Miguel Thiago Alvarenga

UTILIZAÇÃO DA FERRAMENTA J48 PARA DESCOBERTA DO CONHECIMENTO EM BASES DE DADOS FITOSSANITÁRIOS, CLIMÁTICOS E ESPECTRAIS

LAVRAS
MINAS GERAIS - BRASIL
Fevereiro-2014

Miguel Thiago Alvarenga

UTILIZAÇÃO DA FERRAMENTA J48 PARA DESCOBERTA DO CONHECIMENTO EM BASES DE DADOS FITOSSANITÁRIOS, CLIMÁTICOS E ESPECTRAIS

Monografia de graduação apresentada ao Departamento de Ciência da Computação da Universidade Federal de Lavras como parte das exigências do curso de Ciência da Computação para obtenção do título de Bacharel em Ciência da Computação.

Orientadora: Dra. Margarete Marin Lordelo Volpato

Co-orientador: Dr. Wilian Soares Lacerda

LAVRAS
MINAS GERAIS - BRASIL
Fevereiro-2014

Miguel Thiago Alvarenga

UTILIZAÇÃO DA FERRAMENTA J48 PARA DESCOBERTA DO CONHECIMENTO EM BASES DE DADOS FITOSSANITÁRIOS, CLIMÁTICOS E ESPECTRAIS

Monografia de graduação apresentada ao Departamento de Ciência da Computação da Universidade Federal de Lavras como parte das exigências do curso de Ciência da Computação para obtenção do título de Bacharel em Ciência da Computação.

APROVADA em 12 de fevereiro de 2014

Dr. Carlos Ramon Pantaleon Dionisio — UFLA

Dr. Helena Maria Ramos Alves — EPAMIG

Dra. Margarete Marin Lordelo Volpato
Orientadora

LAVRAS
MINAS GERAIS - BRASIL
Fevereiro-2014

DEDICATÓRIA

A minha mãe, Anice Maria Fachini Alvarenga, que com seus zelos de mãe largou tudo pelos filhos, sempre apoiando e fazendo superar meus limites.

Ao Graziano Mateus Valacio, que apesar de não se encontrar mais entre nós, sempre será meu grande amigo, e seu 486 dx4 com o qual começamos a aprender sobre esse mundo da informática.

Aos amigos que fiz nessa caminhada rumo a formatura, Francisco (vulgo Chicão), Vitor Grutner, Jeferson (Jéh), Marlon, Wilson (Rei das Prosa), Rennan (XD), entre todos outros da nossa turma de 2009-1.

Aos meus irmãos Rafael (Jesuis), João Pedro e Marcos Fernando por me suportarem nos momentos de ansiedade e aflição.

Ao meu Pai, Geraldo Alvarenga, pelos seus esforços, sacrifícios e ajudas. Que com muito zelo cuidou de nós sempre tentando, com todas limitações, nos dar o melhor.

Por último, mas não menos especial, a minha grande companheira, Livia Barbosa Franzoso, por seu carinho, amor, dedicação e companheirismo.

AGRADECIMENTOS

A equipe da EPAMIG (Empresa de Pesquisa Agropecuária de Minas Gerais), a equipe do laboratório Geossolos, pelos momentos e em especial a minha orientadora Margarete, pela paciência, compreensão e ajuda.

A Universidade Federal de Lavras e ao Departamento de Ciência da Computação (DCC), assim como todos os funcionários que fazem essa máquina funcionar.

Aos professores pelo conteúdo passado, suas exigências, rigores e compreensão.

RESUMO

O presente trabalho utiliza do Weka como ferramenta de mineração dos dados da praga bicho mineiro contido nos banco de dados da EPAMIG, avaliando se a utilização dos dados espectrais, EVI-2, é eficiente na predição da ocorrência da praga.

Foram gerados diversos modelos em forma de árvore de decisão, os quais foram validados. Para isso, foi feito um processo interativo entre modificações dos parâmetros de entrada do algoritmo, da base de dados e validações dos modelos.

Obteve uma taxa de acerto de aproximadamente 85% utilizando-se do algoritmo árvore de decisão J48 o que aprovaria sua utilização em projetos futuros da Empresa de Pesquisa Agropecuária de Minas Gerais.

Palavras-chave: Mineração de dados, Cafeicultura, Bicho Mineiro, Árvores de decisão.

LISTA DE FIGURAS

LISTA DE TABELAS

LISTA DE SIGLAS

AGNES	*Agglomerative Nesting*
API	*Application Programming Interface*
ARFF	*Attribute-Relation File Format*
CSV	*Comma-Separated Values*
DNA	*DeoxyriboNucleic acid*
EPAMIG	*Empresa de Pesquisa Agropecuária de Minas Gerais*
EVI	*Enhanced Vegetation Index*
FCM	*Fuzzy CMeans*
FMLE	F*uzzy Maximum Likelihood*
GNU	*General Public License*
ID3	*Inductive Decision 3 (tree)*
INMET	*Instituto Nacional de Meteorologia*
INPE	*Instituto Nacional de Pesquisas Espaciais*
JAR	*Java ARchive*
KDD	*Knowledge Discovery in Database*
MD	*Mineração de Dados*
MODIS	*Moderate Resolution Imaging Spectroradiometer*
NASA	*National Aeronautics and Space Administration*
RNA	Redes Neurais Artificiais
ROCK	*Robust Clustering Using Links*
SGBD	*Sistema Gerenciador de Banco de Dados*
SVM	*Support Vector Machines*
WEKA	*Waikato Environment for Knowledge Analysis*

SUMÁRIO

1. Introdução

O Bicho mineiro do cafeeiro *Leucoptera coffeella* (Guérin-Méneville) é uma praga importante no Brasil devido a sua ocorrência generalizada nos cafezais que gera grandes prejuízos a economia cafeeira. Os fatores que afetam o ataque dessa praga ao cafeeiro relacionam-se principalmente aos elementos meteorológicos (clima/tempo). Portanto, o conhecimento da dinâmica desses elementos é fundamental para o manejo da cafeicultura. De acordo com especialistas da Empresa de Pesquisa Agropecuária de Minas Gerais (EPAMIG) a temperatura do ar e a precipitação pluviométrica são os principais fatores relacionados à dinâmica populacional dessa praga.

Apesar da grande importância do monitoramento agrometeorológico da cafeicultura são escassos estudos de modelos que descrevam com exatidão a ocorrência dessa praga e sua relação com as variações climáticas. Para o desenvolvimento desse tipo de modelagem é necessária à utilização de metodologias que permitam a análise de grande número de dados.

Uma alternativa é a utilização de técnicas de mineração de banco de dados que podem resultar na identificação de padrões desconhecidos e potencialmente úteis para entender a ocorrência dessa praga.

Visto que a obtenção de dados fitossanitários e climáticos é muito cara e demandam muito tempo e esforços, uma alternativa é a utilização de dados de sensoriamento remoto coletado por satélites ambientais como o sensor MODIS a bordo dos satélites TERRA e AQUA. Esse sensor capta dados espectrais altamente correlacionados com o desenvolvimento e vigor da vegetação e com as condições climáticas, além de possuir a vantagem de monitorar extensas áreas com grande frequência de imageamento.

1.1. Justificativa

O monitoramento da ocorrência do Bicho mineiro em cafeeiro é de grande importância na economia e na saúde. Economicamente, por reduzir o custo de produção por meio de aplicações oportunas de medidas de controle, geralmente inseticidas altamente tóxicos. Saúde, por envolver não somente o modo de cultivar, como também reduzir o efeito tóxico sobre as plantas e ao ambiente externo, diminuindo a exposição de agrotóxicos sobre as plantas, operários e consumidores.

1.2. Objetivo

Diante deste problema, o presente trabalho tem como objetivo utilizar técnicas de mineração de dados para desenvolver modelos de infestação que possam basear-se na variação de dados meteorológicos e espectrais de sensoriamento remoto orbital.

2.Referencial Teórico

2.1. Descoberta de Conhecimento em Bancos de Dados

Banco de dados pode ser definido com um conjunto de dados integrados com objetivo de atender uma comunidade de usuários (Heuser, 1998). A forma como etes dados são organizados pode variar do mais simples modo de guardar dados, ou seja, uma planilha contendo uma matriz com colunas e linhas, onde as colunas são entidades e as linhas são os dados das entidades, até os mais complexos como SGBD (Sistema Gerenciador de Banco de Dados).

Nos últimos anos, o progresso na captura de dados em formato digital e das tecnologias de armazenamento contribuiu para o desenvolvimento de grandes bases de dados. Para Bucene (2002) cada vez mais, o volume de informações excede a capacidade de sua análise pelos métodos tradicionais. Podem-se gerar relatórios a partir dos dados, mas não se consegue analisá-los sob o enfoque do conhecimento. Como resultado desse aumento efetivo, o processamento dessas informações tornou-se cada vez mais complexo e difícil, e, normalmente, os dados ficam armazenados nas bases de dados sem que sejam utilizados de uma forma realmente eficiente.

A Descoberta de Conhecimento em Bancos de Dados (Knowledge Discovery in Database - KDD) é o processo de identificar em dados, padrões que sejam válidos, previamente desconhecidos, potencialmente úteis e compreensíveis, visando melhorar o entendimento de um problema ou processo de tomada de decisão (Fayyad, 1996). Para Silva (2006), o processo de KDD é iterativo (feito com diversas repetições), cognitivo e exploratório.

A Mineração de Dados (MD) é a etapa em KDD responsável pela seleção dos métodos a serem utilizados para localizar padrões nos dados, seguida da efetiva busca por padrões de interesse numa forma particular de representação, juntamente com a busca pelo melhor ajuste dos parâmetros do algoritmo para a tarefa em questão (Silva, 2006).

Rezende et al. (2003) explicam que a mineração de dados foca em como transformar dados armazenados em conhecimento, expresso em termos de formalismos de representação, tal como regras e relações entre dados. Existe conhecimento que pode ser extraído diretamente de dados sem o uso de qualquer técnica, entretanto, existe também muito conhecimento que está intrínseco na Base de Dados (A dados, informações

gera conhecimento), na forma de relações existentes entre itens de dados que, para ser extraído, requerem o desenvolvimento de técnicas especiais.

Enquanto as ferramentas tradicionais de banco de dados são capazes de mostrar "o que" está na base de dados, os softwares analíticos ajudam o usuário a descobrir o "porquê". Em um pacote estatístico, o usuário formula as prováveis hipóteses, para então testar suas validades. A mineração de dados estende a capacidade de gerar e validar hipóteses e por isso se diz que se pode descobrir conhecimento novo (inesperado), útil e interessante (Bucene, 2002) A informação e o conhecimento obtidos podem ser utilizados para aplicações que abrangem desde gerenciamento de negócios, controle de produção e análise de marketing, à engenharia, projeto e exploração científica (Han e Kamber, 2006).

Atualmente, diferentes técnicas analíticas podem ser empregadas na mineração de dados, variando das abordagens descritivas e gráficas mais básicas até as técnicas multivariadas mais sofisticadas (por exemplo, análise de agrupamentos ou regressão múltipla ou logística) e modelos de aprendizado mais novos (como redes neurais e algoritmos genéticos) (Hair et al., 2005). Mas Rezende et al. (2003) ressaltam que as técnicas utilizadas em MD não devem ser vistas como substitutas de outras formas de análises, mas, como práticas para melhorar os resultados das explorações feitas com as ferramentas atualmente utilizadas.

Todo o processo de MD é orientado em função do domínio de aplicação e dos repositórios de dados inerentes aos mesmos. Para usar os dados é necessário que estes estejam estruturados de forma a serem consultados e analisados adequadamente. Por outro lado, as informações descobertas precisam ser verdadeiras e relevantes para o contexto da exploração daquilo que se está procurando. A regra final é que a descoberta possa ser trabalhada, passível de uma ação estratégica e que traga benefícios para a organização (Murayama, 2002; Rezende et al., 2003).

Contextos que podem parecer pouco comuns, como a identificação de genes específicos, codificadas em moléculas de DNA, caracterização das propriedades de organismos como os vírus, têm utilizado as técnicas de mineração de dados, que frequentemente toma a forma de reconhecimento de padrões e determinação sobre se esses padrões são significativas ou meras coincidências.

Não existe uma forma única de tratar ou classificar a tarefa de classificar os dados para encontrar uma relação, não apenas classificando-os, mas obtendo um conhecimento capaz de prever seu comportamento no futuro. Cortês et al. (2002) apresentam algumas dessas classificações. A mais comum delas é dividir a funcionalidade em mineração de

dados com Análise Descritiva e Análise Preditiva. As descritivas se concentram em encontrar padrões que descrevam os dados de forma interpretável pelos seres humanos. As preditivas realizam interferência nos dados para construir modelos que serão usados para predições do comportamento de novos dados.

2.1.1. Análise de regras de associação

As regras de associações identificam os grupos de dados que apresentam co-infestação entre si. As regras de associação encontram itens que determinam a presença de outros em uma mesma transação e estabelecem regras que correlacionam a presença de um conjunto de itens com um outro intervalo de valores para um outro conjunto de variáveis. Uma associação é normalmente representada por uma regra de associação do tipo X ı Y, que implica numa relação de dependência entre os conjuntos de dados X e Y. Assim, se X ocorre na base de dados, então Y também ocorre (com alguma relação a X) (Hipp et al, 2000; Pierrakos et al., 2003).

A mineração de dados utilizando-se de regras de associação parte do princípio de que, se um conjunto de itens não é frequente, qualquer combinação de itens que inclua este conjunto também será infrequente. Em outras palavras, se a contagem de um determinado conjunto de itens revela que ele não é frequente, então não será mais necessário contar nenhum dos conjuntos que o incluírem, economizando assim bastante tempo. A decisão do que deve ser considerado frequente depende de uma série de fatores, muitos deles subjetivos. Cabe ao usuário informar, através de um parâmetro, a frequência mínima que um conjunto de itens deve apresentar para ser considerado frequente.

A regra de associação pode ser usada sempre que se deseja encontrar relações entre eventos em uma base de dados. Assim, poderíamos usar regras de associação para identificar, por exemplo, correlações entre o número de dias sem chuva e a infestação de pragas no cafeeiro.

Os algoritmos mais populares para regras de associação são o Apriori (AgrawalandSrikant, 1994), o Predictive Apriori (Schefferand et al., 2001) e o Tertius (FlachandLachiche, 2001) e suas variantes.

Hipp et al. (2000) apresentam uma análise e comparação dos algoritmos de Regras de Associação. Tais autores concluíram que, em seus experimentos, os algoritmos mostraram um comportamento similar em tempos de execução e que não existe um

algoritmo que seja essencialmente pior que os outros, mas que as vantagens e desvantagens identificadas dizem respeito à estratégia utilizada.

2.1.2. Análise de clusters (agrupamentos)

Uma tarefa de análise de agrupamentos consiste em identificar classes de itens em uma base de dados de acordo com alguma medida de similaridade. Cada grupo, chamado cluster, consiste de objetos que são similares entre eles e diferentes dos objetos dos outros grupos. Sendo assim, a diferença entre os grupos é dada por um valor, como medidas de distância ou similaridade/dissimilaridade. Diferentemente da classificação e predição, em que os dados estão previamente classificados, a análise de clusters trabalha sobre dados nos quais as classes não estão definidas.

Existem duas abordagens gerais para a tarefa de agrupamento (Pierrakos et al., 2003), o método de particionamento que cria, iterativamente, k grupos a partir de um conjunto de dados, onde cada grupo representa um cluster; e os métodos hierárquicos que decompõe o conjunto de dados, criando uma estrutura hierárquica de clusters. A estratégia usada pode ser *bottom-up*, que cria pequenos grupos juntando as instâncias, repetindo até atingir um critério; ou *top-down*, que considera todas as instâncias como pertencentes a um grande grupo e subdivide recursivamente este grupo.

O grande conjunto de dados e a diversidade de atributos e domínios presentes em mineração de dados adicionam complicações à tarefa de clusterização, à medida que impõe requisitos computacionais. Diversos algoritmos têm surgido desses requisitos e são aplicados com sucesso em problemas reais na mineração de dados (Berkhin, 2006).

Os algoritmos mais clássicos de agrupamento particionais são baseados no K-Media, como o Forgy's (Forgy, 1965). Dos particionais baseados em lógica nebulosa, pode-se citar o FCM - Fuzzy cmeans (Bezdek, 1981) e o FMLE - fuzzy maximum likelihood (Gath and Geva, 1989). Dos métodos hierárquicos de agrupamento, cita-se o Single Link (Florek et al., 1957), o AGNES – Agglomerative Nesting (Kaufman, 1990) e, mas recentemente, o ROCK - Robust Clustering Using Links (Guha et al., 2000).

2.1.3. Classificação

Ao contrário da clusterização, o objetivo da classificação é identificar características distintas de classes pré-definidas, baseadas num conjunto de instâncias.

Essa informação pode ser usada tanto para entender a existência dos dados, quanto para predizer como novas instâncias se comportarão (Phyu, 2009). Em outras palavras, em uma tarefa de classificação, deve ser identificado o conjunto mínimo das características conhecidas de um determinado objeto que sejam suficientes para prever uma característica desconhecida.

Sendo assim, criam-se modelos (funções) que descrevem e distinguem classes ou conceitos, baseados em dados conhecidos, com o propósito de utilizar estes modelos para predizer a classe de objetos que ainda não foram classificados. Como a classificação é um processo de aprendizado supervisionado, o modelo construído baseia-se na análise prévia de um conjunto de dados de amostragem ou dados de treinamento, contendo objetos corretamente classificados (Han e Kamber, 2006).

Os algoritmos de classificação incluem métodos que utilizam árvores de decisão, redes bayesianas, vizinhos mais próximos, algoritmos genéticos, lógica nebulosa, classificação baseadas em regras, entre outros. Como exemplo de algoritmos de classificação, pode-se citar: Árvore de Decisão: ID3 (Quinlan, 1986), Classification and Regression Trees (CART) (Breiman et al., 1984), C4.5 (Quinlan, 1993); Redes Neurais: Back-Propagation (Rumelhart et al., 1986), SVM (support vector machines) - (Goh et al., 2001), Vizinho mais próximo, k-Nearest Neighbors (k-NN) (Cover and Hart., 1967).

Phyu (2009) revisou diferentes técnicas de classificação para mineração de dados. A conclusão foi que as árvores de decisão e as redes bayesianas normalmente têm perfis operacionais opostos.

Quando uma técnica é muito acurada, a outra não o é, e vice-versa. Por outro lado, árvores de decisão e classificação por regras possuem perfis operacionais similares.

2.1.3.1. Árvore de Decisão

As árvores de decisão permitem derivar regras de produção, de decisão ou de classificação, destas regras é gerada uma árvore. As regras são os caminhos de um nó a outro da árvore saindo do nó raiz (atributo de maior relevância) até os nós folha (atributo de interesse). Assim para verificar os padrões obtidos basta percorrer o trajeto do nó raiz até um nó folha nesta árvore.

2.1.3.1.1. ID3

A classificação ID3 (Árvore de Decisão Indutiva) é baseada em árvore de decisão, método de aprendizado simbólico. Um modo de abordar a tarefa de indução é gerar toda

e possível decisão pela árvore, classificando corretamente o conjunto de treinamento e selecionar o mais simples. Sua utilização seria mais eficiente, como qualquer algoritmo de aprendizagem em máquina, sobre banco de dados que possuam um grande número de atributos e do conjunto de treinamento possuindo muitos objetos, retornando um resultado razoável com pouco poder computacional.

Mais importante é a construção da árvore tendo a possibilidade de testá-la, e para isso é necessário um conjunto de dados, que não seja o de treinamento, para efetuar os testes. Esses dados de teste não podem ter sido passados pelo algoritmo antes, para não haver um vício na resposta. Nessa etapa verifica-se como a árvore generaliza os dados e se adapta a novas situações, podendo, também, se estimar proporções de acertos e erros na construção da árvore.

Para classificar os atributos o algoritmo leva em consideração o ganho de informação (Medida do quando um dado atributo separa a base de dados em classes diferentes.) e o grau de entropia (Medida de cálculo do grau de homogeneidade contido no conjunto de dados.).

Pseudo código do ID3:

ID3 (Exemplos, Atributo-objetivo, Atributos)
// ID3 retorna uma árvore de decisão que classifica corretamente os Exemplos
//determinados
// Exemplos são os exemplos de treinamento.
// Atributo-objetivo é o atributo cujo valor deve ser predito pela árvore.
// Atributos são uma lista de outros atributos que podem ser testados pela árvore de
//decisão.
Início
Crie um nó da Raiz para a árvore
Se todos os Exemplos são positivos
Então retorna a Raiz da árvore com o rótulo = sim
Se todos os Exemplos são negativos
Então retorna a Raiz da árvore com o rótulo = não
Se Atributos for vazio
Então retorna a Raiz da árvore com o rótulo = valor mais comum do Atributo-objetivo em Exemplos
Senão
A ← um atributo de Atributos que melhor classifica Exemplos (atributo de decisão)

Raiz ← A (rótulo = atributo de decisão A)
Para cada possível valor vi de A faça
Acrescenta um novo arco abaixo da Raiz, correspondendo à resposta A = vi
Seja Exemp1os vi o subconjunto de Exemplos que têm valor vi para A Se Exemplos vi for vazio
Então acrescenta na extremidade do arco um nó da folha com rótulo = valor mais comum do Atributo-objetivo em Exemplos
Senão acrescenta na extremidade do arco a sub árvore
ID3(Exemplos vi, Atributo-objetivo, Atributos -{A})
Retorna Raiz (aponta para a árvore)
Fim

2.1.3.1.1. Árvore de Decisão J48 (C4.5)

A árvore de decisão C4.5 é uma extensão do algoritmo de classificação ID3 sendo uma melhoria deste, o que o tornou um dos mais clássicos algoritmos de árvore de decisão, que trabalha tanto com atributos discretos quanto contínuos (Quilan, 1993). Também permite a utilização de atributos desconhecidos, representados por "?".

É um algoritmo voraz que utiliza o método divisão e conquista para aumentar a capacidade de predição das árvores de decisão. Assim, sempre usa o melhor passo avaliado localmente, sem se preocupar se esse passo vai produzir a melhor solução, pega um problema e o divide em vários subproblemas sendo criadas sub-árvores entre a raiz e as folhas.

2.2. Weka (Waikato Environment for Knowledge Analysis)

O Weka é uma ferramenta de mineração de dados com uma vasta coleção de algoritmos de aprendizado em máquina e ferramentas de pré-processamento de dados. Foi projetado com intuito de que seus métodos sejam executados em conjunto de dados na forma mais flexível possível. Incluindo dezenas de ferramentas com interface fácil e de rápida comparação entre seus métodos, com grande facilidade na geração e comparação dos modelos mais adequados para a solução do problema. No algoritmo J48, é possível plotar a árvore gerada com suas regras, o que foi de grande ajuda para validação dos modelos junto a um especialista da área.

Seu desenvolvimento pela Universidade de Waikato, na Nova Zelândia, seguiu os termos da GNU (General Public License), com implementação em Linguagem Java, o que facilita a implementação de softwares de terceiros nesta linguagem utilizando suas bibliotecas. (Ian H. Wtten-Data Mining - 2011).

2.3. Mineração de dados associados a problemas fitossanitários

Baker et al. (1993) desenvolveram uma árvore de decisão para predizer a extensão de mortalidade de Pinus (*Pinnus elliottii* e *Pinnus taeda*) em decorrência de podridão das raízes causada por *Heterobasidion annosum*. Os dados foram obtidos de plantações inoculadas com *H. annosum*. Após cinco anos, contou-se a quantidade de árvores mortas mais a de árvores com infecções letais, em cada um dos 152 talhões de 16 locais selecionados em quatro estados norte-americanos. E uma árvore, com apenas dois atributos de teste, classificou corretamente o risco de mortalidade para 85% dos casos. O modelo revelou que o percentual de silte e o pH, ambos do horizonte A do solo, foram as variáveis importantes na predição da classe de risco, indicando que solos com baixo teor de silte ou alto pH são, geralmente, de alto risco. Por meio de validação cruzada (10-fold crossvalidation), a estimativa de acurácia da árvore de decisão foi de 80%.

Pinto et al. (2002) avaliaram o potencial das redes neurais para descrever epidemias da ferrugem do cafeeiro. Eles empregaram as redes neurais para estabelecer relações entre variáveis climáticas e produção, com a incidência da ferrugem do cafeeiro. A rede que melhor descreveu a epidemia da ferrugem do cafeeiro incluiu as variáveis temperatura mínima, umidade relativa do ar, produção e insolação, referentes a 30 dias antes da data de avaliação da incidência da ferrugem. As redes elaboradas a partir das séries temporais também foram adequadas para descrever a epidemia da ferrugem, sendo que a melhor delas incluiu as observações da incidência da doença das quatro quinzenas anteriores à data de avaliação. A escolha dos melhores modelos baseou-se nos menores valores do quadrado médio do desvio e do erro médio de previsão avaliada para as redes.

Meira (2008) utilizou o método de árvore de decisão para desenvolver e analisar a epidemia da ferrugem do cafeeiro no estado de São Paulo. Utilizando dados de estações meteorológicas, ele demonstrou o potencial dessa técnica como modelo simbólico e interpretável, que permite a identificação das fronteiras de decisão e da lógica contidas nos dados, auxiliando na compreensão de quais variáveis e como as interações dessas variáveis condicionaram o progresso da doença no campo. As variáveis explicativas mais

importantes foram: temperatura média nos períodos de molhamento foliar, carga pendente de frutos, média das temperaturas máximas diárias no período de incubação e umidade relativa do ar.

2.4. O sensor MODIS

MODIS (Moderate Resolution Imaging Spectroradiometer) é um instrumento a bordo da terra por via de dois satélites, o TERRA e o AQUA, que é capaz de cobrir a superfície da terra de 1 a 2 dias. Possui um papel importante nos desenvolvimentos de modelos validados capaz de prever mudanças globais com precisão suficiente para ajudar tomar decisões sobre o que ocorre no meio ambiente. Produzindo imagens com resolução espacial de até 250 metros. Esse sensor foi lançado em dezembro de 1999 e 2002 pela NASA com intuito de preencher lacunas na disponibilidade efetiva de dados de sensoriamento remoto de alta resolução temporal e espectral. Trata-se de um espectro radiômetro imageador de resolução espacial, variada, composto por um scanner óptico de varredura transversal e um conjunto de elementos detectores individuais capaz de fornecer imagens da superfície terrestre em 36 bandas espectrais distribuídas entre o visível e o infravermelho termal 0,4 a 14,3 mm.

EVI (Enhanced Vegetation Index) é o valor do índice de vegetação oriundo de operações matemáticas entre bandas espectrais coletadas pelo sensor MODIS. O cálculo do EVI utiliza imagens com resolução espacial de 250 m e resolução temporal de 16 dias disponibilizado pela NASA. Variando entre os valores 0 e 1, foi desenvolvido para otimizar o sinal de resposta da vegetação, melhorando a sensibilidade em regiões com maiores densidades de biomassa, além de propiciar o monitoramento da vegetação através de uma ligação do sinal de fundo do dossel e a redução das influências atmosféricas.

O índice de vegetação melhorado (EVI) foi desenvolvido para otimizar o sinal da vegetação, melhorando a sensibilidade de sua detecção em regiões com maiores densidades de biomassa, e para reduzir a influência do sinal do solo e da atmosfera sobre a resposta do dossel. Nesse sentido, o EVI é calculado através da seguinte equação (Justice e al., 1998):

EVI = G (NIR – Vermelho) / (L + NIR + C1 vermelho – C2 azul) 23

Onde:

L é fator de ajuste para o solo; G é o fator de ganho e C1 e C2 são coeficientes de ajuste para efeito de aerossóis da atmosfera. Os valores dos coeficientes adotados pelo algoritmo do EVI são: L=1, C1=6, C2=7,5 e G= 2,5 (Huete et al., 1997; Justice et al., 1998).

O cálculo do EVI-2 utiliza imagens do MODIS do produto MOD13 (Jiang et al., 2008), com resolução espacial de 250 m e resolução temporal de 16 dias, disponibilizados pela NASA. O EVI-2 apresenta resultados semelhantes ao EVI e foi desenvolvido para otimizar o sinal de resposta da vegetação, diminuindo a influência do solo e da atmosfera. Valores do EVI-2 mais próximos de 1 indicam maior vigor vegetativo (Freitas et al. 2011).

3. Metodologia de Desenvolvimento

Para realização do presente projeto foi seguido o seguinte fluxograma.

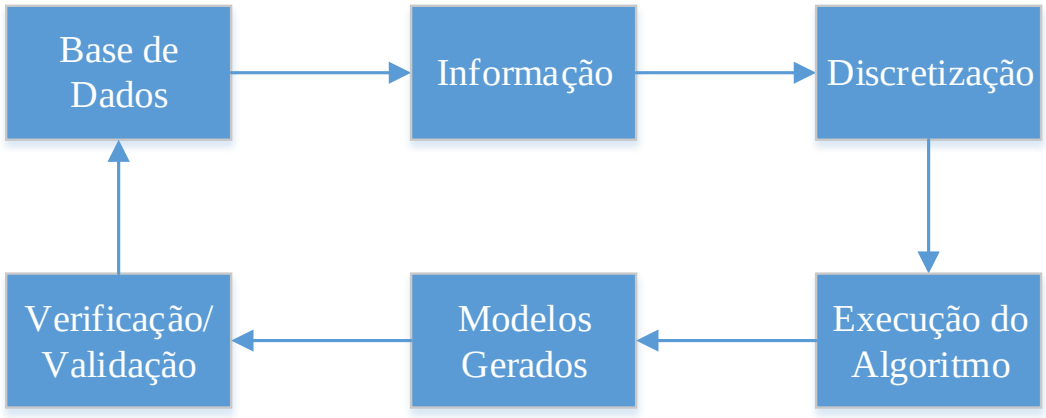

Figura 1: Fluxograma do projeto.

Assim, fica fácil a visualização da transformação dos dados de uma base de dados (Dados Bruto) primeiramente em informação, obtenção dos atributos relevantes. Junto ao estudo do problema pode-se transformar os dados em faixas de informações mais interessantes (Discretização). O algoritmo pode ser executado tanto nos dados discretizados, quanto nos dados não discretizados. A execução do algoritmo gera modelos, regras, padrões da base de dados. Esses padrões são interpretados e levados para verificação de um especialista do domínio. Se não validados, retorna-se a Base de Dados, ficando claro que o processo de KDD é iterativo. Esse processo foi executado até ser obtido um ou mais modelos aceitáveis (Validados).

Para realizar o presente estudo foi utilizado o banco de dados da praga Bicho Mineiro em lavouras cafeeiras, de uma fazenda da EPAMIG (Empresa de Pesquisa Agropecuária de Minas Gerais), com histórico de trinta anos de monitoramento fitossanitário (banco de dados fitossanitário).

A fazenda está situada no município São Sebastião do Paraíso, sul de Minas Gerais e possui uma estação meteorológica (rede oficial do INMET).

Os dados espectrais do local de estudo foram adquiridos do banco de dados espectrais disponibilizado gratuitamente pelo laboratório virtual do Instituto Nacional de Pesquisas Espaciais (INPE) (<http://www.dsr.inpe.br/laf/series />). Para avaliar o vigor vegetativo do cafeeiro utilizou-se dados EVI2 já processadas conforme Freitas et al., (2011).

Para avaliar o índice de vegetação espectral de áreas cafeeiras no município de São Sebastião do Paraíso utilizou-se os dados de três pixels do período de tempo estudado de 2000 a 2010 de imagens do EVI-2 já processadas conforme (Freitas et al., 2011). Os valores de EVI-2 são imagens disponibilizados gratuitamente pelo laboratório virtual do Instituto Nacional de Pesquisas Espaciais (INPE) (<http://www.dsr.inpe.br/laf/series />).

Para consolidar o banco de dados do projeto foram realizados ajustes temporais e de formato dos dados oriundos das três fontes de dados (dados fitossanitários, climáticos e espectrais).

Tomando como base o processo de descoberta de conhecimento (KDD) proposto por Fayyad et al. (1996) foi necessário o retorno e avanço entre as etapas do processo, repetindo operações para se obter um melhor resultado. Por outro lado, para compreensão dos resultados obtidos contamos com especialista, com domínio do problema em todas as fases do processo de KDD, o que correspondeu ao caráter interativo do mesmo.

Sendo assim, primeiramente o algoritmo de mineração de dados foi efetivamente executado, estudado e aplicado a dados do Bicho mineiro. Para a execução de alguns desses algoritmos, o J48-árvore de decisão, utilizou-se dos dados apenas normalizados contrapondo com os dados discretizados, levando em consideração estudos sobre padrões de infestação da praga Bicho Mineiro e dos dados espectrais EVI2, juntamente com o auxílio de especialistas das áreas de pragas dos cafeeiros, clima e sensoriamento remoto.

Foram realizados testes sobre o conjunto de funções (rotinas e padrões) do Weka, para verificar se sua implementação daria o mesmo resultado da ferramenta, e se ao utilizar esse conjunto de funções (API) felicitaria o desenvolvimento de softwares de prescrição sobre problemas semelhantes. Essa etapa é necessária, pois justificaria a utilização desse método para implementação de softwares agrícolas com inteligência artificial sem a necessidade da implementação dos algoritmos de aprendizagem em máquina.

Para o presente projeto cinco atividades foram desempenhadas:

1. Discretização da base de dados do projeto;
2. Execução do algoritmo de Mineração de Dados;
3. Interpretação, seleção e validação dos Modelos;
4. Utilizando a API do Weka.

3.1. Discretização da base de dados

Antes mesmo que os dados fossem discretizados, houve a necessidade de uma padronização dos dados e o acréscimo dos dados espectrais. A ferramenta Weka possui um tipo de arquivo padrão de entrada, o arff, porém também há a possibilidade de entrar com outros formatos como o formato csv (Comma-Separated Values, valores delimitados por vírgulas).

3.1.1. Dados

O dados da base de dados da EPAMIG que foram utilizados nesse projeto foram coletados durante dez anos (de 2000 a 2010), sendo que da lavoura foram selecionadas 10 plantas aleatoriamente e dessas foram retiradas (coletadas) 20 folhas por terço (divisão da planta em 3 partes, superior, media e inferior). Totalizando em 200 folhas por terço (contando todas as 10 plantas.), selecionadas para a verificação de minas (locais "comidos" pela larva) causadas pelo Bicho mineiro.

Foram coletadas 790 amostras (instâncias) contendo 23 atributos, sendo que 10 destes atributos são as plantas de onde foram coletadas as amostras (nomeadas de P seguido de um numero, até 10), sendo transformado em apenas um atributo composto por folhas minadas por planta (Tabela 1).

Tabela 1: Parte da planilha dos dados disponibilizados pela EPAMIG

		P1		P2		P3		P4		P5		P6		
º de planta	folhas por	f. minadas	minas	f. minadas	minas	f. minadas	minas	f. minadas	minas	f. minadas	minas	f. minadas	minas	f. mina
10	20	6	6	1	1	1	1	4	5	2	2	0	0	
10	20	2	5	1	1	2	2	0	0	1	1	0	0	
10	20	1	1	0	0	0	0	0	0	1	1	0	0	
10	20	1	1	1	1	3	3	2	2	4	4	0	0	
10	20	0	0	0	0	0	0	1	1	3	3	0	0	
10	20	0	0	0	0	0	0	0	0	0	0	0	0	
10	20	3	8	1	1	1	1	0	0	4	4	2	2	
10	20	0	0	2	2	2	2	1	1	0	0	1	1	
10	20	0	0	2	2	0	0	0	0	0	0	0	0	
10	20	2	2	4	4	0	0	2	2	2	2	1	1	
10	20	0	0	1	1	0	0	1	1	1	1	0	0	
10	20	2	2	1	1	1	1	3	5	0	0	0	0	
10	20	2	2	6	6	1	1	2	2	0	0	1	1	

Primeiramente foi feito uma soma de folhas minadas, obtendo-se o atributo TotalFolhasMinadas, a data foi substituída apenas por seus respectivos meses. Também houve o acréscimo do atributo EVI2 (dados espectrais) e dos dados climáticos, obtendo a seguinte base de dados (Tabela 2).

Tabela 2: Base de dados padronizados

	A	B	C	D	E	F	G	H	I	J
1									Media EVI-2	
2	coletas	Data da coleta (dd/mm/aaaa)	Terço	PrecipitacaoSOMA	PrecipitacaoMEDIA	TempMedia	TotalFolhasMinadas	Umid. Media	Sim	Não
3	Jan_1	3/1/2000	s	129,6	8,1	23,5	30,00	79,19		
4	Jan_2	3/1/2000	m	129,6	8,1	23,5	11,00	79,19		
5	Jan_3	3/1/2000	i	129,6	8,1	23,5	7,00	79,19		
6	Jan_4	14/1/2000	s	351,5	23,4	24,2	17,00	79,19		
7	Jan_5	14/1/2000	m	351,5	23,4	24,2	5,00	79,19		
8	Jan_6	14/1/2000	i	351,5	23,4	24,2	5,00	79,19		
9	Jan_7	31/1/2000	s	169,9	10,6	25,3	23,00	81,93		
10	Jan_8	31/1/2000	m	169,9	10,6	25,3	1,00	81,93		
11	Jan_9	31/1/2000	i	169,9	10,6	25,3	3,00	81,93		
12	Fev_1	15/2/2000	s	184,3	12,3	24,2	6,00	88,33	0,46568	0
13	Fev_2	15/2/2000	m	184,3	12,3	24,2	0,00	88,33	0,46568	0
14	Fev_3	15/2/2000	i	184,3	12,3	24,2	1,00	88,33	0,46568	0
15	Mar_1	1/3/2000	s	73,7	5,3	24,0	2,00	76,14	0,46153	0,39407
16	Mar_2	1/3/2000	m	73,7	5,3	24,0	0,00	76,14	0,46153	0,39407
17	Mar_3	1/3/2000	i	73,7	5,3	24,0	3,00	76,14	0,46153	0,39407
18	Mar_4	16/3/2000	s	134,4	9,0	23,2	0,00	77,13	0,45628	0,44704
19	Mar_5	16/3/2000	m	134,4	9,0	23,2	0,00	77,13	0,45628	0,44704
20	Mar_6	16/3/2000	i	134,4	9,0	23,2	0,00	77,13	0,45628	0,44704
21	Abr_1	3/4/2000	s	65,8	4,1	23,7	3,00	79,87	0,44938	0,42568
22	Abr_2	3/4/2000	m	65,8	4,1	23,7	1,00	79,87	0,44938	0,42568
23	Abr_3	3/4/2000	i	65,8	4,1	23,7	0,00	79,87	0,44938	0,42568
24	Abr_4	14/4/2000	s	13,1	0,9	22,9	9,00	68,07	0,43912	0,40936
25	Abr_5	14/4/2000	m	13,1	0,9	22,9	3,00	68,07	0,43912	0,40936
26	Abr_6	14/4/2000	i	13,1	0,9	22,9	1,00	68,07	0,43912	0,40936
27	Mai_1	2/5/2000	s	21,8	1,5	19,6	6,00	71,60	0,4209	0,41692

Os dados da tabela 2 foram minerados obtendo-se os modelos não discretizados, já que na literatura pressupõe-se que os dados discretizados perdem informações.

Com intuito de deixar a base de dados mais fácil de ser interpretada, os dados expostos na tabela 2 foram discretizados. Para isso, o atributo Mês foi transformado de forma a sugerir o estágio do segundo ano fenológico da planta, Segundo Camargo & Camargo (2001) (Figura 1).

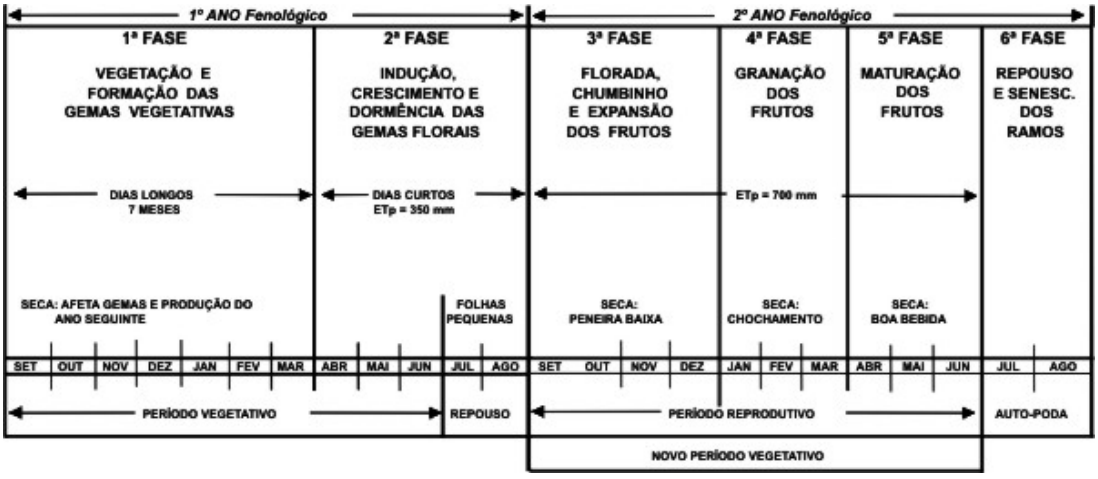

Figura 2: Fenologia do Cafeeiro. Fonte Camargo & Camargo (2001).

Os dados climáticos foram coletados em estação meteorológica convencional presente na fazenda experimental da EPAMIG, em parceria com a INMET. Foram

analisados os seguintes dados: precipitação, temperatura mínima, temperatura máxima e umidade relativa do ar.

Como os dados climáticos eram diários e os de doença eram mensais, foi necessário reorganizar os dados. Os dados de temperatura mínima e máxima foram transformados em atributos de temperaturas médias. Os dados de precipitação soma foram divididos, maior ou menor que 60 mm mensal. Segundo a literatura e a sugestão de especialistas foi estabelecida a tabela 3 para discretização da base de dados (Tabela 3).

Tabela 3: Tabela de discretização da base de dados do Bicho Mineiro.

Discretização dos dados – **Bicho Mineiro**

Total de 200 folhas por terço

Infestação	Folhas minadas	Temperatura média	Umidade Relativa	Precipitação
Baixa	< 40 folhas	< 18°C	< 65%	<60 mm mensal
Média	40 – 60 folhas	18° - 22°C	----------	
Alta	> 60 folhas	> 22°C	> 65%	> 60 mm mensal

Com base nos dados da tabela 3 foi preparada a base de dados para mineração (Tabela 4). Essa base de dados apresenta-se no formato CSV, e também foi minerada com intuito de verificar se há distinção entre esse formato e o formato ARFF (Formato Padrão do Weka). O algoritmo ID3 só minera dados discretos (Tabela 4).

Tabela 4: Base de dados discretizada e convertida para formato csv.

Fenologia	Infestacao	UmidRelativa	Temperatura	PrecipitacaoMedia	PrecipitacaoSoma	Terco	Media EVI-2	
							Sim	Não
Granação dos Frutos	BAIXA	>=65	>=22	<10	>=60	SUPERIOR	<0,3	<0,3
Granação dos Frutos	BAIXA	>=65	>=22	<10	>=60	MEDIO	<0,3	<0,3
Granação dos Frutos	BAIXA	>=65	>=22	<10	>=60	INFERIOR	<0,3	<0,3
Granação dos Frutos	BAIXA	>=65	>=22	>=10	>=60	SUPERIOR	<0,3	<0,3
Granação dos Frutos	BAIXA	>=65	>=22	>=10	>=60	MEDIO	<0,3	<0,3
Granação dos Frutos	BAIXA	>=65	>=22	>=10	>=60	INFERIOR	<0,3	<0,3
Granação dos Frutos	BAIXA	>=65	>=22	>=10	>=60	SUPERIOR	<0,3	<0,3
Granação dos Frutos	BAIXA	>=65	>=22	>=10	>=60	MEDIO	<0,3	<0,3
Granação dos Frutos	BAIXA	>=65	>=22	>=10	>=60	INFERIOR	<0,3	<0,3
Granação dos Frutos	BAIXA	>=65	>=22	>=10	>=60	SUPERIOR	0,4-0,5	<0,3
Granação dos Frutos	BAIXA	>=65	>=22	>=10	>=60	MEDIO	0,4-0,5	<0,3
Granação dos Frutos	BAIXA	>=65	>=22	>=10	>=60	INFERIOR	0,4-0,5	<0,3
Granação dos Frutos	BAIXA	>=65	>=22	<10	>=60	SUPERIOR	0,4-0,5	0,3-0,4
Granação dos Frutos	BAIXA	>=65	>=22	<10	>=60	MEDIO	0,4-0,5	0,3-0,4
Granação dos Frutos	BAIXA	>=65	>=22	<10	>=60	INFERIOR	0,4-0,5	0,3-0,4
Granação dos Frutos	BAIXA	>=65	>=22	<10	>=60	SUPERIOR	0,4-0,5	0,4-0,5

Também criada uma base de dados em formato ARFF, que é o padrão de entrada da ferramenta Weka. Neste arquivo os campos são separados por vírgula, e os atributos

são especificados entre colchetes, ou seja, um atributo infestação da praga tem que possuir seus valores especificados entre colchetes, Exemplo: "@attribute infestação {BAIXA, MEDIA, ALTA}". A figura 2 ilustra como deve ser criado o formato ARFF (Figura 2).

BichoMineiroDiscretizacao.arff

```
@relation weather

@attribute Fenologia {GranacaoDosFrutos,MaturacaoDosFrutos,RepousoeSenescencia,Florada}
@attribute UmidRelativa {>=65,<65}
@attribute Temperatura {>=22,>=18<22,<18}
@attribute PrecipitacaoMedia {<10,>=10}
@attribute PrecipitacaoSoma {<60,>=60
@attribute Terco {SUPERIOR,MEDIO,INFERIOR}
@attribute EVI_Sim {<0.3,0.3-0.4,0.4-0.5,>=0.5}
@attribute Infestacao {BAIXA,MEDIA,ALTA}

@data
GranacaoDosFrutos,>=65,>=22,<10,>=60,SUPERIOR,<0.3,BAIXA
GranacaoDosFrutos,>=65,>=22,<10,>=60,MEDIO,<0.3,BAIXA
GranacaoDosFrutos,>=65,>=22,<10,>=60,INFERIOR,<0.3,BAIXA
GranacaoDosFrutos,>=65,>=22,>=10,>=60,SUPERIOR,<0.3,BAIXA
GranacaoDosFrutos,>=65,>=22,>=10,>=60,MEDIO,<0.3,BAIXA
GranacaoDosFrutos,>=65,>=22,>=10,>=60,INFERIOR,<0.3,BAIXA
GranacaoDosFrutos,>=65,>=22,>=10,>=60,SUPERIOR,<0.3,BAIXA
GranacaoDosFrutos,>=65,>=22,>=10,>=60,MEDIO,<0.3,BAIXA
GranacaoDosFrutos,>=65,>=22,>=10,>=60,INFERIOR,<0.3,BAIXA
GranacaoDosFrutos,>=65,>=22,>=10,>=60,SUPERIOR,0.4-0.5,BAIXA
GranacaoDosFrutos,>=65,>=22,>=10,>=60,MEDIO,0.4-0.5,BAIXA
GranacaoDosFrutos,>=65,>=22,>=10,>=60,INFERIOR,0.4-0.5,BAIXA
GranacaoDosFrutos,>=65,>=22,<10,>=60,SUPERIOR,0.4-0.5,BAIXA
GranacaoDosFrutos,>=65,>=22,<10,>=60,MEDIO,0.4-0.5,BAIXA
GranacaoDosFrutos,>=65,>=22,<10,>=60,INFERIOR,0.4-0.5,BAIXA
GranacaoDosFrutos,>=65,>=22,<10,>=60,SUPERIOR,0.4-0.5,BAIXA
```

Figura 3: Base de dados discretizada e convertida para formato arff.

3.2. Execução do Algoritmo de Mineração de Dados

O algoritmo J48 foi testado tanto na ferramenta Weka, como também em um API do Weka em linguagem Java no ambiente de programação Netbeans.

Esse estudo teve como objetivo de estudar o comportamento do algoritmo de mineração de dados J48 da ferramenta Weka como ferramenta de mineração de dados. Assim poderia verificar se o algoritmo em questão é eficiente na descoberta de conhecimentos sobre base de dados fitossanitários, junto a dados espectrais. Porém, além de executar esse algoritmo J48, também foram testados outros dois, para comparação do desempenho e precisão. Como critério de melhor precisão foi escolhido Correctly Classified Instances (Instancias classificada como corretas, em porcentagem) e a partir desse valor, foram calculadas as acurácias.

A execução do algoritmo Árvore de Decisão J48 foi feita com diversas configurações em seus parâmetros. Para isso foi utilizanda uma parcela dos dados tanto para treinamento teste (Exemplo, 80% dos dados para treinamento e 20% para teste),

também foi utilizado o cross-validation (validação cruzada), que obteve melhores resultados, além de diminuir a possibilidade de um possível vício do algoritmo.

3.3. Interpretação, Seleção e Validação dos Modelos

Esta etapa está associada com a etapa anterior, havendo pré-seleção dos diversos modelos, gerados pela ferramenta Weka, os quais foram levados a um especialista da área do conhecimento de pragas de cafeeiros visando a para validação do algoritmo. Esse processo é iterativo, portanto, o que nos leva a etapa anterior a cada modelo não aceito, ou seja, esse procedimento se repetiu inúmeras vezes até que se encontrou um modelo dito como aceitável. Abaixo são mostrados os 4 modelos de maior sucesso.

Modelo 1: Modelo gerado pela base de dados completa e discretizada, nessa base de dados foram testados diversos parâmetros do algoritmo. O modelo apresentado na figura 6, foi o modelo validado como o melhor pelo especialista da área de pragas do cafeeiro.

O modelo apresentou 85,9671% no grau de instâncias classificadas como corretas, o que é um índice muito bom, visto que o custo no processamento desse algoritmo é extremamente pequeno.

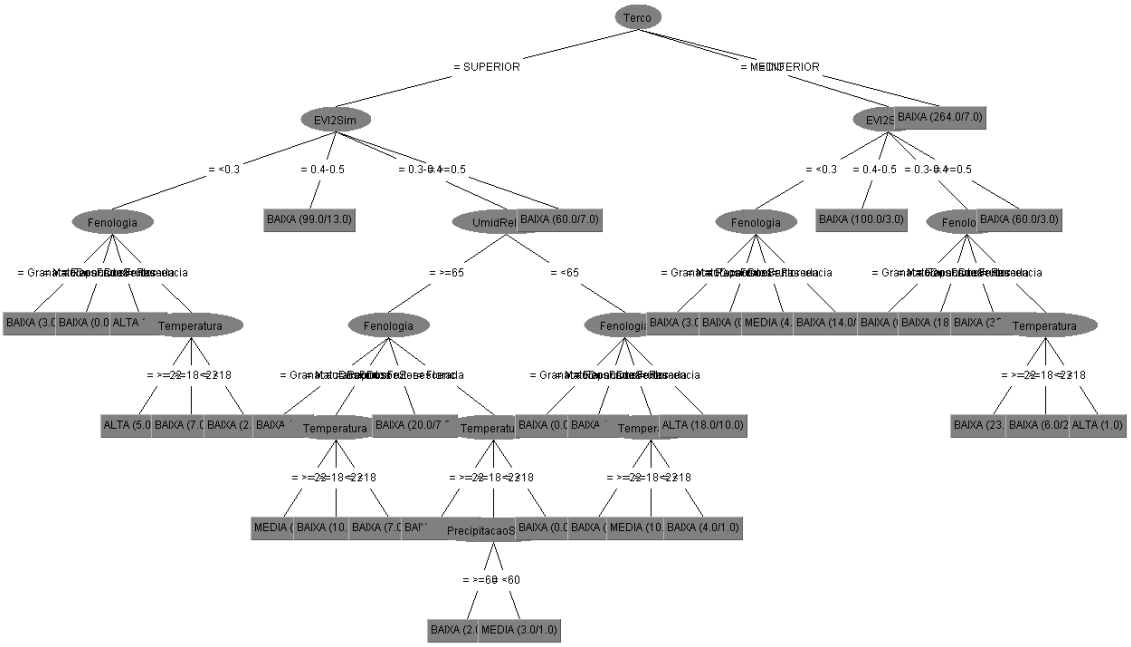

Figura 4: Modelo gerado pela base de dados completa e discretizada.

Modelo 2 : Modelo gerado pela base dados completa e não discretizada. Vale ressaltar que nesse modelo é o algoritmo quem divide os dados conforme sua infestação, já que os dados não são discretizados. Esse modelo obteve uma acurácia de quase 2% maior que o modelo anterior, sendo que o atributo mais significativo foi o EVI2 nó raiz.

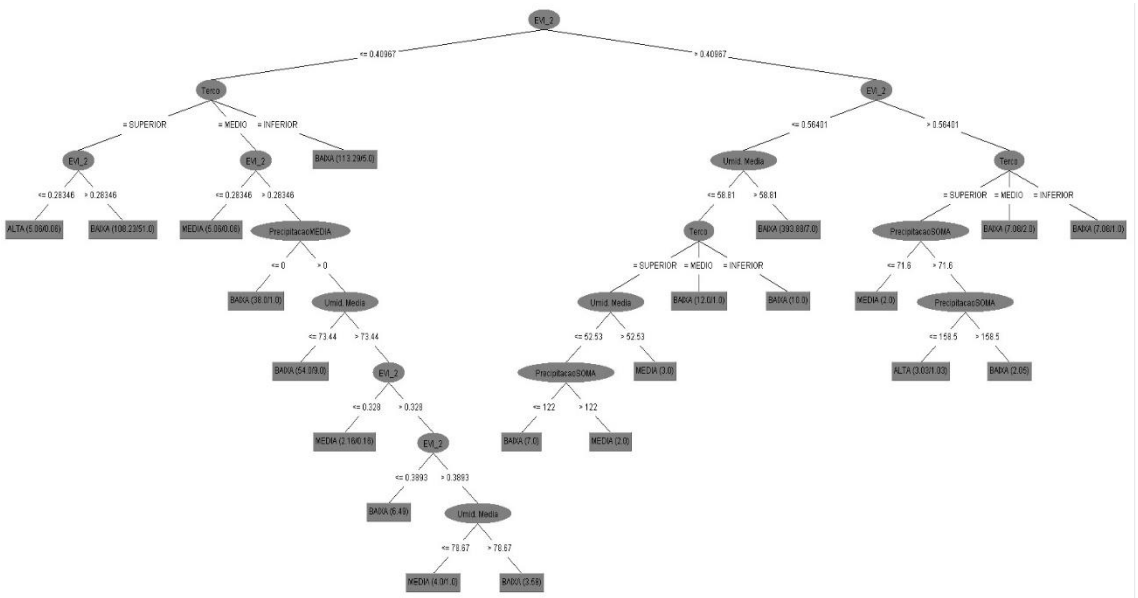

Figura 5: Modelo gerado pela base de dados completa e não discretizada.

Modelo 3: Modelo gerado pela nova base de dados sugerida pelos especialistas em pragas de cafeeiros, onde foi levado em consideração somente os dados do terço superior da planta, porém sem discretização.

Esse modelo obteve 73,4848% de instâncias classificadas como corretas, o que causou perda de 10% de precisão no modelo gerado com o terço superior da planta. Uma possível explicação para essa perda de precisão é um menor número de instâncias na amostra, ou seja, só as instâncias que tinham o atributo terço superior da planta. Diminuindo, assim, em um terço a base de dados do modelo 1 e 2.

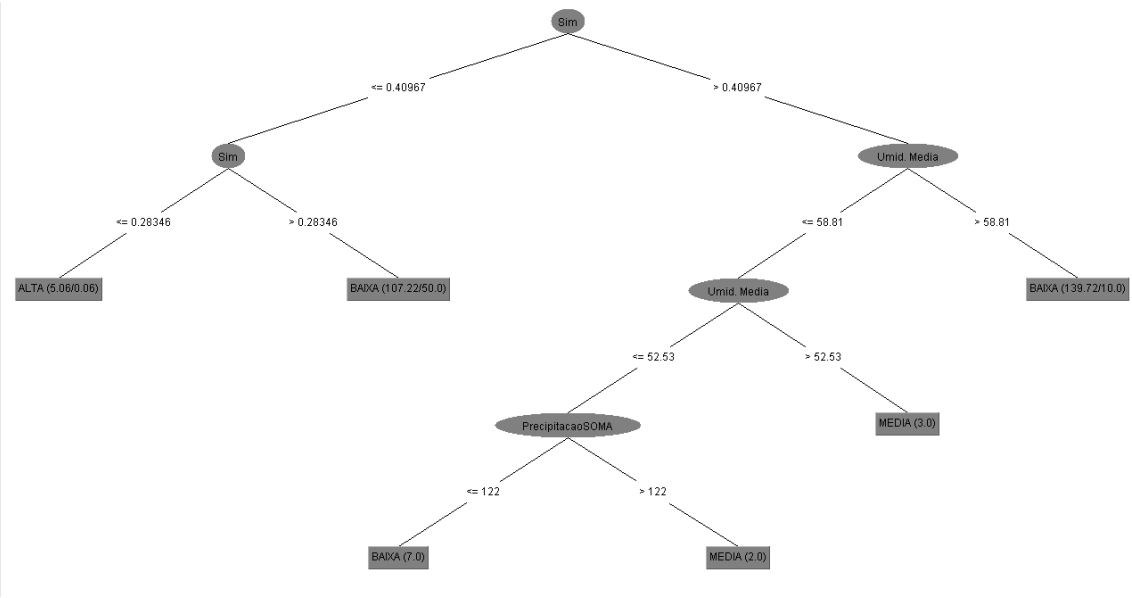

Figura 6: Modelo gerado pela base de dados não discretizada e somente terço superior.

Modelo 4: Modelo gerado pela base de dados discretizados e somente o terço superior da planta, modelo sugerido pelo especialista da área. Esse modelo obteve 67,67% de acerto (tendo Correctly Classified Instances como referência de acerto) o que o torna inviável, visto que o primeiro modelo obteve aproximadamente 85%. Já tendo essa mesma base de dados, contendo somente as instancias do terço superior da planta (modelo figura 5), sem a discretização obteve aproximadamente 73%, comprovando ser verdadeiro a literatura encontrada dizendo que a discretização daria resultados inferiores por perder muita informação.

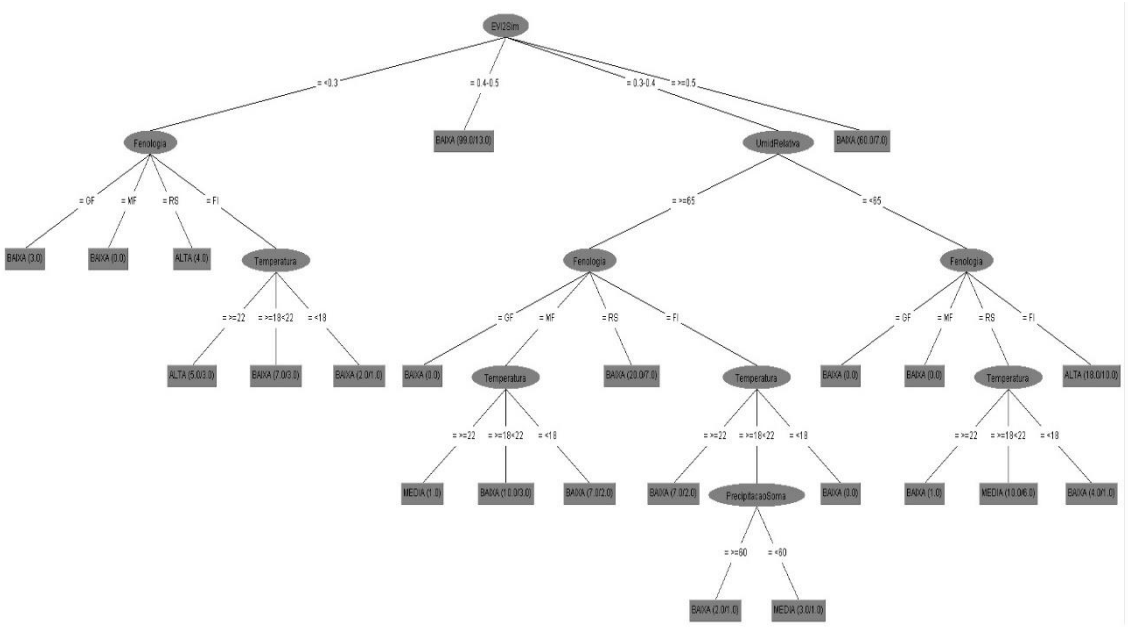

Figura 7: Modelo gerado pela base de dados discretizada e somente terço superior.

3.4. Utilizando a API do Weka

Visto que o Weka é um software livre em linguagem Java e possui um vasto conjunto de algoritmos de aprendizado em máquinas para Mineração de Dados (MD), o conjunto de dados do projeto pode ser utilizado por meio de sua biblioteca (API – *Application Programming Interface*) inserida ao código Java.

Tanto utilizando o Linux, quanto o Windows, a instalação foi bem simples e rápida. Foi adquirida a versão do Software Weka compactado, formato zip, e executando o arquivo weka.jar, quando importada sua biblioteca utilizando o NetBeans ou eclipse. No caso do Windows foi possível instalar o weka (versão de instalador executável, baixada do software), e navegar até a pasta onde se encontra o Weka para executar o weka.jar.

Passos para utilização da API no ambiente de programação NetBeans:

1- Criar o projeto Java com nomeDoSoftware ;
2- Na opção *arquivo*, selecionar *Propriedade do Projeto (nomeDoSoftware). P*ara isso também pode-se clicar com o botão direito do mouse no projeto e selecionar propriedade*;*
3- Na janela que abriu, escolher a opção *biblioteca*, clicar no botão *adicionar JAR/Pasta*, navegar até a pasta onde se encontra o Weka e selecionar o arquivo *weka.jar*, que se encontra no diretório da ferramenta Weka.

A implementação seguiu utilizando-se as bibliotecas do Weka e um exemplo contendo a utilização do algoritmo J48.

Importando as bibliotecas:

```
import java.util.Random;
import weka.core.Instances;
import weka.core.Instance;
import weka.core.converters.ConverterUtils.DataSource;
import weka.classifiers.trees.J48;
import weka.classifiers.Evaluation;
```

Foi utilizado um código exemplo, o que deu o mesmo resultado da ferramenta Weka. Foi necessário aprender como entrar com os parâmetros no código. Por exemplo, o algoritmo J48 por default efetua uma poda da árvore gerada, tornando assim necessária a utilização do parâmetro não poda para a base de dados do projeto. Para isso, foi feito um estudo e efetuada as modificações necessárias no código exemplo, protótipo de software da EPAMIG.

Exemplo da entrada de opção dos parâmetros no código.

```
// Imprmir chamada dos parâmetros como se estivesse usando a interface gráfica
System.out.println("Chamada de linha de código: \n");
String[] opcao = new String[10];
opcao[0] = "-U";  //Não podar a arvore.
opcao[1] = "";
opcao[2] = "";
opcao[3] = "";  //Estipula um Número mínimo de instancias.
opcao[4] = "";   //Implementa o número mínimo de instancias.
opcao[5] = "";
opcao[6] = "";
opcao[7] = "";
opcao[8] = "-t";  //Arquivo de treinamento
opcao[9] = caminhoDados;  //Caminho do arquivo de treinamento.
System.out.println(opcao+"Esta eh a opcao: ");
System.out.println(Evaluation.evaluateModel(new J48(), opcao));
```

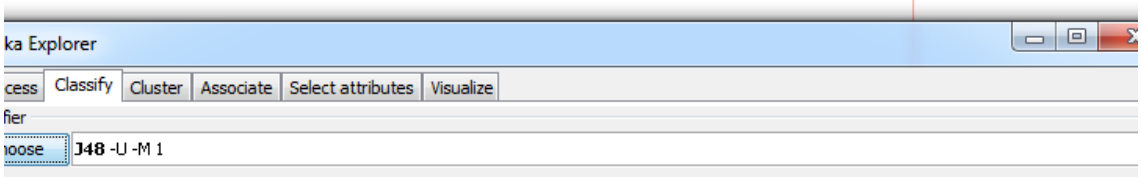

Figura 8:Pedaço do código implementado onde a variável opção recebe os parâmetros de configuração do algoritmo, e entrada dos padrões pela interface da Ferramenta Weka.

4. Discursão dos resultados

Todos os modelos gerados foram testados pelo algoritmo ID3 e MultilayerPerceptron.

Árvore de Decisão ID3: O algoritmo ID3 conseguiu o mesmo índice de acerto do J48, obtendo desempenho muito semelhante, porém só aceita dados discretos como entrada, e isso seria um possível problema visto que a mineração sobre dados não discretizados obteve um resultado melhor. Espera-se que com o acréscimo de mais instâncias na base de dados, melhore a eficiência da mineração de dados que visa a predição da infestação da praga Bicho mineiro em lavoura cafeeira.

MultilayerPerceptron: foi obtido utilizando-se redes neurais artificiais (RNA), esse algoritmo é uma modificação do Perceptron. O modelo obteve uma acurácia 2% acima do algoritmo J48. Para isso foi utilizado com entrada no algoritmo taxa de aprendizado de 0,3 e momento de 0,2. Diferenciando do padrão desse algoritmo na quantidade de camadas ocultas, possuindo seis camadas ocultas nas seguinte configuração: 8, 16, 32, 18, 9, 3.

Vale ressaltar que o problema encontrado nesse algoritmo foi sua demora na execução, demorando mais de 10 minutos em uma base de dados relativamente pequena, o que pode tornar inviável a aplicação da técnica em larga escala.

4.1 Resultado Obtido:

A leitura da árvore gerada é feita de cima para baixo, do nó raiz (atributo de maior significância) aos nós folhas (atributo estudado). Abaixo segue como seria interpretada a árvore do modelo 3.

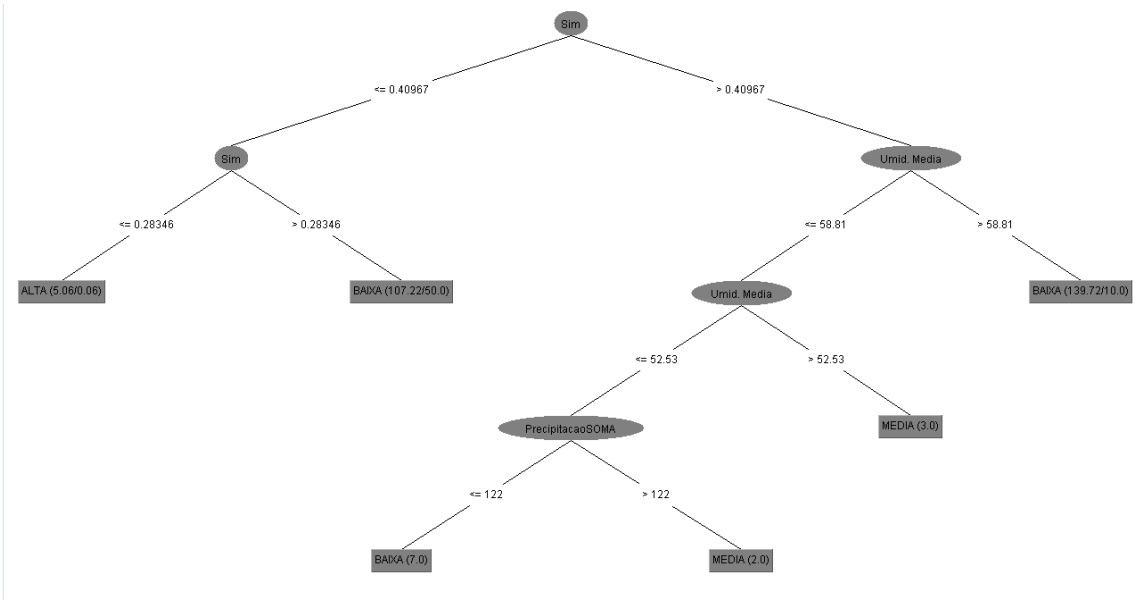

Figura 9: Modelo gerado pela base de dados não discretizada e somente terço superior.

Foram testadas diversas configurações, sendo que a mais eficiênte foi utilizando-se configurações padrões e sem poda.

A árvore (figura 5) possui uma configuração simples, tendo como atributo de maior significância o EVI2. Os valores de EVI2 menores que 0,3 necessitam de outro atributo, a fenologia, para poder então selecionar o seu grau de infestação, sendo que nas fases de granação dos frutos e maturação possui baixa infestação e no repouso e senescência uma alta infestação. Já na florada, leva-se em consideração o atributo temperatura, apresentando baixa infestação em todas as temperaturas testada, exceto, com temperatura maior que 22º C.

O EVI2 maior ou igual a 0,4 apresenta baixa infestação e maior ou igual a 0,3 e menor que 0,4 necessita do atributo umidade relativa (UR). Umidade relativa maior ou igual a 65 por cento teve o atributo fenologia, sendo fenologia repouso e senescência e granação dos frutos com baixa infestação, a fenologia maturação dos frutos, teve o atributo temperatura para diferenciar, sendo a temperatura menor que 22º C possui baixa infestação e maior ou igual a 22º C possui grau de infestação médio. Na fenologia florada, e com atributo temperatura, verifica-se que as temperaturas extremas, ou seja, igual ou maior que 22° C e inferiores a 18° C possuem baixa infestação. Temperaturas maiores ou iguais a 18° C até menor que 22° C necessitam da soma das precipitações para diferenciar, sendo que essas menores que 60 mm, apresentam média infestação e outros resultados apresentam baixa infestação.

Com EVI2 entre 0,3 e 0,4 e umidade relativa menor que 65 mm, necessitou do atributo fenologia para diferenciar, sendo granação dos frutos e maturação dos frutos possuem baixa infestação. No repouso e senescência, precisa do atributo temperatura, sendo as temperaturas maiores e iguais a 22° C apresentam baixa infestação, entre 18 e 22° C, apresentam média infestação e menores que 18 ° C baixa infestação.

EVI2 entre 0,4 e 0,5 com fenologia florada apresentaram alta infestação.

5. Conclusão

Conclui-se que o algoritmo J48 do software Weka é eficiente para solucionar o problema de obtenção do conhecimento sobre a infestação do Bicho mineiro em cafeeiros do Sul de Minas Gerias, a partir do banco se dados fitossanitários, climáticos e espectrais adquiridos pela pesquisa da EPAMIG em 10 anos.

Os resultados foram validados por especialistas da área de pragas dos cafeeiros, sendo necessárias várias reuniões para discussão dos modelos.

Foram produzidos vários modelos com boa precisão.

Para a validação dos dados espectrais, há a necessidade de mais estudos. Porém, a princípio, podem ser utilizados como um atributo de decisão, ainda com a necessidade dos dados fitossanitários e climáticos.

A facilidade de se utilizar a API (Application Programing Interface) do Weka, demonstrou ser muito interessante para a implementação de banco de dados inteligentes, e ou, softwares de previsão da infestação de pragas nas lavouras, o que seria inovador e importante no processo de otimização da produção do café.

6. Referências Bibliográficas

AGRIOS, G.M. Plant pathology. 3ed. London: Academic Press, 1988. 803p.

AKUTSU, M. **Relações de funções climáticas e bióticas com a taxa de infecção da ferrugem do cafeeiro (*Hemileia vastatrix*)**. Viçosa: UFV, 1981. 67p. (Dissertação – Mestrado em Fitopatologia).

ALMEIDA, S. R. Doenças do cafeeiro. In: RENA, A. B.;MALAVOLTA, E., *et al* (Ed.). **Cultura do cafeeiro: Fatores que afetam a produtividade**. Piracicaba: Associação para Pesquisa da Potossa e do Fosfato, 1986. p.391-400.

BUCENE, L. C.; RODRIGUES, L. H. A.; MEIRA, C. A. A. Mineração de dados climáticos para previsão de geada e deficiência hídrica para as culturas do café e da cana-de-açúcar para o Estado de São Paulo - Documento 20. **Documentos**, Campinas/SP, p.41, 2002. Disponível em: < http:// www.cnptia.embrapa.br/files/doc20.pdf > . Acesso em: 15 junho 2012.

CARDOSO, O. N. P. **Gestão do conhecimento usando data mining**: estudo de caso na UFLA. Lavras/MG: Universidade Federal de Lavras, 2005.(Mestrado em Administração), 2005.

CARVALHO, V. L.; CUNHA, R. L.; GUIMARAES, P.T. G.; CARVALHO, J..P. F.; Influência do zinco na incidência de doenças do cafeeiro. **Ciência e Agrotecnologia**, Lavras, v. 32, n. 3, p. 804-808, maio/jun, 2008.

CARVALHO, V. L.; CHALFOUN, S. M. Manejo integrado das principais doenças do cafeeiro. **Informe Agropecuário,** v. 19, n.3, p. 27-35, 1998.

CARVALHO, V.L. de; CHALFOUN, S.M. Doenças do cafeeiro: diagnose e controle. **Boletim Técnico,** v. 58, p. 44, 2000.

CARVALHO, V.L.; CHALFOUN, S.M.; CUNHA, R.L. Manejo de doenças do cafeeiro, p.689-756. In: Reis, P.R.; Cunha, R.L. (Eds) **Café Arábica do plantio a colheita**. Lavras:URESM, v.1, 2010. 896p.

CHALFOUN, S. M. Doenças do cafeeiro : importância, identificação e métodos de controle. Lavras: UFLA/FAEPE, 1997. 96p.

CHALFOUN, S.M.; CARVALHO, V.L. de; PEREIRA, M.C. Efeito de alterações climáticas sobre o progresso da ferrugem (*Hemileia vastratrix* Berk. & Br.) do cafeeiro (*Coffea arabica* L.). **Ciência e agrotecnologia**, Lavras, v.25, n.5, p.1248-1252, set/out., 2001.

CORTÊS, S. D. C.; PORCARO, R. M.; LIFSCHITZ, S. **Mineração de Dados** - Funcionalidades, Técnicas e Abordagens.Rio de
Janeiro/RJ: PUC, 2002.Monografia, Ciência da Computação, 2002.

FREITAS, R. M. D., ARAI, E., ADAMI, M., FERREIRA, A. S., SATO, F. Y., SHIMABUKURO, Y. E., ROSA, R. R., et al. (2011). Virtual laboratory of remote sensing time series : visualization of MODIS EVI2 data set over South America, 2, 57-68.

FREITAS, R. M.; ARAI, E.; ADAMI, M.; SOUZA, A. F.; SATO, F. Y.; SHIMABUKURO, Y. E.; ROSA, R. R.; ANDERSON, L. O.; RUDORFF, B. F. T. Virtual laboratory of remote sensing time series: visualization of MODIS EVI2 data set over South America. **Journal of Computational Interdisciplinary Sciences**. v. 2, n.1, p. 57-68, 2011.

FAYYAD, U.; PIATESKY-SHAPIRO, G.; SMYTH, P.; UTHURUSAMY, R. **Advances in knowledge discovery and data mining**. Cambridge: MIT Press, 1996. 560p.

HAIR Jr., J.F. et al. Análise Multivariada de Dados. 5. ed. Porto Alegre: Bookman, 2005.

HAN, J.; KAMBER, M. **Data Mining** - Concepts and Techniques. San Francisco - USA: Morgan Kaufmann, 2006.INSTITUTO BRASILEIRO DO CAFÉ. **Doenças do cafeeiro**. In.: Cultura de café no Brasil; Manual de Recomendações. Rio de Janeiro: IBC/GERCA, 1981. p. 331-378.

HEUSER, C. A. **Projeto de Banco de Dados**- 4ª Edição –Instituto de Informática da UFRGS (Universidade Federal do Rio Grande do Sul), 1998.

INSTITUTO BRASILEIRO DO CAFÉ. **Doenças do cafeeiro**. In.: Cultura de café no Brasil; Manual de Recomendações. Rio deJaneiro: IBC/GERCA, 1981. p. 331-378.

JIANG Z, HUETE AR, DIDAN K & MIURA T. 2008. Development of a two-band Enhanced Vegetation Index without a blue band. **Remote Sensing of Environment**, v. 112, n. 10, p. 3833–3845, 2008.

JUSTICE, C.O.; VERMOTE, E.; TOWNSHEND, J.R.G.; DEFRIES, R.; ROY, D.P.; HALL, D.K.; SALOMONSON, V.V.; PRIVETTE, J.L.; RIGGS, G.; STRAHLER, A. The moderate resolution imaging spectroradiometer (MODIS): land remote sensing for global change research. **IEEE Transactions on Geoscience and Remote Sensing**, v.36, n. 4, p. 1228-1249, 1998.

MATIELLO, J.B. **O Café:** do cultivo ao consumo. São Paulo: Editora Globo S.A., 1991. cap. 24, p. 345-363. (Coleção do Agricultor – Grãos).

MATIELLO, J. B.; ALMEIDA, S.R. Controle associado de doenças do cafeeiro. **Correio Agrícola**, São Paulo: Bayer SA:, p.25-27, 1997. (Murayama, 2002).

MEIRA, C. A. A.; RODRIGUES, L. H. A.; MORAES, S. A. Análise da epidemia da ferrugem do cafeeiro com árvore de decisão. **Tropical Plant Pathology**, v. 33, n. 2, p. 114-124, Mar./Apr. 2008.

REZENDE, S. O.; PUGLIESI, J. B.; MELANDA, E. A.; PAULA, M. F., 2003. **Mineração de dados.** In: REZENDE, S. O., ed., Sistemas Inteligentes: Fundamentos e Aplicações: Barueri/SP, Manole, p. 307-335.

SILVA, M. P. D. S. **Mineração de padrões de mudança em imagens de sensoriamento remoto.** São José dos Campos/SP: INPE, 2006. (Mestrado em Computação Aplicada).

SOUZA, S.M.C. **Importância da chuva e da temperatura do ar na incidência da ferrugem (*Hemileia vastatrix* Berk. & Br.) em cafeeiros, de três localidades do estado de Minas Gerais**. Lavras: ESAL, 1980. 50p. (Dissertação – Mestrado em Fitotecnia).

Witten, I. H.; Eibe Frank, E.; Hall, M. A. Data Mining Practical Machine Learning Tools and Techniques Third Edition, Burlington, USA, 2011.

ZAMBOLIM, L.; RIBEIRO do VALE, F. X.; PEREIRA, A. A.; CHAVES, G. M. **Café (*Coffea arábica* L.), controle de doenças**. In: Ribeiro do Vale, F. X.; Zambolim, L. Controle de doenças de plantas: grandes culturas. Viçosa: Departamento de Fitopatologia; Brasília: Ministério da Agricultura e Abastecimento, v.2, p.83—Metodologia.