UNIVERSIDADE ESTADUAL DE CAMPINAS

Faculdade de Engenharia Agrícola

THIAGO TOSHIYUKI THAMADA

DESENVOLVIMENTO DE *ENSEMBLES* PARA ALERTA DA FERRUGEM DO CAFEEIRO EM PERÍODO CRÍTICO DE PROGRESSO DA DOENÇA

CAMPINAS

2016

THIAGO TOSHIYUKI THAMADA

DESENVOLVIMENTO DE *ENSEMBLES* PARA ALERTA DA FERRUGEM DO CAFEEIRO EM PERÍODO CRÍTICO DE PROGRESSO DA DOENÇA

Dissertação apresentada à Faculdade de Engenharia Agrícola da Universidade Estadual de Campinas como parte dos requisitos exigidos para a obtenção do título de Mestre em Engenharia Agrícola, na Área de Gestão de Sistemas na Agricultura e Desenvolvimento Rural.

Orientador: Prof. Dr. CARLOS ALBERTO ALVES MEIRA
Coorientador: Prof. Dr. LUIZ HENRIQUE ANTUNES RODRIGUES

ESTE EXEMPLAR CORRESPONDE À VERSÃO FINAL
DA DISSERTAÇÃO DEFENDIDA PELO ALUNO
THIAGO TOSHIYUKI THAMADA, E ORIENTADO PELO
PROF. DR. CARLOS ALBERTO ALVES MEIRA.

CAMPINAS
2016

Agência(s) de fomento e nº(s) de processo(s): CNPq, 131038/2014-1

Ficha catalográfica
Universidade Estadual de Campinas
Biblioteca da Área de Engenharia e Arquitetura
Elizangela Aparecida dos Santos Souza - CRB 8/8098

T329d

Thamada, Thiago Toshiyuki, 1986-
Desenvolvimento de ensembles para alerta da ferrugem do cafeeiro em período crítico de progresso da doença / Thiago Toshiyuki Thamada. – Campinas, SP : [s.n.], 2016.

Orientador: Carlos Alberto Alves Meira.
Coorientador: Luiz Henrique Antunes Rodrigues.
Dissertação (mestrado) – Universidade Estadual de Campinas, Faculdade de Engenharia Agrícola.

1. Mineração de dados (Computação). 2. Hemileia vastatrix. 3. Café - Doenças e pragas. 4. Modelagem. 5. Doenças - Controle. I. Meira, Carlos Alberto Alves. II. Rodrigues, Luiz Henrique Antunes,1959-. III. Universidade Estadual de Campinas. Faculdade de Engenharia Agrícola. IV. Título.

Informações para Biblioteca Digital

Título em outro idioma: Development of ensembles to coffee rust warning in critical period of disease progress
Palavras-chave em inglês:
Data mining
Hemileia vastatrix
Coffee
Modeling
Disease control
Área de concentração: Gestão de Sistemas na Agricultura e Desenvolvimento Rural
Titulação: Mestre em Engenharia Agrícola
Banca examinadora:
Carlos Alberto Alves Meira [Orientador]
Flávia Rodrigues Alves Patrício
Stanley Robson de Medeiros Oliveira
Data de defesa: 29-02-2016
Programa de Pós-Graduação: Engenharia Agrícola

Este exemplar corresponde à redação final da **Dissertação de Mestrado** defendida por **Thiago Toshiyuki Thamada**, aprovada pela Comissão Julgadora em 29 de fevereiro de 2016, na Faculdade de Engenharia Agrícola da Universidade Estadual de Campinas.

FEAGRI

Dr. Carlos Alberto Alves Meira– Presidente e Orientador
CNPTIA/EMBRAPA

Dr.ª Flávia Rodrigues Alves Patrício – Membro Titular
Instituto Biológico/Campinas

Dr. Stanley Robson de Medeiros Oliveira – Membro Titular
CNPTIA/EMBRAPA

Faculdade de Engenharia Agrícola Unicamp

A Ata da defesa com as respectivas assinaturas dos membros encontra-se no processo de vida acadêmica do discente.

DEDICATÓRIA

À minha irmã Andressa e meus pais Valdemar e Chieko, pelo apoio, compreensão e amor incondicional.

AGRADECIMENTOS

À Faculdade de Engenharia Agrícola (FEAGRI/UNICAMP) e à Empresa Brasileira de Pesquisa Agropecuária, pela disponibilização de suas infraestruturas.

À Fundação PROCAFÉ por ceder os dados usados nesta dissertação.

Ao Prof. Dr. Carlos Alberto Alves Meira pela orientação, ensinamento, paciência, conselhos, conversas e apoio durante o mestrado.

Ao Prof. Dr. Luiz Henrique Antunes Rodrigues pela co-orientação, ensinamento, conselhos e apoio.

Ao Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq) pelo apoio financeiro.

Ao Prof. Dr. Stanley Robson de Medeiros Oliveira pelos conselhos na área de mineração de dados.

Ao Prof. Dr. Paulo Cesar Sentelhas e à Dra. Flávia Rodrigues Alves Patrício pelos conselhos na àrea de fitopatologia.

Ao Me. Fernando Dill Hinnah pelo auxílio na verificação e correção dos dados.

Aos meus amigos e demais professores e funcionários da FEAGRI/UNICAMP, que direta ou indiretamente contribuíram para a realização do curso.

Aos meus familiares pelo apoio, carinho e compreensão.

RESUMO

A ferrugem, causada pelo fungo *Hemileia vastatrix,* é a principal doença do cafeeiro. Em média, causa perdas de 35% na produção. Seu controle é realizado por meio de fungicidas seguindo um calendário fixo, iniciado em dezembro. A curva de progresso padrão da ferrugem começa entre dezembro e janeiro, atinge o pico por volta de junho e decresce. Variações no clima podem adiar sua epidemia ou manter altos seus índices ao fim de seu ciclo, em agosto (ferrugem tardia). Nessas situações o controle tradicional é ineficiente, sendo preciso revê-lo e readaptá-lo.

O correto posicionamento das aplicações de fungicidas permite controlar eficientemente a doença durante seu ciclo. Uma ferramenta que pode auxiliar neste posicionamento são os sistemas de alerta de doenças de plantas que, por meio de modelos preditivos baseados em técnicas de mineração de dados, predizem quando uma doença atingirá níveis críticos. Modelos preditivos para ferrugem do cafeeiro foram criados para lavouras com alta e baixa carga pendente de frutos. Os modelos prediziam se a taxa de progresso da doença seria maior ou igual a um limiar. Modelos para baixa carga quando avaliados com novos dados apresentaram desempenho inferior em relação à obtida em sua criação.

O *ensemble*, comitê de modelos preditivos, é uma nova abordagem que pode criar modelos com desempenho superior aos atuais e auxiliar na predição da ferrugem tardia. Para isso, foi proposto usar dados do período crítico para epidemia da ferrugem, dezembro a junho, e definir um novo limiar para baixa carga.

Ensembles desenvolvidos mostraram acurácia acima de 70,00% para alta (78,00% – limiar de 5 pontos percentais (p.p). / 73,33% – 10 p.p.) e baixa carga (72,64% – 5 p.p.). Novo limiar para baixa carga foi definido em 3 p.p. Ocorrências da ferrugem tardia foram detectadas em 40,00% e 66,67% dos casos nos *ensembles* de alta e baixa carga, respectivamente. Atributos ligados ao molhamento foliar foram mais relevantes para ferrugem tardia em lavouras com alta carga, em cafeeiros com baixa carga a temperatura máxima e precipitação foram as mais relevantes.

Ensembles com valor de incidência da ferrugem no mês anterior apresentaram, na média, melhor desempenho preditivo. Variáveis relacionadas ao número de dias favoráveis, em um mês, à ferrugem não auxiliaram no desenvolvimento de *ensembles* melhores. Os

ensembles criados nesse estudo mostraram desempenho superior aos modelos preditivos atuais.

Palavras-chave: mineração de dados, café, ensemble, floresta aleatória, modelagem, *Hemileia vastatrix.*

ABSTRACT

Coffee leaf rust is caused by the fungus *Hemileia vastatrix* and it is the main coffee disease. Usually, it causes losses of 35%. Its control is done by fungicides through fixed calendar starting in december. The typical coffee rust progress curve starts in december/january, reaches its maximum around june and decreases. Climate variations can delay its epidemy or maintain high disease rates until the end of its cycle, in august (late rust). In these situations the typical disease control is inefficient and must be reviewed and readapted.

Fungicide applications in proper time can efficiently control the disease during its cycle. Plant disease warning systems are tools capable to assist in these applications and can, through predictive models based on data mining techniques, predict when a disease will reach critical levels. Predictive models for coffee rust were developed in crops with high and low fruit load. The models predicted if the rust progress rate is equal or greater than a threshold. Models for low fruit load showed lower performance when evaluated with new data compared to that obtained in its creation.

Ensembles, committee of predictive models, are a new approach that can create models with better performance than the current ones and assist in predicting late rust. This study proposed the use of data in critical period to coffee rust epidemy, december to june and find a new threshold for crops with low fruit load.

Ensembles developed in this work showed accuracy up to 70.00% for high (78.00% – 5 percentage points (pp) as threshold / 73.33% – 10 pp) and low fruit load (72.64% – 5 p.p.). The new threshold found for low fruit load data was 3 p.p. Occurrences of late rust in high and low fruit load crop were identified in 40.00% and 66.67% cases, respectively. Attributes based on leaf wetness were most relevant for predicting late rust in high fruit load crops. Maximum temperature and rainfall were the most relevant for late rust predictions in low fruit load crops.

Ensembles with incidence value of the previous month presented, on average, better predictive performance. Variables related to the number of favorable days to coffee rust incidence, in a month, did not help to develop bettter ensembles. Ensembles created in this study showed better results in comparison to actual predictive models.

Keywords: data mining, coffee, ensemble, random forest, modeling, *Hemileia vastatrix.*

LISTA DE FIGURAS

LISTA DE TABELAS

SUMÁRIO

1 INTRODUÇÃO

O Brasil é o maior produtor e exportador de café em grão do mundo e o segundo maior consumidor (USDA, 2015). Sua produção estimada para a safra 2015 é de, aproximadamente, 43,24 milhões de sacas de 60 quilos de café beneficiado (CONAB, 2015).

A ferrugem do cafeeiro, causada pelo fungo *Hemileia vastatrix* Berk. et Br., está presente em todas as regiões produtoras de café do Brasil. É a principal doença da cultura e causa, em média, perdas de 35% na produção quando as condições climáticas favorecem sua epidemia. Sob determinadas condições pode ocorrer perdas superiores a 50% na produção (ZAMBOLIM et al., 2005).

O cafeeiro alterna anos de alta e baixa carga pendente de frutos, caracterizando um ciclo bienal (AVELINO et al., 2015). A ferrugem ataca com maior intensidade em anos de alta carga por motivos ainda desconhecidos. Entretanto, acredita-se que a drenagem de fotossintetizados das folhas para os frutos aliada ao alto índice de enfolhamento do cafeeiro no início do período chuvoso e condições climáticas favoráveis à epidemia da ferrugem estejam entre os motivos da maior intensidade da doença em lavouras de alta carga (ZAMBOLIM et al., 2005; LÓPEZ-BRAVO et al., 2012).

A curva de progresso padrão da doença em lavouras com alta carga pendente inicia-se entre dezembro e janeiro, aumenta de forma logarítmica de março a abril, atinge seu máximo por volta de junho e decresce devido às baixas temperaturas e queda de folhas por decorrência da ferrugem e da colheita de grãos. A doença pode ser controlada através de aplicações de fungicidas, seguindo calendário fixo, com início em dezembro (CHALFOUN et al., 2001; AVELINO e SAVARY, 2002; ZAMBOLIM et al., 2005).

Variações climáticas podem mudar a curva de progresso da ferrugem. A elevação da temperatura e chuvas intensas em dezembro e janeiro, que promovem a lavagem do inóculo da superfície foliar, atrasam o início do desenvolvimento da doença. Simultaneamente, o aumento da temperatura média anual e ocorrência de chuvas esporádicas, entre abril e julho, permitem altos índices da doença ao fim de seu ciclo, em agosto (CHALFOUN e ZAMBOLIM, 1985; CHALFOUN et al., 2001).

Em anos de desenvolvimento tardio da ferrugem o controle tradicional da doença é ineficiente (CHALFOUN e CARVALHO, 1999). Nesses casos é necessário rever e readaptar as medidas de controle às novas condições de progresso da doença (CHALFOUN et

al., 2001) como, por exemplo, realizar aplicações adicionais de fungicidas para evitar prejuízos.

Entretanto, o uso indiscriminado de agroquímicos pode causar contaminação de alimentos, solo e água, reduzir a biodiversidade (BETTIOL e GHINI, 2001) e aumentar a resistência do patógeno aos fungicidas (BOSCH et al., 2011).

Posicionar adequadamente as medidas de controle permite manter a intensidade da doença em níveis baixos ao longo de seu ciclo e evita os problemas do uso excessivo de fungicidas. Uma ferramenta que pode auxiliar neste posicionamento são os sistemas de alerta de doenças de plantas (HARDWICK, 2006; AVELINO et al., 2015). Estes tipos de sistemas de alerta buscam predizer, geralmente por meio de modelos preditivos, quando uma doença poderá atingir níveis críticos na lavoura.

Modelos preditivos em árvore de decisão, uma técnica de mineração de dados, para a ferrugem foram desenvolvidos usando dados mensais meteorológicos e a incidência da doença em cafeeiros com alta (MEIRA et al., 2009) e baixa carga pendente de frutos (MEIRA e RODRIGUES, 2009). Esses modelos predizem se a taxa de progresso da ferrugem, diferença entre a incidência da doença em dois meses subsequentes, será maior ou igual a 5 ou 10 pontos porcentuais (p.p.). Nesse sentido, também foram desenvolvidos modelos em máquinas de vetores suporte, redes neurais artificiais e florestas aleatórias (GIROLAMO NETO et al., 2014).

É natural a procura por novas técnicas para obter modelos preditivos com desempenho superior aos atuais. O *ensemble* combina as predições de um conjunto de modelos na predição de novos registros. Geralmente, o *ensemble* apresenta acurácia superior a qualquer modelo que o compõe (OPITZ e MACLIN, 1999). Dentre as técnicas *ensemble* destacam-se: floresta aleatória, *bagging*, *boosting* e *stacking*. Estudos mostram que *boosting* pode apresentar melhor desempenho preditivo que florestas aleatórias (ROKACH, 2010) e *stacking* ser superior ao *boosting* (TING e WITTEN, 1999). Já o *bagging* é mais robusto a ruídos em relação ao *boosting* (KOTSIANTIS, 2011).

Os modelos preditivos mencionados foram desenvolvidos e avaliados usando dados coletados ao longo do ano agrícola, setembro a agosto. Entretanto, para questão do controle eficiente da ferrugem de desenvolvimento tardio, dados do período mais crítico para o desenvolvimento da epidemia, de dezembro a junho, e a correta predição nesse período são mais interessantes para estudo. Além disso, os dados brutos foram preparados de maneira idêntica em todos os estudos apresentados, resultando em um conjunto de dados final

contendo os mesmos atributos. Dados de incidência serviram apenas na elaboração da taxa de progresso da doença.

Modelos para lavoura com baixa carga não apresentaram bom desempenho preditivo (MEIRA e RODRIGUES, 2009) ou não apresentaram desempenhos satisfatórios quando avaliados com dados de anos agrícolas recentes (MEIRA et al., 2014). Sendo assim, pode-se cogitar um limiar para a taxa de progresso da ferrugem inferior a 5 p.p., já que a epidemia da ferrugem é menos acentuada em anos de baixa carga. O valor do limiar pode ser identificado por meio da técnica de discretização, onde se associa cada valor numérico a um intervalo. A quantidade de intervalos é finita e estes não se sobrepõem (GARCÍA et al., 2013). Dessa forma, é possível verificar qual intervalo melhor divide o conjunto de dados para estabelecer um novo limiar.

A hipótese deste trabalho é:

- A criação de *ensembles* para lavouras de café com alta e baixa carga pendente de frutos, a partir de dados meteorológicos e de incidência da ferrugem referentes ao período crítico para a evolução da doença, que determinem aumentos em sua taxa de progresso mensal pode melhorar a tomada de decisão sobre o controle da doença, detectando meses críticos para progresso da doença, especialmente se as condições climáticas forem propícias à ferrugem tardia.

2 OBJETIVOS

2.1 Objetivo Geral

Desenvolver *ensembles* gerados por técnicas de mineração de dados para predizer aumentos na taxa de progresso da ferrugem do cafeeiro no período crítico para evolução da doença, de dezembro a junho, e com relação a limiares de referência para seu controle, a partir de dados meteorológicos e avaliações mensais da incidência da doença no campo.

2.2 Objetivos Específicos

Os objetivos específicos são:

- Realizar uma preparação de dados distinta da efetuada por Meira et al. (2009) e Girolamo Neto et al. (2014), utilizando dados mensais de atributos meteorológicos (temperatura, umidade relativa do ar e precipitação) e de incidência da ferrugem na lavoura.

- Obter um novo limiar para o atributo meta por meio da discretização dos valores de incidência mensal da ferrugem em cafeeiros com baixa carga pendente de frutos.

- Desenvolver *ensembles* para cafeeiros em alta e baixa carga com bons desempenhos preditivos, avaliados por meio de métricas relevantes.

- Avaliar se os *ensembles* desenvolvidos são capazes de predizer corretamente o primeiro mês em que a taxa de progresso da ferrugem atingirá determinado limiar e/ou os meses de maior desenvolvimento da doença.

- Avaliar os desempenhos preditivos de Girolamo Neto et al. (2014) durante o período crítico na evolução da ferrugem, dezembro a junho, e comparar com o desempenho dos *ensembles* desenvolvidos.

3 REVISÃO BIBLIOGRÁFICA

3.1 A Cultura do Café e a Ferrugem do Cafeeiro

O cafeeiro pertence à família Rubiaceae e é classificado em dois gêneros, *Coffea*, nativo de florestas intertropicais da África, Madagascar e ilhas do Oceano Índico, e *Psilanthus*, originário da Ásia ou África (CROS et al., 1998). Dentre as espécies do gênero *Coffea*, as principais são *Coffea arabica* (café arábica) e *Coffea canephora* (café robusta ou conilon) (ALVES et al., 2006). A ferrugem afeta de modo mais severo, dentre as espécies de café cultivadas, o café arábica (AVELINO et al., 2015).

O Brasil é o maior produtor e exportador de café em grão e o segundo maior consumidor de café do mundo (USDA, 2015). Os principais estados brasileiros produtores de café são Minas Gerais, Espírito Santo e São Paulo. A produção estimada para a safra 2015 é de, aproximadamente, 43,24 milhões de sacas, sendo 32,00 milhões de sacas (74,1% da produção total) de café arábica e 11,19 milhões de café conilon. A área total da cultura do café (arábica e conilon) no Brasil, na safra 2015, totalizou 2.248,9 mil hectares, sendo 326,8 mil hectares (14,5%) em formação e 1.922,1 mil hectares (85,5%) em produção (CONAB, 2015).

A ferrugem é a principal doença do cafeeiro. Seu agente causal é o fungo *Hemileia vastatrix* Berk. et Br., que pode ser encontrado em todas as regiões do mundo onde o café é cultivado. Na média, a ferrugem provoca 35% de perdas na produção se as condições climáticas favorecerem a epidemia da doença. Em caso de estiagem prolongada durante os períodos de maior severidade da ferrugem, pode-se perder mais de 50% da produção (ZAMBOLIM et al., 2005). No Brasil, a doença foi identificada pela primeira vez no sul da Bahia em janeiro de 1970 e após quatro meses pôde ser encontrada em cafeeiros de quase todos os estados brasileiros (WALLER, 1982; ZAMBOLIM et al., 1997).

O cafeeiro possui ciclo de produção bienal, ou seja, alterna anos de alta e baixa carga pendente de frutos (AVELINO et al., 2015). A ferrugem ataca o cafeeiro com maior intensidade em anos de alta carga. Os motivos para esse comportamento ainda são desconhecidos, mas acredita-se que a drenagem de fotossintetizados das folhas aos frutos aliada ao alto índice de enfolhamento da planta no começo do período chuvoso e clima favorável à epidemia da ferrugem sejam motivos para maior intensidade da doença nos cafeeiros em alta carga (ZAMBOLIM et al., 2005; LÓPEZ-BRAVO et al., 2012).

A doença causa desfolha e seca dos ramos e, consequentemente, perdas na produção atual e do próximo ano. A desfolha natural, somada à provocada pela ferrugem antes do florescimento, prejudica o desenvolvimento dos botões florais e da frutificação. Se a desfolha ocorrer no período de produção dos frutos, pode ocorrer formação de grãos defeituosos e/ou de má qualidade e provocar o chochamento dos frutos. Seguidos ataques da doença à lavoura podem diminuir a longevidade das plantas e causar perdas nos ramos laterais (AGRIOS et al., 2005, CARVALHO et al., 2010).

Os sintomas da ferrugem na face inferior das folhas infectadas aparecem na forma de manchas circulares, de coloração amarelo-pálida e pequenas (entre 1 e 3 mm de diâmetro), podendo alcançar 2 cm de diâmetro. Na parte superior das folhas, há a presença de manchas amareladas que futuramente necrosam (AGRIOS, 2005; ZAMBOLIM et al., 2005; AVELINO et al., 2015).

Aplicação de fungicidas é a forma mais comum de controle da ferrugem do cafeeiro. Uma das principais causas da expansão descontrolada da ferrugem é, provavelmente, a aplicação inapropriada de fungicida. Para controlar eficientemente a doença, é preciso realizar as aplicações de forma preventiva (AVELINO et al., 2015).

Usar indiscriminadamente agroquímicos na lavoura pode contaminar alimentos, solo e água, desenvolver doenças iatrogênicas (que ocorrem devido ao uso de agroquímicos) e reduzir a biodiversidade (BETTIOL e GHINI, 2001). Também, pode aumentar a resistência do patógeno aos fungicidas (BOSCH et al., 2011), acarretando em aumento nos níveis de intensidade e, consequentemente, no número de aplicações e na quantidade de agroquímicos utilizada no controle da doença.

Zambolim et al. (2005) recomendam, para controlar a ferrugem do cafeeiro, pulverizações de fungicidas com base na incidência da doença (porcentagem de folhas infectadas) na lavoura. Caso a incidência seja igual ou inferior a 5%, a recomendação é para iniciar a aplicação de fungicida protetor; se for entre 5 e 12%, é indicado a pulverização de fungicida sistêmico (ou curativo). Kushalappa et al. (1984) recomendaram o limiar de 10% de incidência da ferrugem na lavoura para aplicação de fungicida.

3.2 Epidemiologia da Ferrugem do Cafeeiro

A manifestação de uma epidemia da ferrugem do cafeeiro está vinculada, principalmente, às condições do ambiente. Temperatura, umidade relativa do ar, duração do

período de molhamento foliar (acúmulo de água na superfície da folha), precipitação e luminosidade estão entre os fatores que interferem no progresso da ferrugem. A altitude da lavoura em relação ao nível do mar pode influenciar na incidência da doença, uma vez que a temperatura diminui conforme a altitude aumenta (CHALFOUN e LIMA, 1986; VALE et al., 2000; ZAMBOLIM et al., 2002).

O espaçamento entre as plantas também pode interferir na proliferação da ferrugem, uma vez que o microclima dentro da lavoura costuma ser alterado conforme a proximidade entre as plantas (AVELINO et al., 2004). O sombreamento nas folhas, quando não é um fator limitante na produção de frutos, também pode afetar na intensidade da doença (LÓPEZ-BRAVO et al., 2012). A ferrugem progride mais na face sul das plantas quando comparada à face norte. Menor exposição ao Sol e maior período de molhamento foliar parecem justificar esse comportamento (CUSTÓDIO et al., 2010).

O ciclo da ferrugem (Figura 1) começa com os uredósporos, esporos que se depositam na face inferior das folhas e, na presença de água líquida, germinam, penetram e infectam, formando a urédia com uredósporos e causando lesões nas folhas. Em raras ocasiões, as condições climáticas permitem, nas lesões, as formações da télia e de um segundo tipo de esporos, os teliósporos, que ao germinarem formam o basídio e os basidiósporos, de função ainda desconhecida para o ciclo da ferrugem (ZAMBOLIM et al., 1997; FERNANDES et al., 2009). Os uredósporos são produzidos a partir das lesões foliares, permitindo que aconteça um novo ciclo de infecção. Assim, a ferrugem ocorre no cafeeiro nos estádios de urédia, télia e basídio (ZAMBOLIM et al., 1997).

A propagação do patógeno na lavoura ocorre com a dispersão dos uredósporos para folhas ainda não infectadas por meio dos respingos da chuva, da ação do vento, de insetos e pelo homem. O ser humano e os insetos ajudam na propagação do patógeno devido à interação destes com uma planta infectada, disseminando os esporos pela lavoura. A dispersão dentro de uma planta ocorre com o respingo da chuva (WALLER, 1982; ZAMBOLIM et al., 2002; AGRIOS et al., 2005).

Os uredósporos germinam apenas com a presença de água na superfície da folha, especialmente no período noturno. O processo de germinação é inibido se a água secar antes da penetração (KUSHALAPPA e ESKES, 1989). Caso a umidade relativa do ar seja alta, o período estimado de germinação dos uredósporos é de 6 a 8 horas (MARTINS, 1988). A temperatura ótima para germinação varia entre 21 e 23 °C (WALLER, 1982), tendo como limites extremos 15,5 °C e 28,5 °C (NUTMAN e ROBERTS, 1970). Temperaturas abaixo de

14 °C e acima de 32,5 °C são limitantes ao processo de infecção (KUSHALAPPA et al., 1983).

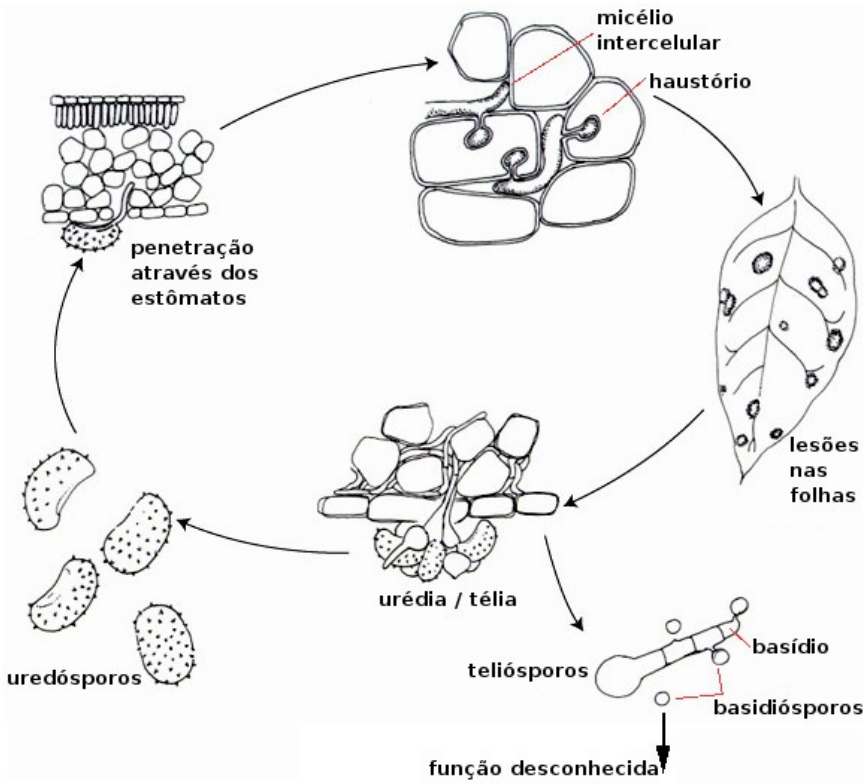

Figura 1. Ciclo da ferrugem do cafeeiro (ARNESON, 2000).

Período de incubação (PI) corresponde ao tempo, em dias, de duração a partir da germinação e penetração nos tecidos da planta até o surgimento dos sintomas. Na ferrugem, o PI costuma durar, em média, 25 a 30 dias, podendo ocorrer entre 18 e 45 dias. O PI tende a aumentar conforme a temperatura, caso esteja acima de 28 °C ou abaixo de 18 °C (ZAMBOLIM et al., 1997).

Moraes et al. (1976) formulou a equação (1) para estimar o PI. Porém, nesse trabalho o período latente, tempo necessário para a formação de 50% das pústulas, foi referenciado como o PI. Os autores observaram de 28 dias (em meses quentes) a 65 dias (em meses frios) como variações na duração do PI.

$$y = 103{,}01 - 0{,}98 * Tmax - 2{,}1 * Tmin \qquad (1)$$

y − estimativa do PI em dias, $Tmax$ − temperatura média máxima e $Tmin$ − temperatura média mínima durante o período.

A curva de progresso padrão da ferrugem tem início entre dezembro e janeiro, aumenta de forma logarítmica entre março a abril, alcança seu máximo ao redor de junho e, a

partir de então, passa a decrescer devido à baixa na temperatura e à desfolha das plantas em consequência da colheita e da própria doença (CHALFOUN et al., 2001; AVELINO e SAVARY, 2002; ZAMBOLIM et al., 2005).

Mudanças nas condições do ambiente são capazes de alterar a curva de progresso padrão da doença. A elevação da temperatura e da precipitação é capaz de atrasar o início do progresso da ferrugem e, simultaneamente, a elevação da temperatura média anual e a presença de chuvas esporádicas entre abril e julho (outono/inverno) permite que a doença mantenha níveis mais elevados ao fim de seu ciclo, em agosto. Dessa forma, o ciclo da ferrugem é prolongado durante o ano agrícola (CHALFOUN et al., 2001). Também foi observado que a ocorrência de um inverno atípico, com chuvas frequentes e altas temperaturas, pode favorecer a infecção e a dispersão do patógeno na lavoura e permitir que a doença se desenvolva tardiamente, atingindo seu máximo entre julho e setembro (TALAMINI et al., 2003).

O desenvolvimento mais tardio da ferrugem em relação à época de aplicação dos tratamentos prejudica a eficiência de diferentes esquemas de controle, permitindo uma elevação da doença ao final do seu ciclo (CHALFOUN e CARVALHO, 1999). Nessas situações é necessário rever e readaptar as aplicações de agroquímicos para controlar eficientemente a doença (CHALFOUN et al., 2001).

O papel da água no desenvolvimento da doença pode parecer contraditório, mas não é. Quando a precipitação é moderada, proporciona o molhamento foliar e a dispersão do patógeno, auxiliando no desenvolvimento da doença. Já a ocorrência de chuvas intensas entre dezembro e janeiro promove a lavagem dos uredósporos da superfície foliar, inibindo a epidemia da ferrugem e podendo atrasar seu início (CHALFOUN e ZAMBOLIM, 1985).

3.3 Modelagem e Alerta da Ferrugem do Cafeeiro

Modelos são simplificações da realidade e tentativas de resumir os principais processos visando testar hipóteses e verificar suas coerências. Na epidemiologia, os modelos buscam auxiliar na compreensão dos aspectos mais determinantes para o desenvolvimento de uma epidemia e assim elaborar estratégias para o controle da doença (VAN MAANEN e XU, 2003).

É possível classificar os modelos em dois grupos dependendo da abordagem usada em seu desenvolvimento: empíricos (ou descritivos) e explanatórios (ou mecanísticos).

A criação de modelos empíricos inicia com a coleta de dados, passa pelo estabelecimento da relação entre os dados e, idealmente, finaliza com uma previsão. Nesse caso não é exigido nenhuma relação de causa-efeito. No desenvolvimento de modelos explanatórios elabora-se um conceito derivado, geralmente, do funcionamento do sistema a ser modelado. Feito isso, ocorre a coleta dos dados e suas relações são pesquisadas. O objetivo do modelo explanatório é compreender melhor o sistema estudado, podendo levar a previsões e inferências (BERGAMIN FILHO e AMORIM, 2011).

Baseando-se em sua experiência, Coakley (1988) sugeriu que um conjunto de dados referentes a um período mínimo de oito a doze anos fosse utilizado para identificar, com maior precisão, os fatores mais influentes na evolução de uma doença. Se o conjunto de dados tiver menos de oito anos de registros, dados de diferentes regiões geográficas podem ser usados para compor o conjunto de dados.

Predições de doenças trazem vantagens econômicas, diminuindo o número de aplicações de agroquímicos e seu desperdício levando a uma redução de custos de produção, e de segurança, reduzindo a contaminação do ambiente e demais seres vivos (HARDWICK, 2006).

Na construção de modelos preditivos de doenças de plantas é comum a utilização de métodos e técnicas de cunho matemático. Montoya e Chaves (1974), Moraes et al. (1976) e Kushalappa et al. (1984) utilizaram técnicas de regressão para estudar a epidemiologia da ferrugem do cafeeiro. Alves et al. (2010) utilizaram regressão linear, regressão não-linear e sistemas de lógica *fuzzy* e *neuro-fuzzy* em seu estudo epidemiológico sobre o processo monocíclico da ferrugem do cafeeiro e da ferrugem asiática da soja.

Técnicas de regressão são as mais utilizadas na criação de modelos preditivos para a ferrugem. Entretanto, recentemente técnicas como rede bayesiana, árvore de decisão (seção 3.4.2.2), máquina de vetor suporte (*SVM*, do inglês *Support Vector Machine*) (seção 3.4.2.3) e rede neural vêm ganhando a atenção de pesquisadores.

Pinto et al. (2002) realizaram uma avaliação do potencial das redes neurais artificiais para descrever epidemias da ferrugem do cafeeiro, estabelecendo relações entre variáveis climáticas e de produção com a incidência da doença. Atributos climáticos foram calculados como somatórios ou médias dos últimos 15, 30, 45 e 60 dias antes da avaliação da incidência da ferrugem. Isoladamente, também foram criadas redes neurais a partir de séries temporais de incidência da doença.

Diversas redes foram desenvolvidas e a melhor apresentou 1,17% de erro médio de previsão e 3,43 para o quadrado médio dos desvios. Essa rede foi elaborada com as

variáveis climáticas e de produção, coletadas 30 dias antes da data de avaliação. A melhor rede criada a partir das séries temporais incluiu observações da incidência referentes às últimas quatro quinzenas anteriores à data de avaliação. Os autores concluíram que as redes foram eficientes para descrever a doença e as séries temporais poderiam facilitar a previsão da ferrugem do cafeeiro.

Sistema simples foi criado por Garçon et al. (2004) para predizer o desenvolvimento da ferrugem na lavoura, identificando o momento propício para início do controle da doença. Dados sobre molhamento foliar diário e temperatura média no período do molhamento foram usados para calcular o valor de severidade da doença, obtidos a partir de uma matriz de valores de severidade semelhante à concebida por Wallin (1962) na questão da requeima da batateira. Sistema foi avaliado por um ano agrícola e recomendou, para lavouras com alta carga pendente, um número de pulverizações de agroquímicos igual ao indicado em calendários fixos (duas aplicações de fungicida sistêmico e quatro de fungicida cúprico). Já em lavoura de média carga, o sistema recomendou uma aplicação de fungicida sistêmico a menos em relação ao calendário fixo. Assim, não houve diferença significativa na eficiência do controle da ferrugem realizado com auxílio do sistema e do calendário.

Meira et al. (2009) criaram modelos preditivos baseados em árvore de decisão para alerta da ferrugem do cafeeiro em plantas com alta carga pendente de frutos. Os modelos foram gerados a partir de dados meteorológicos e de avaliações da incidência da doença. O conjunto de dados correspondia ao período de outubro de 1998 a outubro de 2006.

No estudo, os modelos prediziam se a taxa de progresso da ferrugem, diferença entre a incidência da doença em dois meses subsequentes, iria ser maior ou igual a 5 ou 10 pontos porcentuais (p.p.) no próximo mês.

A avaliação do poder preditivo dos modelos foi feita por meio de medidas como a acurácia, sensitividade, especifidade, entre outras. As acurácias dos modelos de 5 e 10 p.p. como limiar para a taxa de progresso foi de, respectivamente, 81% e 79%. O modelo de 5 p.p. ainda obteve 80% de sensitividade e 83% de especificidade. Para o modelo de 10 p.p., a sensitividade foi de 70% e a especificidade ficou em 83%. Os autores concluíram que as árvores de decisão para alerta da ferrugem do cafeeiro em lavouras com alta carga podem auxiliar na indicação do momento oportuno para controle da doença. Já um modelo para baixa carga pendente gerado para lavouras com 5 p.p. como limiar da taxa de progresso apresentou 72% de acurácia e não conseguiu resultados satisfatórios nas demais medidas avaliadas (MEIRA e RODRIGUES, 2009).

Cintra et al. (2011) uniram árvores de decisão e sistemas *fuzzy* na criação de modelos preditivos para a taxa de progresso mensal da ferrugem do cafeeiro. As árvores de decisão *fuzzy* apresentaram taxa de erro, entre 1,15% e 6,47%, menor que as árvores tradicionais dependendo do conjunto de atributos e o limiar do atributo meta usados.

Luaces et al. (2011) realizaram uma abordagem não-determinística no desenvolvimento de um sistema de alerta para a incidência da ferrugem. Desse modo, foram criados modelos baseados em *SVM*, que emitem três tipos de saída: "alarme", se prever aumento na incidência da doença acima de um valor pré-determinado (*threshold*); "não-alarme", se não prever aumento na incidência da doença acima do *threshold*; "atenção", se o modelo não for capaz de identificar um alarme ou não-alarme. Os autores apresentaram uma coleção de classificadores capazes de predizer um intervalo de possibilidades ao invés de um valor determinado.

A implementação do sistema de alerta foi realizada de forma que, na predição, o custo de falsos negativos fosse maior que o custo dos falsos positivos. Luaces et al. (2011) concluem que a abordagem não-determinística pode ser generalizada e usada em outras tarefas preditivas. Também apontam para a praticidade de implementação dos modelos preditores, necessitando apenas de uma estação meteorológica capaz de registrar as mesmas variáveis do trabalho.

Pérez-Ariza et al. (2012) criaram modelos preditivos, baseados em redes bayesianas, para a ferrugem do cafeeiro. Os autores utilizaram 4,5% como limiar para confirmação da infecção da doença, valor tido como referência para os agricultores no controle da ferrugem. Foi realizado uma discretização nos valores de incidência para se determinar três diferentes níveis de infecção: negativo, alerta e positivo. Dois modelos simples foram gerados, sem aprendizado com o conjunto de dados meteorológicos, o primeiro apresentou 37,33% de taxa de erro e o segundo errou em 34,66% dos casos. Modelos utilizando dados meteorológicos em seu aprendizado chegaram a ter 8,97% de taxa de erro. Ainda, foram gerados novos modelos em árvore de decisão para comparar com as redes bayesianas. As árvores de decisão demonstraram melhor desempenho preditivo.

Girolamo Neto et al. (2014), continuando a pesquisa de Meira et al. (2009), desenvolveram modelos preditivos em *SVM*, redes neurais artificiais, árvores de decisão e florestas aleatórias. O conjunto de dados possuía registros para lavouras com alta e baixa carga pendente. Um balanceamento de classes foi realizado no conjunto de dados relativos à baixa carga pendente. Técnicas de seleção de atributos como *CFS*, InfoGain, GainRatio, Qui-

Quadrado e *Wrapper* foram usadas. Modelos com 5 p.p. como limiar da taxa de progresso mensal da ferrugem foram criados para cada tipo de carga pendente de frutos.

O desenvolvimento e a avaliação dos modelos, como feito por Meira et al. (2009), foram realizados por meio de dados registrados ao longo do ano agrícola, setembro a agosto. Além de medidas de avaliação, os autores utilizaram gráficos do tipo *ROC* (do inglês, *Receiver Operating Characteristic*) para avaliar e selecionar os modelos. O modelo de melhor desempenho preditivo desenvolvido para alta carga obteve 85,3% de acurácia, 85,4% de sensitividade e 85,2% de especificidade e foi baseado em *SVM*. Esse desempenho superou os modelos gerados por Meira et al. (2009) e Cintra et al. (2011). O melhor modelo de baixa carga foi criado com floresta aleatória e obteve 88,9% de acurácia, 86,3% de sensitividade e 89,9% de especificidade. Em modelos com 10 p.p. de taxa de progresso e alta carga, obteve-se 89,9% de acurácia, 90% de sensitividade e 89,9% de especificidade (GIROLAMO NETO, 2013).

De modo geral, os melhores modelos foram criados por *SVM*s e florestas aleatórias. Os autores concluem que os modelos desenvolvidos poderiam fornecer subsídios para o controle da doença em lavouras com alta carga e prover a possibilidade de monitoramento em lavouras com baixa carga.

Thamada et al. (2013) implementaram um sistema *Web* para emissão de alerta para ferrugem do cafeeiro, predizendo, a partir de dados meteorológicos, se a taxa de progresso da doença no próximo mês deverá ser maior ou igual a um limiar. Modelos preditivos de 5 p.p. (GIROLAMO NETO et al., 2014) e 10 p.p. (GIROLAMO NETO, 2013) como limiar para a taxa de progresso foram incorporados ao sistema Web. Assim, ocorreu a inserção de três modelos para cada combinação possível entre carga pendente (alta ou baixa) e espaçamento entre as plantas na lavoura (adensado ou largo).

A emissão do alerta é o resultado de uma votação simples, onde a predição de cada modelo possui o mesmo peso. O alerta emitido é positivo caso a maioria dos modelos indique um aumento na taxa de progresso da ferrugem maior ou igual o limiar, ou negativo caso contrário.

Os modelos desenvolvidos por Meira et al. (2009) e Meira e Rodrigues (2009) foram validados para dados inéditos, que não foram utilizados em sua modelagem. O critério usado para validar um modelo é que a taxa de acerto deveria ser igual ou maior à obtida em sua construção. Nesse quesito, nove de doze modelos foram rejeitados e os demais não tiveram seu uso recomendado devido aos baixos valores de sensitividade (GIROLAMO NETO, 2013). Além disso, os modelos de Girolamo Neto (2013) para baixa carga pendente

não apresentaram desempenhos preditivos satisfatórios quando validados com dados inéditos de três anos agrícolas, setembro a agosto (MEIRA et al., 2014).

3.4 Descoberta de Conhecimento em Bases de Dados (*KDD*)

3.4.1 Visão Geral

A convergência entre a computação e a comunicação tem produzido informações em larga escala, acumuladas em grandes bancos de dados. São necessárias ferramentas e versáteis para transformar essa enorme quantidade de dados em informações úteis e relevantes (WITTEN et al., 2011). Dessa necessidade surgiu o processo de descoberta de conhecimento em bases de dados (*KDD*, do inglês *Knowledge Discovery in Databases*).

O processo *KDD* (Figura 2) é composto por fases iterativas, sendo possível voltar para fases anteriores e refazer o processo (FAYYAD et al., 1996):

Fase de Seleção: Conjuntos ou subconjuntos dos dados disponíveis são selecionados, baseados em critérios pré-definidos.

Fase de Pré-Processamento: Limpeza e pré-processamento de dados, por meio do tratamento de dados incorretos ou ausentes, exclusão de informações desnecessárias e outros processamentos.

Fase de Transformação: Configuração dos dados para que atendam às exigências de determinada técnica de mineração de dados, conversão de dados e criação de novos atributos.

Fase de Mineração de Dados: Extração dos padrões de interesse contidos nos dados, utilizando uma técnica de mineração de dados pré-determinada.

Fase de Interpretação: Entendimento dos padrões encontrados e geração de conhecimento, auxiliando na tomada de decisão humana.

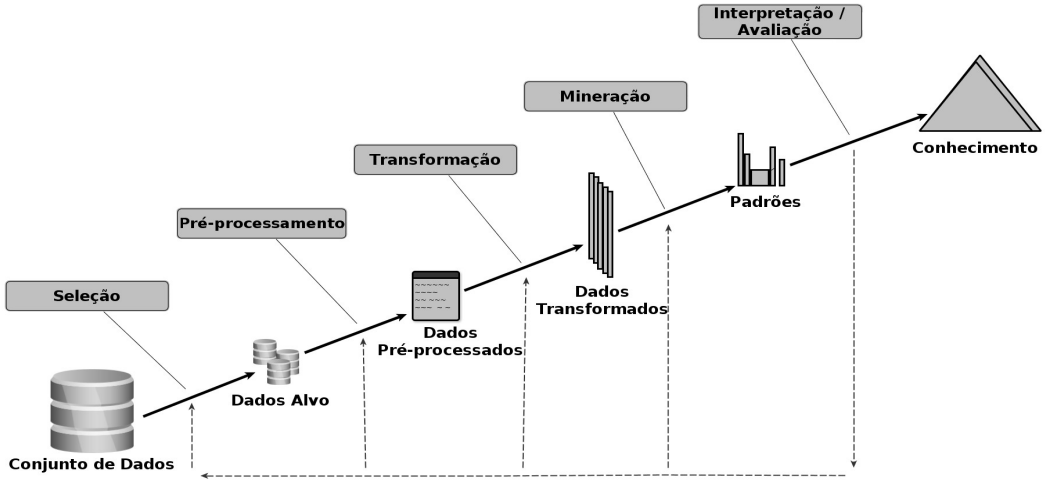

Figura 2. Fases do *KDD* (FAYYAD et al., 1996).

3.4.2 Tarefas e Técnicas de Mineração de Dados

Uma tarefa em mineração de dados consiste no ato de descobrir um tipo de padrão em um conjunto de dados utilizando determinadas técnicas. Existem dois tipos principais de tarefas (Figura 3), as preditivas e as descritivas (FAYYAD et al., 1996).

As tarefas preditivas procuram, a partir de um conjunto de variáveis com valores conhecidos (atributos ou variáveis independentes), predizer o valor de uma variável de interesse (atributo meta ou variável dependente). Os dados (exemplos) costumam ser divididos por classes no atributo meta. Por exemplo, se o atributo meta é o nome de uma doença que ocorre na cultura do café, pode-se definir como classes: ferrugem, cercosporiose, rhizoctoniose, antracnose e phoma.

A principal diferença entre as tarefas preditivas está no tipo do atributo meta:

Classificação: Consiste na classificação (predição) de um atributo meta categórico (discreto e não ordenado). Os modelos (classificadores), criados a partir de um conjunto de exemplos, aprendem sobre o comportamento dos dados. Assim, o modelo é capaz de predizer a classe do atributo meta em novos e desconhecidos exemplos (APTÉ e WEISS, 1997; WITTEN et al., 2011).

Regressão: Similar à classificação, consiste na predição de um atributo meta numérico contínuo. Há diversas formas de regressão, como linear, múltipla, com peso, polinomial, não parametrizada e robusta (APTÉ e WEISS, 1997; WITTEN et al., 2011).

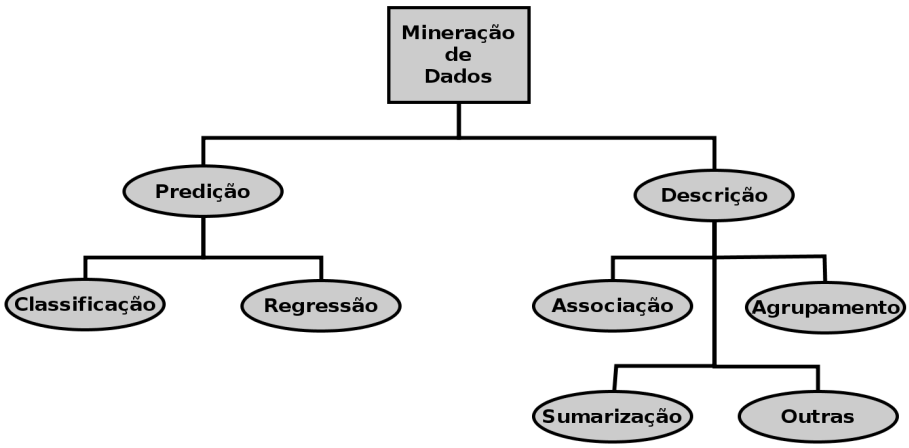

Figura 3. Principais tarefas de mineração de dados (REZENDE et al., 2002).

Tarefas descritivas caracterizam propriedades encontradas no conjunto de dados a fim de encontrar informações interpretáveis pelo ser humano. Essas tarefas não costumam ter um atributo meta definido. Dentre as tarefas descritivas, destacam-se: associação, agrupamento (*clustering*) e sumarização (FAYYAD et al., 1996).

Técnicas de mineração de dados são utilizadas para atingir o objetivo das tarefas. As técnicas são, geralmente, baseadas em algoritmos computacionais, métodos estatísticos, aprendizado de máquina e reconhecimento de padrões (FAYYAD et al., 1996).

A escolha de qual tarefa a ser utilizada depende do objetivo e do conjunto de dados do problema, assim como as técnicas, que podem ser mais de uma (FAYYAD et al., 1996). Não existe a melhor técnica, todas possuem vantagens e desvantagens.

3.4.2.1 Ensembles

Ensemble é uma técnica de mineração de dados que corresponde a diversos modelos treinados individualmente que são combinados a fim de predizer novos registros. A ideia é construir um modelo que possua desempenho preditivo superior a qualquer modelo simples que componha o *ensemble* (OPITZ e MACLIN, 1999). Em uma tarefa de classificação, para inferir a classe de um novo registro cada modelo simples classifica o registro e todas as predições são combinadas por meio de um esquema de votação, simples ou com peso. A partir do resultado da votação sabe-se qual a classe predita pelo *ensemble* (HAN et al., 2011).

Estudos teóricos e empíricos mostram que um bom *ensemble* é composto por modelos eficientes e que erram diferentes registros do conjunto de dados (OPITZ e MACLIN,

1999). Espera-se gerar modelos simples que são, ao mesmo tempo, diferentes entre si e que possuam o melhor desempenho preditivo possível quando analisados individualmente (SENI e ELDER, 2010). Dentre as técnicas de *ensemble*, destacam-se: *bagging*, *boosting* e *stacking*.

3.4.2.1.1 *Bagging*

Bagging é um método que gera múltiplos modelos simples baseados em uma técnica de mineração de dados e os utiliza para compor um modelo final. A predição inferida pelo modelo final é realizada por meio de uma votação, simples ou com peso, entre os resultados dos modelos simples. As múltiplas versões dos modelos são desenvolvidas por meio da aplicação do *bootstrap* (BREIMAN, 1996; QUINLAN, 1996).

O *bootstrap* é um tipo de amostragem aleatória com reposição. Considerando todos os registros do conjunto de dados original, o método escolhe aleatoriamente um registro para compor a amostra de dados que será usada para desenvolvimento de modelos. Posteriormente, esse registro é reintegrado ao conjunto de dados e pode ser selecionado novamente. Assim, a cada seleção todos os registros têm a mesma probabilidade de serem escolhidos. Alguns registros podem ser escolhidos mais de uma vez ou nunca serem selecionados. Cada amostragem bootstrap possui a mesma quantidade de registros do conjunto de dados original (BREIMAN, 1996; HASTIE et al., 2009). Cerca de 2/3 dos dados são selecionados para amostragem *bootstrap* a cada iteração, o restante dos dados é conhecido como observação *out-of-bag* (*OOB*) e pode ser usado para avaliar o *bagging* desenvolvido (JAMES et al., 2013).

A Figura 4 mostra um exemplo da amostragem feita por *bootstrap.* O número *5* do conjunto de dados original não foi selecionado em nenhum conjunto de treinamento, enquanto outros números foram selecionados mais de uma vez. Por exemplo, *3* e *7* foram escolhidos duas vezes no Conjunto de Treinamento-1. Observa-se que um modelo treinado utilizando todos os dados provavelmente errará menos que o modelo gerado com o Conjunto de Treinamento-1, uma vez que seu treinamento abrange maior número de registros distintos. Na realidade, todos os modelos criados com os quatro conjuntos obterão taxas de erros significativas. Porém, quando combinados, esses modelos produzem uma taxa de erro menor que qualquer modelo simples (OPITZ e MACLIN, 1999).

Conjunto de Dados (Original)
1, 2, 3, 4, 5, 6, 7, 8

	Amostragem *Bootstrap*
Conjunto de Treinamento-1	2, 7, 8, 3, 7, 6, 3, 1
Conjunto de Treinamento-2	7, 8, 3, 6, 4, 2, 7, 1
Conjunto de Treinamento-3	3, 6, 2, 7, 4, 6, 2, 2
Conjunto de Treinamento-4	4, 1, 1, 4, 6, 4, 3, 8

Figura 4. Exemplo da amostragem realizada por *bootstrap* (adaptado de OPITZ e MACLIN, 1999).

Bagging pode obter ganhos no desempenho preditivo quando aplicado em conjunto com técnicas de aprendizado instável, onde pequenas mudanças no conjunto de treinamento promove grandes alterações no modelo gerado, como árvores de decisão, por exemplo (BREIMAN, 1996). Quanto ao desempenho, *bagging* é mais robusto que o *boosting* para lidar com conjuntos de dados com ruídos de classificação, ou seja, atributo meta com valor incorreto (DIETTERICH, 2000; KOTSIANTIS, 2011) e o uso de amostragens *bootstrap* diminui a variância – quanto as predições de um modelo variam entre si – do *ensemble* (BAUER e KOHAVI, 1999; JAMES et al., 2013).

3.4.2.1.2 Floresta Aleatória

A combinação de *bagging* com modelos simples baseados em árvore de decisão (seção 3.4.2.2) permitiu o surgimento da floresta aleatória. A técnica reúne diversas árvores para realizar uma tarefa de classificação. As florestas aleatórias e o *bagging* com árvore de decisão são técnicas diferentes. Na floresta, a escolha do atributo a ser usado na ramificação das árvores é aleatória e não são considerados todos os atributos do conjunto de dados, apenas um subconjunto deles (parâmetro *m*). Essa característica permite o desenvolvimento de árvores heterogêneas e, consequentemente, de uma floresta com maior poder de generalização (BREIMAN, 2001; JAMES et al., 2013), além de uma tendência de descarte dos atributos irrelevantes uma vez que a base da floresta aleatória são modelos em árvore de decisão.

Diversos critérios podem ser usados para selecionar o atributo para o nó da árvore como: grau de impureza, ganho de informação e índice Gini (ROKACH e MAIMON, 2005). Entretanto, não há diferença significativa na variação da acurácia obtida por florestas aleatórias geradas utilizando diferentes critérios de seleção de atributos (KULKARNI et al., 2014).

Cada árvore de decisão que forma a floresta é construída a partir de uma amostragem *bootstrap*. A classificação de um novo exemplo na floresta aleatória ocorre por meio de uma votação simples (QUINLAN, 1996; BREIMAN, 2001).

Na Figura 5 é possível verificar a taxa de erro obtida por modelos em *bagging* com árvore de decisão e floresta aleatória (primeira e segunda linha de cima para baixo, respectivamente) conforme o número de árvores aumenta, bem como os erros obtidos em classificações de observações *OOB* (terceira e quarta linha de cima para baixo, respectivamente). Verifica-se que a partir de uma determinada quantidade de árvores, os erros tendem a estabilizar.

Florestas aleatórias evitam o sobreajuste (*overfitting*), que impede o modelo de classificar adequadamente novos exemplos uma vez que possui um aprendizado excessivo sobre um determinado conjunto de dados (BREIMAN, 2001; VERIKAS et al., 2011). Também são robustas a ruídos (*outliers*) e eficientes para se trabalhar com bases de dados grandes e números de atributos altos, uma vez que usam um subconjunto dos atributos na criação das árvores (HAN et al., 2011). Seu desempenho preditivo, às vezes, pode ser superior ao *boosting*, fornece informações sobre a correlação e importância dos atributos e seu tempo de processamento é menor que *bagging* e *boosting* (BREIMAN, 2001).

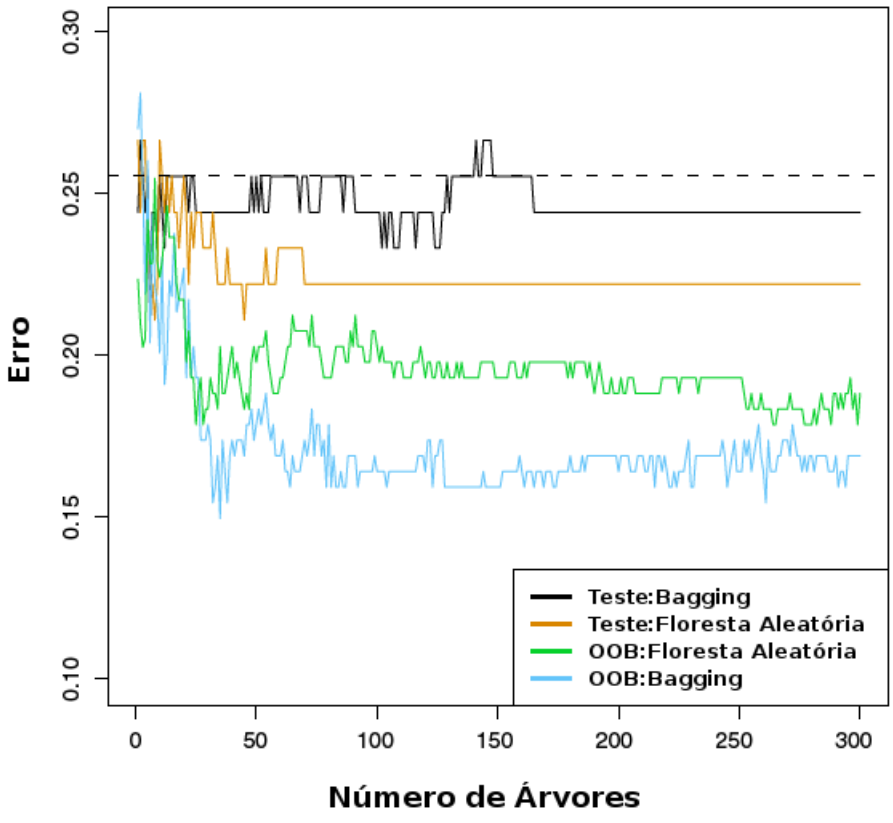

Figura 5. Resultados de uma simulação usando *bagging* com árvores de decisão e floresta aleatória (JAMES et al., 2013).

3.4.2.1.3 *Boosting*

O *boosting* corresponde a outro método *ensemble*. A motivação para desenvolver o *boosting* é produzir um modelo que combine a saída de classificadores fracos, cujas taxas de erro superam levemente à obtida em classificações aleatórias (FREUND e SCHAPIRE, 1996; HASTIE et al., 2009). Inicialmente, um modelo fraco é criado a partir dos dados originais. Então, um novo conjunto de dados é formado dando um peso maior para os registros em que o modelo anterior errou para desenvolver um novo modelo (HAN et al., 2011). Os passos são repetidos conforme a necessidade da modelagem. Ao fim, são atribuídos pesos para as saídas de cada modelo desenvolvido para inferir, por meio de uma votação, a predição final (QUINLAN, 1996). Esses dois aspectos, método de montagem de novos conjuntos de dados e esquema de votação, diferenciam o *boosting* do *bagging* (KOTSIANTIS, 2011).

O *boosting* procura criar classificadores mais eficientes na predição de registros em que os modelos atuais do *ensemble* têm fraco desempenho. Modelos são criados em série, tendo seus erros diminuídos. Entretanto, o método pode apresentar acurácia inferior em relação ao modelos simples, especialmente quando combinado com rede neural artificial e pode ser suscetível ao sobreajuste (MACLIN e OPITZ, 1997; OPITZ e MACLIN, 1999; GALAR et al., 2013).

A Figura 6 mostra um exemplo de montagem dos conjuntos de treinamento. Suponha que o registro *1* seja de difícil aprendizado. Verifica-se, com as iterações, uma maior frequência de *1* nos conjuntos de treinamento. Isso ocorre devido à dificuldade em seu aprendizado e aos erros apresentados em modelos anteriores em sua classificação (OPITZ e MACLIN, 1999; JAMES et al., 2013).

Conjunto de Dados (Original)
1, 2, 3, 4, 5, 6, 7, 8

	Amostragem *Boosting*
Conjunto de Treinamento-1	2, 7, 8, 3, 7, 6, 3, 1
Conjunto de Treinamento-2	1, 4, 5, 4, 1, 5, 6, 4
Conjunto de Treinamento-3	7, 1, 5, 8, 1, 8, 1, 4
Conjunto de Treinamento-4	1, 1, 6, 1, 1, 3, 1, 5

Figura 6. Exemplo da amostragem realizada por *boosting* (adaptado de OPITZ e MACLIN, 1999).

Todos os modelos são gerados por uma mesma técnica de mineração de dados como, por exemplo, árvore de decisão e regras de classificação (ROKACH, 2010). Assim, na teoria, o *boosting* pode ser usado para reduzir significativamente o erro de qualquer algoritmo

de aprendizado que constantemente gera classificadores fracos (FREUND e SCHAPIRE, 1996; BRIEM et al., 2001). Em relação aos modelos simples, o método diminui a variância e o viés, medida de quão perto a média das predições de um determinado algoritmo de aprendizado está correta (FREUND e SCHAPIRE, 1996; BAUER e KOHAVI, 1999).

Comparando com o *bagging*, o *boosting* é mais robusto para trabalhar com dados sem ruído (KOTSIANTIS, 2011) e, geralmente, apresenta desempenho preditivo superior (MACLIN e OPITZ, 1997; BANFIELD et al., 2007). O *ensemble* criado a partir do *boosting* e árvore de decisão se mostra efetivo quanto ao desempenho (FREUND e SCHAPIRE, 1996; BAUER e KOHAVI, 1999).

3.4.2.1.4 *Stacking*

Stacking ou *Stacked Generalization,* trata-se de uma composição de modelos gerados a partir de diferentes técnicas de mineração de dados. A ideia é reunir as vantagens de diferentes técnicas, minimizar a taxa de erro dos modelos e criar um meta-classificador, que combina as saídas (predições) de diversos modelos e atua como substituto aos sistemas de votação (WOLPERT, 1992; TING e WITTEN, 1997a; WITTEN et al., 2011).

Tipicamente, diferentes algoritmos produzem modelos com aprendizado distinto para uma mesma tarefa de mineração de dados. A forma mais comum de *stacking* é coletar as saídas de cada modelo que compõe o *ensemble* para formar um novo conjunto de dados (Figura 7). Segundo a terminologia de Wolpert (1992), os modelos construídos a partir do conjunto de dados original são chamados de modelos nível-0 e o meta-classificador é o generalizador (ou modelo) nível-1. O *stacking* reduz o viés em relação aos modelos nível-0 (GEURTS et al., 2006).

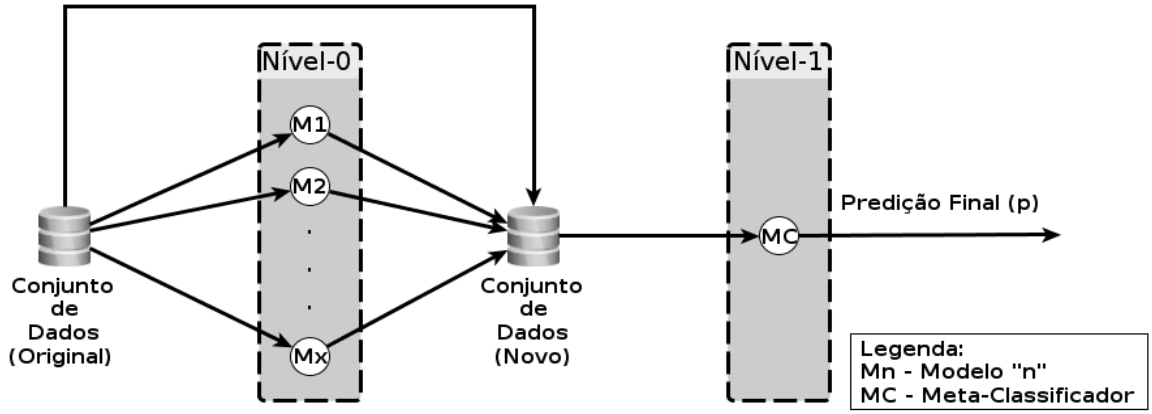

Figura 7. Esquema básico de aplicação do *stacking*.

Para cada registro do conjunto original, o novo conjunto representa a predição de cada modelo para aquela classe do registro, junto com sua classificação verdadeira. Durante esse passo, deve-se assegurar que os modelos que servirão de base para o meta-classificador sejam gerados a partir de um conjunto de dados que não possui o registro mencionado, apenas um subconjunto dos dados é usado na criação dos modelos base (TING e WITTEN, 1999).

Esses subconjuntos podem ser escolhidos a partir da validação cruzada (do inglês, *cross-validation*). A técnica divide os dados em *k* partes mutuamente exclusivas e, geralmente, de mesmo tamanho. O modelo é gerado com base em *k-1* partes e testado na parte restante. Esse procedimento se repete *k* vezes, ou seja, até que todas as partes tenham sido utilizadas para treinar e testar o modelo (WITTEN et al., 2011; MARTON et al., 2013). Uma variação é a validação cruzada estratificada, nela é verificado a proporção que cada classe possui no conjunto de dados e essa proporção, na medida do possível, é mantida em cada *k* parte. O uso da validação cruzada estratificada com k=10 é a mais recomendada para seleção do modelo (KOHAVI, 1995).

Assim, o conjunto de dados novos, que servirá como conjunto de treinamento para desenvolver o meta-classificador, é formado por predições dos modelos nível-0 para a parte de teste. A predição *p* inferida pelo meta-classificador é tida como a saída final do *stacking*. (WOLPERT, 1992; TING e WITTEN, 1999).

A melhor forma para desenvolver o *stacking*, em tarefas de classificação, é usar como saída dos modelos nível-0 a probabilidade do registro ser classificado em determinada classe ao invés de predizer a classe do registro (TING e WITTEN, 1999; SIGLETOS et al., 2005).

Apesar de ser possível aplicar qualquer técnica na construção do meta-classificador, Ting e Witten (1997b) recomendam o uso de uma regressão linear simples (método dos mínimos quadrados), já que o processamento (aprendizado) mais pesado foi feito no nível-0 (WITTEN et al., 2011). Outros modelos lineares também podem ser usados como: *SVM* com *kernel* linear (3.3.2.3), regressão logística, regressão ridge, *perceptron* e passivo-agressivo.

A regressão logística é uma técnica que relaciona o atributo meta categórico, frequentemente binário, com variáveis independentes. Dessa forma, todas variáveis do conjunto de dados estão associadas (possuem relevância) à predição do atributo meta. A técnica também pode ser usada para calcular a probabilidade do valor do atributo meta ser 0 ou 1 (AGRESTI, 2013).

Regressão ridge é uma transformação linear do método de mínimos quadrados, que obtém menor variância, para conjuntos de dados não ortogonais (HOERL e KENNARD, 1970). O método diminui os valores dos coeficientes impondo penalidades em seu tamanho, visando minizá-los por meio do método de mínimos quadrados (HASTIE et al., 2009).

O *perceptron* busca encontrar um hiperplano que classifique eficientemente os registros, para isso o algoritmo percorre o conjunto de dados diversas vezes e, para cada iteração, há um reajuste nos parâmetros (pesos) de cada atributo do hiperplano (FREUND e SCHAPIRE, 1999; WITTEN et al., 2011). Hiperplanos são sub-espaços de dimensão *n-1* em um espaço *n* dimensional. Por exemplo, se um determinado espaço possui duas dimensões, um hiperplano é um sub-espaço plano de dimensão um, ou seja, uma linha (JAMES et al., 2013).

O método passivo-agressivo realiza, em uma classificação binária, predições usando uma função de classificação de forma sequencial. A cada iteração o modelo prediz +1 ou -1. A partir da resposta real o algoritmo pode sofrer uma perda instantânea que reflete no grau de erro da predição. Ao fim de cada iteração o registro predito pelo modelo junto com sua resposta real são usados para aperfeiçoar a função de classificação nas próximas iterações (CRAMMER et al., 2006).

Stacking pode apresentar desempenho superior ao *bagging* e *boosting* (TING e WITTEN, 1999), entretanto não é muito utilizado porque é difícil de analisá-lo teoricamente e não há uma versão aceita sobre qual a melhor maneira de implementação (WITTEN et al., 2011).

3.4.2.2 Árvores de Decisão

Árvore de decisão é uma técnica de mineração de dados usada na descoberta de regras de classificação por meio da divisão dos dados contidos no conjunto de treinamento. As árvores proporcionam um nível de interpretabilidade único em modelos simbólicos (APTÉ e WEISS, 1997).

A técnica de árvores de decisão divide o conjunto de dados, criando um modelo gráfico representado por nós, ramos e folhas, similar a uma árvore invertida. Os algoritmos de árvore de decisão buscam avaliar e selecionar, dentre todos os atributos disponíveis, o que melhor separa os exemplos em subconjuntos distintos (BREIMAN et al., 1984; QUINLAN,

1986) e, consequentemente, o algoritmo faz uma seleção de atributos durante esse processo (HASTIE et al., 2009).

A construção de uma árvore de decisão começa pelo nó raiz e se expande até as folhas. Os nós internos, inclusive o nó raiz, são testes sobre um ou mais atributos. Normalmente, os testes comparam o valor do atributo a uma constante e dividem a árvore em dois ou mais ramos. Ao fim de cada ramo há um nó folha (ou nó terminal), que representa uma classe do atributo meta (BREIMAN et al., 1984; QUINLAN, 1986).

Ao seguir um caminho do nó raiz, passando pelos ramos, até uma folha encontra-se uma regra de decisão (HAN et al., 2011). Não é recomendável que uma árvore cresça indefinidamente, uma vez que isso aumenta a chance de ocorrer um sobreajuste. Para evitar ou diminuir essa desvantagem, é possível podar a árvore, diminuindo seus ramos, folhas e regras (BREIMAN et al., 1984; APTÉ e WEISS, 1997; JAMES et al., 2013).

A Figura 8 representa a árvore de classificação para identificar a presença ou ausência de orvalho em dados climáticos horários e indicar uma fórmula a ser usada para estimar a duração do período de orvalho. Um exemplo de regra para essa árvore seria: se a diferença nos valores do ponto de orvalho forem menores que 3,7 ºC e a velocidade do vento for menor que 2,5 m/s, então use a fórmula 1 para estimar o período de duração de orvalho.

Após a modelagem, a classificação de novos exemplos acontece percorrendo-se a árvore de decisão, a partir do nó raiz, utilizando os valores dos atributos do novo exemplo até chegar a uma folha (QUINLAN, 1987; HAN et al., 2011).

As árvores geram regras facilmente interpretáveis devido suas representações visuais, provêm conhecimento sobre os padrões contidos no conjunto de exemplos e permitem entender sobre suas fronteiras de decisão (SAFAVIAN e LANDGREBE, 1991; APTÉ e WEISS, 1997).

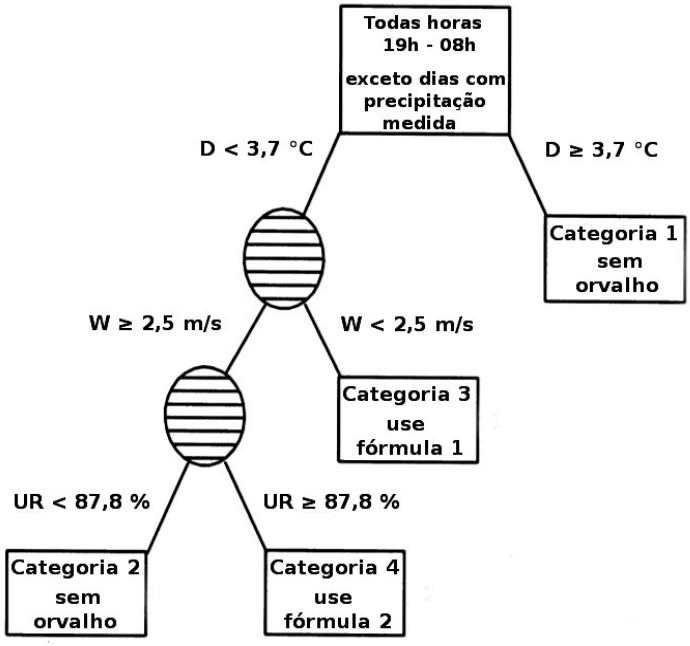

Figura 8. Exemplo de uma árvore de decisão. Em que D – diferença nos valores do ponto de orvalho; W – velocidade do vento e UR – umidade relativa do ar (GLEASON et al., 1994).

3.4.2.3 Máquinas de Vetores Suporte

As máquinas de vetores suporte, ou *SVMs*, realizam transformações espaciais no conjunto de dados, aumentando sua dimensão por meio de um produto interno (*kernel*), de forma que facilite a classificação dos dados. Os principais *kernels* envolvem funções lineares, polinomiais, de base radial (*RBF*, do inglês *Radial Basis Function*) e sigmóides (BURGES, 1998; HASTIE et al., 2009).

A função linear trata-se de um *kernel* linear em que os dados são transformados seguindo a função de grau = 1. A função polinomial, RBF e sigmóide são *kernels* utilizados para transformações não lineares nos dados. Na função polinomial o grau da função usada para aumentar a dimensionalidade dos dados é maior que 1 (JAMES et al., 2013).

No *RBF* são determinados pontos centrais em que a partir destes serão traçadas áreas de influência que os vetores suportes possuirão sobre os dados. Isso significa que o RBF possui um comportamento local, em que apenas registros dentro da área de influência são classificados (HASTIE et al., 2009; JAMES et al., 2013). A função sigmóide é contínua, crescente e limita os valores de entrada a um intervalo específico (WITTEN et al., 2011).

A dimensionalidade dos registros é aumentada por meio da multiplicação destes com um *kernel*. A seguir, *SVM* procura identificar vetores suporte, retas paralelas e que

separam as classes do conjunto de dados eficientemente. *SVM*s selecionam vetores suporte cuja distância (margem) entre si é máxima (JAMES et al., 2013).

O próximo passo é identificar o hiperplano ideal, uma vez que há infinitos hiperplanos na margem. Deve-se escolher o hiperplano que estiver equidistante dos vetores suporte (JAMES et al., 2013). O hiperplano identificado pela *SVM* é paralelo aos vetores suporte e capaz de separar as classes existentes no conjunto de dados (HAN et al., 2011).

Uma vez que o valor da distância entre os vetores suporte está maximizado, o hiperplano selecionado também estará a uma distância máxima dos vetores, isso permite um maior poder de generalização ao modelo com relação a novos exemplos a serem classificados (WU et al., 2008).

Na Figura 9 há duas classes, bolas brancas e quadrados pretos. Vetores suporte (linhas pontilhadas) passam pelos pontos *p1*, *p2* e *p3*, separados a uma distância *d* (margem). O hiperplano paralelo e equidistante (*d/2*) aos vetores suporte é definido como separador das classes.

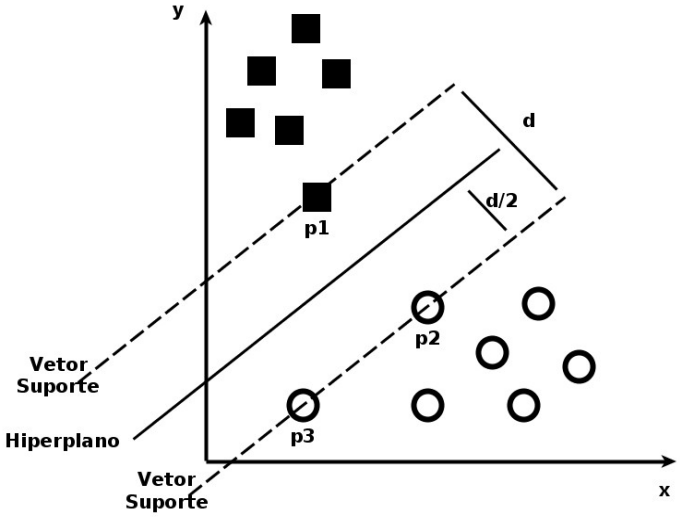

Figura 9. Vetores suporte e detecção de hiperplano.

Não existe uma regra geral que indique qual tipo de *kernel* produz modelos de melhor desempenho preditivo. Na prática, as diferenças entre as acurácias dos modelos gerados por diferentes *kernels* não são significativas (HAN et al., 2011).

Modelos em *SVM* são menos propensos ao sobreajuste, mais robustos e costumam apresentar acurácia maior do que outras técnicas. Os vetores suporte encontrados fornecem uma descrição compacta sobre o modelo desenvolvido e estão relacionados com a complexidade do modelo desenvolvido (WU et al., 2008; HAN et al., 2011).

3.4.3 Métodos de Discretização de Dados

A discretização de dados numéricos substitui seus valores por rótulos de intervalos (0-10, 11-20 e 21-30, por exemplo) ou rótulos conceituais (criança, adulto, idoso, por exemplo) (HAN et al., 2011). Dessa forma, a discretização diminui o número de valores do atributo e auxilia na identificação de limiares contidos dentro do atributo.

Dentre os métodos de discretização de dados destacam-se: número pré-determinado de intervalos iguais (*equal-width*), número uniforme de amostras por intervalo (*equal-frequency*) e discretização por agrupamento (*clustering*). *Equal-width* divide os registros em *k* intervalos de tamanhos iguais. No *equal-frequency* ocorre a divisão dos registros em *k* intervalos contendo, aproximadamente, o mesmo número de ocorrências (DOUGHERTY et al., 1995). O propósito do agrupamento é procurar por valores similares e juntá-los em *clusters*. Os registros que formam um *cluster* devem ser os mais parecidos possíveis. Não há um número ideal de *clusters* a serem formados. No início, escolhe-se aleatoriamente *k* registros para formarem *k clusters*. A seguir, agrupa-se os demais registros de acordo com sua proximidade com os registros iniciais (BLAJDO et al., 2008).

Exemplificando os métodos de discretização, suponha que se deseja discretizar os seguintes valores: 17, 20, 23, 31, 32, 36, 42, 43, 44 em três intervalos utilizando o *equal-width*, *equal-frequency* e agrupamento. Para o método *equal-width* é calculado a diferença dos valores dos registros máximo e mínimo, 44 - 17 = 27. Divide-se a diferença pelo número de intervalos, 27 / 3 = 9, encontrando o tamanho que cada intervalo terá.

No *equal-frequency* deve-se definir quantas amostras se deseja ter por intervalo. Supondo três amostras por intervalo, os três primeiros valores irão fazer parte do intervalo *1*, os próximos três registros corresponderão ao intervalo *2* e assim por diante.

Considerando que deverão ser formados três *clusters* e como o agrupamento seleciona aleatoriamente os registros iniciais, uma possível seleção inicial de amostras seria: 20, 32 e 43. Agrupando esses registros com as demais amostras, procurando as mais similares, teremos ao fim do processamento os seguintes *clusters*: (17, 20, 23), (31, 32, 36) e (42, 43, 44). Os resultados podem ser visualizados na Tabela 1.

Tabela 1. Exemplo de discretização usando *equal-width*, *equal-frequency* e agrupamento.

Valores	Intervalo	Equal-width	Equal-frequency	Agrupamento
{17, 20, 23, 31, 32, 36, 42, 43, 44}	1	[17, 26)	[17, 20, 23]	[17, *20, 23]
	2	[26, 35)	[31, 32, 36]	[31, *32, 36]
	3	[35, 49)	[42, 43, 44]	[42, *43, 44]

***** = amostras selecionadas aleatoriamente

3.4.4 Métodos de Avaliação de Desempenho

O desempenho de um classificador pode ser avaliado de diversas maneiras, por exemplo, por meio da matriz de confusão, base para o cálculo de diversas medidas de desempenho e curva *ROC* (do inglês, *Receiver Operating Characteristics*), para análise gráfica (HANLEY e MCNEIL, 1982; MONARD e BARANAUSKAS, 2003).

3.4.4.1 Matriz de Confusão

A matriz de confusão (Tabela 2) em um problema com duas classes, positiva e negativa, mostra as quatro possibilidades que podem acontecer em predições inferidas por modelos (MONARD e BARANAUSKAS, 2003).

Verdadeiros Positivos (VP): Se referem aos exemplos de valor real "SIM" (classe positiva) preditos como "SIM".

Falsos Negativos (FN): Se referem aos exemplos de valor real "SIM" preditos como "NÃO" (classe negativa).

Verdadeiros Negativos (VN): Se referem aos exemplos de valor real "NÃO" preditos como "NÃO".

Falsos Positivos (FP): Se referem aos exemplos de valor real "NÃO" preditos como "SIM".

A partir da matriz de confusão podem ser derivadas medidas de desempenho como: taxa de acerto (acurácia), taxa de erro, sensitividade, especificidade e precisão, utilizadas nas avaliações dos *ensembles*.

Tabela 2. Matriz de confusão para duas classes: "SIM" e "NÃO".

	Predição	
Valor Real	**SIM**	**NÃO**
SIM	VP	FN
NÃO	FP	VN

Sensitividade: É a proporção de exemplos positivos preditos corretamente como positivos.

Especificidade: É a proporção de exemplos negativos preditos corretamente como negativos.

Precisão: É a proporção com que todos exemplos preditivos como positivos foram corretamente preditos.

As equações referentes às medidas de avaliação apresentadas são:

$$\text{Acurácia} = \frac{VP+VN}{n} \quad (2)$$

$$\text{Taxa de erro} = \frac{FP+FN}{n} \quad (3)$$

$$\text{Sensitividade} = \frac{VP}{VP+FN} \quad (4)$$

$$\text{Especificidade} = \frac{VN}{VN+FP} \quad (5)$$

$$\text{Precisão} = \frac{VP}{VP+FP} \quad (6)$$

n é o número total de exemplos.

3.4.4.2 Gráfico *ROC*

O gráfico *ROC* é uma ferramenta de análise visual útil para comparação entre dois ou mais modelos, pois mostra o *trade-off*, situação onde perde-se em um ponto para ganhar em outro, que existe entre a taxa de verdadeiros positivos (sensitividade) e a taxa de falsos positivos – equação (7) – variando de 0 a 1 (normalizados) ou de 0 a 100% (HANLEY e MCNEIL, 1982).

$$\text{Taxa de Falsos Positivos} = \frac{FP}{FP+VN} \quad (7)$$

A montagem do gráfico é realizada colocando a taxa de falso positivo no eixo das ordenadas e a sensitividade no eixo das abscissas (Figura 10).

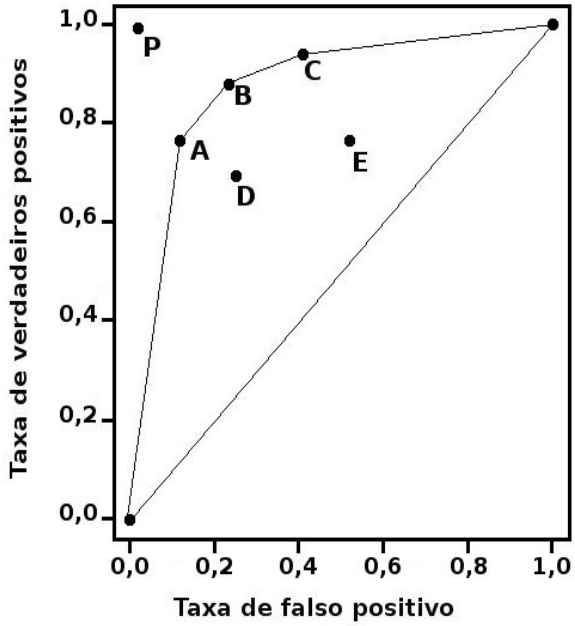

Figura 10. Exemplo de gráfico ROC contendo envelope convexo (*convex hull*).

A Figura 10 mostra um exemplo de gráfico *ROC*. Os pontos *A*, *B*, *C*, *D*, *E* e *P* são modelos selecionados para avaliação de desempenho. O ponto *P* representa o modelo perfeito, que classifica qualquer dado corretamente. A linha diagonal indica um modelo que classifica novos exemplos aleatoriamente. Também é possível observar um envelope externo convexo (*convex hull*), curva que passa sobre os pontos *A*, *B* e *C*. Os modelos presentes no envelope podem ser considerados ótimos, considerando a relação sensitividade e taxa de falso positivo. Já os modelos ausentes no envelope (Pontos *D* e *E*) podem ser descartados (HANLEY e MCNEIL, 1982; FAWCETT 2006).

Uma métrica derivada do gráfico ROC é a área embaixo da curva (*AUC*, do inglês *Area Under the Curve*). O valor do *AUC* varia de 0 a 1, uma vez que se trata de uma porção de um quadrado de tamanho 1 por 1. Isso significa que um modelo que classifica exemplos aleatoriamente (linha diagonal na Figura 10) possui *AUC* igual a 0.5 (FAWCETT, 2006). Esta medida representa o desempenho médio de um modelo, quanto maior seu valor, maior a probabilidade de sua classificação estar correta (BRADLEY, 1997).

3.4.5 Metodologias para Processo de Mineração de Dados

Diversas metodologias com fases (ou estágios) e tarefas bem definidas foram desenvolvidas para serem aplicadas em um processo de descoberta de conhecimento em bases

de dados (*KDD*, do inglês *Knowledge Discovery in Databases*) (MARISCAL et al., 2010). Nesse contexto pode-se destacar as metodologias *SEMMA*, 5A e *CRISP-DM*.

SEMMA (do inglês, *Sample, Explore, Modify, Model, Assess*) foi densevolvido pelo Instituto SAS como uma organização lógica para o conjunto de ferramentas *SAS Enterprise Miner*, responsável pelas tarefas de mineração de dados. Entretanto, é improvável a aplicação do *SEMMA* com outra ferramenta de mineração de dados (MARISCAL et al., 2010). Constitui de ciclos com 5 estágios (Figura 11):

Amostragem: Conjunto de dados a serem analisados. Deve ser grande o suficiente de modo que uma pequena proporção do conjunto: contenha uma quantidade de informações que permita uma análise; possa ser analisado de cada vez.

Exploração: Descoberta, de forma estatística e visual, de relações óbvias e tendências inesperadas no conjunto de dados, fornecendo subsídios para melhor conhecimento dos dados.

Modificação: Estágio de criação, seleção e transformação das variáveis do conjunto de dados.

Modelo: Aplicação de diversas técnicas de mineração de dados, criando modelos preditivos.

Avaliação: Estágio de avaliação do desempenho obtido pelo modelo desenvolvido (MATIGNON, 2007).

Metodologia é focada nos aspectos de desenvolvimento de modelo para projetos de mineração de dados e oferece um entendimento do processo, permitindo desenvolvimento organizado e adequado de projetos de mineração de dados (AZEVEDO e SANTOS, 2008).

Figura 11. Estágios da metodologia *SEMMA* (MARISCAL et al., 2010).

5A constitui de uma metodologia que possui as fases de avaliação, acesso, análise, ação e automatização (Figura 12). Trata-se de uma visão geral da análise de dados e processos de mineração de dados (MARISCAL et al., 2010).

Metodologia é mais próxima de uma tendência no desenvolvimento de projetos de mineração de dados do que modelo de processos, assim o 5A não descreve como desenvolver um projeto de mineração de dados. O 5A propõe tarefas, não como implementá-las. A

metodologia é a precursora do *CRISP-DM*, tanto que o ciclo de vida de ambas é similar (MARBÁN et al., 2009).

Vantagem da metodologia está em suas etapas, que buscam automatizar o processo. Assim, usuários sem conhecimento em mineração de dados podem aplicar modelos anteriores em novos conjuntos de dados. Desvantagem do 5A é a ausência de uma fase de conhecimento dos dados que serão usados no processo e, consequentemente, não há uma fase de teste da qualidade dos dados (MARISCAL et al., 2010).

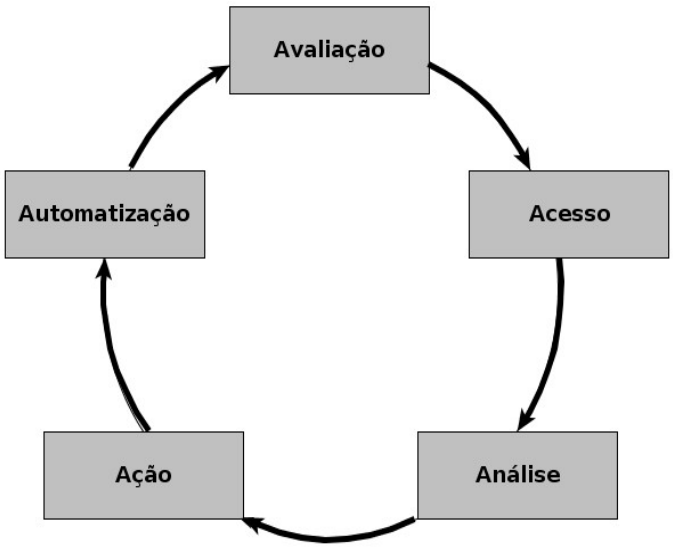

Figura 12. Estágios da metodologia 5A (MARISCAL et al., 2010).

Metodologia *CRISP-DM* (do inglês, *CRoss Industry Standard Process for Data Mining*) (CHAPMAN et al., 2000) é a mais difundida e utilizada, tornando-se a metodologia padrão para aplicar um processo KDD (MARISCAL et al., 2010). O ciclo de vida de um projeto é dividido em seis fases no *CRISP-DM*: compreensão do domínio, entendimento dos dados, preparação dos dados, modelagem, avaliação e distribuição.

A Figura 13 apresenta um modelo de referência, contendo uma visão geral das fases do *CRISP-DM*. A metodologia é iterativa e a sequência lógica das fases não é rígida. Podendo ser necessário avançar e voltar, em um projeto de *KDD*, pelas fases. As setas indicam as dependências mais importantes e frequentes entre as fases. O círculo externo simboliza a natureza cíclica da mineração de dados. As lições aprendidas durante o processo de mineração de dados podem encadear um novo projeto.

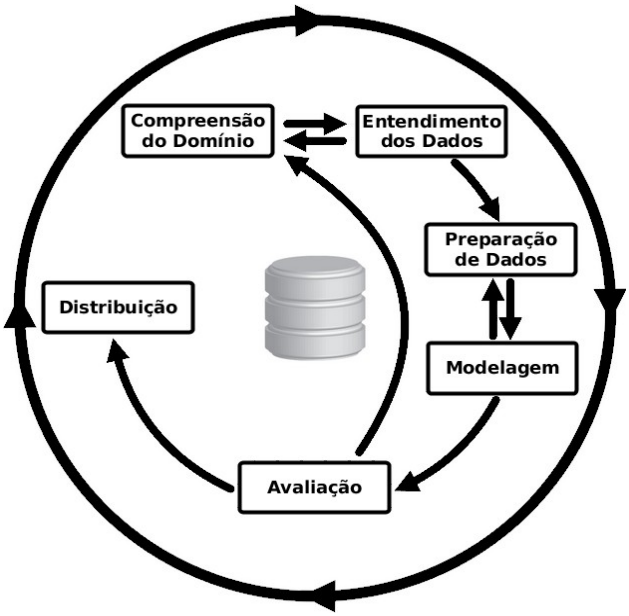

Figura 13. Visão geral das fases do processo *CRISP-DM* (CHAPMAN et al., 2000).

Abaixo descreve-se sucintamente as fases do *CRISP-DM*.

Compreensão do Domínio: A fase inicial foca no entendimento dos objetivos e requisitos do projeto, a partir da perspectiva do domínio da aplicação. Em seguida, converte-se esse conhecimento em um problema de mineração de dados e num plano preliminar, designado para alcançar tais objetivos.

Entendimento dos Dados: A fase começa com um conjunto de dados inicial e avança com atividades de familiarização dos dados, identificação de problemas na qualidade dos dados, descoberta das primeiras compreensões (*insights*) ou detecção de subconjuntos interessantes para o projeto.

Preparação dos Dados: A fase contém todas as atividades relacionadas à construção do conjunto de dados final, que será usado na fase de modelagem, a partir dos dados iniciais. As atividades podem ser realizadas diversas vezes e sem uma ordem específica. Essa fase inclui a transformação de dados e criação de novos atributos para o conjunto de dados.

Modelagem: Nessa fase, diversas técnicas de mineração de dados são selecionadas e aplicadas no conjunto de dados e seus parâmetros são calibrados para valores ótimos. Normalmente, existem diversas técnicas que podem ser usadas para um mesmo conjunto de dados e tarefa de mineração de dados. Algumas técnicas possuem requisitos específicos quanto à formatação dos dados. Então, às vezes é necessário voltar para a fase de preparação de dados.

Avaliação: Ao chegar nessa fase, o modelo construído aparenta ter bom desempenho na perspectiva da análise de dados. Antes de continuar para a fase de distribuição, é importante avaliar o modelo com maior rigor e revisar os passos executados durante sua construção para certificar que os objetivos traçados na primeira fase foram alcançados.

Distribuição: A criação do modelo, geralmente, não é o fim do projeto. Mesmo que o objetivo seja a criação do modelo para adquirir conhecimento sobre o conjunto de dados, o aprendizado ganho deve ser organizado e apresentado de forma que possa ser usado por terceiros.

A Tabela 3 mostra quais fases se equivalem em cada metodologia.

Tabela 3. Fases das metodologias SEMMA, 5A e *CRISP-DM*.

Metodologia	**Fases**						
SEMMA		Amostragem Exploração	Exploração Modificação	Modelagem Avaliação	Avaliação		
5A	Avaliação		Acesso	Análise	Ação		Automatização
CRISP-DM	Compreensão Do Domínio	Entendimento Dos Dados	Preparação Dos Dados	Modelagem	Avaliação	Distribuição	

4 MATERIAL E MÉTODOS

Nesse estudo foi utilizada a metodologia CRISP-DM (seção 3.4.5) para realizar o processo de mineração de dados por conta de sua maior abrangência em relação às demais metodologias, utilização em trabalhos análogos anteriores, de Meira et al. (2009) e Girolamo Neto et al. (2014), e maior relevância/adoção em projetos de mineração de dados (KDDNUGGETS, 2015).

As próximas seções abordam as fases de entendimento de dados (seção 4.1), preparação dos dados (seção 4.2), modelagem (seção 4.3), avaliação dos *ensembles* quanto a identificação dos momentos críticos da epidemia da ferrugem (seção 4.4) e avaliação dos modelos preditivos de Girolamo Neto (2013) (seção 4.5).

4.1 Entendimento dos Dados

4.1.1 Conjunto de Dados Brutos

Este trabalho utilizou dados meteorológicos e de acompanhamento mensal da incidência da ferrugem do cafeeiro, coletados em fazendas experimentais da Fundação Procafé. Os dados referem-se às fazendas localizadas nos municípios mineiros de Boa Esperança (latitude sul de 21º 03' 59", longitude oeste de 45º 34' 37" e altitude de 830 m), Carmo de Minas (latitude sul de 22º 10' 31", longitude oeste de 45º 09' 03" e altitude de 1180 m) e Varginha (latitude sul de 21º 34’ 00’’, longitude oeste de 45º 24’ 22’’ e altitude de 940 m). Em todas as lavouras selecionadas para amostragem foi cultivada a espécie *Coffea arabica* (café arábica).

As lavouras continham entre 6 e 20 anos de idade e dois tipos de plantio: espaçamento largo (cerca de 3,5m entre linhas e 0,7m entre plantas e densidade média de 4.000 plantas/ha) e adensado (cerca de 2,5m entre linhas e 0,5m entre plantas e densidade média de 8.000 plantas/ha). Havia lavouras com alta e baixa carga pendente de frutos e a avaliação da doença foi realizada em plantas das cultivares Catuaí (Vermelho e Amarelo) e Mundo Novo. A ferrugem não foi controlada durante o ano agrícola nos talhões escolhidos.

Dados meteorológicos, como temperatura, umidade relativa do ar e precipitação, foram registrados em intervalos de trinta minutos ao longo de um mês por meio de uma estação meteorológica automática.

Para a cidade de Boa Esperança foram coletados dados meteorológicos entre março de 2010 e junho de 2014 e o tipo de plantio entre plantas era largo. Em Carmo de Minas os registros meteorológicos se referem ao período de março de 2006 a junho de 2014, com tipo de plantio adensado. Boa Esperança e Carmo de Minas tinham duas lavouras com alta carga e duas com baixa carga.

O conjunto de dados meteorológicos de Varginha compreende o período de outubro de 1998 a junho de 2014. Em Varginha havia quatro lavouras com plantio adensado e quatro com espaçamento largo. Cada tipo de plantio possuía duas lavouras com alta carga e duas de baixa carga.

Arquivos

O conjunto de dados, para cada mês, é composto por três tipos de arquivos:

Arquivo Texto (.txt): Contém valores dos atributos meteorológicos, registrados em intervalos de trinta minutos durante um mês por uma estação meteorológica automática.

Planilha (.xls): Contém valores, em escala diária, de alguns atributos meteorológicos, calculados a partir do arquivo texto mencionado.

Documento (.doc/.docs/.pdf): Referente ao boletim mensal de aviso fitossanitário emitido pela Fundação Procafé, que contém informações climáticas e fenológicas sobre a cultura do café, além de dados relacionados à ocorrência de pragas e doenças, como a ferrugem do cafeeiro. No documento há um valor referente ao percentual de ataque (incidência) de determinada doença em cada uma das combinações possíveis entre espaçamento e carga pendente dos cafeeiros.

4.1.2 Descrição dos Dados

Os dados meteorológicos obtidos para Varginha, entre outubro de 1998 e dezembro de 2006, contêm 24 atributos. Demais dados possuem 34 atributos. Apenas alguns atributos, mais relevantes para o progresso da ferrugem do cafeeiro, foram usados para compor o conjunto de dados final, como temperaturas média, máxima e mínima, precipitação e umidade relativa do ar. Além disso, foram utilizados dados que identificam o momento da coleta, como data e hora.

No conjunto de dados brutos, as temperaturas média, máxima e mínima são valores numéricos e medidos em graus Celsius (°C). A precipitação é dada em milímetros

(mm), valor numérico, múltiplo de 0,2 e corresponde à somatória da precipitação durante o período de medição. A umidade relativa do ar é um número inteiro tendo a porcentagem (%) como unidade de medida. A data está no formato dia, mês e ano (dd/mm/aaaa) e a hora está como (hora:min), compreendendo os valores ([0-23]:[00|30]).

Alguns atributos extraídos dos boletins mensais de avisos fitossanitários fizeram parte do conjunto de dados final. Os atributos foram: carga (pendente de frutos) e incidência, que representa o porcentual de folhas com lesões causadas pela ferrugem do cafeeiro em uma lavoura. O atributo carga é categórico e pode assumir os valores *alta* ou *baixa* e a incidência possui dados numéricos e sua unidade de medida é a porcentagem (%). Os valores de incidência variam de 0 a 100.

4.1.3 Verificação da Qualidade dos Dados

Antes de iniciar a fase de preparação dos dados, foi preciso verificar a qualidade do conjunto de dados original. Aplicar procedimentos para verificação da qualidade dos dados de incidência mensal e meteorológicos garantem que as informações tenham sido geradas corretamente, auxiliam na identificação de dados inválidos e detectam/consertam problemas na manutenção das estações meteorológicas e na calibração de seus sensores (DORAISWAMY et al., 2000).

Os três tipos de arquivos foram analisados. Os boletins mensais de avisos fitossanitários foram conferidos para detecção de possíveis erros em suas informações como:

- Falta de dados e/ou valores inconsistentes para incidência mensal.
- Ausência de informações para determinada cidade.

Gráficos de incidência da doença pelo tempo foram criados, para cada ano agrícola, como ferramenta visual para identificação de possíveis valores anormais. Além disso, houve uma contagem na quantidade de boletins disponíveis para confirmar a existência de arquivos para todas as cidades durante o período de estudo.

Não houve ocorrências de falta de boletins, entretanto inconsistências foram verificadas em lavouras de Varginha com alta carga/adensamento largo e baixa carga/adensamento adensado para o ano agrícola 1999/2000. Esses dados foram analisados e eliminados do conjunto de dados brutos (seção 4.2.1.1).

Em relação aos dados climáticos, há, ao lado da estação meteorológica da Fundação Procafé em Varginha, uma estação do Instituto Nacional de Meteorologia (INMet).

Por meio do site do INMet foram obtidos dados de temperatura, precipitação e UR de janeiro de 2007 a junho de 2012, que foram utilizados em uma análise comparativa com registros de mesmo período da Fundação Procafé.

Os dados do INMet foram coletados de hora em hora, assim para que a análise comparativa pudesse ser feita, os dados da Fundação Procafé tiveram de ser transformados de trinta em trinta minutos para horários. Para isso foi calculada a média da temperatura e UR e somatória da precipitação nos registros coletados às [XX]:30 e [XX+1]:00. Por exemplo, os registros das 9:30h e 10:00h foram transformados no registro das 10:00h.

Durante a comparação dos dados foi constatado a ausência, em todos os anos, de dois registros seguidos (02:00h e 02:30h) durante a madrugada em outubro nos dados da Fundação Procafé. A suspeita, que posteriormente foi confirmada junto ao técnico da Fundação, é que a estação meteorológica estava considerando o horário de verão em suas medições. Ao começar o horário de verão a estação pulou dois registros, que equivalem a uma hora de coleta. Esse padrão foi identificado para os dados de Boa Esperança e Carmo de Minas. Em Varginha o horário de verão foi considerado apenas para coletas a partir de 2007. Os períodos de horário de verão foram identificados e corrigidos antes de realizar a análise comparativa com os dados do INMet (seção 4.2.2.4).

Na análise comparativa os dados de ambas estações de Varginha foram divididos por ano de coleta. *Boxplots* (Figura 14), sumarização dos dados (Tabela 4), gráficos de valor da variável pelo tempo (Figura 15) e testes de correlação foram gerados para cada uma das três variáveis por meio de *scripts* na linguagem de programação R.

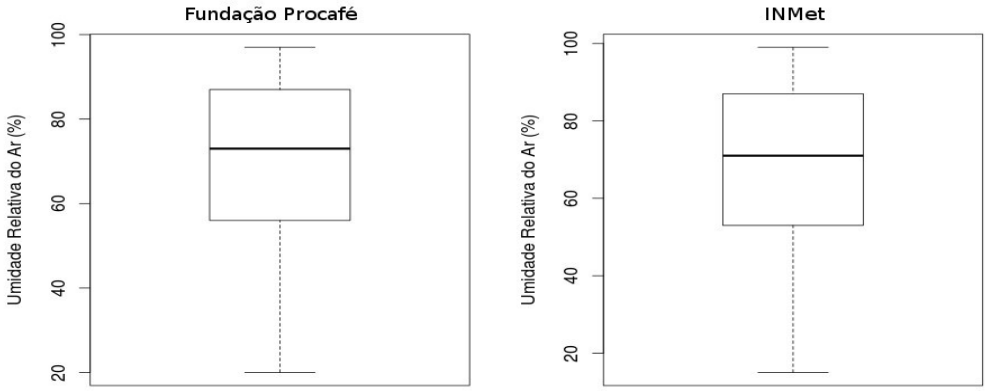

Figura 14. Boxplots dos dados de UR de Varginha registrados pela Fundação Procafé e INMet em 2007.

Tabela 4. Sumário dos dados de UR de Varginha registrados pela Fundação Procafé e INMet em 2007.

	Procafé	InMet
Mínimo	20,00	15,00
1° quartil	56,00	53,00
Mediana	73,00	71,00
Média	70,52	68,67
3° quartil	87,00	87,00
Máximo	97,00	99,00
Variância	344,43	420,81
Desvio padrão	18,56	20,51

O teste de correlação mede o grau de correlação entre duas variáveis e varia de 0 a 1, sendo 1 o valor que indica uma correlação perfeita entre as variáveis. Considerando todo o período de amostragem, a correlação entre os valores de temperatura foi de 0,98, precipitação foi 0,76 e para UR foi 0,97. Posteriormente, com auxílio de especialistas na área de agrometeorologia e estatística foi confirmada alta associação entre os valores de temperatura e UR, além da precipitação para registros de julho 2012 a dezembro 2014.

Os arquivos texto das estações meteorológicas foram verificados mediante exploração visual e testes de consistência dos dados. O primeiro teste abordado foi a verificação no número de registros em cada arquivo. Por exemplo, para os dados registrados em intervalos de trinta minutos ao longo de um mês com trinta dias espera-se um arquivo contendo 48 (registros/dia) x 30 (dias) = 1440 registros. A planilha deve ter o número de registros igual ao número de dias do mês que esta representa.

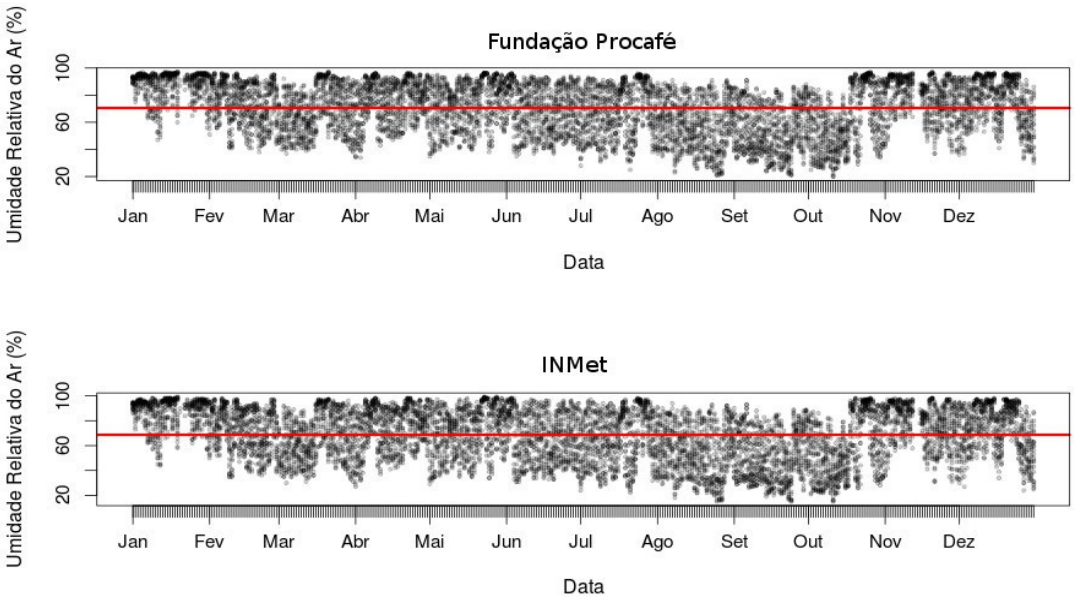

Figura 15. Gráfico de UR pelo tempo para Varginha em 2007. Registros da Fundação Procafé e INMet. A linha horizontal indica o valor médio.

Após a verificação, alguns meses tiveram que ser descartados devido ao alto número de registros faltantes em seu arquivo texto (seção 4.2.1.2). Além disso, foram identificados períodos sem dados para Varginha, que foram completados com os registros do INMet (seção 4.2.2.3). Também constatou-se, em todas as estações, que houve anos em que a UR não atingiu 100% (seção 4.2.2.2).

O intuito da aplicação dos testes de consistência (Tabela 5) é verificar a consistência dos dados e detectar registros defeituosos. Foram testadas, para conjuntos de dados das três cidades, as seguintes variáveis meteorológicas: temperatura, umidade relativa do ar (UR) e precipitação. Testes de qualidade foram selecionados com base em Estévez et al. (2011) e aplicados mediante *scripts* elaborados em *R*.

Por meio desses testes detectou-se registros de UR com valores baixos em sequências com UR=100% para Varginha, nos registros de 1998 a 2006, que também precisaram ser ajustados (seção 4.2.2.1). Não foram identificados valores inconsistentes para temperatura e precipitação.

Tabela 5. Testes de consistência aplicados no conjunto de dados meteorológicos original.

Atributo	**Teste**
data	DATA = DD/MM/AAAA
hora	HORA = HH:[00|30]
temperatura	$-30 < TEMP < 50$
	$TMIN < TEMP < TMAX$
	$\lvert TEMP_i - TEMP_{i-2} \rvert < 4$
	$\lvert TEMP_i - TEMP_{i-4} \rvert < 7$
	$\lvert TEMP_i - TEMP_{i-6} \rvert < 9$
	$\lvert TEMP_i - TEMP_{i-12} \rvert < 15$
	$\lvert TEMP_i - TEMP_{i-24} \rvert < 25$
	$TMAX_d > TEMP_d > TMIN_d$
	$TEMP_d > TEMP_{d-1}$
	$TMIN_d \leq TMAX_{d-1}$
precipitação	$0 \leq PRECIP \leq 120$
	$0 \leq PRECIP_D \leq 508$
	PRECIP múltiplo de 0,2
umidade relativa do ar	$0{,}8 < UR < 103$
	$\lvert UR_i - UR_{i-1} \rvert < 45$

$TEMP_i$: valor de temperatura média registrado no momento *i;*

$TEMP_{i-2}$: valor de temperatura média registrado uma hora antes do momento *i;*

$TEMP_{i-4}$: valor de temperatura média registrado duas horas antes do momento *i* e assim por diante.

$TEMP_d$: valor de temperatura média registrado ao longo do dia *d;*

$TEMP_{d-1}$: valor de temperatura média registrado ao longo do dia *d-1;*

$TMAX_d$: maior valor de temperatura máxima registrado ao longo do dia *d;*

$TMAX_{d-1}$: maior valor de temperatura máxima registrado ao longo do dia *d-1;*

$TMIN_d$: menor valor de temperatura mínima registrado ao longo do dia *d;*

$PRECIP_d$: valor do somatório da precipitação registrado ao longo do dia *d;*

UR_i: valor da umidade relativa do ar registrado no momento *i*.

UR_{i-1}: valor da umidade relativa do ar registrado trinta minutos antes do momento *i*.

4.2 Preparação dos Dados

4.2.1 Eliminação de Dados

4.2.1.1 Eliminação de Dados de Incidência Mensal da Ferrugem

Por meio de gráficos de incidência mensal por tempo criados para cada ano agrícola e estação foram constatadas inconsistências em Varginha para lavouras com alta carga pendente e adensamento largo e baixa carga carga pendente e adensamento adensado em 1999/2000 (Figura 16). Nesse ano agrícola os valores de incidência para baixa/adensado foram maiores que alta/largo e próximos dos números de alta/adensado.

Valores de alta/largo e baixa/largo também apresentaram semelhanças. Esses fatos são incomuns uma vez que a tendência é que cafeeiros com alta carga pendente de frutos sofram ataques mais intensos e severos da ferrugem (ZAMBOLIM et al., 2005).

Nesse caso não foi possível confirmar, junto aos técnicos da Fundação Procafé, se os dados estão corretos. Assim os registros de incidência para alta/largo e baixa/adensado para safra 1999/2000 de Varginha foram removidos do conjunto de dados brutos.

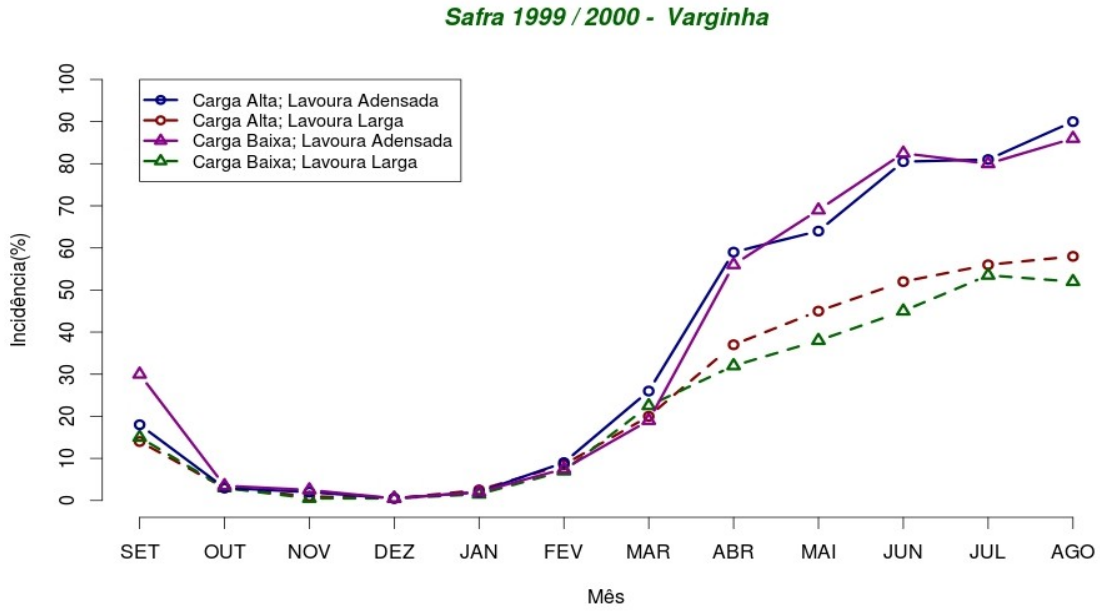

Figura 16. Dados de incidência mensal da ferrugem do cafeeiro na safra 1999/2000 coletados na fazenda experimental da Fundação Procafé em Varginha.

4.2.1.2 Eliminação de Dados Meteorológicos

Conforme a verificação na quantidade de registros nos arquivos texto (seção 4.1.3) foram identificados três meses de Boa Esperança e Carmo de Minas e dois meses de Varginha com quantidade de dados considerado insuficiente (menos de 10% dos registros esperados). Esses oito meses foram eliminados do conjunto de dados brutos (Tabela 6).

Tabela 6. Meses eliminados do conjunto de dados brutos por falta de dados.

mês/ano	cidade
11/2012	Boa Esperança
12/2012	
01/2013	
09/2006	Carmo de Minas
01/2014	
02/2014	
10/2000	Varginha
11/2000	

Dados das três estações apresentaram registros considerados nulos, onde não há valores para temperatura (mínima, média e/ou máxima), precipitação e/ou UR. Nos conjuntos de dados originais ainda continham registros medidos em horário incorreto, ou seja, a coleta não foi realizada em intervalos de trinta (XX:[00|30]). Registros nulos ou coletados em hora indevida foram eliminados do conjunto de dados brutos.

4.2.2 Correções nos Dados Meteorológicos

Esta seção trata da metodologia aplicada para análise e correção de dados meteorológicos, identificados por meio de análise visual e testes de consistência (seção 4.1.3) e contou com auxílio de um especialista em agrometeorologia. A variável que apresentou erros de consistência foi a UR. Alguns registros de temperatura e precipitação foram detectados como inconsistentes. Porém com uma análise mais aprofundada por parte do especialista, avaliou-se que esses registros estavam dentro de um limite aceitável e plausível ao se considerar o contexto do clima no mês ou ano no qual os registros foram coletados.

Foram aplicados procedimentos para corrigir a UR nas três estações (seções 4.2.2.1 e 4.2.2.2) e complementar dados de Varginha (seção 4.2.2.3) nos conjuntos de dados brutos, ou seja, sem considerar a questão do horário de verão (seção 4.2.2.4).

Corrigir os valores de UR, para esta pesquisa, é fundamental pois a partir da UR foi derivado o período de molhamento foliar, que considerou apenas registros com UR ≥ 90% (seção 4.2.4). Todo o processo de correção de dados foi aplicado por meio de *scripts* em R.

4.2.2.1 UR em Varginha (1998 a 2006)

Por meio dos testes de consistência identificou-se a presença de valores baixos de UR dentre sequências UR=100% e em períodos noturnos, chuvosos e/ou com molhamento foliar. Meira (2008) identificou valores de UR=33% em sequências de UR=100%. Todos os registros citados são de Varginha e foram alterados do valor original para 100.

Ainda ocorreram, em sete dias de setembro de 2002 a janeiro de 2003 na estação de Varginha, sequências de UR=100%, independentemente da hora, temperatura e ocorrência de chuva. A UR desses registros foi ajustada com base nos dados da estação meteorológica do INMet localizada em Lavras, que possui características climáticas similares a Varginha.

As coletas em Lavras aconteceram nos horários das 9h, 15h e 21h. A UR nos intervalos entre esses horários foram avaliados e, quando necessários, corrigidos. Para avaliação da UR foi observado a temperatura baseando-se nos registros de Lavras. Sempre que a temperatura em Varginha fosse menor que em Lavras, a UR devia ser maior e vice-versa, uma vez que a temperatura e UR são inversamente proporcionais (TORRES et al., 2009).

Por vezes não foi possível ter uma boa ideia da UR, então os dados foram completados observando a curva padrão da UR nos dias com dados meteorológicos consistentes entre setembro de 2002 e janeiro de 2003. Nesse caso não foi utilizado a base de dados do INMet para Varginha por indisponibilidade dos registros nesse período.

4.2.2.2 UR em Boa Esperança, Carmo de Minas e Varginha (2007 a 2014)

Durante a verificação da consistência dos dados observou-se poucos registros de UR=100% em Boa Esperança, Carmo de Minas e Varginha (período de 2007 a 2014). Isso pode ter ocorrido devido a um sensor descalibrado ou característica do sensor, alguns modelos admitem até 3% de erro na medição. De qualquer forma é esperado que a UR chegue a 100% com maior frequência da constatada nos dados.

A Tabela 7 mostra a quantidade de vezes, nas três estações, que foram observadas UR com valores de 95 a 100%. Nota-se, por exemplo, que em 2010 não foi medido registro com UR=100%.

A partir dos valores observados na Tabela 7 determinou-se qual seria o fator de correção de UR em cada ano e estação, quando necessário. O intuito foi elevar todos os registros de UR, multiplicando-os por um mesmo fator de modo a ajustá-los e, consequentemente, apresentar maior frequência de UR=98, 99 e 100%.

Por exemplo, nos dados de Boa Esperança em 2010 não há UR=100%, ocorre baixa frequência de UR=99% (20) e o número de UR=98% (295) é razoável. Assim, foi considerado UR=98% como novo parâmetro a ser ajustado até UR=100%. Ou seja, buscou-se um fator de correção que multiplicado aos valores de UR fizessem com que registros UR=98% (295) passassem a ser UR=100%, registros UR=97% (416) seriam UR=99% e assim por diante. O fator de correção foi 100/98 = 1,020408. O valor de UR ajustado é o resultado da multiplicação do valor original pelo fator de correção. Os valores ajustados foram arredondados para baixo e caso seu valor excedesse 100%, este era alterado para UR=100%.

Nos anos de correção de valores de 99% para 100%, por exemplo em 2011 e 2013 para Boa Esperança, o fator de correção (100/99=1,010101) não modifica os dados, pois 1,010101*89=89,898989%. Dessa forma, nesses anos o fator de ajuste foi 1,012, uma vez que 1,012*89=90,068%, que era arredondado para 90%.

Tabela 7. Análise da quantidade de vezes que UR chega a 100%, em cada ano, nos dados de Boa Esperança.

		Boa Esperança								
		ANO								
		2006	2007	2008	2009	2010	2011	2012	2013	2014
UR	100%	-	-	-	-	0	2	3	3	1
	99%	-	-	-	-	20	171	54	181	18
	98%	-	-	-	-	295	751	560	717	375
	97%	-	-	-	-	416	724	556	858	438
	96%	-	-	-	-	495	662	590	813	502
	95%	-	-	-	-	455	630	602	748	435

		Carmo de Minas								
		ANO								
		2006	2007	2008	2009	2010	2011	2012	2013	2014
UR	100%	0	0	0	15	19	294	552	1291	703
	99%	0	0	19	98	133	404	420	558	318
	98%	0	0	117	322	255	250	390	441	259
	97%	6	106	247	494	323	318	402	415	243
	96%	101	492	534	675	526	502	511	576	359
	95%	268	392	524	617	458	421	374	374	219

		Varginha								
		ANO								
		2006	2007	2008	2009	2010	2011	2012	2013	2014
UR	100%	-	0	0	0	0	1	0	0	0
	99%	-	0	0	0	0	2	0	0	0
	98%	-	0	0	28	6	113	0	0	0
	97%	-	31	120	412	140	420	0	0	0
	96%	-	392	599	1013	561	481	0	1	1
	95%	-	717	794	982	639	586	31	114	63

Posteriormente ao processo de correção, foram conferidos mês a mês, para cada ano e estação, se o maior valor de UR chegava a 100%. Nos poucos meses em que isso não ocorreu geralmente era um mês da estação seca, onde a umidade relativa é muito baixa, chegando apenas próximo aos 90% em muitas manhãs.

4.2.2.3 Complementação dos dados meteorológicos de Varginha

Em Varginha notou-se a ausência de dados meteorológicos para alguns dias durante o período de 2007 a 2014. Os dados ausentes de temperatura, precipitação e UR foram completados com registros do INMet a fim de evitar o descarte de dados de um mês inteiro e melhorar a qualidade dos dados brutos. Demais variáveis climáticas não foram complementadas por não serem relevantes para esse estudo.

4.2.2.4 Horário de Verão

Identificar o começo do horário de verão foi simples, bastou identificar a falta dos registros das 2:00h e 2:30h nos meses de outubro. Entretanto, o fim da marcação do horário de verão não pôde ser detectado com precisão uma vez que a estação atrasava seu relógio em uma hora e sobrescrevia os registros já coletados, não houve controle por parte da Fundação Procafé sobre essa questão e o período de horário de verão, aparentemente, foi estendida em relação ao período oficial.

A análise sobre o fim da coleta de dados no horário de verão ficou restrita às variáveis radiação solar média (W/m²) e energia solar média (ly), coletadas por todas estações. Observou-se um deslocamento nos valores de radiação solar e energia solar durante a manhã ou noite (nascer ou pôr do sol) em uma hora de um dia para outro. Por exemplo, em Carmo de Minas até o dia 4/4/2009 foram coletados valores positivos para radiação e energia solar até o registro das 19h, a partir do dia 5 a última marcação positiva para essas variáveis foi às 18h. Assim, determinou-se como fim do horário de verão de 2008/2009 o dia 4/4/2009 às 19h (Tabela 8).

Tabela 8. Exemplo da identificação do horário de verão 2008/2009 para Carmo de Minas.

03/04/2009			04/04/2009			05/04/2009		
Hora	Radiação Solar Média (W/m²)	Energia Solar Média (ly)	Hora	Radiação Solar Média (W/m²)	Energia Solar Média (ly)	Hora	Radiação Solar Média (W/m²)	Energia Solar Média (ly)
17:00	183	7,87	17:00	171	7,35	17:00	113	4,86
17:30	112	4,82	17:30	160	6,88	17:30	36	1,55
18:00	26	1,12	18:00	87	3,74	18:00	4	0,17
18:30	18	0,77	18:30	33	1,42	18:30	0	0
19:00	6	0,26	19:00	3	0,13	19:00	0	0
19:30	0	0	19:30	0	0	19:30	0	0
20:00	0	0	20:00	0	0	20:00	0	0

Ainda ocorreram anos em que os dados meteorológicos de alguns meses como fevereiro, março, outubro estavam incompletos a ponto de não permitir a identificação do período do horário de verão. Dessa forma definiu-se os períodos do horário de verão (Tabela 9). Os horários dos registros coletados durante o período de horário de verão foram atrasados em uma hora, por meio de *scripts* em R, antes para que fosse possível realizar a análise comparativa com dados do INMet e utilizá-los no processo de transformação dos dados.

Tabela 9. Períodos do horário de verão identificados para as três estações da Fundação Procafé.

Boa Esperança		
Ano	**Início**	**Fim**
2010/2011	07/11/2010 – 03:00h	28/02/2011 – 19:30h
2011/2012	06/11/2011 – 03:00h	10/03/2012 – 18:30h
2012/2013	-	-
2013/2014	03/11/2013 – 03:00h	28/02/2014 – 20:00h
2014/2015	02/11/2014 – 03:00h	01/01/2015 – 00:00h

Carmo de Minas		
Ano	**Início**	**Fim**
2006/2007	29/10/2006 – 03:00h	31/03/2007 – 19:00h
2007/2008	-	-
2008/2009	26/10/2008 – 03:00h	04/04/2009 – 19:00h
2009/2010	25/10/2009 – 03:00h	02/04/2010 – 19:00h
2010/2011	31/10/2010 – 03:00h	18/02/2011 – 19:00h
2011/2012	30/10/2011 – 03:00h	18/02/2012 – 19:00h
2012/2013	28/10/2012 – 03:00h	25/02/2013 – 19:00h
2013/2014	27/10/2013 – 03:00h	27/02/2014 – 12:00h
2014/2015	26/10/2014 – 03:00h	01/01/2015 – 00:00h

Varginha		
Ano	**Início**	**Fim**
2006/2007	01/01/2007 – 00:30h	01/03/2007 – 10:00h
2007/2008	28/10/2007 – 03:00h	05/04/2008 – 18:00h
2008/2009	26/10/2008 – 03:00h	02/03/2009 – 01:00h
2009/2010	25/10/2009 – 03:00h	04/04/2010 – 06:00h
2010/2011	31/10/2010 – 03:00h	03/03/2011 – 02:00h
2011/2012	-	-
2012/2013	28/10/2012 – 03:00h	06/04/2013 – 15:00h
2013/2014	27/10/2013 – 03:00h	05/04/2014 – 18:00h
2014/2015	26/10/2014 – 03:00h	01/01/2015 – 00:00h

4.2.3 Novo Limiar do Atributo Meta para Baixa Carga Pendente

Nesse trabalho, buscou-se identificar um novo limiar para o atributo meta em lavouras com baixa carga pendente, visto que os modelos preditivos desenvolvidos para a incidência da ferrugem não apresentaram a mesma performance quando usados registros novos, que não estavam contidos no conjunto de treinamento desses modelos (seção 3.3).

Valores sobre TP em lavouras com baixa carga pendente foram discretizados de forma a encontrar um valor menor que 5 p.p. e ser utilizado como limiar do atributo meta. Três abordagens foram usadas para discretização: número pré-determinado de intervalos iguais (*equal-width*), número uniforme de amostras por intervalo (*equal-frequency*) e discretização por agrupamento (*clustering*).

Equal-width e *equal-frequency* foram testadas por meio de *scripts* em R, Python e Weka, enquanto o *clustering* foi aplicado usando o Weka. Nas três abordagens o conjunto de dados foi dividido em dois intervalos (*bins*) de modo a identificar um valor candidato a limiar (Tabela 10), uma vez que este estudo trabalha com atributo meta binário. Assim, foram encontrados os valores: 1,0; 2,25; 2,9; 4,0 e 4,5. Os dois últimos valores foram considerados muito próximos do limiar de 5 p.p. e foram descartados. Os valores 1,0 e 2,25 eram baixos e haveria o risco do modelo predizer somente aumentos na TP se esses limiares fossem utilizados. Escolheu-se 2,9 como novo limiar para o atributo meta, porém este valor foi arredondado para 3,0.

Tabela 10. Candidatos a novo limiar encontrados após discretização dos dados de incidência em lavouras com baixa carga.

Discretização	Weka	R	Python
Equal-frequency (2 bins)	2,25	2,9	1,0
Equal-width (2 bins)	4,5	4,0	4,0
K-means (k=2)	2,25	-	-

4.2.4 Atributos do Conjunto de Dados

Atributo Meta

O atributo meta (ou variável dependente) utilizado foi a taxa de progresso mensal da ferrugem do cafeeiro (TP), gerado a partir dos valores de incidência da doença. A TP, dada em pontos porcentuais (p.p.), é a diferença entre os valores de incidência mensal da ferrugem na lavoura em dois meses subsequentes – equação (8).

$$\mathrm{TP} = \frac{I_x - I_{x-1}}{\Delta t} \qquad (8)$$

TP – taxa de progresso, I_x – incidência da ferrugem no mês x, I_{x-1} – incidência da ferrugem no mês x-1, Δt – período de um mês.

O intuito foi utilizar os valores de 3, 5 e 10 p.p. como limiares para TP. Limiares de 3 e 10 p.p. foram utilizados somente nos *ensembles* desenvolvidos a partir de dados de cafeeiros em baixa e alta carga, respectivamente. Já o limiar de 5 p.p. foi usado nos *ensembles* de ambas cargas pendentes.

Uso de 3 p.p. foi definido conforme a discretização dos valores da TP em lavouras em baixa carga pendente (seção 4.2.3). Limiares de 5 e 10 p.p. foram baseados em Zambolim

et al. (1997) e Kushalappa et al. (1984), que recomendaram o início do controle da ferrugem quando a porcentagem das folhas doentes na lavoura chegasse a 5% e 10%, respectivamente. O limiar de 10% também é próximo ao limite máximo de folhas doentes, de 12%, recomendado na aplicação de fungicidas sistêmicos para controle da ferrugem (ZAMBOLIM et al., 1997).

Limiares de 3 e 5 p.p. foram usados como limiar do atributo meta (TP) no desenvolvimento de *ensembles* para cafeeiros em baixa carga. Já para alta carga, os limiares de 5 e 10 p.p. foram usados para o atributo meta. O valor da TP é binário, atribuindo-se *1* para taxas de progresso maiores ou iguais ao limiar e *0* caso contrário.

Atributos Preditivos

A criação de atributos preditivos (ou variáveis independentes) meteorológicos abordou variáveis de disponibilidade mais ampla e relevantes para desenvolvimento da doença, como temperatura, umidade relativa do ar (UR) e precipitação (ZAMBOLIM et al., 2002). Isso resulta em uma simplificação nos atributos em relação Meira (2008) e Girolamo Neto (2013), que ainda utilizaram dados sobre lavoura, vento e radiação solar.

Alguns atributos meteorológicos foram gerados com base em Meira (2008) e Girolamo Neto (2013) para derivar atributos a partir do período de incubação (PI) e do período de infecção (PINF). Outros foram calculados a partir do período de molhamento foliar prolongado, uma vez que a presença de água é necessária para a germinação do fungo na folha (ZAMBOLIM et al., 2002). Períodos com alta UR (maior ou igual a 90%) foram usados como medida indireta de molhamento foliar contínuo.

A criação de atributos a partir de registros coletados no período noturno também foi implementada, já que a infecção ocorre com pouca ou na ausência de luminosidade (MONTOYA e CHAVES, 1974).

Sob determinadas condições de molhamento foliar (M) e temperatura durante o período de molhamento (T) pode-se inferir se um dia é favorável para o patógeno da ferrugem infectar o cafeeiro.

Neste trabalho buscou-se aperfeiçoar e simplificar as regras inferidas por Meira (2008). Foi criada uma matriz de condições diárias de infecção (Tabela 11) baseando-se em Nutman e Roberts (1970), Montoya e Chaves (1974), Kushalappa et al. (1983) para escolher os valores limites de temperatura e molhamento foliar. Assim, foi possível determinar se um dia seria desfavorável (valor *0*) ou favorável (valor *1*) à ferrugem do cafeeiro.

Tabela 11. Matriz de infecção diária.

M \ T	T < 18	18 ≤ T < 21	21 ≤ T ≤ 24	24 < T ≤ 28	T > 28
NHCUR90 < 6	0 (Desfavorável)	0 (Desfavorável)	0 (Desfavorável)	0 (Desfavorável)	0 (Desfavorável)
6 ≤ NHCUR90 < 12	0 (Desfavorável)	0 (Desfavorável)	1 (Favorável)	0 (Desfavorável)	0 (Desfavorável)
NHCUR90 >= 12	0 (Desfavorável)	1 (Favorável)	1 (Favorável)	1 (Favorável)	0 (Desfavorável)

NHCUR90 – número máximo de horas com UR ≥ 90% considerando quebra de 1h
T = TCUR90 – temperatura média que envolve os registros referentes ao NHCUR90

4.2.5 Transformação dos Dados

Os atributos preditivos (ou variáveis independentes) meteorológicos foram construídos a partir dos dados horários (coletados em intervalos de trinta minutos), registrados pelas estações meteorológicas, passando para o nível diário e, posteriormente, mensal. Essa transformação foi dividida em passos (Figura 17), permitindo, ao fim do procedimento, analisar o relacionamento das variáveis meteorológicas com o atributo meta.

O processo de transformação na granularidade dos dados foi realizado com base em *scripts* em Python. Estes foram modificados a partir dos *scripts* de Meira (2008) a fim de atender a criação dos novos atributos propostos neste trabalho (seção 4.2.4).

Passo 1- Reunião dos Dados Brutos e Criação do Atributo Meta

O Passo 1 da transformação dos dados juntou (concatenou) todos os arquivos com dados meteorológicos (.txt) que possuem uma qualidade mínima em seus registros, verificada na seção 4.1.3. Dessa forma, foi criado um arquivo para cada estação.

Nesse passo também foram criados atributos referentes ao dia epidemiológico (DATA_EPID e HORA_EPID). O período de 24h entre dois dias, começando às 12:00h foi considerado como um dia epidemiológico. Por exemplo, o período das 12:00h às 24:00h do dia 08/07/2014 acrescido do período das 00:00h às 12:00h do dia 09/07/2014 é o dia epidemiológico 08/07/2014. Esse procedimento visou evitar quebra no período noturno, responsável pelos maiores valores de molhamento foliar.

A seguir, os atributos retirados dos boletins de avisos (incidência da ferrugem e carga pendente) foram reunidos em uma planilha (.csv). Os valores do atributo meta foram calculados e adicionados à planilha. A planilha foi retomada no Passo 4.

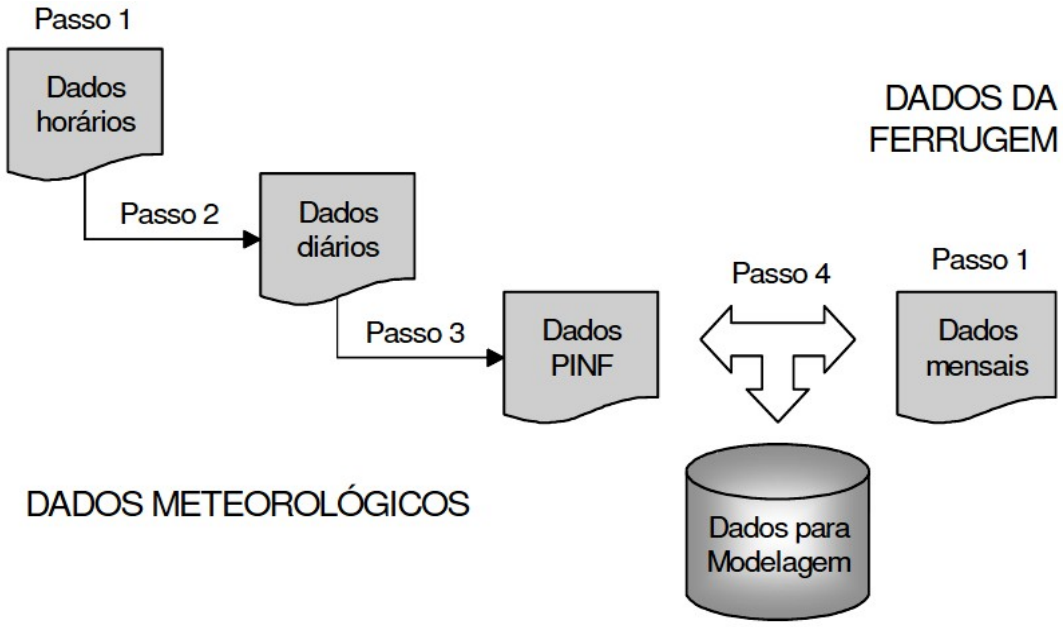

Figura 17. Esquema geral da transformação dos dados para modelagem (MEIRA, 2008).

Passo 2 – Criação de Atributos no Nível Diário

O segundo passo foi o mais trabalhoso e transformou dados horários para diários. Para isso, foram usadas subtarefas para execução do passo.

Subtarefa 1 – Cálculo de Estatísticas Descritivas Diárias: o arquivo gerado no Passo 1 serviu como base para cálculo de estatísticas diárias dos atributos meteorológicos, como temperaturas média, máxima e mínima do dia, precipitação acumulada do dia e umidade relativa média do dia.

Subtarefa 2 – Criação de Atributos Relacionados com Molhamento Foliar: a partir do arquivo gerado no Passo 1 derivou-se, em nível diário, atributos relacionados ao molhamento foliar contínuo. Considerou-se como período de molhamento foliar contínuo o intervalo de tempo no qual a UR foi maior ou igual a 90%. Foram consideradas duas situações: tolerância ou não de uma quebra (UR<90%) durante o intervalo de, no máximo, uma hora. Assim, criou-se os atributos

- NHDUR90 – número de horas do período com molhamento foliar (UR ≥ 90%) considerando quebra;
- TDUR90 – média da temperatura ao longo do período de molhamento foliar (UR ≥ 90%) considerando quebra;
- SMTNHDUR90 – somatório do número de horas de molhamento foliar (UR ≥ 90%) considerando quebra;

- NHCUR90 –número de horas do período de molhamento foliar (UR ≥ 90%) sem considerar quebra;
- TCUR90 – média da temperatura ao longo do período de molhamento foliar (UR ≥ 90%) sem considerar quebra.

NHDUR90 e TDUR90 serviram de base para criação de outros atributos que consideravam o período de infecção (Passo 3). NHCUR90 e TCUR90 foram usadas para verificar se o dia era favorável para infecção do patógeno no cafeeiro (Tabela 11).

Subtarefa 3 – Criação de Atributos Relacionados com Período de Incubação: baseando-se no arquivo gerado na Subtarefa 1, foram gerados atributos relacionados com o período de incubação (PI), calculado para cada dia do conjunto de dados. A partir do PI foram estimados os prováveis dia, mês e ano de infecção que, provavelmente, contribuíram à incidência da ferrugem na lavoura (Meira, 2008). O intuito era conseguir relacionar todos os dias (prováveis para infecção do patógeno no hospedeiro) com as taxas mensais de progresso da incidência da ferrugem do cafeeiro.

Subtarefa 4 – Junção dos arquivo gerados: todos os arquivos gerados nas Subtarefas 1 a 3 foram juntados em um arquivo, usado no Passo 3.

Passo 3 - Criação de Atributos para o PINF

No Passo 3 foram criados atributos de cada período de infecção (PINF), a partir do arquivo gerado na Subtarefa 4 do Passo 2, convertendo os registros diários para mensais. A seguir, foram adicionados os valores de incidência mensal da ferrugem no mês anterior.

Passo 4 – Integração dos Dados

No quarto passo houve a integração entre a planilha com dados retirados dos boletins de aviso do Passo 1 e o arquivo resultante do Passo 3, contendo dados meteorológicos relacionados ao PINF. A integração ocorreu por meio de dois atributos comuns à planilha e ao arquivo: mês e ano.

Passo Adicional – Integração dos Dados das Cidades

Os Passos 1 a 4 foram realizados isoladamente para cada cidade (Boa Esperança, Carmo de Minas e Varginha), resultando em três arquivos ao fim do Passo 4. Nesse Passo Adicional foi realizada a integração (junção) entre os registros mensais das três cidades. Isso produziu o conjunto de dados parcial, contendo todas as variáveis criadas (Tabela 12), dessa forma os *ensembles* foram desenvolvidos a partir de registros coletados nas três cidades. Esse

conjunto de dados ainda foi dividido em cenários distintos, que se distinguiram conforme a carga pendente da lavoura, limiar da TP e atributos selecionados (seção 4.2.6).

Tabela 12. Todos atributos derivados a partir do conjunto de dados brutos.

Atributo	Significado	Tipo	Formato	Unidade de Medida
incidencia_anterior	Valor de incidência na lavoura no mês anterior	Numérico	Decimal	%
tmed_pinf	Média das temperaturas médias diárias no PINF	Numérico	Decimal	° C
tmax_pinf	Média das temperaturas máximas diárias no PINF	Numérico	Decimal	° C
tmin_pinf	Média das temperaturas mínimas diárias no PINF	Numérico	Decimal	° C
tmed_pi_pinf	Média das temperaturas médias diárias no PI para os dias do PINF	Numérico	Decimal	° C
tmax_pi_pinf	Média das temperaturas máximas diárias no PI para os dias do PINF	Numérico	Decimal	° C
tmin_pi_pinf	Média das temperaturas mínimas diárias no PI para os dias do PINF	Numérico	Decimal	° C
ur_pinf	Média da UR diária no PINF	Numérico	Inteiro	-
med_nhdur90_pinf	Média do número de horas diárias com UR ≥ 90 no PINF	Numérico	Decimal	h
smt_nhdur90_pinf	Somatório do número de registros diários com UR ≥ 90 no PINF	Numérico	Decimal	h
tdur90_pinf	Média das temperaturas médias dos registros diários com UR ≥ 90 no PINF	Numérico	Decimal	° C
med_precip_pinf	Média das precipitações diárias no PINF	Numérico	Decimal	mm
smt_precip_pinf	Somatório das precipitações diárias no PINF	Numérico	Decimal	mm
dchuv_pinf	Número de dias que houve precipitação (≥ 1 mm) durante o PINF	Numérico	Inteiro	dias
ddi_pinf	Número de dias desfavoráveis à infecção no PINF	Numérico	Inteiro	dias
dfi_pinf	Número de dias favoráveis à infecção no PINF	Numérico	Inteiro	dias
taxa_inf_M3	Taxa de progresso da ferrugem – 3 p.p.	Booleano	0 ou 1	-
taxa_inf_M5	Taxa de progresso da ferrugem – 5 p.p.	Booleano	0 ou 1	-
taxa_inf_M10	Taxa de progresso da ferrugem – 10 p.p.	Booleano	0 ou 1	-

4.2.6 Descrição dos Cenários de Modelagem

Neste trabalho não foram utilizados dados meteorológicos e de incidência da doença referentes a todo o ano agrícola. O intervalo de estudo considerado como interessante para a questão da ferrugem tardia foi de dezembro a junho.

O conjunto de dados brutos (arquivo .txt e planilha .xls para cada mês avaliado) foi separado por carga pendente (alta ou baixa) dos cafeeiros. Havia para cada carga dados de 49, 97 e 179 meses para Boa Esperança, Carmo de Minas e Varginha, respectivamente, já considerando os meses descartados por falta de dados (seção 4.2.1.2). Inicialmente, todos os registros do conjunto de dados original foram transformados de horários para mensais (seção 4.2.5). Feito isso, foram selecionados os meses de interesse para estudo (dezembro a junho).

Em Boa Esperança foram selecionados os registros de março a junho de 2010 e de dezembro a junho para demais anos agrícolas, totalizando 28 registros. Analogamente, Carmo de Minas e Varginha tiveram 57 e 111 registros escolhidos. Entretanto, há valores de incidência para lavouras com espaçamento adensado e largo em Varginha, dobrando o número de registros para 222. Com a eliminação dos valores inconsistentes de incidência (seção 4.2.1.1), Varginha ficou com 215 registros. Assim, tanto o conjunto de treinamento parcial para baixa quanto para alta carga continham 28+57+215=300 registros.

Após a formação do conjunto de dados parcial (seção 4.2.5) contendo 300 registros, dividiu-se os dados entre registros para plantas com alta e baixa carga pendente devido à característica da ferrugem, que atinge de forma mais acentuada a lavoura em anos de alta carga pendente de frutos. Cada tipo de carga pendente ainda foi separado conforme um limiar do atributo meta e uma seleção de atributos de maneira subjetiva. Foram selecionados três (01, 02 e 03) conjuntos distintos de atributos (Tabela 13).

Tabela 13. Lista de atributos para cada conjunto selecionado.

Atributos	01	02	03
incidencia_anterior	*		
tmed_pinf	*	*	*
tmax_pinf	*	*	*
tmin_pinf	*	*	*
tmed_pi_pinf	*	*	*
tmax_pi_pinf	*	*	*
tmin_pi_pinf	*	*	*
ur_pinf	*	*	*
med_nhdur90_pinf	*	*	*
smt_nhdur90_pinf	*	*	*
tdur90_pinf	*	*	*
med_precip_pinf	*	*	*
smt_precip_pinf	*	*	*
dchuv_pinf	*		*
ddi_pinf	*		*
dfi_pinf	*		*
taxa_inf_Mx	*	*	*

taxa_inf_Mx - taxa de progresso com limiar x

No conjunto de atributos 01 estão presentes todos os atributos do conjunto de dados (Tabela 12). O conjunto 02 possui apenas os registros climáticos derivados dos dados brutos, sem os atributos gerados a partir da matriz de infecção diária (Tabela 11). O uso desses atributos foi para aprofundar o estudo da interferência do clima sobre a epidemia da ferrugem a partir de variáveis de obtenção simples. A diferença do conjunto de atributos 03 para 01 é a ausência do atributo “incidencia_anterior”.

A formação dos três conjuntos de atributos ocorreu para que fosse possível analisar a influência da incidência da ferrugem no mês anterior ao que se quer predizer e dos atributos derivados a partir da matriz de infeção diária no desempenho preditivo dos *ensembles*, além de ser uma forma subjetiva para selecionar atributos contidos no conjunto de dados. Assim, doze cenários distintos (Tabela 14) foram usados na modelagem (seção 4.3).

Tabela 14. Cenários de modelagem.

Carga Pendente	Limiar do Atributo Meta (p.p.)	Conjunto de Atributos
Alta	5	01 02 03
Alta	10	01 02 03
Baixa	3	01 02 03
Baixa	5	01 02 03

4.3 Modelagem

Os modelos foram criados utilizando os conjuntos de treinamento preparados na fase anterior utilizando algoritmos de aprendizado de máquina (seção 3.4.2). As técnicas usadas como base para os *ensembles* foram: árvore de decisão e *SVM*. *Ensembles* foram desenvolvidos por meio das técnicas: *bagging*, *boosting* (algoritmo *AdaBoost*), floresta aleatória e *stacking*. A modelagem foi realizada para cada cenário (seção 4.2.6) e por meio de *scripts* em Python utilizando, principalmente, algoritmos implementados no módulo scikit-learn.

Inicialmente, foi realizado uma busca em *grid* (Apêndice A) para identificar os melhores valores de cada parâmetro para as combinações entre árvore de decisão e *SVM* com *bagging* e *boosting*, além da floresta aleatória.

Nos ensembles baseados em *SVM* os dados foram normalizados antes da busca em *grid* e geração de modelos. Foram testados quatro tipos de kernel: linear, polinomial, *rbf* (base radial) e sigmóide.

A busca em *grid*, para cada cenário e técnica, funcionou da seguinte maneira: para cada combinação possível entre os parâmetros testados, avaliou-se a acurácia do modelo por meio da validação cruzada estratificada (*stratified cross-validation*) em 10 intervalos (k=10). Foram usados 10 partições ($k=10$) pois costuma produzir modelos com melhor desempenho (KOHAVI, 1995).

Após a verificação para todos os casos, selecionou-se os onze modelos de melhor acurácia. Os parâmetros que resultaram nesses melhores modelos foram aplicados em todo conjunto de treinamento para gerar os modelos definitivos e salvá-los em arquivo de extensão .pkl para que possam ser reutilizados. Também foi realizada uma avaliação mais

detalhada dos onze modelos por meio da validação cruzada estratificada (k=10) para verificar demais métricas de desempenho do modelo e gerar gráficos *ROC* e histogramas.

Procurando melhorar o desempenho preditivo em relação às demais técnicas, os modelos do nível-0 no *stacking* foram compostos pelos demais *ensembles* gerados (*bagging*, *boosting* e floresta aleatória), que totalizaram 121 modelos para cada cenário de modelagem. A partir desses modelos foi gerado um novo conjunto de dados utilizando suas predições em relação à probabilidade de um registro avaliado ser de determinada classe. Para criar o meta-classificador no *stacking* foram usadas as técnicas: perceptron, passivo-agressivo, ridge, regressão logística e *SVM* (kernel: linear e polinomial).

Foram criados dois cenários para o *stacking*, primeiro (tudo) continha todos os *ensembles* no nível-0 e o segundo (simples) teve apenas os modelos em floresta aleatória, *boosting* e *bagging* com árvore de decisão e *boosting* e *baggging* com *SVM* (kernel: polinomial), totalizando 55 modelos. Os modelos, no segundo cenário, foram selecionados por apresentarem os melhores desempenhos preditivos. A formação do segundo cenário ocorreu uma vez que foi constatado que alguns modelos de *boosting* e *bagging* com *SVM* (kernel: linear, rbf e sigmóide) classificavam todos os registros em apenas uma classe. Ao fim foram desenvolvidos, para cada cenário (seção 4.2.6), 253 *ensembles* (Tabela 15).

Desse modo foram gerados 759 *ensembles* (3 cenários x 253 modelos) para uma das quatro combinações carga pendente e limiar da TP (alta/5 p.p., alta/10 p.p., baixa/3 p.p. e baixa/5 p.p.), totalizando 3036 *ensembles*.

Tabela 15. Técnicas usadas na criação dos *ensembles*.

Algoritmo	**Kernel**	**Meta-classifcador**	**Quantidade de *ensembles***
Floresta Aleatória	-	-	11
Boosting + Árvore de Decisão	-	-	11
Boosting + *SVM*	linear	-	11
Boosting + *SVM*	polinomial	-	11
Boosting + *SVM*	rbf	-	11
Boosting + *SVM*	sigmóide	-	11
Bagging + Árvore de Decisão	-	-	11
Bagging + *SVM*	linear	-	11
Bagging + *SVM*	polinomial	-	11
Bagging + *SVM*	rbf	-	11
Bagging + *SVM*	sigmóide	-	11
Stacking (tudo)	-	regressão logística	11
Stacking (tudo)	-	passivo-agressivo	11
Stacking (tudo)	-	perceptron	11
Stacking (tudo)	-	ridge	11
Stacking (tudo)	linear	svm	11
Stacking (tudo)	polinomial	svm	11
Stacking (simples)	-	regressão logística	11
Stacking (simples)	-	passivo-agressivo	11
Stacking (simples)	-	perceptron	11
Stacking (simples)	-	ridge	11
Stacking (simples)	linear	svm	11
Stacking (simples)	polinomial	svm	11
		total de *ensembles*	253

4.4 Avaliação de *ensembles* quanto a detecção de períodos críticos para epidemia da ferrugem do cafeeiro

Os *ensembles* gerados foram avaliados quanto sua capacidade de predizer momentos críticos para epidemia da ferrugem. Neste estudo foram determinados como períodos críticos: primeiro mês de aumento da taxa de progresso mensal (TP) da ferrugem, ocorrência da ferrugem tardia e meses de maior aumento da TP.

A intenção de avaliar esses meses foi para verificar, de modo mais detalhado, o comportamento dos *ensembles* perante meses importantes para o progresso da ferrugem entre dezembro e junho.

Foi considerado como primeiro aumento da TP da ferrugem o primeiro mês entre dezembro e junho em que a TP superou ou igualou um limiar, definido conforme a carga pendente de frutos. Um exemplo para a lavoura de Carmo de Minas na safra 2009/10 – alta carga pendente e 5 p.p. – se encontra na Tabela 16. A partir dessa tabela verifica-se que o primeiro mês com TP $\geq$ 5 p.p. foi janeiro de 2010.

Tabela 16. Valores de TP da ferrugem para safra 2009/10 em Carmo de Minas nas lavouras em alta carga pendente.

Mês	Ano	TP
DEZ	2009	2
JAN	2010	11,5
FEV	2010	5,5
MAR	2010	1,5
ABR	2010	-3,5
MAI	2010	26
JUN	2010	29,5

A Figura 18, que mostra a curva de progresso da ferrugem em cafeeiros com alta carga pendente em Carmo de Minas na safra 2009/10, auxiliou na análise visual sobre o primeiro mês de aumento na TP.

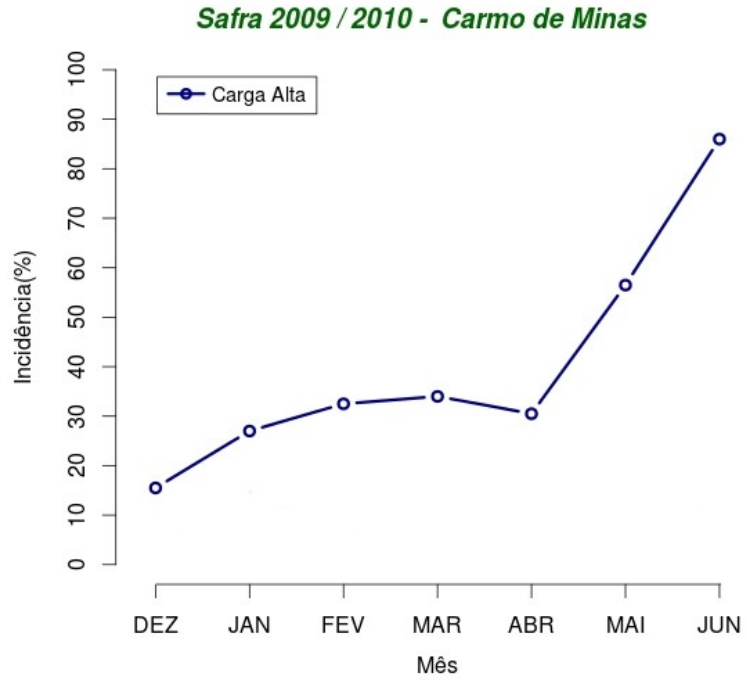

Figura 18. Curva de progresso da ferrugem nas plantas em alta carga de Carmo de Minas para safra 2009/10.

Para considerar que o *ensemble* detectou o primeiro mês de aumento na TP, este deve ter saída igual a *1* apenas no mês correto. Voltando ao exemplo de Carmo de Minas (safra 2009/10), o primeiro mês de aumento foi janeiro de 2010. Para que o *ensemble* detecte o primeiro mês, este deve apresentar como saída: *0* em dezembro de 2009 e *1* para janeiro de 2010. Qualquer outra combinação de saídas configura erro na predição do *ensemble* nesse quesito.

Uma vez que o controle da doença ocorre, principalmente, por aplicações de fungicidas, caso o *ensemble* consiga predizer corretamente o primeiro mês em que o atributo meta será *1*, este poderá ser uma ferramenta auxiliar na tomada de decisão em relação ao posicionamento adequado das primeiras medidas de controle entre dezembro e junho.

No total, o conjunto de dados final era composto por registros referentes a 29 anos agrícolas (5 para Boa Esperança, 8 para Carmo de Minas e 16 para Varginha) e, portanto, foram identificados 29 primeiros meses de aumento da TP.

Também verificou-se a eficácia dos *ensembles* na identificação do desenvolvimento tardio da ferrugem. Sua ocorrência foi definida como sendo os anos cujo primeiro mês em que o valor da TP foi maior ou igual ao limiar aconteceu de março em diante. A escolha por março foi baseada em Chalfoun e Carvalho (1999). Os autores constataram que esse foi o mês inicial do período de maior desenvolvimento da ferrugem onde não foi possível controlá-la seguindo um calendário fixo de aplicações de fungicida cúprico.

Um exemplo de ocorrência da ferrugem tardia é a safra 2011/12 de Boa Esperança em lavouras com alta carga pendente (Figura 19). Nestes dados o aumento da TP em março foi superior a 5 p.p. e, assim, foi considerada nesta safra que ocorreu um desenvolvimento tardio da ferrugem.

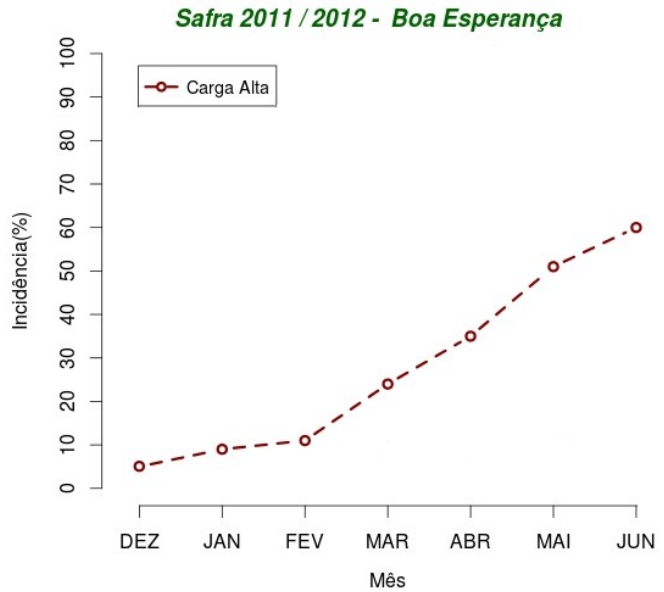

Figura 19. Ocorrência de ferrugem tardia nos cafeeiros com alta carga em Boa Esperança na safra 2011/12.

A Tabela 17 apresenta os valores da TP para lavouras em alta carga de Boa Esperança durante a safra 2011/12. Observa-se que a TP ultrapassou 5 p.p. pela primeira vez em março de 2012. A questão da ocorrência do desenvolvimento tardio da ferrugem foi analisada apenas nos dados de cafeeiros em alta carga/5 p.p. e baixa carga/3 p.p.

Tabela 17. Valores de TP da ferrugem para safra 2011/12 em Boa Esperança nas lavouras em alta carga pendente.

Mês	Ano	TP
DEZ	2011	2
JAN	2012	4
FEV	2012	2
MAR	2012	13
ABR	2012	11
MAI	2012	16
JUN	2012	9

Ensembles foram analisados em relação aos períodos de maior desenvolvimento da TP a fim de verificar seus desempenhos preditivos quanto ao rápido crescimento da epidemia que ocorre, geralmente, de março a abril (ZAMBOLIM et al., 2005). Para isso os meses em que a TP igualou ou superou 12 e 6 p.p. para lavouras em alta e baixa carga, respectivamente. Esses valores foram selecionados, em caráter exploratório, por serem 20% acima do maior limiar de cada carga (10 p.p. para alta e 5 p.p. para baixa carga) e por análise visual dos gráficos de curva de progresso da ferrugem.

A análise visual indicou que as inclinações mais acentuadas na curva ocorriam quando a TP era maior ou igual a 12 p.p. e 6 p.p nas cargas alta e baixa, respectivamente. Assim foi possível identificar em quais meses ocorreu os maiores aumentos na TP da ferrugem e verificar se os *ensembles* eram capazes de predizer tais meses.

Exemplificando a análise visual realizada, a curva de progresso da ferrugem na safra 2002/03 em Varginha para lavouras em alta carga pendente (Figura 20) mostra que entre março e maio de 2003 a curva de progresso foi mais inclinada. Valores de TP superiores a 12 p.p., em alta carga, e 6 p.p., para baixa carga, coincidiam com os períodos em que a inclinação da curva de progresso era mais acentuada.

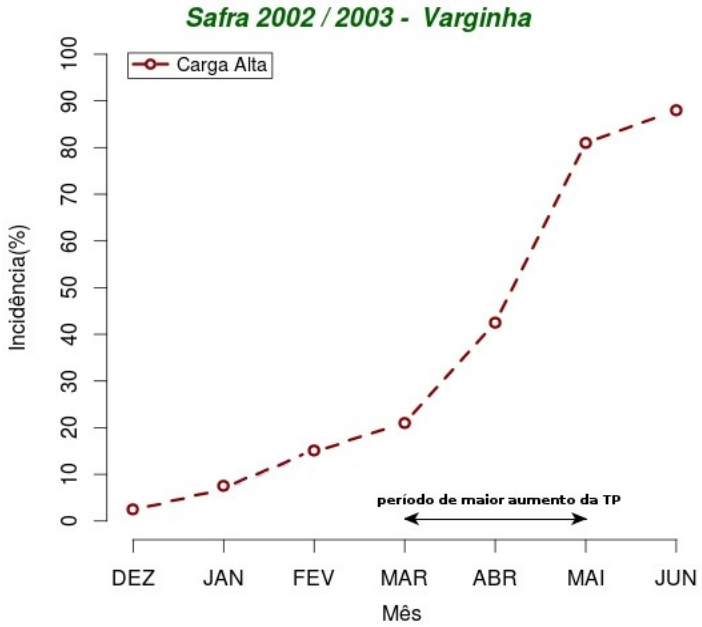

Figura 20. Curva de progresso da ferrugem em cafeeiros de Varginha com alta carga durante a safra 2002/03.

Ao analisar os valores de TP (Tabela 18) constatou-se que nos meses de abril e maio de 2013 a TP foi superior a 12 p.p., logo esses meses foram considerados como de maior aumento da TP na safra 2002/03 de Varginha.

Tabela 18. Valores de TP da ferrugem para safra 2002/03 em Varginha nas lavouras em alta carga pendente.

mes	ano	TP
DEZ	2002	2
JAN	2003	4,5
FEV	2003	8
MAR	2003	6
ABR	2003	21,5
MAI	2003	38,5
JUN	2003	7

A verificação da eficácia dos *ensembles* em identificar esses períodos críticos para epidemia da ferrugem foi realizada por meio da validação cruzada estratificada (k=10). Os meses que fizeram parte do período crítico para epidemia da doença variaram de acordo com a carga pendente e o limiar da TP. Detalhes em relação aos meses do período crítico estão no Apêndice B (primeiro mês e ferrugem tardia) e C (maiores aumentos da TP).

4.5 Avaliação de modelos de Girolamo Neto (2013)

Os modelos de Girolamo Neto (2013) para ferrugem do cafeeiro foram avaliados por meio do Weka. Ao conjunto de dados brutos, preparados conforme a seção 4.2, foi aplicado o mesmo processo de preparação de dados realizado por Girolamo Neto (2013) uma vez que a preparação dos dados realizada pelo autor difere da empregada nesta dissertação.

Assim, o conjunto de dados final conteve 300 registros e foi separado em três cenários conforme carga pendente e limiar do atributo meta: alta/5 p.p., alta/10 p.p. e baixa/5 p.p.. Para cada uma das três estações, Boa Esperança (BE), Carmo de Minas (CM) e Varginha (VG), foram selecionados os três melhores modelos desenvolvidos por Girolamo Neto (2013) em cada cenário, totalizando 27 modelos. A avaliação dos modelos foi realizada por meio da validação cruzada estratificada (k=10).

4.6 Configurações de Software

Todos os *scripts* na linguagem de programação R (versão 3.1.1) utilizados foram desenvolvidos no ambiente de desenvolvimento integrado (*IDE,* do inglês *Integrated Development Environment*) RStudio (versão 0.98.507). Também foram utilizadas funções presentes nos pacotes:

- plyr (versão 1.8.1): pacote com implementações úteis para manipulação de dados como, por exemplo, em forma de matrizes (data.frames);
- stringr (0.6.2): possui funções para manipulação de *strings*.

Para os *scripts* na linguagem Python (versão 3.4) utilizou-se como *IDE* o Spyder (versão 2.3.4). Além dos módulos padrões da linguagem, foram usados os seguintes módulos:

- NumPy (versão 1.8.2): módulo para computação científica, permite operações em *arrays* e matrizes multidimensionais;
- SciPy (versão 0.13.3): trabalha com matrizes, fornece rotinas numéricas e possui módulos para otimização, álgebra linear, integração, dentre outras;
- pandas (versão 0.13.1): fornece suporte para estruturas de dados e ferramentas de análise de dados;
- matplotlib (versão 1.3.1): módulo para visualização de dados, como criação de gráficos e histogramas;

- scikit-learn (versão 0.16.1): módulo para aprendizado de máquina, possui implementações de algoritmos para execução de processos de mineração de dados.

Criação de planilhas e processamento simples foram realizados no LibreOffice Calc (versão 4.2.8.2). Uma das ferramentas utilizadas para a discretização de dados de incidência mensal da ferrugem em cafeeiros com baixa carga pendente foi o Weka (versão 3.7.10) que contém uma coleção de algoritmos de aprendizado de máquina para tarefas de mineração de dados (HALL et al., 2009). RStudio, Spyder, Calc e Weka são *softwares* gratuitos e livres (pode-se adaptar ou modificar seu código fonte sem ter a necessidade de solicitar permissão ao seu proprietário). O sistema operacional utilizado foi o Ubuntu (versão 14.04 LTS – 64-bits).

5 RESULTADOS E DISCUSSÃO

Este capítulo mostra o desempenho preditivo dos *ensembles* desenvolvidos conforme os cenários apresentados na seção 4.2.6. Os *ensembles* foram avaliados, utilizando todos os registros do conjunto de dados, por meio das métricas relatadas anteriormente (seção 3.4.4). Também foram avaliadas as capacidades dos *ensembles* em predizer os períodos críticos para epidemia da doença (seção 4.4). Essa avaliação contou somente com dados relacionados aos períodos críticos.

Seção 5.1 mostra e discute os resultados obtidos na avaliação dos modelos de Girolamo Neto (2013). As seções 5.2, 5.3, 5.4 e 5.5 apresentam os desempenhos dos *ensembles.* As seções possuem gráfico ROC com destaque para os modelos presentes no envelope convexo. Uma vez que para cada cenário de modelagem foram criados 253 *ensembles* distintos (seção 4.3), um gráfico ROC com dados de todos os modelos é de difícil visualização. Sendo assim, foram selecionados os dez melhores *ensembles*, em termos de acurácia, baseados em cada conjunto de atributos (01, 02 e 03), totalizando 30 *ensembles*.

Além dessses modelos, presentes no gráfico ROC, houve em alguns cenários casos em que outros *ensembles* apresentaram performance superior em relação aos modelos padrão sobre os períodos críticos.

A seção 5.6 traz uma compilação de informações obtidas por meio dos resultados das seções anteriores.

5.1 Desempenho dos modelos de Girolamo Neto (2013)

Os desempenhos dos modelos de Girolamo Neto (2013), considerando toda a base de dados (300 registros), estão dispostos nas tabelas 19 a 21, onde TFP é a taxa de falsos positivos.

Tabela 19. Resultado da avaliação dos modelos para alta carga e 5 p.p. de Girolamo Neto (2013).

Modelo	Técnica	Acurácia (%)	Sensitividade (%)	TFP (%)	Precisão (%)	Especificidade (%)	*AUC*
BE-1	Árvore de Decisão	69,00%	69,00%	66,00%	61,70%	34,00%	0,516
BE-2	Floresta Aleatória	73,33%	73,30%	47,60%	71,30%	52,40%	0,705
BE-3	Rede Neural	72,33%	72,30%	60,50%	68,70%	39,50%	0,602
CM-1	Árvore de Decisão	71,66%	71,70%	68,40%	67,80%	31,60%	0,523
CM-2	Floresta Aleatória	72,33%	72,30%	52,20%	69,60%	47,80%	0,692
CM-3	Floresta Aleatória	70,00%	70,00%	55,90%	66,50%	44,10%	0,682
VG-1	Árvore de Decisão	71,66%	71,70%	68,40%	67,80%	31,60%	0,515
VG-2	Floresta Aleatória	71,33%	71,30%	53,30%	68,40%	46,70%	0,699
VG-3	*SVM*	75,33%	75,30%	50,90%	73,40%	49,10%	0,622

Considerando a acurácia, o melhor modelo para alta carga e 5 p.p. (Tabela 19) foi o terceiro modelo de Varginha ("VG-3") com 75,33%, uma vantagem de 2 p.p. em relação ao segundo melhor modelo, de Boa Esperança ("BE-2"). "VG-3" também apresenta os maiores valores para sensitividade (75,30%) e precisão (73,40%). Os modelos "CM" se referem a Carmo de Minas.

Dentre os modelos para alta/10 p.p. (Tabela 20) destaca-se "VG-6", que apresenta os maiores valores para acurácia (70,66%), sensitividade (70,70%), precisão (70,60%) e especificidade (69,90%). Nota-se que tanto "VG-3" quanto "VG-6" são modelos baseados em *SVM*, que foi a técnica com melhor desempenho preditivo, considerando a validação cruzada estratificada e o conjunto de dados brutos corrigidos neste estudo (seções 4.1 e 4.2), para os dados referentes a lavouras com alta carga pendente.

Tabela 20. Resultado da avaliação dos modelos para alta carga e 10 p.p. de Girolamo Neto (2013).

Modelo	Técnica	Acurácia (%)	Sensitividade (%)	TFP (%)	Precisão (%)	Especificidade (%)	*AUC*
BE-4	Floresta Aleatória	70,00%	70,00%	30,70%	70,00%	69,30%	0,736
BE-5	Floresta Aleatória	68,66%	68,70%	32,30%	68,70%	67,70%	0,722
BE-6	Rede Neural	58,33%	58,30%	43,80%	58,20%	56,20%	0,593
CM-4	Árvore de Decisão	55,67%	55,70%	48,50%	55,90%	51,50%	0,515
CM-5	Floresta Aleatória	70,00%	70,00%	30,70%	70,00%	69,30%	0,736
CM-6	Floresta Aleatória	69,00%	69,00%	31,90%	69,00%	68,10%	0,736
VG-4	Árvore de Decisão	57,66%	57,70%	45,00%	57,60%	55,00%	0,566
VG-5	Floresta Aleatória	70,00%	70,00%	30,70%	70,00%	69,30%	0,736
VG-6	*SVM*	70,66%	70,70%	30,10%	70,60%	69,90%	0,703

Nos modelos de baixa carga (Tabela 21), "VG-7" apresentou maior acurácia (64,33%), sensitividade (64,30%), especificidade (48,60%) e segundo maior valor para precisão (a 0,1 p.p. do melhor modelo nessa métrica). Entretanto, "VG-7" não demonstrou alto valor para *AUC*, sendo o antepenúltimo nesse quesito. O modelo é baseado em floresta aleatória.

Tabela 21. Resultado da avaliação dos modelos para baixa carga e 5 p.p. de Girolamo Neto (2013).

Modelo	Técnica	Acurácia (%)	Sensitividade (%)	TFP (%)	Precisão (%)	Especificidade (%)	*AUC*
BE-7	Floresta Aleatória	61,33%	61,30%	54,80%	58,20%	45,20%	0,582
BE-8	Floresta Aleatória	61,33%	61,30%	54,80%	58,20%	45,20%	0,582
BE-9	Floresta Aleatória	64,00%	64,00%	53,40%	60,80%	46,60%	0,584
CM-7	Floresta Aleatória	60,66%	60,70%	63,30%	52,40%	36,70%	0,490
CM-8	Floresta Aleatória	61,33%	61,30%	54,80%	58,20%	45,20%	0,582
CM-9	*SVM*	64,00%	64,00%	52,00%	61,20%	48,00%	0,560
VG-7	Floresta Aleatória	64,33%	64,30%	51,40%	61,70%	48,60%	0,555
VG-8	Floresta Aleatória	61,33%	61,30%	54,80%	58,20%	45,20%	0,582
VG-9	*SVM*	63,00%	63,00%	53,50%	59,90%	46,50%	0,548

As Tabelas 22 a 24 mostram a performance dos modelos de Girolamo Neto (2013) em relação à detecção dos períodos críticos para epidemia da ferrugem (seção 4.4).

No cenário alta/5 p.p. (Tabela 22), "VG-3" não apresentou as melhores performances na detecção do primeiro mês de taxa de progresso (TP) da ferrugem maior ou igual a 5 p.p. (51,72%) e dos meses de grandes aumentos na TP (87,67%). Entretanto foi o único modelo capaz de identificar a ferrugem tardia em dois anos agrícolas distintos (40%).

Tabela 22. Desempenhos dos modelos de Girolamo Neto (2013) nos períodos críticos para epidemia (alta/5 p.p.).

Modelo	Técnica	Primeiro mês			Ferrugem tardia			Maior aumento De TP		
		Total	Acertos	Acertos (%)	Total	Acertos	Acertos (%)	Total	Acertos	Acertos (%)
BE-1	Árvore de Decisão	29	16	55,17	5	1	20,00	73	68	93,15
BE-2	Floresta Aleatória	29	14	48,28	5	1	20,00	73	59	80,82
BE-3	Rede Neural	29	14	48,28	5	1	20,00	73	65	89,04
CM-1	Árvore de Decisão	29	15	51,72	5	1	20,00	73	70	95,89
CM-2	Floresta Aleatória	29	15	51,72	5	1	20,00	73	65	89,04
CM-3	Floresta Aleatória	29	13	44,83	5	1	20,00	73	58	79,45
VG-1	Árvore de Decisão	29	15	51,72	5	1	20,00	73	70	95,89
VG-2	Floresta Aleatória	29	12	41,38	5	1	20,00	73	57	78,08
VG-3	*SVM*	29	15	51,72	5	2	40,00	73	64	87,67

O modelo "VG-6", de melhor desempenho geral (Tabela 20) para cenário de alta carga/10 p.p., apresentou baixa performance quanto aos períodos críticos (Tabela 23) predizendo corretamente 17,24% do primeiro mês de aumento na TP e 36,99% dos maiores aumentos. Isso inviabiliza seu uso para as questões relacionadas aos períodos críticos para doença. Para os modelos de alta/10 p.p. o destaque foi "BE-5" com 31,03% de predição correta do primeiro mês e 46,58% de detecção de grandes aumentos na incidência da doença.

Tabela 23. Desempenhos dos modelos de Girolamo Neto (2013) nos períodos críticos para epidemia (alta/10 p.p.).

Modelo	Técnica	Primeiro mês			Maior aumento De TP		
		Total	Acertos	Acertos (%)	Total	Acertos	Acertos (%)
BE-4	Floresta Aleatória	29	5	17,24	73	36	49,32
BE-5	Floresta Aleatória	29	9	31,03	73	34	46,58
BE-6	Rede Neural	29	5	17,24	73	28	38,36
CM-4	Árvore de Decisão	29	1	3,45	73	13	17,81
CM-5	Floresta Aleatória	29	5	17,24	73	36	49,32
CM-6	Floresta Aleatória	29	5	17,24	73	33	45,21
VG-4	Árvore de Decisão	29	2	6,90	73	22	30,14
VG-5	Floresta Aleatória	29	5	17,24	73	36	49,32
VG-6	*SVM*	29	5	17,24	73	27	36,99

Dentre os modelos para baixa/5 p.p. (Tabela 24), "VG-7" obteve melhor desempenho na identificação dos meses com maior aumento na TP da doença (33,33%). Por também possuir melhor desempenho preditivo geral (Tabela 21), "VG-7" foi o melhor modelo para baixa carga e 5 p.p..

Tabela 24. Desempenhos dos modelos de Girolamo Neto (2013) nos períodos críticos para epidemia (baixa/5 p.p.).

Modelo	Técnica	Início da epidemia			Grande aumento De incidência		
		Total	Acertos	Acertos (%)	Total	Acertos	Acertos (%)
BE-7	Floresta Aleatória	29	11	37,93	45	12	26,67
BE-8	Floresta Aleatória	29	11	37,93	45	12	26,67
BE-9	Floresta Aleatória	29	11	37,93	45	12	26,67
CM-7	Floresta Aleatória	29	4	13,79	45	4	8,89
CM-8	Floresta Aleatória	29	11	37,93	45	12	26,67
CM-9	*SVM*	29	13	44,83	45	14	31,11
VG-7	Floresta Aleatória	29	12	41,38	45	15	33,33
VG-8	Floresta Aleatória	29	11	37,93	45	12	26,67
VG-9	*SVM*	29	12	41,38	45	13	28,89

5.2 *Ensembles* para alta carga pendente e 5 p.p.

Esta seção trata das lavouras com alta carga e 5 p.p. como limiar para a taxa de progresso mensal (TP) da ferrugem. Foram encontradas cinco ocorrências de ferrugem tardia nestes dados – quatro em Boa Esperança (BE) e uma em Varginha (VG) (Apêndice B.1). Meses de maior aumento na TP (≥ 12 p.p.) totalizaram 73, juntando as três estações (Apêndice C).

A Figura 21 mostra o gráfico ROC contendo os melhores *ensembles*, em termos de acurácia, desenvolvidos para cada cenário. Os modelos de "1" a "10" foram gerados com o conjunto de atributos 01, de "11" a "20" com atributos 02 e de "21" a "30" com atributos 03. Esses trinta *ensembles* foram denominados de modelos padrão e os detalhes dos valores de seus parâmetros se encontram no Apêndice A.5.

No gráfico ROC ainda estão representados os modelos "31" a "34" e "35" a "38", de acurácia superior aos *ensembles* "1" a "10" e "11" a "20", respectivamente, na predição dos períodos críticos. Os modelos baseados nos atributos 03 ("21" a "30") já possuem as melhores acurácias em termos de período crítico. Dessa forma, o número de *ensembles*

avaliados para alta/5 p.p. foi 38. Maiores detalhes sobre os parâmetros dos modelos “31” a “38” estão no Apêndice A.6.

Pelo gráfico ROC é possível verificar que os ensembles “1”, “2”, “4”, “5”, “6” e “10” fazem parte do envelope convexo (curva). O gráfico ROC não mostra os 38 pontos referentes a todos os modelos avaliados pois em alguns casos os valores de sensitividade e 1-especificidade (taxa de falsos positivos) de dois ou mais modelos foram os mesmos. Por exemplo, na legenda da Figura 21 mostra 3 = 3, 7, 8, 9 e 33, o que significa que os modelos “3”, “7”, “8”, “9” e “33” apresentaram mesmos valores para sensitividade e 1-especificidade e no gráfico ROC esses *ensembles* estão representados apenas pelo ponto 3.

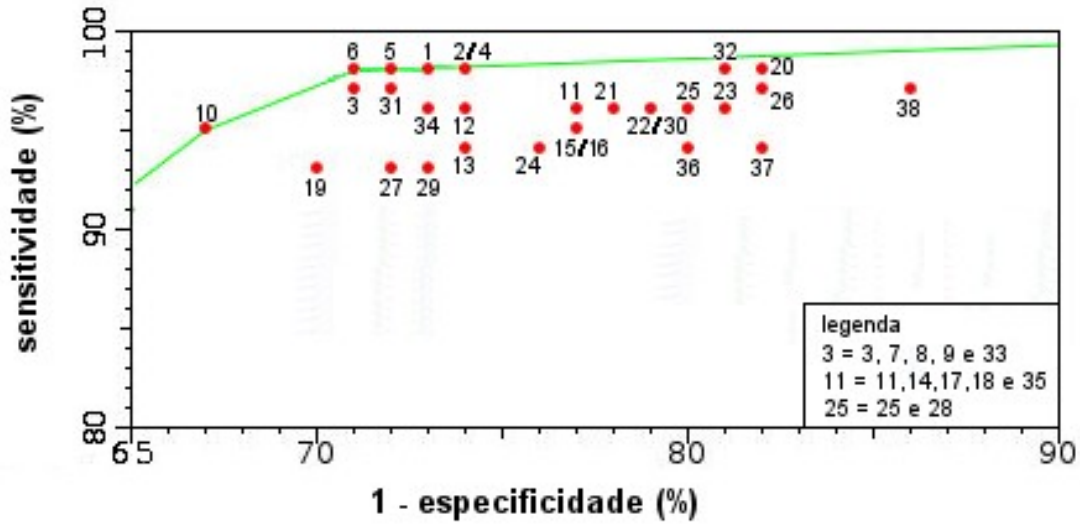

Figura 21. Gráfico ROC para limiar de 5 p.p. e alta carga.

A Tabela 25 mostra os desempenhos dos *ensembles* pertencentes ao envelope convexo quando avaliados com todo conjunto de dados. As performances preditivas dos *ensembles* foram próximas para acurácia e *AUC*. Destacam-se também os altos valores de sensitividade, acima de 95% para todos os *ensembles*. Em relação à técnica usada em seus desenvolvimentos, os modelos selecionados no envelope convexo são baseados em árvore de decisão. Isso pode indicar uma maior facilidade de aprendizado desse tipo de modelo.

Tabela 25. Desempenho preditivo dos *ensembles* selecionados no envelope convexo (alta/5 p.p.).

Ensemble	**Técnica**	**Acurácia (%)**	**Sensitividade (%)**	**TFP (%)**	**Precisão (%)**	**Especificidade (%)**	***AUC***
1	FA	77,96	98,13	72,77	77,57	27,23	0,713
2	FA	77,66	98,11	73,88	77,57	26,12	0,688
4	FA	77,62	97,18	71,38	77,72	28,62	0,684
5	*Boosting* + AD	77,93	97,64	71,67	77,76	28,33	0,684
6	*Boosting* + AD	78,00	97,64	71,38	78,02	28,62	0,671
10	*Bagging* + AD	77,68	95,36	66,94	78,63	33,06	0,712

TFP – taxa de falsos positivos / FA – floresta aleatória / AD – árvore de decisão

Todos os *ensembles* selecionados no envelope convexo foram criados utilizando o conjunto de atributos 01, cuja principal diferença para os demais conjuntos é a presença do valor da incidência da ferrugem no mês anterior (“incidencia_anterior”).

A importância de “incidencia_anterior” em “6”, melhor modelo em termos de acurácia (Tabela 25), é mostrada na Figura 22. Atributos derivados da umidade relativa do ar (UR) no período de infecção (PINF) – “ur_pinf” e “smt_nhdur90_pinf” – tiveram maior relevância em relação à temperatura média e máxima.

Dentre os atributos derivados da temperatura, “tmin_pinf” foi mais relevante, indicando maior importância da temperatura mínima no PINF para o progresso da ferrugem em relação às temperaturas média e máxima. O atributo menos relevante para “6” foi o número de dias desfavoráveis à infecção (“ddi_pinf”), derivado a partir da matriz de infecção diária (Tabela 11).

Uma vez que a altura das árvores foi limitada a 3, nem todos os atributos de 01 foram relevantes para o *ensemble* “6”. Nesse sentido, destaca-se a ausência do número de dias favoráveis à infecção (“dfi_pinf”), também criado a partir da matriz de infecção diária.

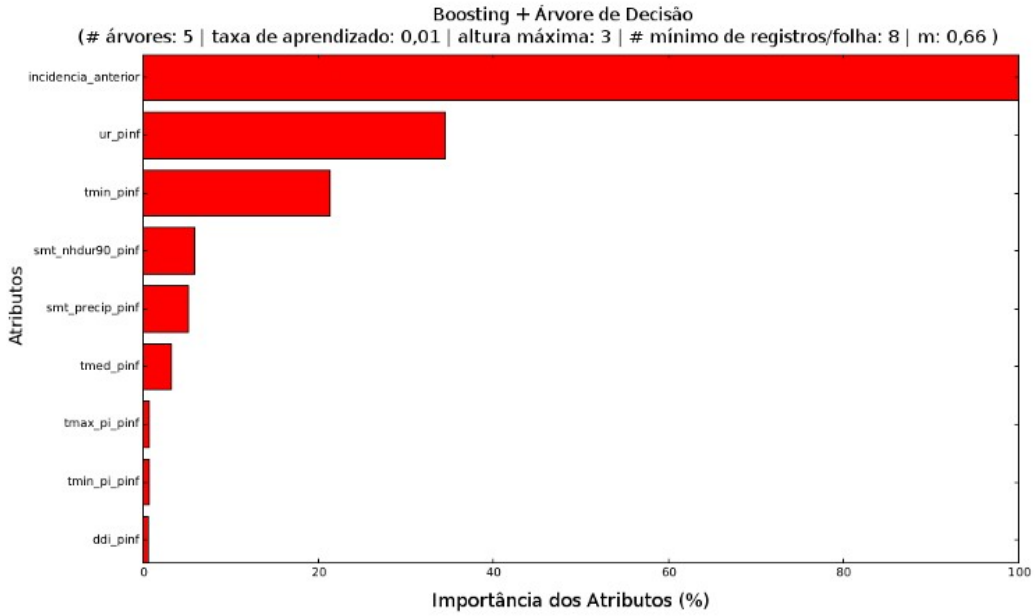

Figura 22. Importância dos atributos no *ensemble* “6” (alta/5 p.p.).

A avaliação dos *ensembles* do envelope convexo quanto à predição dos períodos críticos para epidemia da ferrugem está apresentada na Tabela 26.

Tabela 26. Detecção dos momentos críticos para epidemia da ferrugem dos *ensembles* contidos no envelope convexo (alta/5 p.p.).

Ensemble	**Técnica**	**Primeiro mês**			**Ferrugem tardia**			**Maior aumento De TP**		
		Total	**Acertos**	**Acertos (%)**	**Total**	**Acertos**	**Acertos (%)**	**Total**	**Acertos**	**Acertos (%)**
1	FA	29	13	44,83	5	2	40,00	73	72	98,63
2	FA	29	14	48,28	5	2	40,00	73	73	100,00
4	FA	29	14	48,28	5	2	40,00	73	71	97,26
5	*Boosting* + AD	29	13	44,83	5	2	40,00	73	72	98,63
6	*Boosting* + AD	29	15	51,72	5	2	40,00	73	71	97,26
10	*Bagging* + AD	29	15	51,72	5	2	40,00	73	72	98,63

FA – floresta aleatória / AD – árvore de decisão

Em relação aos momentos críticos existe, o desempenho dos modelos “31” a “38” estão demonstrados na Tabela 27.

Tabela 27. Desempenho preditivo dos melhores *ensembles* quanto à detecção dos períodos críticos para epidemia (alta/5 p.p.).

Cenário	***Ensemble***	**Técnica**	**Primeiro mês**			**Ferrugem tardia**			**Maior aumento De TP**		
			Total	**Acertos**	**Acertos (%)**	**Total**	**Acertos**	**Acertos (%)**	**Total**	**Acertos**	**Acertos (%)**
01	31	FA	29	17	58,62	5	2	40,00	73	73	100,00
01	32	*Boosting* + *SVM*(p)	29	17	58,62	5	2	40,00	73	72	98,63
01	33	Stack-t + Ridge	29	19	65,52	5	2	40,00	73	71	97,26
01	34	Stack-t + *SVM*(l)	29	19	65,52	5	2	40,00	73	72	98,63
02	35	FA	29	21	72,41	5	1	20,00	73	69	94,52
02	36	*Boosting* + AD	29	20	68,97	5	1	20,00	73	71	97,26
02	37	*Boosting* + *SVM*(l)	29	20	68,97	5	1	20,00	73	68	93,15
02	38	Stack-s + *SVM*(p)	29	23	79,31	5	2	40,00	73	70	95,89

FA – floresta aleatória / AD – árvore de decisão / *SVM*(l) – *SVM* com kernel linear / *SVM*(p) – *SVM* com kernel polinomial
Stack-t – *stacking* utilizando todos os *ensembles* / Stack-s – *stacking* simples / Ridge – classificador Ridge

Seleção de *ensembles*

A acurácia foi a principal métrica utilizada para avaliar e selecionar os *ensembles*. A partir desta sabe-se a taxa de acerto das predições. Sensitividade, que informa como o modelo lida com registros positivos (TP $\geq$ limiar), e especificidade também possuíram relevância para seleção dos modelos.

Para a questão da ferrugem do cafeeiro, uma predição correta de registros positivos significa que o *ensemble* alertará corretamente o usuário (qualquer pessoa com interesse na epidemia da ferrugem do cafeeiro) sobre os períodos de aumento na epidemia da doença. Assim, o *ensemble* pode auxiliar o usuário a tomar decisões relacionadas ao controle da ferrugem e, consequentemente, aplicar fungicidas em momentos oportunos, controlar eficientemente a epidemia da doença e, talvez, diminuir o número de aplicações em relação ao calendário fixo.

A especificidade indica como o modelo trabalha com registros negativos (TP < limiar), quanto maior seu valor, maior é a porcentagem de registros classificados corretamente. Isso concede ao modelo maior certeza de que não haverá aumentos significativos na TP quando este predizer *0*, evitando aplicações de fungicidas na lavoura em períodos irrelevantes para o desenvolvimento da ferrugem.

Uma predição incorreta de um registro negativo é o pior cenário possível. Nele, o modelo não prediz o aumento na TP da doença, o usuário pode não tomar medidas necessárias para controlar a doença e acaba ocorrendo aumento na epidemia. Uma vez que a ferrugem é capaz de provocar danos superiores a 50% na produção (ZAMBOLIM et al., 2005) e evoluir rapidamente em meses quentes (MORAES et al., 1976), perder uma oportunidade para aplicar agroquímicos na lavoura pode resultar em grandes perdas na produção.

Quanto mais próximos forem os valores de sensitividade e especificidade, melhor e mais equilibrado é o modelo (FAWCETT, 2006) uma vez que prediz corretamente os registros positivos e negativos com uma porcentagem similar.

A seleção dos melhores *ensembles* também considerou suas performances em relação aos períodos críticos (seção 4.4). Pode, por exemplo, ocorrer de um *ensemble* conseguir identificar, com maior precisão, apenas a ferrugem tardia em relação aos demais. Neste caso, poderiam ser utilizados diferentes *ensembles* em cada tarefa como, por exemplo: detectar o primeiro mês de aumento da TP ou identificar se, para a safra vigente, ocorrerá a ferrugem tardia.

Verifica-se nos modelos do envelope convexo (Tabela 25) a proximidade nos valores de acurácia dos *ensembles* “1”, “5” e “6” (diferença máxima de 0,07 p.p.), com vantagem para “6” (78,00%). Especificidade e precisão em “6” também foram maiores, além da melhor performance quanto ao período crítico para ferrugem (Tabela 26). O *boosting* “6” indicou corretamente o primeiro mês de aumento da TP em 15 anos (51,72%), ferrugem tardia em 2 anos (40%) e detectou 97,26% (71 dentre 73) dos meses em que a TP ≥ 12 p.p.

Com relação aos *ensembles* da Tabela 27, os modelos baseados nos atributos 02 apresentaram melhores performances quanto à detecção do primeiro mês. O modelo “38”, um *stacking* simples (seção 4.3), identificou corretamente 79,31% desses casos. Sobre a predição dos meses de maior aumento da TP destacaram-se “2” e “31”, com 100% de acerto.

Ferrugem tardia teve, no máximo, 40% das ocorrências identificadas. Considerando todos os 38 modelos, nenhum *ensemble* detectou corretamente os inícios da epidemia em março de 2012 e março de 2014 para BE e março de 2008 em VG. Nos dois

últimos casos (Figura 23), todos *ensembles* apontaram dezembro do ano anterior como o mês inicial para epidemia.

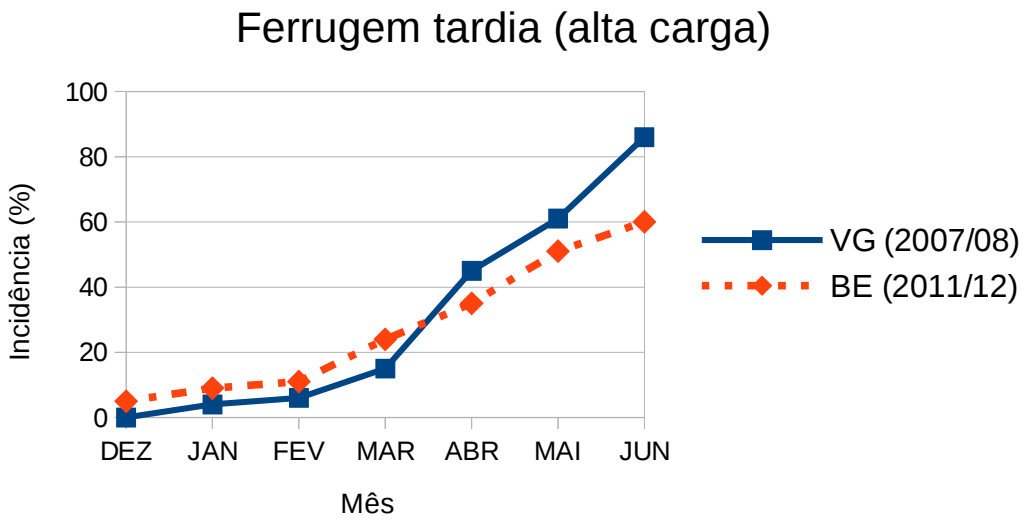

Figura 23. Ocorrência de ferrugem tardia em BE (2011/12) e VG (2007/08) em lavouras de alta carga pendente.

Para os casos de março de 2012 em BE e março de 2008 em VG, os *ensembles* predizeram dezembro como primeiro mês com TP ≥ 5 p.p. Para analisar o possível motivo para o erro dos modelos foram verificados os dados de dezembro/2011 em BE e dezembro/2007 em VG. Constatou-se que valores de temperatura média nos períodos de incubação (PI) e infecção (PINF) foram próximos do ideal para germinação do fungo, 21 a 23 °C (WALLER, 1982) (Tabela 28). Além disso, a temperatura se manteve dentro dos limites extremos para germinação, 15,5 e 28,5 °C (NUTMAN e ROBERTS, 1970). Isso pode ter contribuído para predição incorreta dos *ensembles* selecionados.

Tabela 28. Dados dos atributos referentes a temperatura em dezembro de 2011 para BE e 2007 para VG.

Mês	Ano	Estação	tmed_pinf	tmax_pinf	tmin_pinf	tmed_pi_pinf	tmax_pi_pinf	tmin_pi_pinf
DEZ	2011	BE	21.3	27.2	16.9	21.8	27.4	18
DEZ	2007	VG	20,8	27,6	16,9	21,3	27,9	17

Na safra de 2013/14 em BE os *ensembles* também predizeram, erroneamente, aumento na TP em dezembro de 2013. Analisou-se dados de dezembro em safras anteriores e foi constatada similaridade nos valores de algumas variáveis de temperatura e dias desfavoráveis à epidemia (“ddi_pinf”) em dezembro de 2010 e 2013 (Tabela 29). Na safra 2010/11 o primeiro mês de desenvolvimento da doença ocorreu em dezembro (“taxa_inf_M5”

= 1), predição que foi acertada por todos *ensembles*. Esse aprendizado anterior pode ter influenciado os modelos ao erro em dezembro/2013.

Tabela 29. Dados dos atributos em dezembro de 2010 e 2013 para Boa Esperança com valores semelhantes.

Mês	Ano	tmax_pinf	tmin_pinf	tmin_pi_pinf	tdur90_pinf	ddi_pinf	taxa_inf_M5
DEZ	2010	28,3	18,1	19,1	19,1	25	1
DEZ	2013	28,5	18,2	18,9	19,2	27	0

Para o cenário de alta/5 p.p não houve um *ensemble* que se destacou em todos os aspectos. Mas, conforme mencionado, é possível usar diferentes modelos para tarefas distintas. Caso seja necessário usar um *ensemble* para monitoramento mensal da doença, entre dezembro e junho, o modelo mais indicado é o "6" pelo equilíbrio nos valores de acurácia, sensitividade e especificidade.

A identificação da ferrugem tardia pode ser feita por meio de diversos modelos, que predizeram corretamente 40,00% dos casos (Tabela 27). A relevância dos atributos nos *ensembles* só foi possível de ser verificada nos modelos baseados em árvore de decisão, devido às funções implementadas no scikit (seção 4.6). Assim, recomenda-se o uso de "31", que é uma floresta aleatória, e teve 100,00% de acurácia nos meses em que a TP foi maior ou igual a 12 p.p.

5.3 *Ensembles* para alta carga pendente e 10 p.p.

A Figura 24 mostra o gráfico ROC para alta carga pendente e 10 p.p. como limiar para a TP. Além dos 30 modelos padrão, no gráfico foram plotados dados referentes a três ("31", "32" e "33"), dois ("34" e "35") e um ("36") *ensembles*, provenientes dos conjuntos de atributos 01, 02 e 03, respectivamente. Esses modelos apresentaram performance superior aos modelos padrão quanto aos períodos críticos. Os *ensembles* "1", "7", "10", "26", "29" e "31" fizeram parte do envelope convexo (curva).

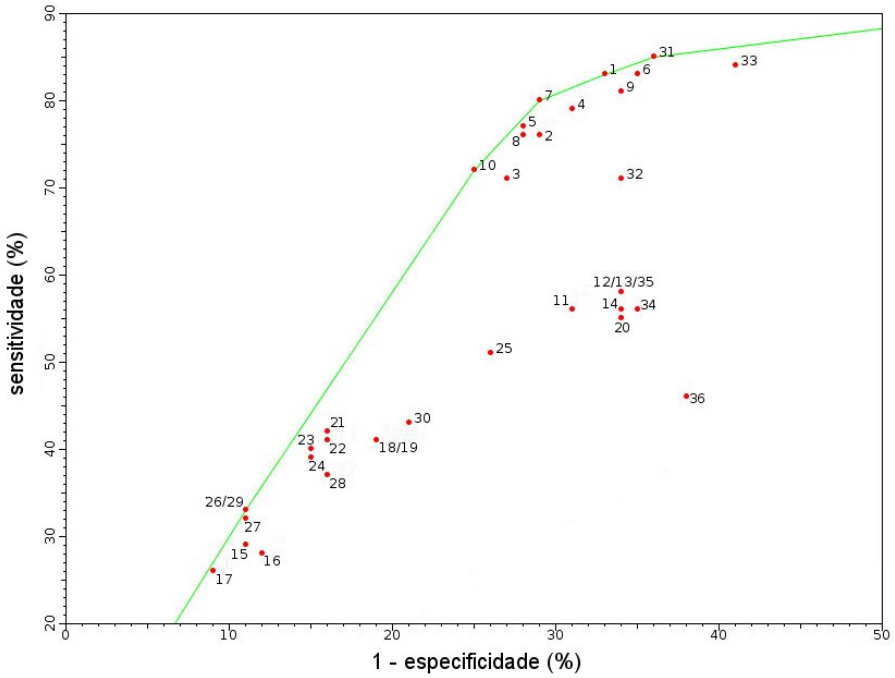

Figura 24. Gráfico ROC para limiar de 10 p.p. e alta carga.

A Tabela 30 contém os valores de desempenho preditivo somente dos *ensembles* presentes no envelope convexo. Na Tabela 31 estão apresentados os desempenhos dos melhores modelos quanto aos períodos críticos para epidemia.

Tabela 30. Desempenho preditivo dos *ensembles* selecionados no envelope convexo (alta/10 p.p.).

Ensemble	**Técnica**	**Acurácia (%)**	**Sensitividade (%)**	**TFP (%)**	**Precisão (%)**	**Especificidade (%)**	***AUC***
1	*Bagging* + AD	74,33	82,85	33,12	70,72	66,88	0,735
7	FA	73,33	79,70	28,75	74,41	71,25	0,751
10	FA	72,33	71,53	25,12	70,07	74,88	0,758
26	Stack-t + *SVM*(p)	62,67	32,82	11,25	73,38	88,75	0,601
29	*Boosting* + *SVM*(p)	62,67	32,85	11,25	74,86	88,75	0,391
31	FA	71,67	84,53	36,11	67,12	63,89	0,764

TFP – taxa de falsos positivos / FA – floresta aleatória / AD – árvore de decisão / *SVM*(p) – *SVM* com kernel polinomial

Stack-t – *stacking* utilizando todos os *ensembles* / TFP – taxa de falsos positivos

Observa-se na Tabela 30 a presença de *ensembles* desenvolvidos a partir dos atributos de 03 (“26” e “29”), diferentemente do cenário alta/5 p.p. Maiores valores de sensitividade e *AUC* e performance quanto aos períodos críticos foram apresentados pelo modelo “31” (Tabela 31).

Tabela 31. Detecção dos momentos críticos para epidemia da ferrugem dos *ensembles* contidos no envelope convexo (alta/10 p.p.).

Ensemble	**Técnica**	**Primeiro mês**			**Maior aumento De TP**		
		Total	**Acertos**	**Acertos (%)**	**Total**	**Acertos**	**Acertos (%)**
1	*Bagging* + AD	29	7	24,14	73	54	73,97
7	FA	29	6	20,69	73	53	72,60
10	FA	29	6	20,69	73	46	63,01
26	Stack-t + *SVM*(p)	29	6	20,69	73	22	30,14
29	*Boosting* + *SVM*(p)	29	7	24,14	73	24	32,88
31	FA	29	10	34,48	73	60	82,19

AD – árvore de decisão / FA – floresta aleatória

SVM(p) – *SVM* com kernel polinomial / Stack-t – *stacking* utilizando todos os *ensembles*

A Tabela 32 mostra o desempenho dos modelos “32” a “36”.

Tabela 32. Desempenho preditivo dos melhores *ensembles* quanto à detecção dos períodos críticos para epidemia (alta/10 p.p.).

Cenário	***Ensemble***	**Técnica**	**Primeiro mês**			**Maior aumento De TP**		
			Total	**Acertos**	**Acertos (%)**	**Total**	**Acertos**	**Acertos (%)**
01	32	FA	29	10	34,48	73	53	72,60
01	33	*Boosting* + AD	29	10	34,48	73	59	80,82
02	34	*Boosting* + *SVM*(p)	29	13	44,83	73	44	60,27
02	35	Stack-t + *SVM*(p)	29	13	44,83	73	41	56,16
03	36	*Bagging* + AD	29	10	34,48	73	33	45,21

AD – árvore de decisão / FA – floresta aleatória

SVM(p) – *SVM* com kernel polinomial / Stack-t – *stacking* utilizando todos os *ensembles*

Seleção de *ensembles*

Analisando os modelos do envelope convexo (Tabela 30), os modelos baseados no conjunto de atributos 01 (“1”, “7”, “10” e “31”) apresentaram valores próximos de acurácia e *AUC,* com diferenças máximas de 2,66 p.p. e 0,029, respectivamente. Valores de sensitividade, precisão e especificidade nas florestas aleatórias “7” e “10” foram mais balanceados, mostrando ligeira vantagem a “7”.

Em “7” todos os atributos contidos em 01 possuíram relevância (Figura 25) e destacou-se “incidencia_anterior”, com maior relevância (100%). Em seguida veio a média do número de horas diárias com UR ≥ 90 no PINF (“med_nhdur90_pinf”) com, aproximadamente, 15%. A diferença na relevância desses dois primeiros atributos mostra o peso que teve “incidencia_anterior”.

Ao contrário do melhor modelo para alta/5 p.p. (“6”), os atributos derivados a partir da matriz de infecção diária (Tabela 11), “ddi_pinf” e “dfi_pinf”, foram importantes para o limiar 10 p.p. e “tmin_pinf” ficou em penúltimo lugar.

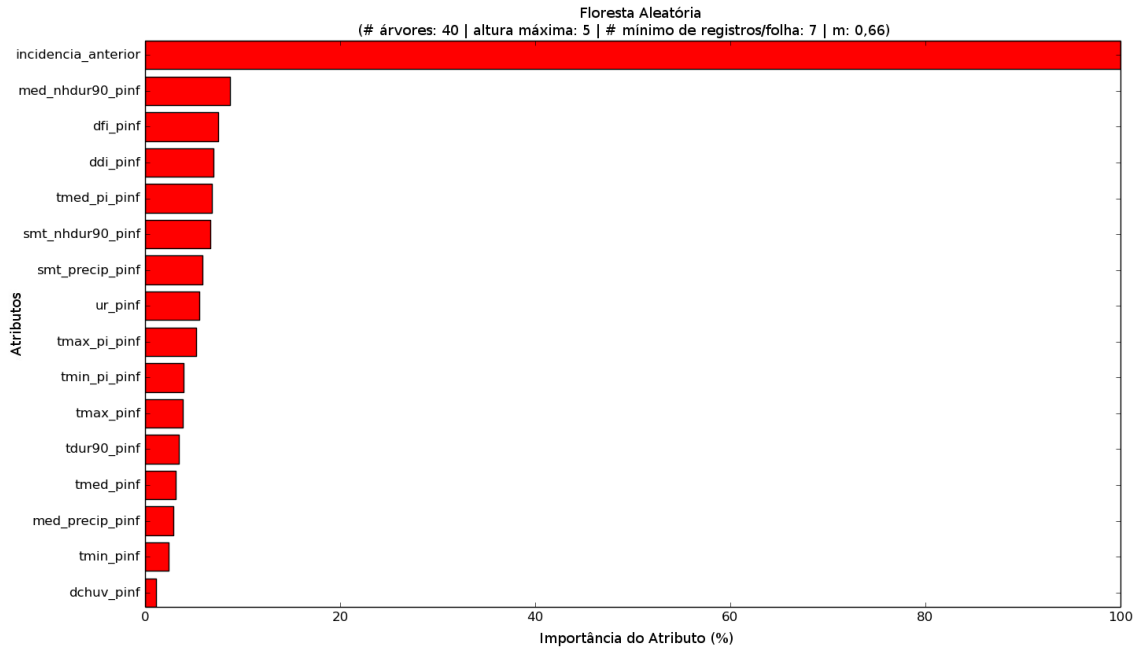

Figura 25. Importância dos atributos no *ensemble* “7” (alta/10 p.p.).

Em relação à Tabela 32, destacaram-se “34” e “35” com maiores percentuais de acerto na identificação do primeiro mês (44,83%). *Ensembles* desenvolvidos com conjunto de atributos 02 apresentaram maior capacidade nesse quesito (10,35 p.p. acima em relação aos demais modelos).

Os *ensembles* “34” e “35” foram baseados em *boosting* e *stacking*, respectivamente. *Stacking* possui desenvolvimento mais complexo: é preciso criar modelos nível-0, gerar conjunto de dados nível-1 e criar meta-classificador. Predizer um novo registro possui alta custo, pois é necessário a predição de todos os modelos nível-0 antes da predição do meta-classificador. Esses fatores desestimulam o uso de “35” uma vez que “34” possui mesmo desempenho preditivo.

Ensembles “31”, “32” e “33”, baseados nos atributos 02, tiveram melhor desempenho nas predições sobre os meses de maior aumento da TP, com 82,19%, 72,60% e 80,82%, respectivamente.

5.4 *Ensembles* para baixa carga pendente e 3 p.p.

No cenário baixa/3 p.p. foram constatadas três ocorrências da ferrugem tardia (Apêndice B.3), todas em Boa Esperança (BE) e para baixa carga houve 45 meses em que a TP foi maior ou igual a 6 p.p. (Apêndice C).

O gráfico ROC da Figura 26 indica que os *ensembles* "14", "17" e "27" pertencem ao envelope convexo (curva). Observa-se a ausência de modelos baseados a partir do conjunto de atributos 01 ("1" a "10").

No cenário baixa/3 p.p. todos os modelos padrão selecionados foram baseados em *stackings*. Seus desempenhos preditivos considerando todo o conjunto de dados estão apresentados na Tabela 33.

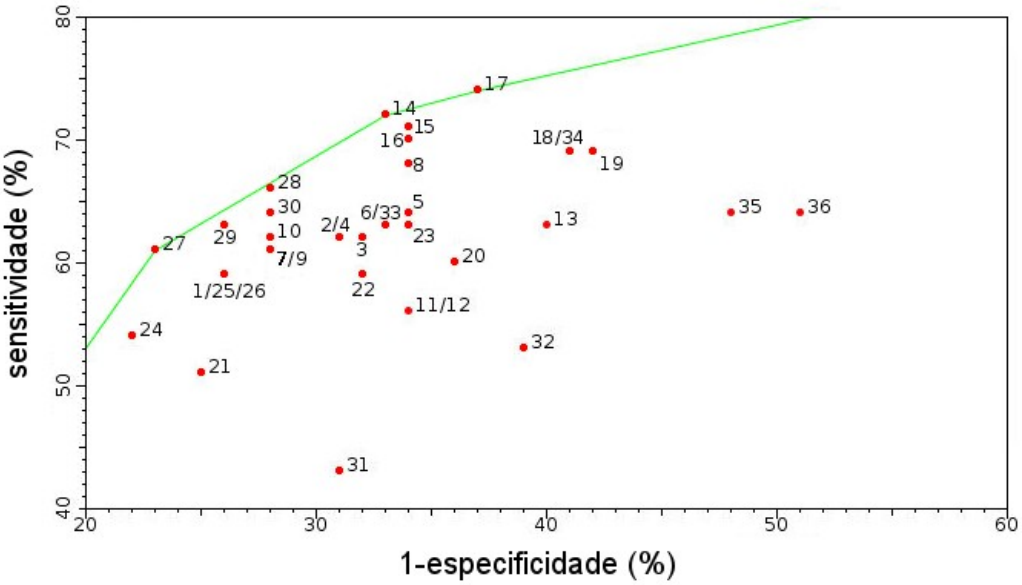

Figura 26. Gráfico ROC para limiar de 3 p.p. e baixa carga.

Tabela 33. Desempenho preditivo dos *ensembles* selecionados no envelope convexo (baixa/3 p.p.).

Ensemble	Técnica	Acurácia (%)	Sensitividade (%)	TFP (%)	Precisão (%)	Especificidade (%)	*AUC*
14	Stack-t + Ridge	69,33	72,14	33,12	70,74	33,12	0,695
17	Stack-t + Ridge	68,33	74,28	36,87	69,07	36,87	0,687
27	Stack-t + *SVM*(p)	69,67	61,42	23,12	70,82	23,12	0,692

SVM(p) – *SVM* com kernel polinomial / Stack-t – *stacking* utilizando todos os *ensembles*
Ridge – classificador Ridge / TFP – taxa de falsos positivos

Na Tabela 34 constam desempenhos preditivos dos *ensembles* do envelope convexo em relação aos períodos críticos para progresso da ferrugem.

Tabela 34. Detecção dos momentos críticos para epidemia da ferrugem dos *ensembles* contidos no envelope convexo (baixa/3 p.p.).

Ensemble	**Técnica**	**Primeiro mês**			**Ferrugem tardia**			**Maior aumento De TP**		
		Total	**Acertos**	**Acertos (%)**	**Total**	**Acertos**	**Acertos (%)**	**Total**	**Acertos**	**Acertos (%)**
14	Stack-t + Ridge	29	19	65,52	3	1	33,33	45	33	73,33
17	Stack-t + Ridge	29	18	62,07	3	1	33,33	45	33	73,33
27	Stack-t + *SVM*(p)	29	16	55,17	3	1	33,33	45	28	62,22

SVM(p) – *SVM* com kernel polinomial / Stack-t – *stacking* utilizando todos os *ensembles*
Ridge – classificador Ridge / TFP – taxa de falsos positivos

Quando a avaliação se restringe somente aos períodos críticos, sobressaíram *ensembles* baseados em árvores de decisão (“31”, “32”, “35” e “36”), criados a partir dos atributos 01 e 03 (Tabela 35).

Tabela 35. Desempenho preditivo dos melhores *ensembles* quanto à detecção dos períodos críticos para epidemia (baixa/3 p.p.).

Cenário	***Ensemble***	**Técnica**	**Primeiro mês**			**Ferrugem tardia**			**Maior aumento De TP**		
			Total	**Acertos**	**Acertos (%)**	**Total**	**Acertos**	**Acertos (%)**	**Total**	**Acertos**	**Acertos (%)**
01	31	*Boosting* + AD	29	21	72,41	3	1	33,33	45	27	60,00
01	32	*Bagging* + AD	29	21	72,41	3	1	33,33	45	27	60,00
01	33	Stack-s + *SVM*(p)	29	19	65,52	3	2	66,67	45	29	64,44
02	34	Stack-t + *SVM*(p)	29	21	72,41	3	2	66,67	45	27	60,00
03	35	*Boosting* + AD	29	21	72,41	3	1	33,33	45	27	60,00
03	36	*Boosting* + AD	29	17	58,62	3	2	66,67	45	31	68,89

AD – árvore de decisão / SVM(p) – SVM com kernel polinomial

Seleção de *ensembles*

Analisando os modelos do envelope convexo (Tabela 33), “27” possui maiores valores de acurácia (69,97%), precisão (70,82%) e *AUC* (0,692). Entretanto, seu uso não é recomendado devido sua baixa sensitividade (61,42%), especificidade (23,12%) e pior desempenho quanto aos momentos críticos (Tabela 34).

Nos demais modelos do envelope convexo, “17” possui maiores valores de sensitividade e especificidade. O *stacking* “14” possui acurácia, precisão e *AUC* maiores em 1 p.p., 1,67 p.p. e 0,008, respectivamente, em relação a “17” (Tabela 33). Além disso, “14” possui melhor desempenho (65,52%) na predição do primeiro mês (Tabela 34).

Em relação aos períodos críticos (Tabela 35), o modelo “36” apresentou melhor desempenho na predição dos meses de maior aumento na TP (68,89%). Quatro modelos (“31”, “32”, “34”, e “35”) predizeram corretamente o primeiro mês em 72,41% dos casos.

A ferrugem tardia foi identificada em 66,67% dos casos em três *ensembles* (“33”, “34” e “36”). Por ser o único *ensemble* baseado em árvore de decisão, neste cenário é

recomendado o uso de apenas um modelo para a tarefa dos meses de maior aumento e ferrugem tardia, “36” (Tabela 35).

O único erro da predição da ferrugem tardia (Figura 27) realizada por “36” foi na safra 2011/12 de BE. Neste ano agrícola, o modelo indicou janeiro/2012 como primeiro mês em que TP ≥ 3 p.p.

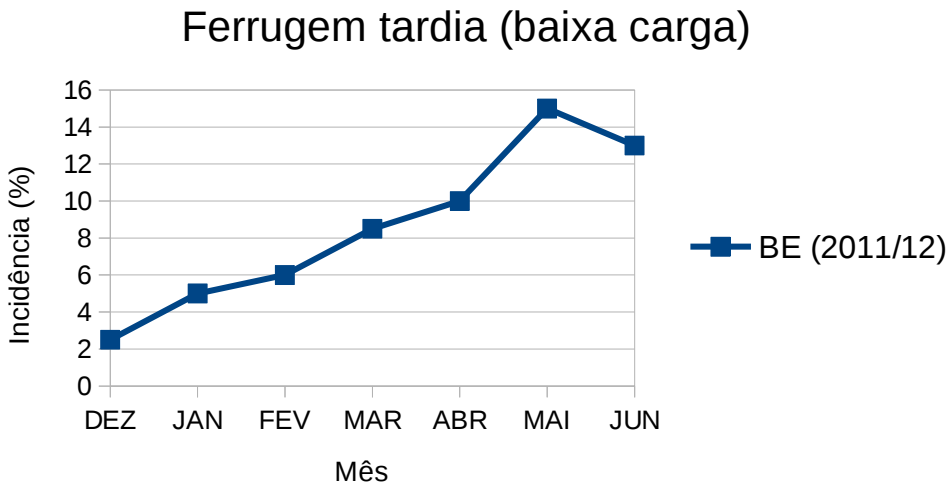

Figura 27. Ocorrência de ferrugem tardia em BE (2011/12) em lavouras de baixa carga pendente.

Analisando os valores dos atributos preditivos para safra 2011/12 (Tabela 36) observa-se que em janeiro/2012 os valores dos atributos relacionados à UR (“ur_pinf”) e período de molhamento foliar (“med_nhdur90_pinf” e “smt_nhdur90_pinf”) estiveram entre os maiores. Além disso, atributos derivados da precipitação (“med_precip_pinf” e “smt_precip_pinf”) foram os mais elevados. Esses atributos estão relacionados com o acúmulo de água na superfície foliar, indispensável para o desenvolvimento da ferrugem na lavoura (ZAMBOLIM et al., 2005) e podem ter influenciado erroneamente o modelo “36”.

Tabela 36. Dados dos atributos referentes a UR e precipitação na safra 2011/12 para BE.

Mês	Ano	ur_pinf	med_nhdur90_pinf	smt_nhdur90_pinf	med_precip_pinf	smt_precip_pinf
DEZ	2011	76	6,1	213	3,5	122,8
JAN	2012	84	11,3	349,5	11,5	357,2
FEV	2012	87	13,7	425,5	7,9	246,4
MAR	2012	80	8,4	251,5	5,3	158,2
ABR	2012	77	6,4	174	3,6	98,2
MAI	2012	82	9	208	2,2	50,2
JUN	2012	82	9,7	251,5	2	51,6

5.5 *Ensembles* para baixa carga pendente e 5 p.p.

O gráfico ROC presente na Figura 28 se refere aos modelos desenvolvidos para baixa/5 p.p.. Além dos modelos padrão houve um modelo (“31”), com atributos 02, e um (“32”), com atributos 03, que apresentaram desempenho superior em relação aos períodos críticos. Fizeram parte do envelope convexo (curva) os *ensembles* “7”, “8”, “9”, “16”, “21”, “22” e “32”. A Tabela 37 mostra o desempenho preditivo dos modelos presentes no envelope convexo.

Assim como em baixa/3 p.p., os modelos padrão de baixa/5 p.p. foram baseados em *stackings*, sendo que 12 utilizaram regressão logística como meta-classificador, 12 em Ridge, 5 usaram *SVM* com kernel polinomial e 1 em perceptron.

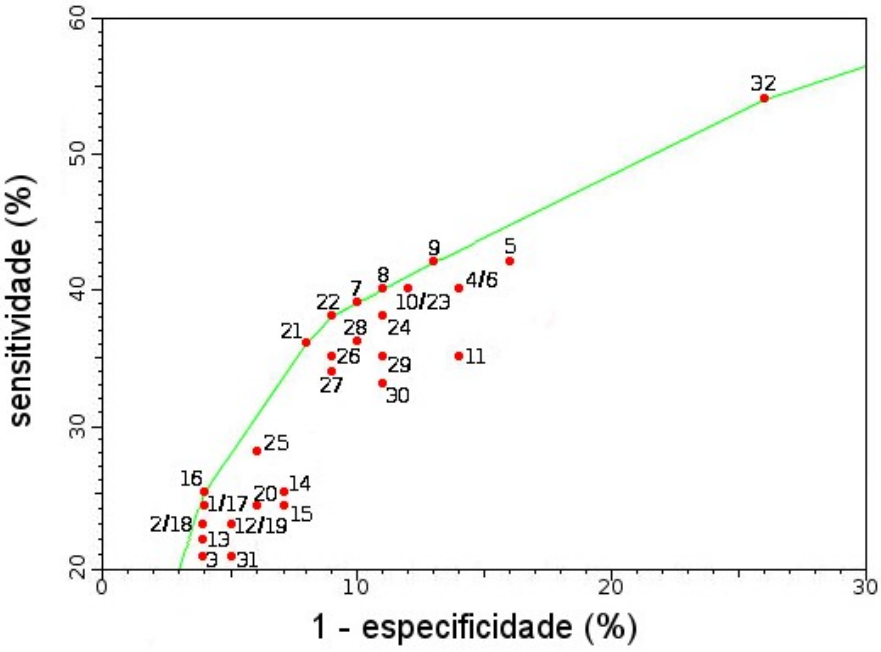

Figura 28. Gráfico ROC para limiar de 5 p.p. e baixa carga.

Tabela 37. Desempenho preditivo dos *ensembles* selecionados no envelope convexo (baixa/5 p.p.).

Ensemble	Técnica	Acurácia (%)	Sensitividade (%)	TFP (%)	Precisão (%)	Especificidade (%)	*AUC*
7	Stack-t + Ridge	72,64	39,27	9,73	74,15	90,27	0,648
8	Stack-t + Ridge	72,29	40,18	10,76	69,65	89,24	0,647
9	Stack-t + Ridge	71,62	42,10	12,81	67,84	87,19	0,646
16	Stack-t + Ridge	71,26	24,72	4,02	49,33	95,98	0,604
21	Stack-t + Ridge	72,88	36,10	7,63	77,23	7,63	0,642
22	Stack-t + Ridge	72,53	38,10	9,21	72,38	9,21	0,644
32	Stack-t + *SVM*(p)	67,12	53,81	25,60	56,79	74,40	0,641

SVM(p) – *SVM* com kernel polinomial / Stack-t – *stacking* utilizando todos os *ensembles*

Ridge – classificador Ridge / TFP – taxa de falsos positivos

Observa-se pela Tabela 37 que, apesar de não estar presente entre os modelos padrão, o *stacking* "32" fez parte do envelope convexo e, consequentemente, obteve melhor desempenho quanto ao período crítico em relação aos demais modelos do envelope convexo (Tabela 38).

Tabela 38. Detecção dos momentos críticos para epidemia da ferrugem dos *ensembles* contidos no envelope convexo (baixa/5 p.p.).

Ensemble	**Técnica**	**Primeiro mês**			**Maior aumento De TP**		
		Total	**Acertos**	**Acertos (%)**	**Total**	**Acertos**	**Acertos (%)**
7	Stack-t + Ridge	29	12	41,38	45	13	28,89
8	Stack-t + Ridge	29	12	41,38	45	14	31,11
9	Stack-t + Ridge	29	10	34,48	45	14	31,11
16	Stack-t + Ridge	29	7	24,14	45	7	15,56
21	Stack-t + Ridge	29	9	31,03	45	12	26,67
22	Stack-t + Ridge	29	8	27,59	45	13	28,89
32	Stack-t + *SVM*(p)	29	16	55,17	45	22	48,89

SVM(p) – *SVM* com kernel polinomial / Stack-t – *stacking* utilizando todos os *ensembles*
Ridge – classificador Ridge

O *stacking* "31", único *ensemble* com performance superior aos modelos padrão e não selecionado no envelope convexo, tem seu desempenho demonstrado na Tabela 39.

Tabela 39. Desempenho preditivo do melhor *ensemble* quanto à detecção dos períodos críticos para epidemia (baixa/5 p.p.).

Cenário	***Ensemble***	**Técnica**	**Primeiro mês**			**Maior aumento De TP**		
			Total	**Acertos**	**Acertos (%)**	**Total**	**Acertos**	**Acertos (%)**
02	31	Stack-t + *SVM*(p)	29	11	37,93	45	22	48,89

SVM(p) – *SVM* com kernel polinomial / Stack-t – *stacking* utilizando todos os *ensembles*

Seleção de *ensembles*

À exceção de "32", os modelos do envelope convexo obtiveram alta acurácia (acima de 71,00%), sendo a maior diferença de, apenas, 1,62 p.p. (Tabela 37). Os valores de *AUC* também foram próximos (diferença máxima de 0,044).

Assim, os fatores diferenciais na escolha do *ensemble* para monitoramento mensal da ferrugem, de dezembro a junho, foram a sensitividade, precisão e especificidade. À exceção de "32", os modelos apresentaram baixos valores de sensitividade (abaixo de 43,00%). Os piores nessa métrica foram "16" (24,72%), "21" (36,10%) e "22" (38,10%). *Stackings* "16" e "32" ainda tiveram os piores índices para precisão (49,33% e 56,79%,

respectivamente). Descartando “16”, os piores valores de especificidade foram de “32”, “9” e “8” com 74,40%, 87,19% e 89,24%, respectivamente. Assim, de forma geral, “7” apresentou valores mais balanceados entre as métricas de avaliação.

Comparando os *ensembles* “31” e “32” sobre seus desempenhos preditivos dos períodos críticos, “32” é melhor em relação ao primeiro mês de TP ≥ 5 p.p. (55,17%) – 17,24 p.p. superior a “31” – e apresentou mesma acurácia quanto aos meses de maior aumento da TP (48,89%). O uso de “32” é mais indicado para predições referentes aos períodos críticos para desenvolvimento da ferrugem.

5.6 Resultados gerais

5.6.1 Comparação dos modelos

A Tabela 40 resume os principais dados dos *ensembles* selecionados nos cenários estudados, conforme carga pendente de frutos e limiar do atributo meta.

Tabela 40. *Ensembles* selecionados para cada tarefa preditiva nos cenários estudados.

Cenário		Alta/5 p.p.	Alta/10 p.p.	Baixa/3 p.p.	Baixa/5 p.p.
Monitoramento mensal (dezembro a junho)	***Ensemble***	6	7	14	7
	Cenário	01	01	02	01
	Acurácia	78,00%	73,33%	69,33%	72,64%
Primeiro mês	***Ensemble***	38	34	34	32
	Cenário	02	02	02	03
	Acurácia	79,31%	44,83%	72,41%	55,17%
Ferrugem tardia	***Ensemble***	31	-	36	-
	Cenário	01	-	03	-
	Acurácia	40,00%	-	66,67%	-
Maior aumento de TP	***Ensemble***	31	31	36	32
	Cenário	01	01	03	03
	Acurácia	100,00%	82,19%	68,89%	48,89%

Os desempenhos preditivos dos *ensembles* selecionados para alta/5 e 10 p.p. e baixa/5 p.p. foram superiores aos modelos de Girolamo Neto (2013) em todos os cenários considerando o monitoramento mensal da ferrugem (Tabela 41). Os *ensembles* apresentaram acurácias, 2,67, 2,67 e 8,31 p.p. superiores em relação aos modelos de Girolamo Neto (2013) para alta carga/5 p.p., alta carga/10 p.p. e baixa carga/5 p.p., respectivamente.

Tabela 41. Compilação dos melhores resultados obtidos nesse estudo (destacado) e pelos modelos de Girolamo Neto (2013) avaliados em todo período de dezembro a junho.

Ano agrícola (dezembro a junho)					
Carga	**Limiar (p.p.)**	**Modelo**	**Técnica**	**Acurácia (%)**	**Diferença (p.p.)**
alta	5	6	*Boosting* + AD	78,00	2,67
		VG-3	*SVM*	75,33	
alta	10	7	FA	73,33	2,67
		VG-6	*SVM*	70,66	
baixa	5	7	Stack-t + Ridge	72,64	8,31
		VG-7	FA	64,33	

AD – árvore de decisão / FA – floresta aleatória

Stack-t – *stacking* utilizando todos os *ensembles* / Ridge – Classificador Ridge

Ensembles também apresentaram melhor desempenho quanto à predição do primeiro mês de taxa de progresso mensal (TP) acima ou igual ao limiar (Tabela 42). Destaque para o *ensemble* "38" (cenário alta/5 p.p.), que foi 24,14 p.p. superior ao modelo BE-1 de Girolamo Neto (2013). A menor diferença foi no cenário baixa/5 p.p., em que o ensemble "32" foi superior em 10,34 p.p. em relação a CM-9.

Tabela 42. Compilação dos melhores resultados obtidos nesse estudo (destacado) e pelo modelo de Girolamo Neto (2013) para detecção do primeiro mês com TP acima ou igual ao limiar.

Primeiro mês					
Carga	**Limiar (p.p.)**	**Modelo**	**Técnica**	**Acurácia (%)**	**Diferença (p.p.)**
alta	5	38	Stack-s + *SVM*(p)	79,31	24,14
		BE-1	AD	55,17	
alta	10	34	*Boosting* + *SVM*(p)	44,83	13,08
		BE-5	FA	31,03	
baixa	5	32	Stack-t + *SVM*(p)	55,17	10,34
		CM-9	*SVM*	44,83	

AD – árvore de decisão / FA – floresta aleatória / Ridge – Classificador Ridge

Stack-t – *stacking* utilizando todos os *ensembles* / Stack-s – *stacking* simples

Em relação à ferrugem tardia, houve empate para o cenário alta/5 p.p. (Tabela 43). Apesar da baixa quantidade de registros (300) usados nos conjuntos de dados deste trabalho, o desempenho preditivo dos *ensembles* superaram ou igualaram os modelos de Girolamo Neto (2013), que foram criados a partir de 738 registros. A restrição dos dados meteorológicos e de incidência ao período crítico para desenvolvimento da ferrugem produziu *ensembles* de desempenho preditivo superior aos modelos atuais, considerando o período de dezembro a junho.

Tabela 43. Compilação dos melhores resultados obtidos nesse estudo (destacado) e pelos modelos de Girolamo Neto (2013) avaliados para questão do desenvolvimento tardio da ferrugem.

ferrugem tardia					
carga	**limiar (p.p.)**	**modelo**	**técnica**	**acurácia (%)**	**diferença (p.p.)**
alta	5	31	FA	40,00	0
		VG-3	*SVM*	40,00	

FA – floresta aleatória

5.6.2 Ferrugem tardia

Em dados de 29 anos agrícolas, distribuídos nas três cidades estudadas, o desenvolvimento tardio da ferrugem ocorreu com baixa frequência – 17,24% para alta carga e 5 p.p. e 10,34% para baixa carga e 3 p.p. – e, portanto, de difícil aprendizado para os *ensembles*. Apesar disso, foram desenvolvidos *ensembles* para baixa carga e 3 p.p., que conseguiram predizer a ferrugem tardia em duas de três ocasiões.

5.6.2.1 Alta carga pendente

Em relação à ferrugem tardia em lavouras de alta carga, o *ensemble* "31" (Tabela 43) foi baseado em floresta aleatória, o que permitiu verificar a relevância dos atributos para o *ensemble.* Na literatura há citações da influência de fatores climáticos na ferrugem (CHALFOUN et al., 2001, AVELINO et al., 2015). Ao analisar a relevância das variáveis presentes em "31", ficou evidente a importância de "incidencia_anterior" (Figura 29).

Outros atributos relevantes foram baseados na umidade relativa do ar (UR) e período de molhamento foliar, "ur_pinf" e "med_nhdur90_pinf", respectivamente. Altos níveis de UR (≥ 90%) também podem servir como base para cálculo do molhamento foliar (SENTELHAS et al., 2008). Períodos prolongados de molhamento foliar (acima de 6 horas) favorece o desenvolvimento da ferrugem do cafeeiro (ZAMBOLIM et al., 2002).

A temperatura mínima e média durante o período de infecção (PINF) – "tmin_pinf" e "tmed_pinf", respectivamente – também foram relevantes, diferentemente dos atributos referentes a temperatura máxima ("tmax_pinf" e "tmax_pi_pinf").

Segundo Kushalappa et al., (1983), temperaturas inferiores a 14,0 °C e superiores a 32,5 °C limitam o processo de infecção do fungo na folha. Avaliando o conjunto de dados

usados para modelagem, os maiores valores de “tmax_pinf” e “tmax_pi_pinf” foram 31,6 °C e 31,4 °C, respectivamente. Logo, em nenhum momento a temperatura máxima atingiu níveis que fossem desfavoráveis à infecção. Isso pode explicar a baixa relevância da temperatura máxima, resultado também encontrado por Girolamo Neto (2013), que utilizou dados meteorológicos de toda safra. Por outro lado, houve casos em que a temperatura mínima (“tmin_pinf” e “tmin_pi_pinf”) ficou abaixo do limite de 14,0 °C.

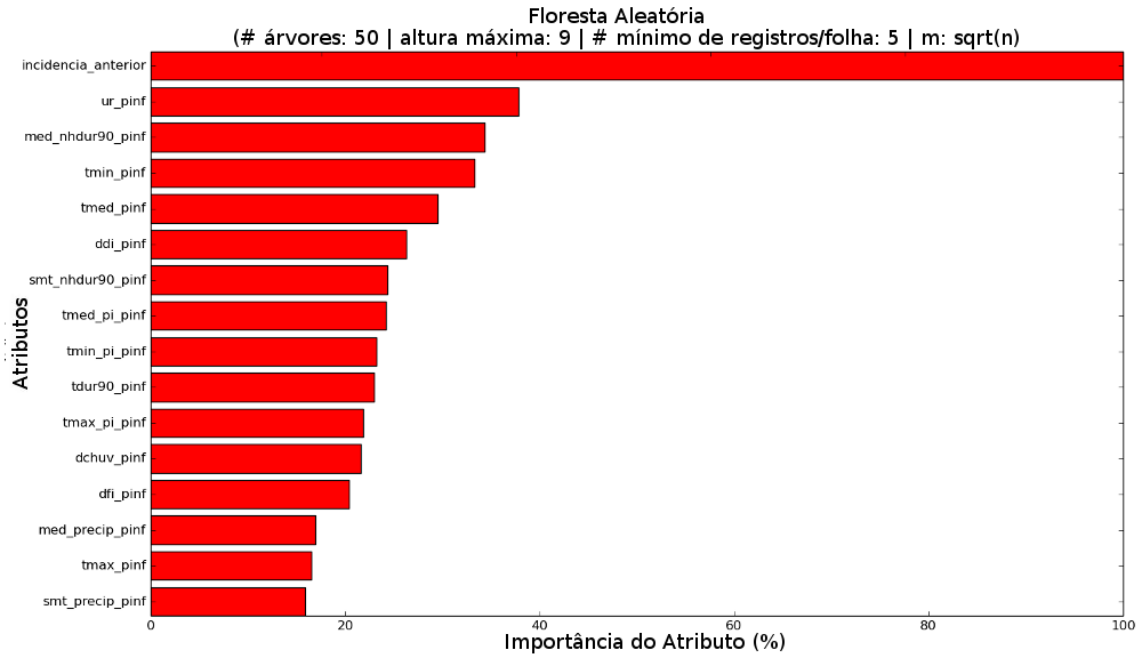

Figura 29. Importância dos atributos na floresta aleatória “31” (alta/5 p.p.).

Dentre os demais atributos pouco relevantes, o número de dias favoráveis à infecção (“dfi_pinf”) se encontrou nas últimas colocações. Atributos derivados a partir da precipitação (“dchuv_pinf”, “med_precip_pinf” e “smt_precip_pinf”) também tiveram pouca importância. Analisando os valores desses três atributos no conjunto de dados não foi possível observar nenhum padrão comportamental que fosse possível diferenciar anos com e sem ferrugem tardia. O que se observou em todas as safras, independentemente da ocorrência da ferrugem tardia, foram maiores valores de precipitação nos primeiros meses, dezembro a fevereiro.

Essa menor importância destes atributos com a ocorrência da ferrugem tardia pode estar relacionado com o duplo papel da chuva no desenvolvimento da doença. Chuvas leves podem auxiliar no acúmulo de água e formar uma lâmina d'água, indispensável para o processo de germinação (ZAMBOLIM et al., 2002). Entretanto, chuvas de maior intensidade, que ocorrem no início do ano, podem promover a lavagem do inóculo da superfície foliar

(CHALFOUN e ZAMBOLIM, 1985), dificultar o processo de infecção do fungo na folha e atrasar o desenvolvimento da epidemia da ferrugem. Assim, como não foi possível identificar uma padrão na precipitação em anos de ferrugem tardia em lavouras de alta carga, seus atributos derivados tiveram pouca importância para "31".

5.6.2.2 Baixa carga pendente

No cenário baixa/3 p.p. o melhor *ensemble* para predição da ferrugem tardia foi o "36" (Tabela 35), um *boosting* com árvore de decisão criado a partir do cenário 03. Verificando a relevância dos atributos em "36" nota-se a ausência de "dfi_pinf" e "ddi_pinf" (número de dias desfavoráveis à infecção no PINF). Considerando os resultados para alta/5 p.p. e baixa/3 p.p., "ddi_pinf" e, principalmente, "dfi_pinf" não tiveram importância destacada na questão da ferrugem tardia (Figura 30).

Com a ausência da "incidencia_anterior", a UR no PINF ("ur_pinf") foi o atributo mais importante em "36". Ao contrário do resultado para alta carga, a temperatura máxima ("tmax_pinf") e precipitação ("med_precip_pinf") tiveram alta relevância para o *ensemble*. Como a ocorrência da ferrugem tardia em lavouras de baixa carga ocorreram na mesma estação e safra que os de alta carga, a análise quanto aos valores para temperatura máxima e precipitação são as mesmas da seção 5.6.2.1.

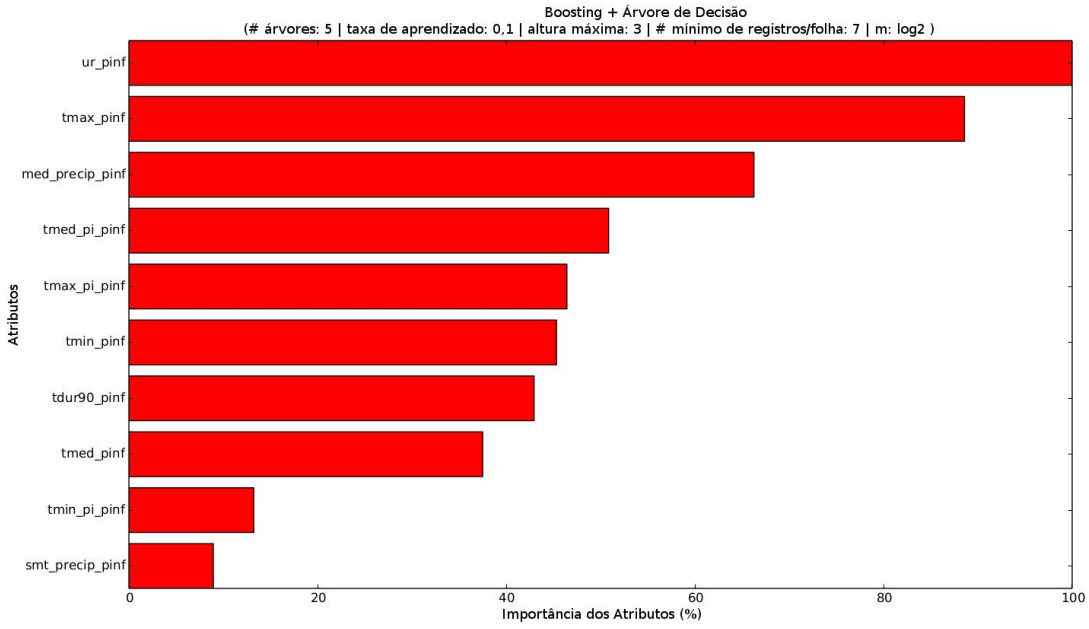

Figura 30. Importância dos atributos no *boosting* com árvore de decisão "36" (baixa/3 p.p.).

5.6.3 Conjuntos de atributos, kernels e meta-classificadores

Na Tabela 44 estão presentes os melhores modelos, em termos de acurácia e relação sensitividade/especificidade, desenvolvidos para cada cenário e conjuntos de atributos. Esses *ensembles* estão contidos nos modelos padrão. O desempenho preditivo é em relação ao monitoramento mensal do progresso da ferrugem, ou seja, os modelos foram avaliados em todo o conjunto de dados (total de 300 registros).

Analisando o cenário alta/5 p.p., modelos criados a partir dos atributos 01 foram ligeiramente superiores a 02, com acurácia (1,33 p.p.), sensitividade (1,76 p.p.) e especificidade (2,88 p.p.) superiores. Em alta/10 p.p. o *ensemble* "1", com atributos 01, teve acurácia superior aos demais por 11,66 p.p. e sensitividade maior por, no mínimo, 27,14 p.p. Mas a especificidade de "26" foi a maior (88,75%).

Para baixa/3 p.p. o *ensemble* "10" (com atributos 01), apresentou números mais balanceados entre as três métricas, seguido pelo modelo "14". Em baixa/5 p.p. os modelos "7" e "16" apresentaram melhor balanceamento entre as métricas em relação a "21", de baixa especificidade (7,63%).

Tabela 44. Melhores *ensembles* criados para cada conjunto de atributos e cenário.

Alta carga pendente – 5 p.p.

Atributos	*Ensemble*	Técnica	Acurácia (%)	Sensitividade (%)	Especificidade (%)
01	6	*Boosting* + AD	78,00	97,64	28,62
02	12	FA	76,67	95,88	25,74
03	21	Stack-t + Ridge	74,62	95,75	21,81

Alta carga pendente – 10 p.p.

Atributos	*Ensemble*	Técnica	Acurácia (%)	Sensitividade (%)	Especificidade (%)
01	1	*Bagging* + AD	74,33	82,85	66,88
02	11	Stack-t + *SVM*(p)	62,67	55,71	31,25
03	26	Stack-t + *SVM*(p)	62,67	32,82	88,75

Baixa carga pendente – 3 p.p.

Atributos	*Ensemble*	Técnica	Acurácia (%)	Sensitividade (%)	Especificidade (%)
01	10	Stack-t + Ridge	67,67	62,14	72,50
02	14	Stack-t + Ridge	69,33	72,14	33,12
03	27	Stack-t + *SVM*(p)	69,67	61,42	23,12

Baixa carga pendente – 5 p.p.

Atributos	*Ensemble*	Técnica	Acurácia (%)	Sensitividade (%)	Especificidade (%)
01	7	Stack-t + Ridge	72,64	39,27	90,27
02	16	Stack-t + Ridge	71,26	24,72	95,98
03	21	Stack-t + Ridge	72,88	36,10	7,63

Verficando os *ensembles* recomendados para predição do primeiro mês e ferrugem tardia (Figura 40), foram selecionados para alta carga modelos baseados nos atributos 02 e 01, respectivamente. Em baixa carga, foi recomendada a predição da ferrugem tardia utilizando um *ensemble* com atributos 03 (“36”) pelo fato deste ser baseado em árvore de decisão e, assim, ser possível verificar os atributos mais relevantes. Porém, considenrao apenas a acurácia, existem modelos com performance superior a “36”, baseados em 01 e 02 (Tabela 35).

De modo geral, os modelos criados com atributos 01 apresentaram melhor performance. Entretanto, para formação deste conjunto de atributos é necessário um acompanhamento mensal da incidência da ferrugem, devido ao atributo “incidencia_anterior”. Nem sempre este atributo estará disponível no conjunto de dados, sendo um ponto negativo para utilização de 01. Como forma de contornar tal questão, pode-se adotar os atributos 02, de cálculo mais simples (não possui “dfi_pinf” e “ddi_pinf”) e sem a necessidade de acompanhamento mensal da incidência da ferrugem na lavoura (“incidencia_anterior”).

Girolamo Neto (2013) usou em seu trabalho uma versão mais complexa da matriz de infecção (Meira, 2008), e derivou atributos mais complexos dos utilizados neste trabalho. Maioria de seus modelos (82%), selecionados por meio do envelope convexo, continham atributos derivados da matriz. Neste trabalho, “ddi_pinf” e “dfi_pinf” (presentes em 03) não tiveram a mesma relevância que demais atributos em relação ao monitoramento mensal e predição do primeiro mês. A criação desses atributos, da forma como realizada neste estudo, não contribuiu significativamente na melhora do desempenho preditivo dos *ensembles* (Tabela 44).

Em relação aos kernels utilizados no *SVM* (linear, polinomial, rbf e sigmóide), *ensembles* baseados no polinomial apresentaram melhores resultados, considerando os parâmetros testados (Apêndice A). Em uma menor proporção, linear também foi eficaz. Rbf e sigmóide chegaram a produzir *ensembles* que não adquiriram nenhum conhecimento do conjunto de dados, ou seja, o modelo classificou todos os registros com o mesmo valor (*0* ou *1*).

De forma geral, os *stackings* com meta-classificador baseado em Ridge e *SVM* com kernel polinomial apresentaram melhor desempenho preditivo em todas as avaliações em relação aos demais algoritmos. Dentre os *stackings* presentes entre os modelos padrão, considerando todos os cenários de modelagem (seção 4.2.6), não houve *stacking* com uso de passivo-agressivo como meta-classificador, apenas um com perceptron e dois baseados em *SVM* de kernel linear.

6 CONCLUSÕES

Diferentes variáveis climáticas foram as mais importantes para predição da ocorrência da ferrugem tardia nos cafeeiros em alta e baixa carga pendente. Atributos referentes ao período de molhamento foliar ("ur_pinf" e "med_nhdur90_pinf") foram mais relevantes nas lavouras em alta carga. Temperatura máxima ("tmax_pinf") e precipitação ("med_precip_pinf") tiveram maior importância em cafeeiros com baixa carga pendente.

Ensembles desenvolvidos neste trabalho apresentaram, de forma geral, desempenho preditivo superior aos demais modelos existentes considerando o período de maior importância para o progresso da ferrugem na lavoura, dezembro a junho.

A utilização do valor de incidência do mês anterior ("incidencia_anterior"), presente apenas no conjunto de atributos 01, foi capaz de produzir *ensembles*, na média, com melhor desempenho preditivo. Porém, o uso dos atributos 01 se limita ao acompanhamento mensal da incidência da ferrugem na lavoura.

Predições quanto à identificação do primeiro mês em que a taxa de progresso da ferrugem atingirá o valor do limiar pré-estabelecido foram melhores em modelos criados a partir do conjunto de atributos 02.

Variáveis derivadas da matriz de infecção ("ddi_pinf" e "dfi_pinf"), presentes no conjunto de atributos 03, não auxiliaram no desenvolvimento de *ensembles* com desempenho preditivo superior aos modelos baseados em outros conjuntos de atributos.

Dentre os tipos de kernel utilizados, o polinomial produziu os melhores *ensembles* baseados em *SVM* (*bagging* e *boosting*), considerando o conjunto de parâmetros testados (Apêndice A).

6.1 Trabalhos Futuros

A ferrugem tardia é um evento raro e para estudá-la mais profundamente talvez seja mais indicado o uso de regras de exceção, principalmente em modelagens que usarem a técnica de árvore de decisão. Essas regras contradizem o senso comum e aparecem poucas vezes no conjunto de dados. Elas se diferem do restante dos registros e devido a isto são interessantes e podem oferecer esclarecimentos para novas descobertas (SUZUKI, 2006).

Além disso, indica-se para trabalhos futuros:

- coletar dados de outras regiões produtores de café, inclusive de outros estados;

- segmentação maior dos dados para modelagem conforme a estação climática (desde que haja dados suficientes): dezembro a fevereiro (verão) / março a junho (outono);
- uso de outras técnicas *ensemble* com, por exemplo, combinar diversos modelos em um esquema de votação com peso;
- inclusão de variáveis relacionados aos fenômenos climáticos *El Niño* e *La Niña*;
- utilizar valor de incidência da ferrugem na lavoura de dois meses anteriores como atributo preditivo, uma vez que o período de incubação da ferrugem pode chegar a 65 dias (MORAES et al., 1976).

7 REFERÊNCIAS BIBLIOGRÁFICAS

AGRESTI, A. Categorical Data Analysis. 3ª edição. New Jersey: John Wiley & Sons, 2013. 714 p.

AGRIOS, G. N. **Plant pathology**. 5ª edição. San Diego: Elsevier Academic Press, 2005. 922 p.

ALVES, S. T.; DIAS, R. C. E.; BENASSI, M. T. Metodologia para análise simultânea de ácido nicotínico, trigonelina, ácido clorogênico e cafeína em café torrado por cromatografia líquida de alta eficiência. **Química Nova**, v. 29, n. 6, p. 1164-1168, 2006.

ALVES, M. de C.; CARVALHO, L. G. de; POZZA, E. A.; ALVES, L. S. A Soft Computing Approach For Epidemiological Studies of Coffee And Soybean Rusts. **International Journal of Digital Content Technology and its Applications (JDCTA)**, v. 4, n. 1, p. 149-154, 2010.

APTÉ, C.; WEISS, S. Data mining with decision trees and decision rules. **Future generation computer systems**, v. 13, n. 2, p. 197-210, 1997.

ARNESON, P. A. Coffee rust. *The Plant Health Instructor*. **doi:10.1094/PHI-I-2000-0718-02**. 2000. Updated 2011.

AVELINO, J.; SAVARY, S. Rational and optimizes chemical control of coffee leaf rust (Hemileia vastatrix). In: BERRY, D. (Ed.) **Research and Coffee Growing**. Edição 2002. Montpellier, CIRAD. 2002. p. 134-143.

AVELINO, J.; WILLOCQUET, L.; SAVARY, S. Effects of crop management patterns on coffee rust epidemics. **Plant pathology**, v. 53, p. 541-547, 2004.

AVELINO, J.; CRISTANCHO, M.; GEORGIOU, S.; IMBACH, P.; AGUILAR, L.; BORNEMANN, G.; LÄDERACH, P.; ANZUETO, F.; HRUSKA, A. J.; MORALES, C. The coffee rust crises in Colombia and Central America (2008–2013): impacts, plausible causes and proposed solutions. **Food Security**, v. 7, p. 303-321, 2015.

AZEVEDO, A.; SANTOS, M. F. KDD, SEMMA and CRISP-DM: A Parallel Overview. In: IADIS European Conference on Data Mining, 2008, Amsterdam, **Proceedings...**, p. 182-185, 2008.

BANFIELD, R. E.; HALL, L. O.; BOWYER, K. Q.; KEGELMEYER, W. P. A comparison of decision tree ensemble creation techniques. **IEEE Transactions on Pattern Analysis and Machine Intelligence**, v. 29, n. 1, p. 173-180, 2007.

BAUER, E.; KOHAVI, R. An empirical comparison of voting classification algorithms: Bagging, boosting, and variants. **Machine learning**, v. 36, p. 105-139, 1999.

BERGAMIN FILHO, A.; AMORIM, L. Sistemas de previsão e avisos. In: AMORIM, L.; REZENDE, J. A. M.; BERGAMIN FILHO, A. (Eds.) **Manual de fitopatologia**: Princípios e Conceitos. 4ª edição. v. 1. São Paulo: Agronômica Ceres. 2011. 704 p.

BETTIOL, W.; GHINI, R. Proteção de plantas em sistemas agrícolas alternativos. In: MICHEREFF, S. J. & BARROS, R. (Eds.) **Proteção de plantas na agricultura sustentável**. Recife: Universidade Federal Rural de Pernambuco, 2001. p. 1-13.

BLAJDO, P.; GRZYMALA-BUSSE, J. W.; HIPPE, Z. S.; KNAP, M.; MROCZEK, T.; PIATEK, L. A Comparison of Six Approaches to Discretization - A Rough Set Perspective. In: International Conference, RSKT, 3., 2008, Chengdu, **Proceedings...** Berlin: Springer, p. 31-38, 2008.

BOSCH, F. van den; PAVELEY, N.; SHAW, M.; HOBBELEN, P.; OLIVER, R. The dose rate debate: does the risk of fungicide resistance increase or decrease with dose? **Plant Pathology**, v. 60, p. 597-606, 2011.

BRADLEY, A. P. The use of the area under the ROC curve in the evaluation of machine learning algorithms. **Pattern Recognition**, v. 30, n. 7, p. 1145-1159, 1997.

BREIMAN, L. FRIEDMAN, J. H.; OLSHEN, R. A.; STONE, C. J. **Classification and regression trees**. 1ª edição. Boca Raton: Chapman & Hall/CRC, 1984. 358 p.

BREIMAN, L. Bagging Predictors. **Machine Learning**, v. 24, p. 123-140, 1996.

BREIMAN, L. Random forests. **Machine Learning Journal**, v. 45, p. 5-32, 2001.

BRIEM, G. J.; BENEDIKTSSON, J. A.; SVEINSSON, J. R. Use of Multiple Classifiers in Classification of Data from Multiple Data Sources. In: International Geoscience and Remote Sensing Symposium, 2001, Sydney, **Proceedings...** Sydney: IEEE, p. 882-884, 2001.

BURGES, C. J. C. A tutorial on support vector machines for pattern recognition. **Data mining and knowledge discovery**, v. 2, p. 121-167, 1998.

CARVALHO, V. L. de; CHALFOUN, S. M.; CUNHA, R. L. da. Manejo de doenças do cafeeiro. In: REIS, P. R.; CUNHA, R. L. da (Eds.). **Café Arábica do plantio à colheita**. 1ª edição. v. 1. Lavras: U. R. EPAMIG SM. 2010. 895 p.

CHALFOUN, S. M.; ZAMBOLIM, L. Ferrugem do cafeeiro. **Informe Agropecuário**, v. 11, n. 126, p. 42-46, 1985.

CHALFOUN, S. M.; CARVALHO, V. L. de. Controle químico da ferrugem (*Hemileia vastatrix* berk & br.) do cafeeiro através de diferentes esquemas de aplicação. **Pesquisa Agropecuária Brasileira**, v. 34, n. 3, p. 363-367, 1999.

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

CHALFOUN, S. M.; LIMA, R. D. de. Influência do clima sobre a incidência de doenças infecciosas. **Informe Agropecuário**, v. 12, n. 138, p. 31-36, 1986.

CHAPMAN, P.; CLINTON, J.; KERBER, R.; KHABAZA, T.; REINARTZ, T.; SHEARER, C.; WIRTH, R. **CRISP-DM 1.0**: step-by-step data mining guide. Illinois: SPSS, 2000.

CINTRA, M. E.; MEIRA, C. A. A.; MONARD, M. C.; CAMARGO, H. A.; RODRIGUES, L. H. A. The use of fuzzy decision trees for coffee rust warning in Brazilian crops. In: International Conference on Intelligent Systems Design and Applications, 11., 2011, Córdoba. **Proceedings...** Córdoba: IEEE, p. 1347-1352, 2011.

COAKLEY, S. M. Variation in climate and prediction of disease in plants. **Annual Review of Phytopathology**, v. 26, p. 163-181, 1988.

CONAB. Companhia Nacional de Abastecimento. **Acompanhamento da Safra Brasileira, Café, SAFRA 2015**. Quarto Levantamento. Dezembro de 2015. Disponível em: <http://www.conab.gov.br/OlalaCMS/uploads/arquivos/15_12_17_09_02_47_boletim_cafe_dezembro_2015_2.pdf>. Acesso em 04/01/2016.

CRAMMER, K.; DEKEL, O.; KESHET, J.; SHALEV-SHWARTZ, S.; SINGER, Y. Online passive-aggressive algorithms. **Journal of Machine Learning Research**, v. 7, p. 551-585, 2006.

CROS, J.; COMBES, M. C.; TROUSLOT, P.; ANTHONY, F.; HAMON, S.; CHARRIER, A.; LASHERMES, P. Phylogenetic analysis of chloroplast DNA Variation in *Coffea* L. **Molecular Phylogenetics and Evolution** v. 9, n. 1, p. 109-117, 1998.

CUSTÓDIO, A. A. de P.; POZZA, E. A.; CUSTÓDIO, A. A. de P.; SOUZA, P. E. de; LIMA, L. A.; LIMA, L. M. de. Intensidade da Ferrugem e da Cercosporiose em Cafeeiro quanto à face de exposição das plantas. **Coffee Science**, v. 5, n. 3, p. 214-228, 2010.

DIETTERICH, T. G. An experimental comparison of three methods for constructing ensembles of decision trees: Bagging, boosting, and randomization. **Machine learning**, v. 40, n. 2, p. 139-157, 2000.

DORAISWAMY, P. C.; PASTERIS, P. A.; JONES, K. C.; MOTHA, R. P.; NEJEDLIK, P. Techniques for methods of collection, database management and distribution of agrometeorological data. **Agricultural and Forest Meteorology**, v. 103, p. 83-97, 2000.

DOUGHERTY, J.; KOHAVI, R.; SAHAMI, M. Supervised and unsupervised discretization of continuous features. In: International Conference on Machine Learning, 12., 1995, Tahoe City, **Proceedings...** Morgan Kaufmann, p. 194-202, 1995.

ESTÉVEZ, J.; GAVILÁN, P.; GIRÁLDEZ, J. V. Guidelines on validation procedures for meteorological data from automatic weather stations. **Journal of Hydrology**, v. 402, p. 144-154, 2011.

FAWCETT, T. An introduction to ROC analysis. **Pattern Recognition Letters**, v. 27, n. 8, p. 861-874, 2006.

FAYYAD, U.; PIATETSKY-SHAPIRO, G.; SMYTH, P. From data mining to knowledge discovery in databases. **AI Magazine**, v. 17, n. 3, p. 37-54, 1996.

FERNANDES, R. C.; EVANS, H. C.; BARRETO, R.W. Confirmation of the occurrence of teliospores of Hemileia vastatrix in Brazil with observations on their mode of germination. **Tropical Plant Pathology**, v. 34, n. 2, p. 108-113, 2009.

FREUND, Y.; SCHAPIRE, R. E. Experiments with a New Boosting Algorithm. In: International Conference on Machine Learning, 13., 1996, Bari, **Proceedings...** Morgan Kaufmann, p. 148-156, 1996.

FREUND, Y.; SCHAPIRE, R. E. Large Margin Classification Using the Perceptron Algorithm. **Machine learning**, v. 37, p. 277-296, 1999.

GALAR, M.; FERNANDÉZ, A.; BARRENECHEA, E.; HERRERA, F. EUSBoost: Enhancing ensembles for highly imbalanced data-sets by evolutionary undersampling. **Pattern Recognition**, v. 46, p. 3460-3471, 2013.

GARCÍA, S.; LUENGO, J.; SÁEZ, J. A.; LÓPEZ, V.; HERRERA, F. A survey of discretization techniques: Taxonomy and empirical analysis in supervised learning. **IEEE Transactions on Knowledge and Data Engineering**, v. 25, n. 4, p. 734-750, 2013.

GARÇON, C. L. P.; ZAMBOLIM, L.; MIZUBUTI, E. S. G.; VALE, F. X. R.; COSTA, H. Controle da ferrugem do cafeeiro com base no valor de severidade. **Fitopatologia Brasileira**, v. 29, n. 5, p. 486-491, 2004.

GEURTS, P.; ERNST, D.; WEHENKEL, L. Extremely randomized trees. **Machine learning**, v. 63, p. 3-42, 2006.

GIROLAMO NETO, C. di. **Desenvolvimento e avaliação de modelos de alerta para a ferrugem do cafeeiro**. 2013. 155 f. Dissertação (Mestrado em Engenharia Agrícola). Universidade Estadual de Campinas, Campinas. 2013.

GIROLAMO NETO, C. di; RODRIGUES, L. H. A.; MEIRA, C. A. A. Modelos de predição da ferrugem do cafeeiro (*Hemileia vastatrix*) por técnicas de mineração de dados. **Coffee Science**, v. 9, n. 3, p. 408-418, 2014.

GLEASON, M. L.; TAYLOR, S. E.; LOUGHIN, T. M.; KOEHLER, K. J. Development and Validation of an Empirical Model to Estimate the Duration of Dew Periods. **Plant Disease**, v. 78, p. 1011-1016, 1994.

HALL, M.; FRANK, E.; HOLMES, G.; PFAHRINGER, B.; REUTEMANN, P.; WITTEN, I. A. The WEKA data mining software: an update. **ACM SIGKDD explorations newsletter**, v. 11, n. 1, p. 10-18, 2009.

HAN, J.; KAMBER, M.; PEI, J. **Data mining**: concepts and techniques. 3ª edição. San Francisco: Morgan Kaufmann Publishers, 2011. 744 p.

HANLEY, J. A.; MCNEIL, B. J. The meaning and use of the area under a receiver operating characteristic (ROC) curve. **Radiology**, v. 143, n. 1, p. 29-36, 1982.

HARDWICK, N. V. Disease forecasting. In: COOKE, B. M.; JONES, D. G.; KAYE, B. (Eds.). **The epidemiology of plant diseases**. 2ª edição. Wageningen: Springer, 2006. p. 239-267.

HASTIE, T.; TIBSHINARI, R.; FRIEDMAN, J. **The Elements of Statistical Learning**: Data Mining, Inference, and Prediction. 2ª edição. New York:Springer, 2009. 739 p.

HOERL, A. F.; KENNARD, R. W. Ridge regression: Biased estimation for nonorthogonal problems. Technometrics, v. 12, n. 1, p. 55-67, 1970.

JAMES, G.; WITTEN, D.; HASTIE, T.; TIBSHIRANI, R. **An Introduction to Statistical Learning**: with Applications in R. New York: Springer, 2013. 426 p.

KDDNUGGETS. **What main methodology are you using for your analytics, data mining, or data science projects?** 2015. Disponível em: <http://www.kdnuggets.com/polls/2014/analytics-data-mining-data-science-methodology.html>. Acesso em: 25/11/2015.

KOHAVI, R. A Study of Cross-Validation and Bootstrap for Accuracy Estimation and Model Selection. In: International Joint Conference on Artificial Intelligence (IJCAI), 14., v. 2, 1995, Montreal. **Proceedings...** San Francisco: Morgan Kaufmann Publishers Inc., p. 1137-1143, 1995.

KOTSIANTIS, S. Combining bagging, boosting, rotation forest and random subspace methods. **Artificial Intelligence Review**, v. 35, p. 223-240, 2011.

KULKARNI, V. Y.; PETARE, M.; SINHA, P. K. Analyzing random forest classifier with different split measures. In: International Conference on Soft Computing for Problem Solving (SocPros), 2., 2012, Jaipur. **Proceedings...** New Delhi: Springer India, p. 691-699, 2014.

KUSHALAPPA, A. C.; AKUTSU, M.; LUDWIG, A. Application of survival ratio for monocyclic process of *Hemileia vastatrix* in predicting coffee rust infection rates. **Phytopathology**, v. 73, p. 96-103, 1983.

KUSHALAPPA, A. C.; AKUTSU, M.; OSEGUERA, S. H.; CHAVES, G. M.; MELLES, C. Equations for predicting the rate of coffee rust development based on net survival ratio for monocyclic process of Hemileia vastatrix. **Fitopatologia Brasileira**, v. 9, p. 255-271, 1984.

KUSHALAPPA, A. C.; ESKES, A. B. Advances in coffee rust research. **Annual Review Phytopathology**, v. 27, p. 503-531, 1989.

LÓPEZ-BRAVO, D. F.; VIRGINIO-FILHO, E. de M.; AVELINO, J. Shade is conducive to coffee rust as compared to full sun exposure under standardized fruit load conditions. **Crop Protection**, v. 38, p. 21-29, 2012.

LUACES, O.; RODRIGUES, L. H. A.; MEIRA, C. A. A.; BAHAMONDE; A.. Using nondeterministic learners to alert on coffee rust disease. **Expert systems with applications**, v. 38, n. 11, p. 14276-14283, 2011.

MACLIN, R.; OPITZ, D. An empirical evaluation of bagging and boosting. In: National Conference on Artificial Intellingence, 14., 1997, Providence. **Proceedings...**, Providence: AAAI Press, 1997.

MARBÁN, O.; MARISCAL, G.; SEGOVIA, J. A Data Mining & Knowledge Discovery Process Model. In: PONCE, J.; KARAHOCA, A. (Eds.) **DATA MINING AND KNOWLEDGE DISCOVERY IN REAL LIFE APPLICATIONS**. Vienna: In-Teh. 2009. 438 p.

MARISCAL, G.; MARBÁN, O.; FERNÁNDEZ, C. A survey of data mining and knowledge discovery process models and methodologies. **The Knowledge Engineering Review**, v. 25 n. 2, p. 137-166, 2010.

MARTINS, E. M. F. **Sequência de eventos primários do desenvolvimento de Hemileia vastatrix em folhas de cafeeiro com suscetibilidade genética, resistência induzida ou resistência genética**. 149 p. Dissertação (Mestrado em Agronomia) – Escola superior de Agricultura “Luiz de Queiroz”, Universidade de São Paulo, Piracicaba. 1988.

MARTON, Z.; SEIDEL, F.; BALINT-BENCZEDI, F.; BEETZ, M. Ensembles of strong learners for multi-cue classification. **Pattern Recognition Letters**, v. 34, p. 754-761, 2013.

MATIGNON, R. **Data Mining Using SAS® Enterprise Miner™**. 1ª edição. New Jersey: John Wiley & Sons. 2007. 564 p.

MEIRA, C. A. A. **Processo de descoberta de conhecimento em bases de dados para a análise e o alerta de doenças de culturas agrícolas e sua aplicação na ferrugem do cafeeiro**. 198p. Tese (Doutorado em Engenharia Agrícola) - Faculdade de Engenharia Agrícola, Universidade Estadual de Campinas, Campinas. 2008.

MEIRA, C. A. A.; RODRIGUES, L. H. A. Modelos em árvore de decisão para alerta da ferrugem do cafeeiro. In: SIMPÓSIO DE PESQUISA DOS CAFÉS DO BRASIL, 6., 2009, Vitória. **Anais...** Vitória: Consórcio Pesquisa Café, 2009.

MEIRA, C. A. A.; RODRIGUES, L. H. A.; MORAES, S. A. de. Modelos de alerta para o controle da ferrugem-do-cafeeiro em lavouras com alta carga pendente. **Pesquisa Agropecuária Brasileira**, v. 44, n. 3, p. 233-242, 2009.

MEIRA, C. A. A.; THAMADA, T. T.; HOLZHAUSEN, P. P. P. Avaliação do SAFCAFE – Sistema de Alerta da Ferrugem do Cafeeiro em três anos agrícolas. In: Congresso Brasileiro de Pesquisas Cafeeiras, 40., 2014, Serra Negra. **Anais...** Varginha: Fundação Procafé, p. 220-221, 2014.

MONARD, M. C.; BARANAUSKAS, J. A. Conceitos sobre Aprendizado de Máquina. In: REZENDE, S. O. (Coo.) **Sistemas Inteligentes**: Fundamentos e Aplicações. Barueri: Manole. 2003. 525 p.

MONTOYA, R. H.; CHAVES, G. M. Influência da temperatura e da luz na germinação, infectividade e período de geração de *Hemileia vastatrix* Berk. e Br. **Experientiae**, v.18, n.11, p. 239-266, 1974.

MORAES, S. A.; SUGIMORI, M. H.; RIBEIRO, I. J. A.; ORTOLANI, A. A.; PEDRO JUNIOR, M. J. Período de incubação de *Hemileia vastatrix* Berk. e Br. em três regiões do Estado de São Paulo. **Summa Phytopathologica**. Piracicaba, v. 2, n. 1, p. 32-38, 1976.

NUTMAN, F. J.; ROBERTS, F. M. Coffee leaf rust. **PANS Pest Articles & News Summaries**, v. 16, n. 4, p. 606-624, 1970.

OPITZ, D.; MACLIN, R. Popular Ensemble Methods: An Empirical Study. **Journal of Artificial Intelligence Research**, v. 11, p. 169-198, 1999.

PÉREZ-ARIZA, C. B.; . Prediction of coffee rust disease using Bayesian networks. In: European Workshop on Probabilistic Graphical Models, 6., 2012, Granada. **Proceedings...** Granada: University of Granada, p. 259-266, 2012.

PINTO, A. C. S.; POZZA, E. A.; SOUZA, P. E.; POZZA, A. A. A.; TALAMINI, V.; BOLDINI, J. M.; SANTOS, F. S. Descrição da epidemia da ferrugem do cafeeiro com redes neuronais. **Fitopatologia Brasileira**, Brasília, v. 27, n. 5, p. 517-524, 2002.

QUINLAN, J. R. Induction of Decision Trees. **Machine learning**, v. 1, p. 81-106, 1986.

QUINLAN, J. R. Simplifying decision trees. **International journal of man-machine studies**, v. 27, n. 3, p. 221-234, 1987.

QUINLAN, J. R. Bagging, Boosting, and C4.5. In: National Conference on Artificial Intelligence, 13., 1996, Portland. **Proceedings...** Portland: AAAI Press, 1996.

REZENDE, S. O.; PUGLIESI, J. B.; MELANDA, E. A.; PAULA, M. F. de. Mineração de dados. In: REZENDE, S. O. (Org.). **Sistemas inteligentes**: fundamentos e aplicações. Barueri: Editora Manole, 2002. p. 307-335.

ROKACH, L. Ensemble-based classifiers. **Artificial Intelligence Review**, v. 33, n. 1-2, p.1-39, 2010.

ROKACH, L.; MAIMON, O. Top-down induction of decision trees classifiers-a survey. **IEEE Transactions on Systems, Man, and Cybernetics - Part C: Applications and Reviews**, v. 35, n. 4, p. 476-487, 2005.

SAFAVIAN, S. R.; LANDGREBE, D. A survey of decision tree classifier methodology. **IEEE Transactions on Systems, Man, and Cybernetics**, v. 21, n. 3, p. 660-674, 1991 .

SENI, G.; ELDER, J. **Ensemble Methods in Data Mining**: Improving Accuracy Through Combining Predictions. MORGAN & CLAYPOOL PUBLISHERS, 2010. 126 p.

SENTELHAS, P. C.; MARTA, A. D.; ORLANDINI, S.; SANTOS, E. A.; GILLESPIE, T. J.; GLEASON, M. L. Suitability of relative humidity as an estimator of leaf wetness duration. **AGRICULTURAL AND FOREST METEOROLOGY**, v. 148, p. 392-400, 2008.

SIGLETOS, G.; PALIOURAS, G.; SPYROPOULOS, C. D. Combining information extraction systems using voting and stacked generalization. **Journal of Machine Learning Research**, v. 6, p. 1751-1782, 2005.

SUZUKI, E. Data Mining Methods for Discovering Interesting Exceptions from an Unsupervised Table. **Journal of Universal Computer Science**, v. 12, n. 6, p. 627-653, 2006.

TALAMINI, V; POZZA, E. A.; SOUZA, P. E. de; SILVA, A. M. da. Progresso da ferrugem e da cercosporiose em cafeeiro (*Coffea arabica* L.) com diferentes épocas de início e parcelamentos da fertirrigação. **Ciência e Tecnologia**, v. 27, n. 1, p. 141-149, 2003.

THAMADA, T. T.; GIROLAMO NETO, C. di; MEIRA, C. A. A. Sistema de alerta da ferrugem do cafeeiro: resultado de um processo de mineração de dados. In: Congresso Brasileiro de Agroinformática, 9., 2013, Cuiabá. **Anais...** Cuiabá: Editora UFMT, 2013.

TING, K. M.; WITTEN, I. H. Stacking bagged and dagged models. In: International Conference on Machine Learning (ICML), 14., 1997, Nashville, **Proceedings...** San Francisco: Morgan Kaufmann, p. 367-375, 1997a.

TING, K. M.; WITTEN, I. H. Stacked Generalization: when does it work? In: International Joint Conference on Artificial Intelligence, 15., 1997, Nagoya, **Proceedings...** San Francisco: Morgan Kaufmann, p. 866-871, 1997b.

TING, K. M.; WITTEN, I. H. Issues in Stacked Generalization. **Journal of Artificial Intelligence Research**, v. 10, p. 271-289, 1999.

TORRES, F. T. P.; DAGNINO, R. S.; JUNIOR, A. O. **CONTRIBUIÇÕES GEOGRÁFICAS**. UBÁ: Editora Geographica, 2009. 542 p.

USDA. Coffee: World Markets and Trade. **United States Department Of Agriculture**. Junho de 2015. Disponível em: <http://www.fas.usda.gov/psdonline/circulars/coffee.pdf>. Acesso em: 19/10/2015.

VALE, F. X. R.; ZAMBOLIM, L.; JESUS JUNIOR, W. C. Efeito de fatores climáticos na ocorrência e no desenvolvimento da ferrugem do cafeeiro. In: Simpósio de pesquisa dos cafés do Brasil, 1, 2000, Poços de Caldas, **Resumos...** Brasília: EMBRAPA, p. 171-174, 2000.

VAN MAANEN, A.; XU, X.-M. Modelling plant disease epidemics. **European Journal of Plant Pathology**, v. 109, n. 7, p. 669-682, 2003.

VERIKAS, A.; GELZINIS, A.; BACAUSKIENE, M. Mining data with random forests: A survey and results of new tests. **Pattern Recognition**, v. 44, p. 330-349, 2011.

WALLER, J. M. Coffee rust - epidemiology and control. **Crop Protection**, v. 1, n. 4, p. 385-404, 1982.

WALLIN, J. R. Summary of recent progress in predicting late blight epidemics in United States and Canada. **American Potato Journal**, v. 39, n. 8, p. 306-312, 1962.

WITTEN, I. H.; FRANK, E.; HALL, M. A. **Data mining**: practical machine learning tools and techniques. 3ª edição. Burlington: Elsevier Science, 2011. 629 p.

WOLPERT, D. H. Stacked generalization. **Neural networks**, v. 5, p. 241-259, 1992.

WU, X.; KUMAR, V.; QUINLAN, J. R.; GHOSH, J.; YANG, Q.; MOTODA, H.; MCLACHLAN, G. J.; NG, A.; LIU, B.; YU, P. S.; ZHOU, Z.; STEINBACH, M.; HAND, D. J.; STEINBERG, D. Top 10 algorithms in data mining. **Knowledge and Information Systems**, v. 14, p. 1-37, 2008.

ZAMBOLIM, L.; VALE, F. X. R.; PEREIRA, A. A.; CHAVES, G. M. Café (Coffea arabica L.): controle de doenças – doenças causadas por fungos, bactérias e vírus. In: VALE, F. X. R.

do; ZAMBOLIM, L. (Eds.) **Controle de doenças de plantas: grandes culturas**. Viçosa: UFV, 1997. p. 83-139.

ZAMBOLIM, L.; VALE, F. X. R.; COSTA, H.; PEREIRA, A. A.; CHAVES, G. M. Epidemiologia e controle integrado da ferrugem-do-cafeeiro. In: ZAMBOLIM, L. (Ed.) **O estado da arte de tecnologias na produção de café**. Viçosa: Suprema Gráfica e Editora, 2002. p. 369-449.

ZAMBOLIM, L.; VALE, F. X. R.; ZAMBOLIM, E. M. Doenças do cafeeiro. In: KIMATI, H.; AMORIM, L.; REZENDE, J. A.M.; BERGAMIN FILHO, A.; CAMARGO, L. E. A. (Eds.) **Manual de Fitopatologia**: Doenças das Plantas Cultivadas. 4ª edição. v. 2. São Paulo: Agronômica Ceres. 2005. 663 p.

APÊNDICE A – Parâmetros selecionados na busca em *grid*

Este apêndice contém os valores dos parâmetros selecionados para cada técnica de mineração de dados. Qualquer parâmetro não mencionado significa que foram usados os valores padrão; pré-configurados no módulo scikit-learn. A variável *n* que aparece em alguns algoritmos (por exemplo floresta aleatória e *boosting* + árvore de decisão) se refere ao número total de atributos presentes no conjunto de dados.

A.1 Floresta aleatória

Seguem abaixo os valores dos parâmetros utilizados para o algoritmo floresta aleatória (Tabela 45).

Tabela 45. Parâmetros utilizados na busca em grid para floresta aleatória.

parâmetro	descrição	valores
n_estimators	número de árvores na floresta	5; 10; 20; 30; 40; 50; 75; 100; 200; 300; 400; 500; 600; 700; 800; 900; 1000; 1250; 1500; 1750; 2000; 2250; 2500; 2750; 3000
criterion	função que mede a qualidade da ramificação	gini
max_features	número de atributos candidatos para próxima ramificação (parâmetro m)	sqrt(n); $\log_2(n)$; 0,33; 0,66; 0,75
max_depth	altura máxima de cada árvore	ilimitado; 3; 5; 7; 9; 11
min_samples_leaf	número mínimo de registros por folha recém criada	3; 4; 5; 6; 7; 8; 9
bootstrap	uso da amostragem *bootstrap*	verdadeiro
oob_score	uso de amostras *oob* para estimar o erro generalizado	verdadeiro

A.2 *Boosting*

Seguem abaixo os valores dos parâmetros utilizados para o algoritmo *boosting*, tendo como base árvore de decisão e *SVM*, com kernel: linear, polinomial, rbf e sigmóide (Tabelas 46 a 50).

Tabela 46. Parâmetros utilizados na busca em grid para *boosting* com árvore de decisão.

parâmetro	descrição	valores
n_estimators	número máximo de iterações sobre o modelo inicial	5; 10; 25; 50; 100; 200; 300; 400; 500; 600; 700; 800; 900; 1000; 1250; 1500; 1750; 2000; 2250; 2500; 2750; 3000
learning_rate	cada modelo gerado tem seu peso diminuído na votação a essa taxa de aprendizado	0,001; 0,01; 0,1
criterion	função que mede a qualidade da ramificação	gini
max_features	número de atributos candidatos para próxima ramificação	sqrt(n); $\log_2(n)$; 0,3; 0,66
max_depth	altura máxima de cada árvore	3; 5; 7; 9
min_samples_leaf	número mínimo de registros por folha recém criada	5; 6; 7; 8; 9

Tabela 47. Parâmetros utilizados na busca em grid para *boosting* com *SVM* (linear).

parâmetro	descrição	valores
n_estimators	número máximo de iterações sobre o modelo inicial	5; 10; 25; 50; 100; 200; 300; 400; 500; 600; 700; 800; 900; 1000; 1100; 1200; 1300; 1400; 1500; 1600; 1700; 1800; 1900; 2000; 2250; 2500; 2750; 3000
learning_rate	cada modelo gerado tem seu peso diminuído na votação a essa taxa de aprendizado	0,001; 0,005; 0,01; 0,05; 0,1; 0,5; 1
c	penalidade para predição errada	1; 5; 10; 20; 30; 40; 50; 60; 70; 80; 90; 100
kernel	tipo de kernel	linear

Tabela 48. Parâmetros utilizados na busca em grid para *boosting* com *SVM* (polinomial).

parâmetro	descrição	valores
n_estimators	número máximo de iterações sobre o modelo inicial	50; 100; 500; 1000; 1500; 2000
learning_rate	cada modelo gerado tem seu peso diminuído na votação a essa taxa de aprendizado	0,001; 0,01; 0,1; 1
c	penalidade para predição errada	1; 5; 10; 100
kernel	tipo de kernel	poly
degree	grau do polinômio	2; 3; 4; 5; 6; 7; 8
gamma	coeficiente do kernel	1/n; 0,001; 0,01; 0,1; 0

Tabela 49. Parâmetros utilizados na busca em grid para *boosting* com *SVM* (rbf).

parâmetro	descrição	valores
n_estimators	número máximo de iterações sobre o modelo inicial	50; 100; 250; 500; 750; 1000; 1250; 1500; 1750; 2000
learning_rate	cada modelo gerado tem seu peso diminuído na votação a essa taxa de aprendizado	0,001; 0,005; 0,01; 0,05; 0,1; 0,5; 1
c	penalidade para predição errada	1; 5; 10; 25; 50; 75; 100
kernel	tipo de kernel	rbf
gamma	coeficiente do kernel	1/n; 0,001; 0,01; 0,1; 0

Tabela 50. Parâmetros utilizados na busca em grid para *boosting* com *SVM* (sigmóide).

parâmetro	descrição	valores
n_estimators	número máximo de iterações sobre o modelo inicial	50; 100; 250; 500; 750; 1000; 1250; 1500; 1750; 2000
learning_rate	cada modelo gerado tem seu peso diminuído na votação a essa taxa de aprendizado	0,001; 0,005; 0,01; 0,05; 0,1; 0,5; 1
c	penalidade para predição errada	1
kernel	tipo de kernel	sigmoid
gamma	coeficiente do kernel	1/n; 0,001; 0,01; 0,1
coef0	termo independente da função do kernel	0; 1; 2; 3; 4; 5; 6; 7; 8; 9

A.3 *Bagging*

Seguem abaixo os valores dos parâmetros utilizados para o algoritmo *bagging*, tendo como base árvore de decisão e *SVM*, com kernel: linear, polinomial, rbf e sigmóide (Tabelas 51 a 55).

Tabela 51. Parâmetros utilizados na busca em grid para *bagging* com árvore de decisão.

parâmetro	**descrição**	**valores**
n_estimators	número de modelos	5; 10; 25; 50; 100; 200; 300; 400; 500; 600; 700; 800; 900; 1000; 1100; 1200; 1300; 1400; 1500; 1750; 2000; 2250; 2500; 3000
max_features	número de atributos candidatos para próxima ramificação	1
bootstrap	uso da amostragem *bootstrap*	verdadeiro
oob_score	uso de amostras *oob* para estimar o erro generalizado	verdadeiro
max_depth	altura máxima de cada árvore	3; 4; 5; 6; 7; 9; 11
min_samples_leaf	número mínimo de registros por folha recém criada	4; 5; 6; 7; 8; 9

Tabela 52. Parâmetros utilizados na busca em grid para *bagging* com *SVM* (linear).

parâmetro	**descrição**	**valores**
n_estimators	número de modelos	5; 10; 25; 50; 75; 100; 150; 200; 250; 300; 350; 400; 450; 500; 550; 600; 650; 700; 750; 800; 850; 900; 950; 1000; 1200; 1300; 1400; 1500; 1600; 1700; 1800; 1900; 2000; 2250; 2500; 2750; 3000
max_features	número de atributos candidatos para próxima ramificação	sqrt(n); 0,1; 0,66; 0,75; 1
bootstrap	uso da amostragem *bootstrap*	verdadeiro
oob_score	uso de amostras *oob* para estimar o erro generalizado	verdadeiro
c	penalidade para predição errada	1; 5; 10; 20; 30; 40; 50; 75; 100; 200; 300; 400; 500; 600; 700; 800; 900; 1000
kernel	tipo de kernel	linear

Tabela 53. Parâmetros utilizados na busca em grid para *bagging* com *SVM* (polinomial).

parâmetro	**descrição**	**valores**
n_estimators	número de modelos	50; 100; 250; 500; 750; 1000; 1500; 2000
max_features	número de atributos candidatos para próxima ramificação	sqrt(n); 0,67; 1
bootstrap	uso da amostragem *bootstrap*	verdadeiro
oob_score	uso de amostras *oob* para estimar o erro generalizado	verdadeiro
c	penalidade para predição errada	1; 5; 10; 25; 100
kernel	tipo de kernel	poly
degree	grau do polinômio	2; 3; 4; 5; 6; 7; 8
gamma	coeficiente do kernel	1/n; 0,001; 0,01; 0,1; 0

Tabela 54. Parâmetros utilizados na busca em grid para *bagging* com *SVM* (rbf).

parâmetro	**descrição**	**valores**
n_estimators	número de modelos	5; 10; 50; 100; 200; 300; 400; 500; 600; 700; 800; 900; 1000; 1500; 2000; 2500; 3000
max_features	número de atributos candidatos para próxima ramificação	sqrt(n); 0,67; 0,75; 1
bootstrap	uso da amostragem *bootstrap*	verdadeiro
oob_score	uso de amostras *oob* para estimar o erro generalizado	verdadeiro
c	penalidade para predição errada	1; 5; 10; 20; 30; 40; 50; 75; 100
kernel	tipo de kernel	rbf
gamma	coeficiente do kernel	1/n; 0,001; 0,01; 0,05; 0,1; 0,5; 0

Tabela 55. Parâmetros utilizados na busca em grid para *bagging* com *SVM* (sigmóide).

parâmetro	descrição	valores
n_estimators	número de modelos	5; 10; 50; 100; 200; 300; 400; 500; 600; 700; 800; 900; 1000; 1250; 1500; 1750; 2000; 2250; 2500; 2750; 3000
max_features	número de atributos candidatos para próxima ramificação	sqrt(n); 0,67; 0,75; 1
bootstrap	uso da amostragem *bootstrap*	verdadeiro
oob_score	uso de amostras *oob* para estimar o erro generalizado	verdadeiro
c	penalidade para predição errada	1
kernel	tipo de kernel	sigmoid
gamma	coeficiente do kernel	1/n; 0,001; 0,01; 0,1; 0
coef0	termo independente da função do kernel	-5; -4; -3; -2; -1; 0; 1; 2; 3; 4; 5

A.4 *Stacking*

Seguem abaixo os valores dos parâmetros utilizados para o algoritmo *stacking* (Tabelas 56 a 60). Os valores usados foram os mesmos para os dois conjuntos de dados nível-1 (tudo e simples) utilizados para modelar o meta-classificador, ou modelo nível-1. Os algoritmos usados para criar o meta-classificador foram: perceptron, passivo-agressivo, ridge, regressão logística e *SVM* (kernel: linear e polinomial).

Tabela 56. Parâmetros utilizados na busca em grid para *stacking* (perceptron).

parâmetro	descrição	valores
penalty	penalidade (termo de regularização)	nenhuma; l2; l1; elasticnet
alpha	constante que multiplica o termo da penalidade (alpha)	0,00001; 0,0001; 0,001; 0,01; 0,1
fit_intercept	uso de constante (viés) nos dados, se falso; assume-se que os dados já estão normalizados	verdadeiro; falso
n_iter	número de iterações sobre o conjunto de dados (*epoch*)	1; 2; 5; 10; 25; 50; 100
eta0	constante que multiplica as atualizações	1; 2; 3; 4; 5

Tabela 57. Parâmetros utilizados na busca em grid para *stacking* (passivo-agressivo).

parâmetro	descrição	valores
c	tamanho máximo do passo (regularização)	0,001; 0,01; 0,1; 1; 10; 100
fit_intercept	uso de constante (viés) nos dados, se falso; assume-se que os dados já estão normalizados	verdadeiro; falso
n_iter	número de iterações sobre o conjunto de dados (*epoch*)	1; 2; 5; 10; 25; 50; 100
loss	função de perda	squared_hinge; hinge

Tabela 58. Parâmetros utilizados na busca em grid para stacking (regressão logística).

parâmetro	descrição	valores
c	quanto menor o valor de c mais forte é regularização	0,1; 1; 10; 25; 50; 100
fit_intercept	uso de constante (viés) na função de decisão	verdadeiro; falso
solver	algoritmo a ser usado no problema de otimização	newton-cg; lbfgs; liblinear
tol	tolerância para critério de parada	0,0001; 0,0005; 0,001; 0,005; 0,01; 0,05; 0,1; 0,5; 1; 5

Tabela 59. Parâmetros utilizados na busca em grid para *stacking* (ridge).

parâmetro	descrição	valores
alpha	pequenos valores positivos para alpha reduz a variância do modelo	0,25; 0,5; 0,75; 1; 1,25; 1,5; 2,5
fit_intercept	uso de constante (viés) nos dados, se falso; assume-se que os dados já estão normalizados	verdadeiro; falso
normalize	normalização dos dados	verdadeiro; falso
solver	uso de rotinas computacionais	cholesky; lsqr; sparse_cg
tol	precisão na solução encontrada	0,0001; 0,0005; 0,001; 0,005; 0,01; 0,05; 0,1; 0,5; 1; 5

Tabela 60. Parâmetros utilizados na busca em grid para *stacking* (*SVM* - kernel: polinomial).

parâmetro	descrição	valores
c	penalidade para predição errada	0,1; 1; 10; 25; 50; 100
kernel	tipo de kernel	poly
degree	grau do polinômio	2; 3; 4
gamma	coeficiente do kernel	1/n; 0,1
coef0	termo independente da função do kernel	0; 1; 2
shrinking	uso de heurística de diminuição	verdadeiro; falso
tol	tolerância para critério de parada	0,001; 0,01; 0,1; 1

Tabela 61. Parâmetros utilizados na busca em grid para *stacking* (*SVM* - kernel: linear).

parâmetro	descrição	valores
c	penalidade para predição errada	0,1; 1; 10; 25; 50; 100
kernel	tipo de kernel	linear
shrinking	uso de heurística de diminuição	verdadeiro; falso
tol	tolerância para critério de parada	0,0001; 0,0005; 0,001; 0,005; 0,01; 0,05; 0,1; 0,5; 1; 5

A Tabela 62 mostra a quantidade total de combinações de parâmetros testadas para cada técnica e de modo geral. Os números para *stacking* foram menores uma vez que os valores se referem à busca em *grid* para o meta-classificador (modelo nível-1), que não necessita de maiores esforços computacionais já que a maior parte do processamento ocorreu no nível-0 (WITTEN et al., 2011). No total foram testadas 42686 combinações distintas.

Tabela 62. Quantidade de combinações de parâmetros testados em cada algoritmo por meio da busca em grid.

algoritmo	combinações
Floresta Aleatória	5250
Boosting + Árvore de Decisão	5280
Boosting + *SVM* (linear)	2352
Boosting + *SVM* (polinomial)	3360
Boosting + *SVM* (rbf)	2450
Boosting + *SVM* (sigmóide)	2800
Bagging + Árvore de Decisão	1008
Bagging + *SVM* (*linear*)	3330
Bagging + *SVM* (polinomial)	4200
Bagging + *SVM* (rbf)	4284
Bagging + *SVM* (sigmóide)	4620
Stacking (perceptron)	1400
Stacking (*passivo-agressivo*)	168
Stacking (*ridge*)	840
Stacking (regressão logística)	360
Stacking (*SVM*)	120
Stacking (*SVM*)	864
TOTAL	42686

A.5 Modelos padrão

A seguir são apresentados os valores de parâmetros dos dez melhores *ensembles* desenvolvidos para cada cenário de modelagem (seção 4.2.6), além dos modelos de melhor desempenho preditivo dos períodos críticos.

Cenário: alta/5 p.p.

Conjunto de atributos 01

modelo 1 – FA – n_estimators: 30; m: 0,75; max_depth: 3; min_samples_leaf: 5

modelo 2 – FA – n_estimators: 75; m: 0,75; max_depth: 3; min_samples_leaf: 8

modelo 3 – FA – n_estimators:75; m: sqrt(n); max_depth: 5; min_samples_leaf: 5

modelo 4 – *Boosting* + AD – n_estimators: 50; learning_rate: 0,001; m: sqrt(n); max_depth: 5; min_samples_leaf:7

modelo 5 – *Boosting* + AD – n_estimators: 25; learning_rate: 0,001; m: 0,66; max_depth: 3; min_samples_leaf: 7

modelo 6 – *Boosting* + AD – n_estimators: 5; learning_rate: 0,01; m: 0,66; max_depth: 3; min_samples_leaf: 8

modelo 7 – *Bagging* + AD – n_estimators: 400; max_depth: 3; min_samples_leaf: 5

modelo 8 – *Bagging* + AD – n_estimators: 100; max_depth: 3; min_samples_leaf: 5

modelo 9 – *Bagging* + AD – n_estimators: 1400; max_depth: 3; min_samples_leaf: 5

modelo 10 – *Bagging* + AD – n_estimators: 100; max_depth: 4; min_samples_leaf: 4

Conjunto de atributos 02

modelo 11 – FA – n_estimators: 300; m: 0,33; max_depth: 9; min_samples_leaf: 3

modelo 12 – FA – n_estimators:10; m: + Log2; max_depth:7; min_samples_leaf: 3

modelo 13 – FA – n_estimators: 50; m: 0,33; max_depth: 9; min_samples_leaf: 3

modelo 14 – FA – n_estimators: 2250; m: 0,33; max_depth: 9; min_samples_leaf: 3

modelo 15 – FA – n_estimators: 2750; m: 0,33; max_depth: 9; min_samples_leaf: 3

modelo 16 – FA – n_estimators: 2250; m: sqrt(n); max_depth: 9; min_samples_leaf: 3

modelo 17 – FA – n_estimators: 300; m: sqrt(n); max_depth: ilimitado; min_samples_leaf: 4

modelo 18 – FA – n_estimators: 1500; m: 0,33; max_depth: 9; min_samples_leaf: 3

modelo 19 – *Boosting* + *SVM(p)* – n_estimators: 100; learning_rate: 1; c: 100; degree:3; gamma: 0,1

modelo 20 – Stack-t + *SVM(p)* – c: 100; degree: 4; gamma: 1/n; coef0: 2; shrinking: verdadeiro; tol: 0,1

Conjunto de atributos 03

modelo 21 – Stack-t + Ridge – alpha: 0,5; fit_intercept: verdadeiro; normalize: verdadeiro; solver: lsqr; tol: 5

modelo 22 – Stack-t + Ridge; alpha: 0,5; fit_intercept: verdadeiro; normalize: verdadeiro; solver: cholesky; tol: 0,0001

modelo 23 – FA – n_estimators: 200; m: sqrt(n); max_depth: 5; min_samples_leaf: 3

modelo 24 – FA – n_estimators: 50; m: sqrt(n); max_depth: 5; min_samples_leaf: 4

modelo 25 – FA – n_estimators: 2000; m: sqrt(n); max_depth: 5; min_samples_leaf: 4

modelo 26 – FA – n_estimators: 30; m: sqrt(n); max_depth: 5; min_samples_leaf: 5

modelo 27 – *Boosting* + *SVM*(p) – n_estimators: 2000; learning_rate: 0,01; C: 1; degree: 5; gamma: 0,1

modelo 28 – *Boosting* + *SVM*(p) – n_estimators: 100; learning_rate: 1; C: 5; degree: 5; gamma: 0,1

modelo 29 – *Boosting* + *SVM*(p) – n_estimators: 1000; learning_rate: 0,01; C: 1; degree: 5; gamma: 0,1

modelo 30 – *Bagging* + AD – n_estimators: 25; max_depth: 3; min_samples_leaf: 4

Cenário: alta/10 p.p.

Conjunto de atributos 01

modelo 1 – *Bagging* + AD – n_estimators: 5; max_depth: 4; min_samples_leaf: 5

modelo 2 – *Bagging* + AD – n_estimators: 10; max_depth: 5; min_samples_leaf: 7

modelo 3 – *Boosting* + AD – n_estimators: 5; learning_rate: 0,001; m: 0,66; max_depth: 7; min_samples_leaf: 7

modelo 4 – *Boosting* + AD – n_estimators: 5; learning_rate: 0,1; m: 0,66; max_depth: 5; min_samples_leaf: 9

modelo 5 – FA – n_estimators: 20; m: 0,66; max_depth: 7; min_samples_leaf: 6

modelo 6 – FA – n_estimators: 100; m: 0,75; max_depth: 3; min_samples_leaf: 6

modelo 7 – FA – n_estimators: 40; m: 0,66; max_depth: 5; min_samples_leaf: 7

modelo 8 – FA – n_estimators: 40; m: 0,75; max_depth: 7; min_samples_leaf: 8

modelo 9 – FA – n_estimators: 30; m: 0,66; max_depth: 3; min_samples_leaf: 9

modelo 10 – FA – n_estimators: 5; m: 0,75; max_depth: ilimitado; min_samples_leaf: 7

Conjunto de atributos 02

modelo 11 – Stack-t + *SVM*(p) – C: 100; degree: 4; gamma: 1/n; coef0: 2; shrinking: verdadeiro; tol: 1

modelo 12 – Stack-t + *SVM*(p) – C: 25; degree: 3; gamma: 1/n; coef0: 2; shrinking: verdadeiro; tol: 1

modelo 13 – Stack-t + *SVM*(p) – C: 50; degree: 4; gamma: 1/n; coef0: 2; shrinking: verdadeiro; tol: 1

modelo 14 – Stack-t + *SVM*(p) – C: 25; degree: 4; gamma: 1/n; coef0: 1; shrinking: verdadeiro; tol: 1

modelo 15 – *Bagging* + *SVM*(p) – n_estimators: 100; m: 1; C: 5; degree: 3; gamma: 1/n

modelo 16 – *Bagging* + *SVM*(p) – n_estimators: 500; m: 1; C: 5; degree: 3; gamma: 0

modelo 17 – *Bagging* + *SVM*(p) – n_estimators: 2000; m: 0,67; C: 10; degree: 3; gamma: 0,1

modelo 18 – Stack-t + Ridge – alpha: 0,5; fit_intercept: verdadeiro; normalize: verdadeiro; solver: lsqr; tol: 0,5

modelo 19 – Stack-t + Ridge – alpha: 0,75; fit_intercept: verdadeiro; normalize: verdadeiro; solver: lsqr; tol: 0,5

modelo 20 – Stack-t + Ridge – alpha: 0,75; fit_intercept: verdadeiro; normalize: verdadeiro; solver: lsqr; tol: 1

Conjunto de atributos 03

modelo 21 – Stack-t + Ridge – alpha: 1,5; fit_intercept: verdadeiro; normalize: verdadeiro; solver: sparse_cg; tol: 0,05

modelo 22 – Stack-t + Ridge – alpha: 1,5; fit_intercept: verdadeiro; normalize: verdadeiro; solver: lsqr; tol: 5

modelo 23 – Stack-t + Ridge – alpha: 1,25; fit_intercept: verdadeiro; normalize: verdadeiro; solver: lsqr; tol: 5

modelo 24 – Stack-t + Ridge – alpha: 1,25; fit_intercept: verdadeiro; normalize: verdadeiro; solver: lsqr; tol: 0,01

modelo 25 – Stack-t + Ridge – alpha: 0,25; fit_intercept: verdadeiro; normalize: verdadeiro; solver: lsqr; tol: 0,05

modelo 26 – Stack-t + *SVM*(p) – C: 0,1; degree: 3; gamma: 0,1; coef0: 2; shrinking: verdadeiro; tol: 0,1

modelo 27 – Stack-t + *SVM*(p) – C: 0,1; degree: 3; gamma: 0,1; coef0: 2; shrinking: verdadeiro; tol: 1

modelo 28 – Stack-t + *SVM*(p) – C: 1; degree: 2; gamma: 0,1; coef0: 2; shrinking: verdadeiro; tol: 1

modelo 29 – *Boosting* + *SVM*(p) – n_estimators: 2000; learning_rate: 0,1; C: 1; degree: 7; gamma: 1/n

modelo 30 – *Boosting* + *SVM*(p) – n_estimators: 1000; learning_rate: 0,01; C: 10; degree: 5; gamma: 1/n

<u>Cenário: baixa/3 p.p.</u>

Conjunto de atributos 01

modelo 1 – Stack-t + *SVM*(p) – C: 1; degree: 3; gamma: 0,1; coef0: 0; shrinking: verdadeiro; tol: 1

modelo 2 – Stack-t + Log – C: 100; fit_intercept: verdadeiro; solver: newton-cg; tol: 1

modelo 3 – Stack-t + Log – C: 100; fit_intercept: falso; solver: newton-cg; tol: 0,1

modelo 4 – Stack-t + Log – C: 100; fit_intercept: verdadeiro; solver: newton-cg; tol: 0,1

modelo 5 – Stack-s + *SVM*(p) – C: 50; degree: 2; gamma: 0,1; coef0: 2; shrinking: verdadeiro; tol: 1

modelo 6 – Stack-s + *SVM*(p) – C: 100; degree: 4; gamma: 1/n; coef0: 2; shrinking: verdadeiro; tol: 1

modelo 7 – Stack-t + Ridge – alpha: 2,5; fit_intercept: verdadeiro; normalize: verdadeiro; solver: cholesky; tol: 0,0001

modelo 8 – Stack-t + Ridge – alpha: 0,75; fit_intercept: verdadeiro; normalize: verdadeiro; solver: lsqr; tol: 0,1

modelo 9 – Stack-t + Ridge – alpha: 2,5; fit_intercept: verdadeiro; normalize: verdadeiro; solver: lsqr; tol: 0,01

modelo 10 – Stack-t + Ridge – alpha: 2,5; fit_intercept: verdadeiro; normalize: verdadeiro; solver: sparse_cg; tol: 0,1

Conjunto de atributos 02

modelo 11 – Stack-s + *SVM*(p) – C: 0,1; degree: 3; gamma: 0,1; coef0: 1; shrinking: verdadeiro; tol: 1

modelo 12 – Stack-t + Log – C: 1; fit_intercept: verdadeiro; solver: lbfgs; tol: 0,5

modelo 13 – Stack-t + Log – C: 100; fit_intercept: verdadeiro; solver: newton-cg; tol: 0,05

modelo 14 – Stack-t + Ridge – alpha: 0,5; fit_intercept: verdadeiro; normalize: verdadeiro; solver: lsqr; tol: 0,05

modelo 15 – Stack-t + Ridge – alpha: 0,25; fit_intercept: verdadeiro; normalize: verdadeiro; solver: lsqr; tol: 0,1

modelo 16 – Stack-t + Ridge – alpha: 0,25; fit_intercept: verdadeiro; normalize: verdadeiro; solver: lsqr; tol: 0,05

modelo 17 – Stack-t + Ridge – alpha: 0,75; fit_intercept: verdadeiro; normalize: verdadeiro; solver: lsqr; tol: 0,05

modelo 18 – Stack-t + *SVM*(p) – C: 100; degree: 2; gamma: 0,1; coef0: 0; shrinking: falso; tol: 0,1

modelo 19 – Stack-t + *SVM*(p) – C: 100; degree: 2; gamma: 0,1; coef0: 0; shrinking: verdadeiro; tol: 1

modelo 20 – Stack-t + *SVM*(p) – C: 0,1; degree: 3; gamma: 0,1; coef0: 0; shrinking: verdadeiro; tol: 0,1

Conjunto de atributos 03

modelo 21 – Stack-s + Log – C: 1; fit_intercept: verdadeiro; solver: lbfgs; tol: 0,5

modelo 22 – Stack-s + Ridge – alpha: 1,25; fit_intercept: verdadeiro; normalize: verdadeiro; solver: lsqr; tol: 5

modelo 23 – Stack-s + *SVM*(l); C: 10; shrinking: verdadeiro; tol: 0,01

modelo 24 – Stack-t + Log – C: 1; fit_intercept: verdadeiro; solver: newton-cg; tol: 0,5

modelo 25 – Stack-t + Log – C: 10; fit_intercept: verdadeiro; solver: newton-cg; tol: 0,5

modelo 26 – Stack-t + Ridge – alpha: 0,75; fit_intercept: verdadeiro; normalize: verdadeiro; solver: sparse_cg; tol: 0,5

modelo 27 – Stack-t + *SVM*(p) – C: 0,1; degree: 3; gamma: 0,1; coef0: 0; shrinking: verdadeiro; tol: 1

modelo 28 – Stack-t + *SVM*(p) – C: 100; degree: 4; gamma: 1/n; coef0: 1; shrinking: verdadeiro; tol: 1

modelo 29 – Stack-t + *SVM*(p) – C: 25; degree: 4; gamma: 1/n; coef0: 1; shrinking: verdadeiro; tol: 1

modelo 30 – Stack-t + *SVM*(p) – C: 1; degree: 2; gamma: 0,1; coef0: 2; shrinking: verdadeiro; tol: 1

Cenário: baixa/5 p.p.

Conjunto de atributos 01

modelo 1 – Stack-t + Log – C: 50; fit_intercept: falso; solver: lbfgs; tol: 1

modelo 2 – Stack-t + Log – C: 25; fit_intercept: falso; solver: lbfgs; tol: 1

modelo 3 – Stack-t + Log – C: 10; fit_intercept: falso; solver: lbfgs; tol: 1

modelo 4 – Stack-t + *SVM*(p) – C: 1; degree: 3; gamma: 0.1; coef0: 0; shrinking: verdadeiro; tol: 0.1

modelo 5 – Stack-t + *SVM*(p) – C: 0,1; degree: 4; gamma: 0.1; coef0: 1; shrinking: verdadeiro; tol: 1

modelo 6 – Stack-t + *SVM*(p) – C: 1; degree: 3; gamma: 0.1; coef0: 0; shrinking: verdadeiro; tol: 0.01

modelo 7 – Stack-t + Ridge – alpha: 1,25; fit_intercept: verdadeiro; normalize: verdadeiro; solver: lsqr; tol: 0,05

modelo 8 – Stack-t + Ridge – alpha: 1; fit_intercept: verdadeiro; normalize: verdadeiro; solver: lsqr; tol: 0,05

modelo 9 – Stack-t + Ridge – alpha: 1; fit_intercept: verdadeiro; normalize: verdadeiro; solver: sparse_cg; tol: 0,1

modelo 10 – Stack-t + Ridge – alpha: 1,25; fit_intercept: verdadeiro; normalize: verdadeiro; solver: cholesky; tol: 0,0001

Conjunto de atributos 02

modelo 11 – Stack-s + Log – C: 100; fit_intercept: verdadeiro; solver: newton-cg; tol: 0,1

modelo 12 – Stack-t + Log – C: 10; fit_intercept: verdadeiro; solver: lbfgs; tol: 1

modelo 13 – Stack-t + Log – C: 10; fit_intercept: falso; solver: lbfgs; tol: 1

modelo 14 – Stack-t + Log – C: 100; fit_intercept: falso; solver: lbfgs; tol: 1

modelo 15 – Stack-t + Log – C: 10; fit_intercept: verdadeiro; solver: newton-cg; tol: 0,5

modelo 16 – Stack-t + Ridge – alpha: 2,5; fit_intercept: verdadeiro; normalize: verdadeiro; solver: 'lsqr'; tol: 0,05

modelo 17 – Stack-t + Ridge – alpha: 1; fit_intercept: verdadeiro; normalize: verdadeiro; solver: 'lsqr'; tol: 0,05

modelo 18 – Stack-t + Ridge – alpha: 0,25; fit_intercept: verdadeiro; normalize: verdadeiro; solver: 'lsqr'; tol: 0,5

modelo 19 – Stack-t + *SVM*(p) – C: 100; degree: 3; gamma: 1/n; coef0: 2; shrinking: verdadeiro; tol: 1

modelo 20 – Stack-t + *SVM*(p) – C: 25; degree: 4; gamma: 1/n; coef0: 2; shrinking: verdadeiro; tol: 1

Conjunto de atributos 03

modelo 21 – Stack-t + Ridge – alpha: 1; fit_intercept: verdadeiro; normalize: verdadeiro; solver: lsqr; tol: 0,05

modelo 22 – Stack-t + Ridge – alpha: 0,75; fit_intercept: verdadeiro; normalize: verdadeiro; solver: lsqr; tol: 0,05

modelo 23 – Stack-t + Ridge – alpha: 1,5; fit_intercept: verdadeiro; normalize: verdadeiro; solver: sparse_cg; tol: 0,1

modelo 24 – Stack-t + Ridge – alpha: 0,5; fit_intercept: verdadeiro; normalize: verdadeiro; solver: lsqr; tol: 0,05

modelo 25 – Stack-t + Ridge – alpha: 0,25; fit_intercept: verdadeiro; normalize: falso; solver: lsqr; tol: 0,1

modelo 26 – Stack-t + Log – C: 100; fit_intercept: verdadeiro; solver: lbfgs; tol: 1

modelo 27 – Stack-t + Log – C: 50; fit_intercept: verdadeiro; solver: lbfgs; tol: 1

modelo 28 – Stack-t + Log – C: 50; fit_intercept: falso; solver: lbfgs; tol: 1

modelo 29 – Stack-t + Log – C: 25; fit_intercept: falso; solver: lbfgs; tol: 1

modelo 30 – Stack-t + Perceptron – penalty: nenhuma; alpha: 0,00001; fit_intercept: verdadeiro: n_iter: 25: eta0: 1

A.6 Melhores *ensembles* para período críticos

Esta seção apresenta os valores de parâmetros dos *ensembles* de desempenho superior aos modelos padrão (Apêndice A.5) quanto à predição dos períodos críticos para desenvolvimento da ferrugem do cafeeiro (seção 4.4). O desempenho preditivos dos *ensembles* desta seção estão apresentados nas seções 5.2 a 5.5.

Alguns casos como, por exemplo, em alta/5 p.p. não foram desenvolvidos *ensembles* a partir do conjunto de atributos 03 com performance superior aos modelos padrão criados a partir dos mesmos atributos.

Cenário: alta/5 p.p.

Conjunto de atributos 01

modelo 31 – FA – n_estimators: 50; m: sqrt(n); max_depth: 9; min_samples_leaf: 5

modelo 32 – *Boosting* + *SVM*(p) – n_estimators: 1500; learning_rate: 0,001; C: 100; degree: 2; gamma: 0,1

modelo 33 – Stack-t + Ridge – alpha: 0,25; fit_intercept: verdadeiro; normalize: verdadeiro; solver: lsqr; tol: 0,5

modelo 34 – Stack-t + *SVM*(l) – C: 25; shrinking: falso; tol: 0,0001

Conjunto de atributos 02

modelo 35 – FA – n_estimators: 400; m: 0,33; max_depth: 9; min_samples_leaf: 3

modelo 36 – *Boosting* + AD – n_estimators: 50; learning_rate: 0,001; m: sqrt(n); max_depth: 5; min_samples_leaf: 7

modelo 37 – *Boosting* + *SVM*(l) – n_estimators: 100; learning_rate: 0,05; C: 40

modelo 38 – Stack-s + *SVM*(p) – C: 100; degree: 4; gamma: 1/n; coef0: 2; shrinking: verdadeiro; tol: 0,1

Cenário: alta/10 p.p.

Conjunto de atributos 01

modelo 31 – FA – n_estimators: 20; m: 0,75; max_depth: 3; min_samples_leaf: 5

modelo 32 – FA – n_estimators: 20; m: 0,66; max_depth: 5; min_samples_leaf: 3

Conjunto de atributos 02

modelo 33 – *Boosting* + *SVM*(p) – n_estimators: 1000; learning_rate: 0,1; C: 100; degree: 5; gamma: 0,1

modelo 34 – Stack-t + *SVM*(p) – C: 50; degree: 4; gamma: 1/n; coef0: 2; shrinking: verdadeiro; tol: 0,1

Conjunto de atributos 03

modelo 35 – *Bagging* + AD – n_estimators: 5; max_depth: 6; min_samples_leaf: 8

<u>Cenário: baixa/3 p.p.</u>

Conjunto de atributos 01

modelo 31 – *Boosting* + AD – n_estimators: 5; learning_rate: 0,001; m: 0,3; max_depth: 5; min_samples_leaf: 7

modelo 32 – *Bagging* + AD – n_estimators: 1750; max_depth: 3; min_samples_leaf: 4

modelo 33 – Stack-s + *SVM*(p) – C: 50; degree: 2; gamma: 0,1; coef0: 1; shrinking: verdadeiro; tol: 1

Conjunto de atributos 02

modelo 34 – Stack-t + *SVM*(p) – C: 100; degree: 2; gamma: 0,1; coef0: 0; shrinking: falso; tol: 1

Conjunto de atributos 03

modelo 35 – *Boosting* + AD – n_estimators: 10; learning_rate: 0,01; m: + Log2; max_depth: 3; min_samples_leaf: 6

modelo 36 – *Boosting* + AD – n_estimators: 5; learning_rate: 0,1; m: + Log2; max_depth: 3; min_samples_leaf: 7

<u>Cenário: baixa/5 p.p.</u>

Conjunto de atributos 02

modelo 31 – Stack-t + *SVM*(p) – C: 25; degree: 4; gamma: 1/n; coef0: 2; shrinking: verdadeiro; tol: 0,1

Conjunto de atributos 03

modelo 32 – Stack-t + *SVM*(p) – C: 1; degree: 4; gamma: 0,1; coef0: 1; shrinking: falso; tol: 0,1

APÊNDICE B – Início da epidemia da ferrugem no cafeeiro

Este apêndice detalha os meses que apresentaram pela primeira vez um valor de taxa de progresso acima ou igual ao limiar relacionado à carga pendente. Também estão representados nas Tabelas 63 a 66 os anos agrícolas em que ocorreu a ferrugem tardia considerando que o início da epidemia foi a partir de março, baseado em Chalfoun e Carvalho (1999).

B.1 Carga pendente: alta – limiar: 5 p.p.

Tabela 63. Meses de primeira ocorrência de TP ≥ 5 p.p. em lavouras de alta carga pendente entre dezembro e junho. Estão destacados os anos de ocorrência da ferrugem tardia.

estação	mês	ano	ano agrícola	taxa de progresso (p.p.)
Boa Esperança	ABR	2010	2009-10	9,5
Boa Esperança	DEZ	2010	2010-11	7,0
Boa Esperança	MAR	2012	2011-12	13,0
Boa Esperança	ABR	2013	2012-13	11,0
Boa Esperança	MAR	2014	2013-14	9,0
Carmo de Minas	JAN	2007	2006-07	6,0
Carmo de Minas	JAN	2008	2007-08	5,0
Carmo de Minas	DEZ	2008	2008-09	17,0
Carmo de Minas	JAN	2010	2009-10	11,5
Carmo de Minas	DEZ	2010	2010-11	7,0
Carmo de Minas	FEV	2012	2011-12	8,0
Carmo de Minas	DEZ	2012	2012-13	12,0
Carmo de Minas	DEZ	2013	2013-14	9,0
Varginha	JAN	1999	1998-99	9,5
Varginha	FEV	2000	1999-00	6,0
Varginha	JAN	2001	2000-01	13,0
Varginha	DEZ	2001	2001-02	5,0
Varginha	JAN	2003	2002-03	6,0
Varginha	DEZ	2003	2003-04	7,5
Varginha	DEZ	2004	2004-05	6,5
Varginha	DEZ	2005	2005-06	13,0
Varginha	DEZ	2006	2006-07	13,0
Varginha	MAR	2008	2007-08	9,0
Varginha	DEZ	2008	2008-09	22,5
Varginha	DEZ	2009	2009-10	5,5
Varginha	DEZ	2010	2010-11	6,0
Varginha	JAN	2012	2011-12	6,5
Varginha	DEZ	2012	2012-13	10,0
Varginha	DEZ	2013	2013-14	6,5

B.2 Carga pendente: alta – limiar: 10 p.p.

Tabela 64. Meses de primeira ocorrência de TP ≥ 10 p.p. em lavouras de alta carga pendente entre dezembro e junho.

estação	mês	ano	ano agrícola	taxa de progresso (p.p.)
Boa Esperança	MAI	2010	2009-10	31,0
Boa Esperança	JAN	2011	2010-11	31,0
Boa Esperança	MAR	2012	2011-12	13,0
Boa Esperança	ABR	2013	2012-13	21,0
Boa Esperança	ABR	2014	2013-14	17,0
Carmo de Minas	FEV	2007	2006-07	17,5
Carmo de Minas	ABR	2008	2007-08	29,0
Carmo de Minas	DEZ	2008	2008-09	17,0
Carmo de Minas	JAN	2010	2009-10	11,5
Carmo de Minas	JAN	2011	2010-11	27,0
Carmo de Minas	MAR	2012	2011-12	13,0
Carmo de Minas	DEZ	2012	2012-13	12,0
Carmo de Minas	ABR	2014	2013-14	18,0
Varginha	MAI	1999	1998-99	10,5
Varginha	MAR	2000	1999-00	11,5
Varginha	JAN	2001	2000-01	13,0
Varginha	JAN	2002	2001-02	16,0
Varginha	ABR	2003	2002-03	21,5
Varginha	JAN	2004	2003-04	12,5
Varginha	FEV	2005	2004-05	16,5
Varginha	DEZ	2005	2005-06	13,0
Varginha	DEZ	2006	2006-07	13,0
Varginha	ABR	2008	2007-08	30,0
Varginha	DEZ	2008	2008-09	22,5
Varginha	FEV	2010	2009-10	18,0
Varginha	JAN	2011	2010-11	40,5
Varginha	FEV	2012	2011-12	23,5
Varginha	DEZ	2012	2012-13	10,0
Varginha	ABR	2014	2013-14	15,0

B.3 Carga pendente: baixa – limiar: 3 p.p.

Tabela 65. Meses de primeira ocorrência de TP ≥ 3 p.p. em lavouras de baixa carga pendente entre dezembro e junho. Estão destacados os anos de ocorrência da ferrugem tardia.

estação	mês	ano	ano agrícola	taxa de progresso (p.p.)
Boa Esperança	MAI	2010	2009-10	5,2
Boa Esperança	DEZ	2010	2010-11	5,0
Boa Esperança	MAI	2012	2011-12	5,0
Boa Esperança	MAI	2013	2012-13	8,0
Boa Esperança	DEZ	2013	2013-14	4,0
Carmo de Minas	JAN	2007	2006-07	5,5
Carmo de Minas	JAN	2008	2007-08	9,5
Carmo de Minas	DEZ	2008	2008-09	14,5
Carmo de Minas	JAN	2010	2009-10	4,0
Carmo de Minas	DEZ	2010	2010-11	4,0
Carmo de Minas	FEV	2012	2011-12	5,0
Carmo de Minas	DEZ	2012	2012-13	4,0
Carmo de Minas	DEZ	2013	2013-14	16,0
Varginha	JAN	1999	1998-99	4,2
Varginha	FEV	2000	1999-00	5,5
Varginha	JAN	2001	2000-01	12,5
Varginha	DEZ	2001	2001-02	5,0
Varginha	JAN	2003	2002-03	4,0
Varginha	DEZ	2003	2003-04	9,5
Varginha	DEZ	2004	2004-05	6,5
Varginha	DEZ	2005	2005-06	8,0
Varginha	DEZ	2006	2006-07	6,0
Varginha	JAN	2008	2007-08	6,5
Varginha	DEZ	2008	2008-09	26,0
Varginha	DEZ	2009	2009-10	8,0
Varginha	JAN	2011	2010-11	16,5
Varginha	JAN	2012	2011-12	8,0
Varginha	DEZ	2012	2012-13	3,0
Varginha	DEZ	2013	2013-14	13,5

B.4 Carga pendente: baixa – limiar: 5 p.p.

Tabela 66. Meses de primeira ocorrência de TP ≥ 5 p.p. em lavouras de baixa carga pendente entre dezembro e junho.

estação	mês	ano	ano agrícola	taxa de progresso (p.p.)
Boa Esperança	MAI	2010	2009-10	5,2
Boa Esperança	DEZ	2010	2010-11	5,0
Boa Esperança	MAI	2012	2011-12	5,0
Boa Esperança	MAI	2013	2012-13	8,0
Boa Esperança	MAI	2014	2013-14	8,0
Carmo de Minas	JAN	2007	2006-07	5,5
Carmo de Minas	JAN	2008	2007-08	9,5
Carmo de Minas	DEZ	2008	2008-09	14,5
Carmo de Minas	MAR	2010	2009-10	8,0
Carmo de Minas	JAN	2011	2010-11	18,0
Carmo de Minas	FEV	2012	2011-12	5,0
Carmo de Minas	MAI	2013	2012-13	5,5
Carmo de Minas	DEZ	2013	2013-14	16,0
Varginha	ABR	1999	1998-99	6,0
Varginha	FEV	2000	1999-00	5,5
Varginha	JAN	2001	2000-01	12,5
Varginha	DEZ	2001	2001-02	5,0
Varginha	FEV	2003	2002-03	5,0
Varginha	DEZ	2003	2003-04	9,5
Varginha	DEZ	2004	2004-05	6,5
Varginha	DEZ	2005	2005-06	8,0
Varginha	DEZ	2006	2006-07	6,0
Varginha	JAN	2008	2007-08	6,5
Varginha	DEZ	2008	2008-09	26,0
Varginha	DEZ	2009	2009-10	8,0
Varginha	JAN	2011	2010-11	16,5
Varginha	JAN	2012	2011-12	8,0
Varginha	JAN	2013	2012-13	7,5
Varginha	DEZ	2013	2013-14	13,5

APÊNDICE C – Meses de maior desenvolvimento da ferrugem no cafeeiro

Este apêndice detalha os meses que apresentaram maiores aumentos no valor de taxa de progresso (TP) da ferrugem na lavoura. Nos dados de alta (Tabelas 67 a 69) e baixa carga pendente (Tabelas 70 e 72) foram considerados aumentos na TP maiores ou iguais aos limiares de 12 e 6 p.p., respectivamente, para cada estação.

C.1 Carga pendente: alta

C.1.1 Boa Esperança

Tabela 67. Meses de maior aumento da taxa de progresso da ferrugem em Boa Esperança para lavouras de alta carga.

mês	ano	ano agrícola	taxa de progresso (p.p.)
MAI	2010	2009-10	31,0
JUN	2010	2009-10	25,0
JAN	2011	2010-11	31,0
FEV	2011	2010-11	19,0
MAR	2011	2010-11	18,0
MAR	2012	2011-12	13,0
MAI	2012	2011-12	16,0
ABR	2013	2012-13	21,0
MAI	2013	2012-13	24,5
JUN	2013	2012-13	18,0
ABR	2014	2013-14	17,0
MAI	2014	2013-14	28,0

C.1.2 Carmo de Minas

Tabela 68. Meses de maior aumento da taxa de progresso da ferrugem em Carmo de Minas para lavouras de alta carga.

mês	ano	ano agrícola	taxa de progresso (p.p.)
FEV	2007	2006-07	17,5
ABR	2008	2007-08	29,0
MAI	2008	2007-08	21,5
JUN	2008	2007-08	20,5
DEZ	2008	2008-09	17,0
FEV	2009	2008-09	23,5
ABR	2009	2008-09	14,5
MAI	2010	2009-10	26,0
JUN	2010	2009-10	29,5
JAN	2011	2010-11	27,0
FEV	2011	2010-11	14,0
MAR	2011	2010-11	20,0
MAI	2011	2010-11	13,5
MAR	2012	2011-12	13,0
MAI	2012	2011-12	21,0
JUN	2012	2011-12	25,5
DEZ	2012	2012-13	12,0
FEV	2013	2012-13	12,0
ABR	2013	2012-13	16,5
MAI	2013	2012-13	18,5
ABR	2014	2013-14	18,0

C.1.3 Varginha

Tabela 69. Meses de maior aumento da taxa de progresso da ferrugem em Varginha para lavouras de alta carga.

mês	ano	ano agrícola	taxa de progresso (p.p.)
MAR	2000	1999-00	17,0
ABR	2000	1999-00	33,0
JAN	2001	2000-01	13,0
JUN	2001	2000-01	20,5
JAN	2002	2001-02	16,0
FEV	2002	2001-02	14,0
ABR	2002	2001-02	23,0
MAI	2002	2001-02	22,0
ABR	2003	2002-03	21,5
MAI	2003	2002-03	38,5
JAN	2004	2003-04	13,0
FEV	2004	2003-04	25,0
MAR	2004	2003-04	22,0
ABR	2004	2003-04	14,0
FEV	2005	2004-05	19,5
MAR	2005	2004-05	15,0
ABR	2005	2004-05	14,0
MAI	2005	2004-05	30,0
MAR	2006	2005-06	38,0
ABR	2006	2005-06	14,0
JAN	2007	2006-07	35,5
FEV	2007	2006-07	20,0
ABR	2008	2007-08	30,0
MAI	2008	2007-08	23,0
JUN	2008	2007-08	25,0
DEZ	2008	2008-09	24,0
JAN	2009	2008-09	21,0
FEV	2009	2008-09	24,0
FEV	2010	2009-10	18,0
MAR	2010	2009-10	18,0
JAN	2011	2010-11	40,5
FEV	2011	2010-11	25,5
MAR	2011	2010-11	22,5
FEV	2012	2011-12	23,5
MAR	2012	2011-12	12,0
ABR	2012	2011-12	37,0
JAN	2013	2012-13	17,5
FEV	2013	2012-13	14,0
ABR	2013	2012-13	28,0
JUN	2014	2013-14	18,0

C.2 Carga pendente: baixa

C.2.1 Boa Esperança

Tabela 70. Meses de maior aumento da taxa de progresso da ferrugem em Boa Esperança para lavouras de baixa carga.

mês	ano	ano agrícola	taxa de progresso (p.p.)
MAI	2011	2010-11	7,0
MAI	2013	2012-13	8,0
MAI	2014	2013-14	8,0

C.2.2 Carmo de Minas

Tabela 71. Meses de maior aumento da taxa de progresso da ferrugem em Carmo de Minas para lavouras de baixa carga.

mês	ano	ano agrícola	taxa de progresso (p.p.)
JAN	2008	2007-08	9,5
ABR	2008	2007-08	17,0
JUN	2008	2007-08	16,0
DEZ	2008	2008-09	14,5
FEV	2009	2008-09	12,5
MAR	2009	2008-09	8,5
MAR	2010	2009-10	8,0
JAN	2011	2010-11	18,0
DEZ	2013	2013-14	16,0
ABR	2014	2013-14	16,0
JUN	2014	2013-14	8,0

C.2.3 Varginha

Tabela 72. Meses de maior aumento da taxa de progresso da ferrugem em Varginha para lavouras de baixa carga.

mês	ano	ano agrícola	taxa de progresso (p.p.)
ABR	1999	1998-99	14,7
MAI	1999	1998-99	10,7
JUN	1999	1998-99	24,0
MAR	2000	1999-00	15,5
ABR	2000	1999-00	37,0
MAI	2000	1999-00	13,0
JUN	2000	1999-00	13,5
JAN	2001	2000-01	14,0
MAR	2001	2000-01	8,5
MAI	2002	2001-02	8,5
MAI	2003	2002-03	12,0
JAN	2004	2003-04	10,0
MAR	2004	2003-04	30,0
FEV	2005	2004-05	17,5
MAI	2005	2004-05	20,5
JAN	2006	2005-06	11,5
JAN	2007	2006-07	15,0
MAR	2008	2007-08	14,5
MAI	2008	2007-08	13,0
JUN	2008	2007-08	26,0
DEZ	2008	2008-09	26,0
MAR	2010	2009-10	20,0
JUN	2010	2009-10	8,5
JAN	2011	2010-11	22,0
FEV	2011	2010-11	12,0
FEV	2012	2011-12	8,5
ABR	2012	2011-12	10,0
JAN	2013	2012-13	7,5
MAI	2013	2012-13	9,0
DEZ	2013	2013-14	17,0
JAN	2014	2013-14	10,0