73 e 20081202 >= 0,77 então soja (62)	0,1	0,094656489	0,094656489
K13	Se 20090101 >= 0,71 e 20080913 >= 0,71 então milho (1)	0,1	0,001526718	0,001526718
F6	Se 20090101 < 0,71 e 20081202 >= 0,62 e 20081202 < 0,81 e 20080913 >= 0,44 então algodão (3)	0,09375	0,004580153	0,004580153
F9	Se 20090101 >= 0,71 e 20090322 >= 0,7 e 20090101 < 0,9 e 20080913 < 0,2 então algodão (3)	0,09375	0,004580153	0,004580153
C7	Se 20090306 <= 0,7888 e 20090423 <= 0,8039 então soja (35)	0,085365854	0,053435115	0,053435115
A1	Se 20090306 <= 0,8648 e 20090101 <= 0,7134 e 20090202 <= 0,7093 e 20090218 <= 0,288509 então algodão (2)	0,083333333	0,003053435	0,003053435
G17	Se 20090101 >= 0,54 e 20090306 >= 0,74 e 20081202 >= 0,67 e 20081202 >= 0,77 então soja (49)	0,07980456	0,07480916	0,07480916
K11	Se 20090101 >= 0,71 e 20080913 <0,71 e 20090322 >= 0,75 e 20090322 < 0,79 e 20080913 >= 0,19 então soja (49/1)	0,079288026	0,07480916	0,076335878
L5	Se 20090101 >= 0,48 e 20090101 < 0,81 e 20090306 < 0,74 então soja (49)	0,078904992	0,07480916	0,07480916
C4	Se 20090407 > 0,7757 e 20090509 > 0,82 então soja (28)	0,068292683	0,042748092	0,042748092
B2	Se 200903022 <= 0,8141 e 20090306 <= 0,7173 então soja (46/1)	0,066957787	0,063186813	0,06456044

F10	Se 20090101 >= 0,71 e 20090322 >= 0,7 e 20090101 < 0,9 e 20080913 >= 0,2 e 20081202 < 0,62 então milho (1)	0,066666667	0,001526718	0,001526718
M11	Se 20090101 >= 0,71 e 20080913 <0,7 e 20090322 >= 0,75 e 20080913 >= 0,71 então milho (1)	0,066666667	0,001526718	0,001526718
G10	Se 20090101 >= 0,54 e 20090306 < 0,74 e 20090322 >= 0,7 e 20080913 >= 0,2 e 20090322 < 0,71 então algodão (2)	0,0625	0,003053435	0,003053435
G16	Se 20090101 >= 0,54 e 20090306 >= 0,74 e 20081202 >= 0,67 e 20081202 < 0,77 e 20081202 >= 0,75 então algodão (2)	0,0625	0,003053435	0,003053435
C10	Se 20090322 <= 0,7906 e 20090423 <= 0,8538 e 20090509 > 0,6142 e 20090101 <= 0,9584 então soja (25)	0,06097561	0,038167939	0,038167939
D17	Se 20090101 >= 0,71 e 20080913 >= 0,2 e 20081202 >= 0,99 então algodão (2)	0,060606061	0,003053435	0,003053435
J5	Se 20090101 < 0,71 e 20081202 > 0,63 e 20080913 <0.44 então soja (37)	0,058917197	0,05648855	0,05648855
J6	Se 20090101 < 0,71 e 20081202 > 0,63 e 20080913 >=0.44 e 20080913 < 0,5 então algodão (1)	0,052631579	0,001526718	0,001526718
J11	Se 20090101 >= 0,71 e 20081202 >= 0,51 e 20090306 >= 0,68 e 20080913 < 0,19 e 20090101 <0,93 então algodão (1)	0,052631579	0,001526718	0,001526718
A3	Se 20090306 <= 0,8648 e 20090101 <= 0,7134 e 20090202 <= 0,7093 e 20090218 > 0,288509 e 20081015 > 0,2626 então soja (32)	0,051446945	0,048854962	0,048854962
H7	Se 20090101 < 0,71 e 20081202 >= 0,63 e 20080913 < 0,46 então soja (29)	0,047933884	0,044274809	0,044274809
K8	Se 20090101 < 0,71 e 20081202 >= 0,63 e 20081202 >= 0,85 então soja (28)	0,045307443	0,042748092	0,042748092
D8	Se 20090101 < 0,71 e 2001202 >= 0,8 então soja (27)	0,043902439	0,041221374	0,041221374
E7	Se 20090101 >= 0,71 e 20080913 < 0,2 e 20080913 < 0,19 então soja (26)	0,04290429	0,039694656	0,039694656
F7	Se 20090101 < 0,71 e 20081202 >= 0,62 e 20081202 >= 0,81então soja (26)	0,042763158	0,039694656	0,039694656
D14	Se 20090101 >= 0,71 e 20080913 >= 0,2 e 20081202 < 0,99 e 20090306 < 0,8 e 20080913 >= 0,41 então soja (25/1)	0,040650407	0,038167939	0,039694656
K7	Se 20090101 < 0,71 e 20081202 >= 0,63 e 20081202 < 0,8 e 20080913 >= 0,44 então algodão (1)	0,037037037	0,001526718	0,001526718
K10	Se 20090101 >= 0,71 e 20080913 <0,71 e 20090322 >= 0,75 e 20090322 < 0,79 e 20080913 < 0,19 então algodão (1)	0,037037037	0,001526718	0,001526718
L2	Se 20090101 < 0,48 e 20090322 < 0,88 e 20090322 >= 0,83 e 20091202 < 0,65 então algodão (1)	0,037037037	0,001526718	0,001526718
L13	Se 20090101 >= 0,48 e 20090101 >= 0,81 e 20090322 >= 0,7 e 20090101 >= 0,88 e 20081202 >= 0,99 então algodão (1)	0,037037037	0,001526718	0,001526718
J8	Se 20090101 >= 0,71 e 20081202 <0,5 e 20080913 <0,58 então soja (23)	0,036624204	0,035114504	0,035114504
D16	Se 20090101 >= 0,71 e 20080913 >= 0,2 e 20081202 < 0,99 e 20090306 >= 0,8 e 20090306 >= 0,81 então soja(22)	0,035772358	0,033587786	0,033587786
10	Se 20090101 >= 0,48 e 20090101 >= 0,81 e 20090322 >= 0,7 e 20090101 < 0,88 e 20080913 < 0,38 então soja (22)	0,035426731	0,033587786	0,033587786
E3	Se 20090101 < 0,71 e 20090101 < 0,7 e 20081202 >= 0,41 e 20090306 < 0,72 então soja (21)	0,034653465	0,032061069	0,032061069
F12	Se 20090101 >= 0,71 e 20090322 >= 0,7 20090101 >= 0,9 e 20090322 < 0,71 então algodão (1)	0,03125	0,001526718	0,001526718
I2	Se 20090309 < 0,73 e 20090322 >= 0,75 e 20090322 < 0,78 e 20090322 < 0,78 e 20080913 < 0,41 então soja (18)	0,029032258	0,027480916	0,027480916
H9	Se 20090101 < 0,71 e 20081202 >= 0,63 e 20080913 >= 0,46 e 20090101 >= 0,59 então algodão (1)	0,028571429	0,001526718	0,001526718
H15	Se 20090101 >= 0,71 e 20080913 < 0,67 e 20080913 >= 0,2 e 20081202 >= 0,99 entâo algodão (1)	0,028571429	0,001526718	0,001526718
H10	Se 20090101 >= 0,71 e 20080913 < 0,67 e 20080913 < 0,2 e 20090322 < 0,76 então soja (17)	0,028099174	0,025954198	0,025954198

M3	Se 20090101 < 0,71 e 20090322 < 0,81 e 20090322 >= 0,73 então soja (17)	0,027732463	0,025954198	0,025954198
B4	Se 20081202 > 0,63206 então soja (19/1)	0,027656477	0,026098901	0,027472527
D12	Se 20090101 >= 0,71 e 20080913 < 0,2 e 20090101 >= 0,93 então soja (17)	0,027642276	0,025954198	0,025954198
L8	Se 20090101 >= 0,48 e 20090101 < 0,81 e 20080306 >= 0,74 e e 20081202 >= 0,63 então soja (17)	0,027375201	0,025954198	0,025954198
E10	Se 20090101 >= 0,71 e 20080913 >= 0,2 e 20081202 >= 0,99 então algodão (1)	0,024390244	0,001526718	0,001526718
E5	Se 20090101 < 0,71 e 20090101 < 0,7 e 20081202 >= 0,41 e 20090306 >= 0,72 e 20081202 >= 0,81 então soja (13)	0,021452145	0,019847328	0,019847328
G6	Se 20090101 < 0,54 e 20081202 >= 0,63 então soja (13)	0,021172638	0,019847328	0,019847328
G8	Se 20090101 >= 0,54 e 20090306 < 0,74 e 20090322 >= 0,7 e 20080913 < 0,2 e 20090306 < 0,54 então soja (13)	0,020833333	0,019548872	0,019548872
M7	Se 20090101 < 0,71 e 20090322 >= 0,81 e 20081202 >= 0,82 então soja (12)	0,019575856	0,018320611	0,018320611
L6	Se 20090101 >= 0,48 e 20090101 < 0,81 e 20080306 >= 0,74 e 20081202 < 0,63 e 20090101 < 0,69 então soja (12/2)	0,019323671	0,018320611	0,021374046
D2	Se 20090101 < 0,71 e 20081202 < 0,8 e 20090322 < 0,81 e 20090101 < 0,69 e 20081202 >= 0,32 então soja (11)	0,017886179	0,016793893	0,016793893
L1	Se 20090101 < 0,48 e 20090322 < 0,88 e 20090322 < 0,83 então soja (11)	0,017713366	0,016793893	0,016793893
A9	Se 20090306 > 0,8648 e 20081202 > 0,6228 então soja (11)	0,017684887	0,016793893	0,016793893
G12	Se 20090101 >= 0,54 e 20090306 >= 0,74 e 20081202 < 0,67 e 20080913 < 0,25 então soja (10)	0,016286645	0,015267176	0,015267176
H12	Se 20090101 >= 0,71 e 20080913 < 0,67 e 20080913 < 0,2 e 20090322 >= 0,76 e 20090101 >= 0,93 então soja (9)	0,014876033	0,013740458	0,013740458
M1	Se 20090101 < 0,71 e 20090322 < 0,81 e 20090322 < 0,73 e 20090306 < 0,71 então soja (9)	0,014681892	0,013740458	0,013740458
K1	Se 20090101 < 0,71 e 20081202 < 0,63 e 20090306 < 0,85 e 20090101 < 0,68 e 20090322 < 0,81 então soja (9/2)	0,014563107	0,013740458	0,016793893
F5	Se 20090101 < 0,71 e 20081202 >= 0,62 e 20081202 < 0,81 e 20080913 < 0,44 então soja (8)	0,013157895	0,01221374	0,01221374
M9	Se 20090101 >= 0,71 e 20090322 < 0,7 e 20090322 < 0,79 e 20080913 < 0,71 e 20080913 < 0,2 então soja (7/3)	0,01141925	0,010687023	0,015267176
G15	Se 20090101 >= 0,54 e 20090306 >= 0,74 e 20081202 >= 0,67 e 20081202 < 0,77 e 20081202 < 0,75 então soja (7)	0,011400651	0,010687023	0,010687023
K6	Se 20090101 < 0,71 e 20081202 >= 0,63 e 20081202 < 0,8 e 20080913 < 0,44 então soja (7)	0,011326861	0,010687023	0,010687023
I6	Se 20090309 >= 0,73 e 20081202 < 0,77 e 20090322 < 0,81 e 20081202 < 0,48 então soja (7)	0,011290323	0,010687023	0,010687023
I8	Se 20090101 >= 0,73 e 20081202 < 0,77 e 20090322 < 0,81 e 20081202 >= 0,48 e 20081202 >= 0,62 então soja (7)	0,011290323	0,010687023	0,010687023
A5	Se 20090306 <= 0,8648 e 20090101 <= 0,7134 e 20090202 > 0,7093 e 20090101 <= 0,6761 e 20081015 <= 0,7411 então soja (7)	0,011254019	0,010687023	0,010687023
J1	Se 20090101 < 0,71 e 20081202 < 0,63 e 20090322 < 0,8 e 20080913 < 0,26 então soja (7)	0,011146497	0,010687023	0,010687023
B5	Se 20081202 <= 0,4775 então soja (7)	0,010189229	0,009615385	0,009615385
H1	Se 20090101 < 0,71 e 20081202 < 0,63 e 20090322 < 0,59 então soja (6)	0,009917355	0,009160305	0,009160305
D9	Se 20090101 >= 0,71 e 20080913 < 0,2 e 20090322 < 0,78 então soja (6)	0,009756098	0,009160305	0,009160305

F2	Se 20090101 < 0,71 e 20081202 < 0,62 e 20090322 < 0,8 e 20081202 < 0,48 e 20081202 >= 0,36 então soja (5)	0,008223684	0,007633588	0,007633588
M6	Se 20090101 < 0,71 e 20090322 >= 0,81 e 20081202 < 0,82 e 20090101 >= 0,21 e 20081202 >= 0,64 então soja (4/2)	0,006525285	0,00610687	0,009160305
D1	Se 20090101 < 0,71 e 20081202 < 0,8 e 20090322 < 0,81 e 20090101 < 0,69 e 20081202 < 0,32 então soja (4/1)	0,006504065	0,00610687	0,007633588
I10	Se 20090101 >= 0,73 e 20081202 < 0,77 e 20090322 >= 0,81 e 20080913 < 0,26 e 20081202 >= 0,44 então soja (4/2)	0,006451613	0,00610687	0,009160305
H2	Se 20090101 < 0,71 e 20081202 < 0,63 e 20090322 >= 0,59 e 20080913 < 0,25 então soja (3)	0,004958678	0,004580153	0,004580153
H8	Se 20090101 < 0,71 e 20081202 >= 0,63 e 20080913 >= 0,46 e 20090101 < 0,59 então soja (3)	0,004958678	0,004580153	0,004580153
G5	Se 20090101 < 0,54 e 20081202 < 0,63 e 20080913 >= 0,29 e 20090101 >= 0,51 então soja (3)	0,004885993	0,004580153	0,004580153
D3	Se 20090101 < 0,71 e 20081202 < 0,8 e 20090322 < 0,81 e 20090101 >= 0,69 e 20090306 < 0,62 então soja (3)	0,004878049	0,004580153	0,004580153
I3	Se 20090309 < 0,73 e 20090322 >= 0,75 e 20090322 < 0,78 e 20090322 < 0,78 e 20080913 >= 0,41 então soja (3/1)	0,00483871	0,004580153	0,00610687
B7	Se 20090306 > 0,5921 então soja (3/1)	0,004366812	0,004120879	0,005494505
G4	Se 20090101 < 0,54 e 20081202 < 0,63 e 20080913 >= 0,29 e 20090101 < 0,51 e 20081202 >= 0,36 então soja (2)	0,003257329	0,003053435	0,003053435
G14	Se 20090101 >= 0,54 e 20090306 >= 0,74 e 20081202 < 0,67 e 20080913 >= 0,25 e 20090101 >= 0,81 então soja (2)	0,003257329	0,003053435	0,003053435
K4	Se 20090101 < 0,71 e 20081202 < 0,63 e 20090306 >= 0,85 e 20090101 < 0,21 então soja (2)	0,003236246	0,003053435	0,003053435
H6	Se 20090101 < 0,71 e 20081202 < 0,63 e 20090322 >= 0,80 e 20080913 >=0,51 então soja (1)	0,001652893	0,001526718	0,001526718
H17	Se 20090101 >= 0,71 e 20080913 >= 0,67 e 20080913 >= 0,77 então soja (1)	0,001652893	0,001526718	0,001526718
M4	Se 20090101 < 0,71 e 20090322 >= 0,81 e 20081202 < 0,82 e 20090101 < 0,21 então soja (1)	0,001631321	0,001526718	0,001526718
G1	Se 20090101 < 0,54 e 20081202 < 0,63 e 20080913 < 0,29 e 20090306 < 0,57 então soja (1)	0,001628664	0,001526718	0,001526718
D5	Se 20090101 < 0,71 e 20081202 < 0,8 e 20090322 >= 0,81 e 20080913 < 0,62 e 20090101 < 0,21 então soja (1)	0,001626016	0,001526718	0,001526718
D7	Se 20090101 < 0,71 e 20081202 < 0,8 e 20090322 >= 0,81 e 20080913 >= 0,62 então soja (1)	0,001626016	0,001526718	0,001526718
D10	Se 20090101 >= 0,71 e 20080913 < 0,2 e 20090322 >= 0,78 e 20090306 < 0,54 então soja (1)	0,001626016	0,001526718	0,001526718
L3	Se 20090101 < 0,48 e 20090322 < 0,88 e 20090322 >= 0,83 e 20091202 >= 0,65 então algodão (1)	0,001610306	0,001526718	0,001526718
J2	Se 20090101 < 0,71 e 20081202 < 0,63 e 20090322 < 0,8 e 20080913 >= 0,26 e 20080322 < 0,59 então soja (1)	0,001592357	0,001526718	0,001526718
J7	Se 20090101 < 0,71 e 20081202 > 0,63 e 20080913 >=0.44 e 20080913 >= 0,5 então soja (1)	0,001592357	0,001526718	0,001526718
J12	Se 20090101 >= 0,71 e 20081202 >= 0,51 e 20090306 >= 0,68 e 20080913 < 0,19 e 20090101 >= 0,93 então soja (1)	0,001592357	0,001526718	0,001526718

Tabela 4.47: Tabela com todas as regras ranqueadas

De acordo com as análises realizadas com as matrizes de contingência com cada selecionador de atributo, têm-se as melhores regras de cada teste. Essas regras são apresentadas na Tabela 4.48. Verificando também quais são as regras em comum dados pelos classificadores:

Algoritmo	Melhores regras separadas por classificadores	
PCA	R4	Se 20090306 <= 0,8648 e 20090101 > 0,7134 então soja (572/72)
	R7	Se 20090306 <= 0,8648 e 20090101 <= 0,7134 e 20090202 > 0,7093 e 20090101 > 0,6761 então milho (6)
	R8	Se 20090306 > 0,8648 e 20081202 <= 0,6228 então algodão (19/1)
CFS A	R1	Se 20090306 <= 0,865337 e 20081218 > 0,7067 então soja (612/7)
	R3	Se 200900101 <= 0,5635 e 20081202 <= 0,6475 então algodão (36/5)
	R6	Se 20090306 > 0,5921 então milho (5)
Gain Ratio	R1	Se 20090101 > 0,8495 e 20081202 <= 0,986 e 20090218 > 0,3052 e
	R2	Se 200900424 > 0,8039 e 20081202 <= 0,7814 e 20090525 <= 0,853 e 20090525 > 0,652 então algodão (88)
	R6	Se 20090101 <= 0,8599 e 20090306 > 0,6535 e 20081031 <= 0,6705 e 20090218 <= 0,8646 e 20090306 <= 0,8013 e 20081116 <= 0,571 então milho (104/1)
Wrapper 1	R4	Se 20090101 < 0,71 e 20081202 < 0,8 e 20090322 < 0,81 e 20090101 >= 0,69 e 20090306 >= 0,62 então milho (6)
	R6	Se 20090101 < 0,71 e 20081202 < 0,8 e 20090322 >= 0,81 e 20080913 < 0,62 e 20090101 >= 0,21 então algodão (26/1)
	R13	Se 20090101 >= 0,71 e 20080913 >= 0,2 e 20081202 < 0,99 e 20090306 < 0,8 e 20080913 < 0,41 então soja (497)
wrapper 2	R4	Se 20090101 < 0,71 e 20090101 < 0,7 e 20081202 >= 0,41 e
	R6	Se 20090101 < 0,71 e 20090101 >= 0,7 então milho (7)
	R9	Se 20090101 >= 0,71 e 20080913 >= 0,2 e 20081202 < 0,99 então soja (546)
Wrapper 3	R3	Se 20090101 < 0,71 e 20081202 < 0,62 e 20090322 < 0,8 e 20081202 >= 0,48 então milho (11)
	R4	Se 20090101 < 0,71 e 20081202 < 0,62 e 20090322 >= 0,8 então algodão (25)
	R8	Se 20090101 >= 0,71 e 20090322 < 0,7 então soja (311)
	R13	Se 20090101 >= 0,71 e 20090322 >= 0,7 20090101 >= 0,9 e 20090323 >= 0,71 então soja (194)

Wrapper 5	**R3**	Se 20090101 < 0,71 e 20081202 < 0,63 e 20090322 >= 0,59 e 20080913 >= 0,25 então milho (13)
	R4	Se 20090101 < 0,71 e 20081202 < 0,63 e 20090322 >= 0,80 e 20080913 <0,51 e 20080913 < 0,22 então algodão (12/1)
	R12	Se 20090101 >= 0,71 e 20080913 < 0,67 e 20080913 < 0,2 e 20090322 >= 0,76 e 20090101 >= 0,93 então soja (9)
	R13	Se 20090101 >= 0,71 e 20080913 < 0,67 e 20080913 >= 0,2 e 20081202 < 0,99 e 20090101 < 0,88 então soja (127/2)
Wrapper 6	**R1**	Se 20090309 < 0,73 e 20090322 < 0,75 então soja (361)
	R7	Se 20090101 >= 0,73 e 20081202 < 0,77 e 20090322 < 0,81 e 20081202 >= 0,48 e 20081202 < 0,62 então milho (11)
	R9	Se 20090101 >= 0,73 e 20081202 < 0,77 e 20090322 >= 0,81 e 20080913 < 0,26 e 20081202 < 0,44 então algodão (8)
	R10	Se 20090101 >= 0,73 e 20081202 < 0,77 e 20090322 >= 0,81 e 20080913 < 0,26 e 20081202 >= 0,44 então soja (4/2)
	R11	Se 20090306 >= 0,73 e 20081202 < 0,77 e 20090322 >= 0,81 e 20080913 >= 0,26 então soja (13)
Wrapper 7	**R3**	Se 20090101 < 0,71 e 20081202 < 0,63 e 20090322 < 0,8 e 20080913 >= 0,26 e 20080322 >= 0,59 então milho (7)
	R4	Se 20090101 < 0,71 e 20081202 < 0,63 e 20090322 >= 0,8 então algodão (17)
	R10	Se 20090101 >= 0,71 e 20081202 >= 0,51 e 20090306 < 0,68 então soja (443)
Wrapper 8	**R3**	Se 20090101 < 0,48 e 20090322 < 0,88 e 20090322 >= 0,83 e 20091202 >= 0,65 então algodão (1)
	R5	Se 20090101 >= 0,48 e 20090101 < 0,81 e 20090306 < 0,74 então soja (49)
	R9	Se 20090101 >= 0,48 e 20090101 >= 0,81 e 20090322 ,< 0,7 então soja (275)
	R12	Se 20090101 >= 0,48 e 20090101 >= 0,81 e 20090322 >= 0,7 e 20090101 >= 0,88 e 20081202 < 0,99 então soja (234/2)
Wrapper 9	**R4**	Se 20090101 < 0,71 e 20081202 < 0,63 e 20090306 >= 0,85 e 20090101 < 0,21 então soja (2)
	R7	Se 20090101 < 0,71 e 20081202 >= 0,63 e 20081202 < 0,8 e 20080913 >= 0,44 então algodão (1)
	R9	Se 20090101 >= 0,71 e 20080913 < 0,71 e 20090322 < 0,75 então soja (352)
	R11	Se 20090101 >= 0,71 e 20080913 <0,71 e 20090322 >= 0,75 e 20090322 < 0,79 e 20080913 >= 0,19 então soja (49/1)
	R12	Se 20090101 >= 0,71 e 20080913 < 0,71 e 20090322 >= 0,75 e 20090322 >= 0,79 então soja (171)

Wrapper 9	R4	Se 20090101 < 0,71 e 20081202 < 0,63 e 20090306 >= 0,85 e 20090101 < 0,21 então soja (2)
	R7	Se 20090101 < 0,71 e 20081202 >= 0,63 e 20081202 < 0,8 e 20080913 >= 0,44 então algodão (1)
	R9	Se 20090101 >= 0,71 e 20080913 < 0,71 e 20090322 < 0,75 então soja (352)
	R11	Se 20090101 >= 0,71 e 20080913 <0,71 e 20090322 >= 0,75 e 20090322 < 0,79 e 20080913 >= 0,19 então soja (49/1)
	R12	Se 20090101 >= 0,71 e 20080913 < 0,71 e 20090322 >= 0,75 e 20090322 >= 0,79 então soja (171)
Wrapper 10	R2	Se 20090101 < 0,71 e 200900233 < 0,81 e 20090322 < 0,73 e 20090306 >= 0,71 então milho (14)
	R5	Se 20090101 < 0,71 e 20090322 >= 0,81 e 20081202 < 0,82 e 20090101 >= 0,21 e 20081202 < 0,64 então algodão (27)
	R8	Se 20090101 >= 0,71 e 20090322 < 0,7 então soja (307)
	R12	Se 20090101 >= 0,71 e 20080922 >= 0,7 e 20090322 >= 0,79 então soja (159)

Tabela 4.48: Melhores regras geradas pelos classificadores

De acordo com a Tabela 4.48, verifica-se que apenas a regra R8, gerada pelo Random Forest com o Wrapper nas árvores 3 e 4 se repete em todo o banco de dados. Ou seja, conseguiu se montar um banco de dados com regras distintas entre si exceto pela regra R8.

4.3 Mapas para interpretação das regras

Diante dos bons resultados que o primeiro teste (J48 com PCA) mostrou, foi gerado um mapa do Estado de Mato Grosso, para verificar se os resultados dados por ele, realmente eram condizentes com a realidade.

Então, a Figura 4.1 mostra o plantio da soja, milho e algodão no Estado de Mato Grosso. As áreas descritas na imagem fazem referência as culturas cultivadas e detectadas palo algoritmo na região. A área em verde faz referência ao plantio de soja, a área em azul refere-se ao algodão e a laranja ao milho. Segundo o especialista da Embrapa, pode ser visto que o modelo consegue descrever com facilidade as áreas em que a soja é cultivada, há também alguns vestigios do algodão. Percebe-se que essa distinção é devido a predominância da cultura soja na região. Pode-se notar também a não detecção do milho, isso se deve ao fato das poucas representações do mesmo no conjunto de dados.

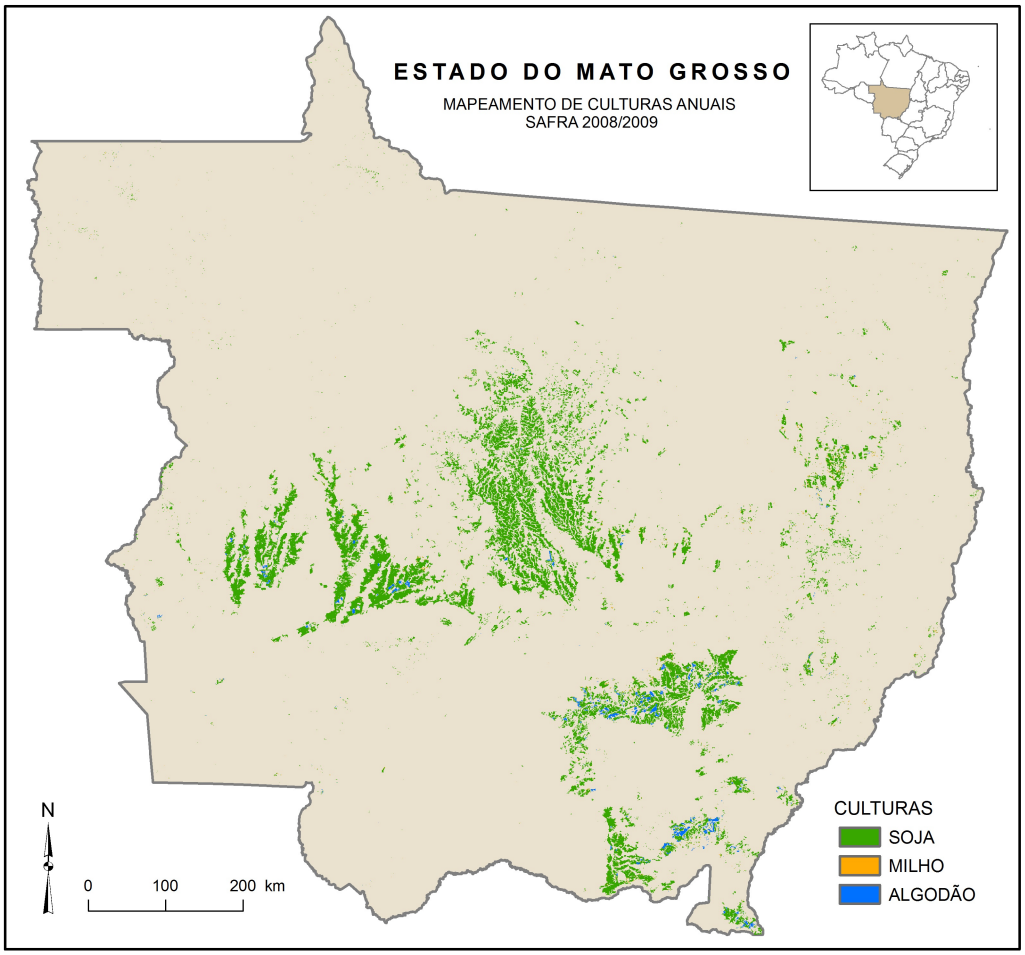

Figura 4.1: Mapa Geral de Mato Grosso

E na Figura 4.2, pode se constatar uma área de plantio de algodão ao sudeste da região. Essa região detectada pelo modelo não fazia parte das áreas estudadas. Então foi necessário que o especialista verifica-se se a região detectada era realmente de cultivo de algodão, o que foi verificado.

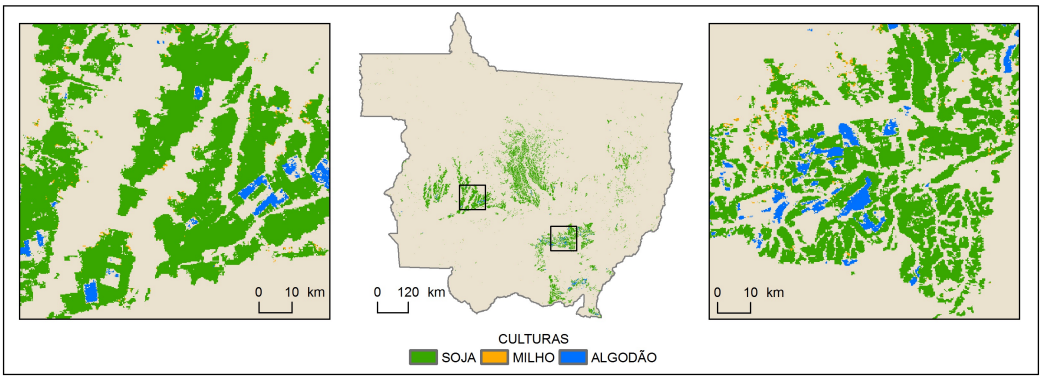

Figura 4.2: Áreas de plantio de algodão em Mato Grosso

4.4 Modelo para interpretação das regras de classificação

A partir do modelo preditivo, a função Visualize Tree do Weka foi utilizado com a finalidade de se obter a árvore de decisaão explicita gerada pelo primeiro modelo. Para analisar o modelo as regras foram analisadas de forma geral, buscando compor um cenário propício de interpretação para as coberturas dos solos através do NDVI.

A Figura 4.3 representa a árvore de decisão gerada pelo J48 em conjunto com o PCA:

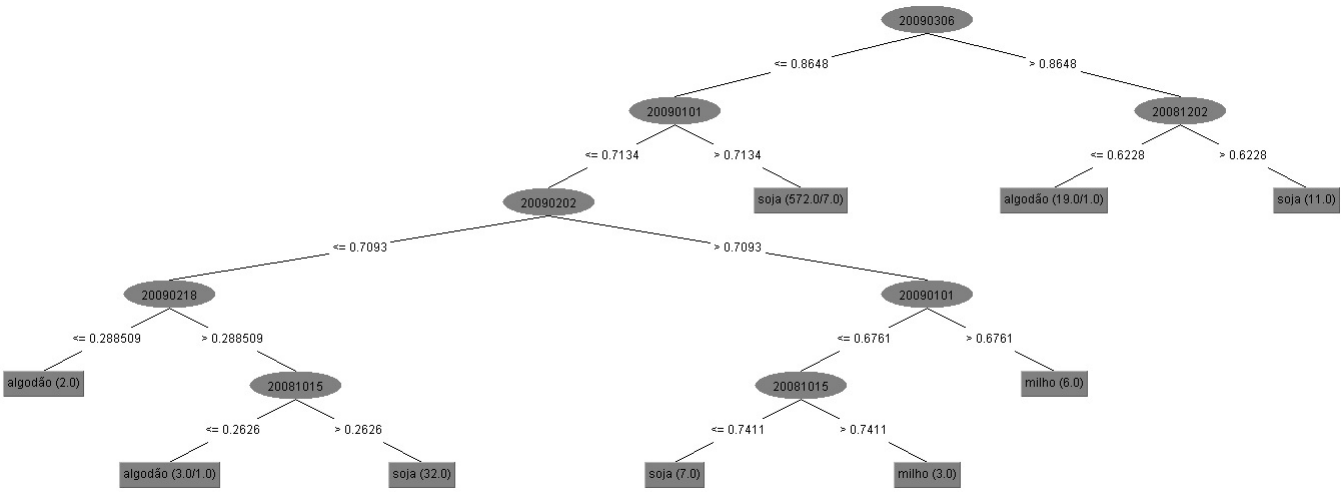

Figura 4.3: Árvore de decisão gerada com o J48 e o PCA

Para a análise das regras, deixa-se claro os períodos de safra cada cultura:

- A soja é plantada entre os meses de Setembro a Outubro e sua colheita é feita entre os meses de Fevereiro a Março.

- O algodão é plantado entre os meses de Novembro e Dezembro e sua colheita é feita entre os meses de Abril e Maio.

- O milho é plantado entre os meses de Janeiro e Fevereiro e sua colheita é feita no mês de Maio (Safrinha).

Em todos os teste realizados, o atributo 20090306 sempre foi selecionado como sendo a raiz da árvore, isso mostra que o atributo possui um alto valor de Ganho de informação. Segundo o especialista de Embrapa, o atributo é uma data importante para a diferenciação do algodão.

Seguindo a árvore para seu ramo da direta, tem-se algodão e soja. Para valores de NDVI maiores que 0,8648 e 0,6228 no mês de Dezembro a regra dá tem na sua folha a soja, onde passaram 11 observações e nenhuma delas foram registradas incorretamente. Já para valores maiores que 0,8648 e menores ou iguais à 0,6228 no mês de Dezembro a regra dá o millho, onde passaram 19 observações e apenas 1 observação foi classificada incorretamente.

Indo para o ramo da esquerda têm-se, para valores de NDVI menores ou iguais à 0,8648 uma folha com soja, onde passaram 572 observações e apenas 7 delas foram registradas incorretamente.

Indo mais para a esquerda, estando na data 20090202 (mês de Fevereiro) para valores de NDVI maiores que 0,7093 e em 20090101 (mês de Janeiro) com NDVI maior que 0,6761 têm-se uma folha com o milho, onde passaram 6 observações sendo todas as classificções corretas. Para a mesma data, mas com NDVI menores ou iguais à 0,6761, têm-se e 20081015 (Outubro) e NDVI maiores que 0,7411, têm-se uma folha com milho, que passaram 3 observações sendo todas corrretas. E para 20081015 com NDVI menores ou iguais à 0,7411 têm-se uma folha com o soja, onde passaram 7 observações sendo também todas as classificções corretas.

Seguindo para o próximo ramo, para a data de 20090210 (Fevereiro) com NDVI maiores que 0,288509 e para a data 20081015(Outubro) com NDVI maiores que 0,2626, existe uma folha com soja, onde passaram 32 observações onde todas foram classificadas corretamente. Para as mesmas datas, mas com NDVI menores ou iguais à 0,2626, existe uma folha com algodão, onde passaram 3 observações sendo 1 classificada incorretamente.

E no último ramo, para a data de 20090210 (Fevereiro) para NDVI menores ou iguais a 0,288509, têm-se uma folha com algodão, onde passaram 2 observações sendo também todas as classificações corretas.

De fato, esses resultados são visivelmente identificados na Tabela 4.49, pois as regras estão restritas ao período de Dezembro a Março. Para essa verificação, as regras foram separadas por datas (ou períodos). A intenção é verificar quais são os meses em que ocorre maior quantidade de classificação.

Analise das regras por data		
Data		quantidade
Setembro	20080913	0
	20080929	0
Outubro	20081015	0
	20081031	0
Novembro	20081116	0
Dezembro	20081202	3
	20081218	1
Janeiro	20090101	141
	20090117	0
Fevereiro	20090202	0
	20090218	0
Março	20090306	15
	20090322	2
Abril	20090407	2
	20090423	0
Maio	20090509	0
	20090525	0
Junho	20090610	0
	20090626	0
Julho	20090712	0
	20090728	0
Agosto	20090813	0
	20090829	0

Tabela 4.49: Análise das regras por data

A não presença de regras de classificaçâo de Abril até Agosto é justificado pelo fato desse banco de dados estar incluso o período da Safrinha. Esse período está diretamente associado as culturas da Safrinha que são cultivadas após o período de Safra, para a proteção e o preparo do solo. Por esse motivo foi fixado o período de Setembro de 2008 a Março de 2009.

A grande quantidade de regras em Janeiro, é pelo fato de ser o pico do cultivo da soja, que por sua vez possui a maior quantidade de informações no conjunto de dados.

Então, de acordo com as Tabelas 4.49 e 4.50, é possível verificar que as melhores regras dadas pela Tabela 4.48 estão realmente dentro do período chamado de ano-safra, como sugeriu o especialista.

Classificador	Verificação do período Safra	
PCA	**R4**	Se março e janeiro então soja
	R7	Se março, janeiro, fevereiro então milho
	R8	Se março, dezembro então algodão
CFS	**R1**	Se março, dezembro então soja
	R3	Se janeiro, dezembro então algodão
	R6	Se março então milho
GR	**R1**	Se janeiro, dezembro, fevereiro, maio então soja
	R2	Se abril, dezembro, maio então algodão
	R6	Se janeiro, março, outubro, fevereiro, novembro então milho
Wrapper 1	**R4**	Se janeiro, dezembro, março então milho
	R6	Se janeiro, dezembro, março, setembro então algodão
	R13	Se janeiro, setembro, dezembro, março então soja
Wrapper 2	**R6**	Se janeiro então milho
	R9	Se janeiro, setembro, dezembro então soja
Wrapper 3	**R3**	Se janeiro, dezembro, março então milho
	R4	Se janeiro, dezembro, março então algodão
	R8	Se janeiro, março então soja
	R13	Se janeiro, março então soja
Wrapper 4	**R2**	Se janeiro, dezembro, setembro, março então algodão
	R7	Se janeiro, março então soja
	R11	Se janeiro, março, setembro então soja
	R13	Se janeiro, março, dezembro, setembro então milho
Wrapper 5	**R3**	Se janeiro, dezembro, março, setembro então milho
	R5	Se janeiro, dezembro, março, setembro então algodão
	R14	Se janeiro, setembro, dezembro então soja

Wrapper 6	**R1**	Se março soja
	R7	Se janeiro, dezembro, março então milho
	R9	Se janeiro, dezembro, março, setembro então algodão
	R11	Se março, dezembro, setembro então soja
Wrapper 7	**R3**	Se janeiro, dezembro, março, setembro então milho
	R4	Se janeiro, dezembro, março então algodão
	R10	Se janeiro, dezembro, março então soja
Wrapper 8	**R3**	Se janeiro, dezembro março então milho
	R5	Se janeiro, dezembro, março então algodão
	R9	Se janeiro, setembro, março então soja
	R12	Se janeiro, setembro, março então soja
Wrapper 9	**R4**	Se janeiro, março então algodão
	R7	Se janeiro, março, dezembro então milho
	R9	Se janeiro, março então soja
	R12	Se janeiro, março, dezembro então soja
Wrapper 10	**R2**	Se janeiro, fevereiro, março então milho
	R5	Se janeiro, março, dezembro então algodão
	R8	Se janeiro, março então soja
	R12	Se janeiro, setembro, março então soja

Tabela 4.50: Verificação da safra por período

Capítulo 5

Conclusões

O trabalho constituiu-se na geração de regras de classificação obtidas a partir de modelos preditivos, desenvolvidos por meio de ténicas de mineração de dados para a classificação de perfis de NDVI das culturas de algodão, soja e milho, no estado de Mato Grosso. O banco de dados era composto de centenas de registros de campo, com o objetivo de identificar coberturas de solo e gerar uma base de conhecimento para cada cultura estudada.

Para cada cultura, vários modelos foram gerados, considerando intervalos de NDVI do ano safra de 2008/2009, confirmando a hipótese inicial do trabalho. Os resultados obtidos através das séries temporais de NDVI geraram informações com potencial de uso em modelos agrometeorológicos-espectrais.

Verificou-se a possibilidade de classificar as culturas de: soja, milho e algodão através dos valores de NDVI, dados pelas regras de classificação, validando os objetivos desse trabalho.

As regras de classifição que deram origem à base de conhecimento foram ranqueadas de acordo com os melhores valores da sensitividade, do suporte e da cobertura; métricas tradicionais recomendadas pela literatura para este fim.

Foi gerado um mapa de Mato Grosso para verificar a compatibilidade das regras dadas pelo modelo, o algoritmo J48. Foi possível verificar a compatibilidade das culturas cultivadas na região (Figura 4.1) com as culturas descritas pelas regras do modelo, principalmente da soja. Com isso, também foi possível identificar a existência de uma nova região de cultivo de algodão no Sudeste de Mato Grosso, conforme ilustrado na Figura 4.2.

Segundo análise de um especialista da Embrapa, a data 20090306 é uma data importante para a diferenciação do algodão. No entanto, é importante considerar que o conjunto de dados possui algumas incertezas, por ser coletado na forma de censo e por não haver registros fiéis.

Da mesma forma, os dados meteorológicos utilizados podem não refletir fielmente as condições de NDVI de cada talhão, visto que o NDVI é dado pelo centróide de cada talhão e não de cada gleba analisada.

Uma das dificuldades para o desenvolvimento do trabalho foi a falta de dados mais antigos para a formação de um grande banco de dados, um dos requisitos essênciais à mineração de dados. Outra dificuldade foi a falta de dados atuais para uma comparação com os resultados obtidos nos modelos gerados.

Não se pode dizer que as regras encontradas e descritas neste trabalho descrevem com exatidão

as culturas existentes nesse período, mas é possível afirmar que elas contribuem com fortes indícios para isso.

Este trabalho representa uma contribuição na utilização de um processo de mineração de dados, para explorar coberturas de solos, sendo inédito nessa linha de pesquisa.

A seguir são listadas algumas possibilidades de continuidade deste trabalho:

- Utilizar os modelos em dados mais recentes de Mato Grosso, para confirmar sua legitimidade.
- Testar esses modelos com dados de outras regiões do Brasil, para verificar sua aderência com outras regiões.
- Analisar conjuntos de dados atuais, considerando a safra e a safrinha juntas.
- A metodologia pode ser aprimorada para a identificação de outros tipos de culturas importantes em outras regiões do país. Isso será possível desde que, exista dados de referência da localização das culturas para o treinamento da árvore de decisão.

Referências Bibliográficas

[1] Ndvi. Disponível em `http://www.earthobservatory.nasa.gov/Features/MeasuringVegetation/measuring_vegetation_2.php`.Acesso em 2 Maio 2013.

[2] Soja. Disponível em `http://www.conab.gov.br/`. Acesso em 23 de Agosto 2013.

[3] Soja. Disponível em `http://www.ibge.gov.br/`. Acesso em 23 de Agosto 2013.

[4] Teste do qui quadrado. Disponível em `http://www.ufpa.br/dicas/biome/biopdf/bioqui.pdf`. Acesso em 22 Abril 2013.

[5] ANTUNES, J. F. G. Aplicação de lógica fuzzy para estimativa de área plantada da cultura de soja utilizando imagens avhrr-noaa. Master's thesis, 2005.

[6] BACKES, K. S. Variações do índices de vegetação por diferença normalizada (ndvi) do sensor modis associadas a variáveis climáticas para o estado do rio grande do sul. Master's thesis, Universidade Federal de Santa Maria.

[7] BRANDÃO, Z. N., BEZERRA, M. V. C., FREIRE, E. C., AND SILVA, B. B. Determinação de índices de vegetação usando imagens de satélite para agricultura de precisão. *V Congresso Brasileiro de algodão*.

[8] CHAPMAN, P., CLINTON, J., KERBER, R., KHABAZA, T., REINARTZ, T., SHEARER, C., AND WIRTH, R. Crisp-dm 1.0: Step-by-step data mining guide.

[9] CÔRTES, C. S. D. C., PORCARO, R. M., AND LIFSCHITZ, S. *Mineração de dados-funcionalidades, técnicas e abordagens*.

[10] DEPPE, F., LOHMANN, M., MARTINI, L., AND FARIA, R. Monitoramento da evolução temporal de cultivos agrícolas através de imagens terra/modis. *SIMPÓSIO BRASILEIRO DE SENSORIAMENTO REMOTO 13* (2007), 145–152.

[11] FONTANA, D. C. E. A. Assessing the relationship between shire winter crop yield and seasonal variability of the modis ndvi and evi images. In *Applied GIS, Victori* (2007), vol. 3.

[12] FRANK, E., AND WITTEN, I. H. Generating accurate rule sets without global optimization. 144–151.

[13] HAN, J., KAMBER, M., AND JIAN, P. *Data Mining: Concepts and Techniques*. Morgan Kaufmann Publishers, San Francisco, CA, USA.

[14] HOFFER, R. M. *Biological and physical considerations in applying computer aided analysis techniques to remote sensor data*. Mc Graw-Hill, 1978.

[15] IAN, J. *Principal Component Analysis*.

[16] JUNGES, A. H., AND FONTANA, D. C. Desenvolvimento das culturas de cereais de inverno no rio grande do sul por meio de perfis temporais do índice de vegetação por diferença normalizada. *Ciência Rural 40*, 11 (2010).

[17] KLERING, E. V. Avaliação do uso de imagens modis na modelagem agrometeorológica-espectral de rendimento do arroz irrigado no rio grande do sul. Master's thesis, Universidade Federal do Rio Grande do Sul, 2007.

[18] MEGETO, G. A. S. Avaliação da influência da temperatura e da precipitação na ocorrência da ferrugem asiática da soja por meio da técnica de árvore de decisão. Master's thesis, - Universidade Estadual de Campinas, Faculdade de Engenharia Agrícola, 2012.

[19] MOREIRA, M. A. *Fundamentos de sensoriamento remoto e metodologia de aplicação*. UFG, 2003.

[20] NASCIMETO, D. S. C. *Configuração Heterogênea de Ensembles de Classificadores: Investigação em Bagging, Boosting e MultiBoosting*. PhD thesis, Dissertação de mestrado, Universidade de Fortaleza. UNIFOR., Fortaleza, CE, 2009.

[21] NOVO, E. L. D. M. *Sensoriamento remoto: princípios e aplicações*. Edgard Blücher.

[22] ROBERTO, R. Introdução ao sensoriamento remoto.

[23] ROCHA, C. H. B. *Geoprocessamento: tecnologia transdisciplinar*. Ed. do autor.

[24] ROSENDO, J. D. S. Índices de vegetação e monitoramento do uso do solo e cobertura vegetal na bacia do rio araguaia -mg-utilizando dados do sensor modis. Master's thesis, Universidade Federal de Uberlândia.

[25] SABINS, F. F. *Remote Sensing: Principles and Interpretation*. New York: W. H. Freeman and Company, 1986.

[26] SANTOS, J. S. Mineração de dados utilizando algoritmos genéticos. *Monografia apresentada ao curso de graduação em Ciência da Computação da Universidade Federal da Bahia* (2008).

[27] TAN, P.-N., STEINBACH, M., AND KUMAR, V. *Introdução ao data mining: mineração de dados*. Rio de Janeiro: Ciência Moderna, 2009.

[28] WITTEN, I. H., AND FRANK, E. *Data Mining: Practical machine learning tools and techniques*. Morgan Kaufmann, 2005.