MARIANA FIGUEIRA RAMOS

TÉCNICA DE MINERAÇÃO DE DADOS NA DISCRIMINAÇÃO SENSORIAL DA QUALIDADE DO CAFÉ ARÁBICA E O MEIO FÍSICO

LAVRAS – MG

2013

MARIANA FIGUEIRA RAMOS

TÉCNICA DE MINERAÇÃO DE DADOS NA DISCRIMINAÇÃO SENSORIAL DA QUALIDADE DO CAFÉ ARÁBICA E O MEIO FÍSICO

Dissertação apresentada à Universidade Federal de Lavras, como parte das exigências do Programa de Pós-Graduação em Estatística e Experimentação Agropecuária, área de concentração Estatística e Experimentação Agropecuária para a obtenção do título de Mestre.

Orientador

Dr. Marcelo Angelo Cirillo

Coorientador

Dr. Flávio Meira Borém

LAVRAS - MG

2013

Ficha Catalográfica Elaborada pela Divisão de Processos Técnicos da Biblioteca da UFLA

Ramos, Mariana Figueira.
Técnica de mineração de dados na discriminação sensorial da qualidade do café arábica e o meio físico / Mariana Figueira Ramos. – Lavras : UFLA, 2013.
70 p. : il.

Dissertação (mestrado) – Universidade Federal de Lavras, 2013.
Orientador: Marcelo Angelo Cirillo.
Bibliografia.

1. CHAID. 2. Data mining. 3. Árvore de decisão. I. Universidade Federal de Lavras. II. Título.

CDD – 006.312

MARIANA FIGUEIRA RAMOS

TÉCNICA DE MINERAÇÃO DE DADOS NA DISCRIMINAÇÃO SENSORIAL DA QUALIDADE DO CAFÉ ARÁBICA E O MEIO FÍSICO

Dissertação apresentada à Universidade Federal de Lavras, como parte das exigências do Programa de Pós-Graduação em Estatística e Experimentação Agropecuária, área de concentração em Estatística e Experimentação Agropecuária, para a obtenção do título de Mestre.

APROVADA em 22 de fevereiro de 2013.

Dr. Flávio Meira Borém	UFLA
Dra. Thelma Sáfadi	UFLA
Dra. Margarete Marin Lordelo Volpato	EPAMIG

Dr. Marcelo Angelo Cirillo

Orientador

LAVRAS - MG

2013

AGRADECIMENTOS

Aos meus pais, Carlos e Clara, a quem devo o maior dos agradecimentos, pelo apoio, confiança, devoção e companheirismo. A distância não foi suficiente para torná-los distantes e, sim, cada vez mais presentes.

Aos meus irmãos, Bruno e Tiago, exemplos de superação e conquista, que me incentivaram a percorrer essa caminhada com garra e perseverança.

Aos amigos, em especial os conquistados na trajetória UFLA, agradeço pelo companheirismo, amizade e a tudo que vivemos juntos.

Aos professores, em particular, ao professor Marcelo Cirillo e ao professor Flávio Borém, pela credibilidade depositada em mim para confeccionar este, bem como outros trabalhos.

À Universidade Federal de Lavras e ao Departamento de Ciências Exatas (DEX), pela oportunidade concedida para a realização do mestrado.

À FAPEMIG, pela concessão da bolsa de estudos, fundamental para a essa conquista.

Aos membros do projeto “Protocolo de identidade, qualidade e rastreabilidade para embasamento da indicação geográfica dos cafés da Mantiqueira de Minas Gerais”.

RESUMO

A mineração de dados - *data mining* - tem sido utilizada nas mais diversas áreas do conhecimento como potencial ferramenta para estratégias de decisões. Na detecção de perfis/hábitos de consumidores, fraudes, riscos e otimização de recursos, a técnica foi bem sucedida, permitindo seu emprego em áreas ainda não utilizadas. Em se tratando da mineração de dados utilizada na pesquisa cafeeira, nota-se uma carência de resultados mencionados na literatura, sugerindo sua aplicabilidade na descrição do perfil sensorial da qualidade do café associado a fatores genéticos e ambientais, bem como tecnológicos. Diante disso, neste trabalho propõe-se a utilização da técnica de *data mining Chi-square automatic iteration detection*, ou CHAID, no banco de dados do projeto intitulado "Protocolo de identidade, qualidade e rastreabilidade para embasamento da indicação geográfica dos cafés da Mantiqueira", para identificar características sensoriais da bebida do café arábica que se associam com o ambiente. Os resultados obtidos permitiram identificar fatores do meio físico associados às características sensoriais consideradas.

Palavras-chave: CHAID. *Data mining.* Árvore de decisão.

ABSTRACT

Data mining has been used in various fields of knowledge as a potential tool for strategy decisions. In detection of profiles / habits of consumers, fraud, risk and resource optimization, the technique was successful, allowing its use in areas not used yet. In the case of data mining used in coffee research, there is a lack of results mentioned in the literature, suggesting its applicability in describing the sensory profile of coffee quality associated with genetic, environmental, and technological. Thus, the paper proposes the use of the technique of data mining CHAID (Chi-square iteration automatic detection) in the database of the project entitled "Identity protocol, quality and traceability to the basement a geographical indication of coffees from Mantiqueira of Minas Gerais", to identify characteristics sensory of coffee drink that are associated with the environment. The results obtained allowed to identify physical factors associated with sensory characteristics considered.

Keywords: CHAID. *Data mining.* Decision tree.

LISTA DE FIGURAS

LISTA DE TABELAS

SUMÁRIO

1 INTRODUÇÃO

Devido à acessibilidade computacional, os métodos tradicionais de inspeção tornaram-se um tanto quanto inviáveis para captar informações. A partir daí, originou-se a mineração de dados, ou *data mining*, que foi uma alternativa eficaz na extração de informações por meio da detecção de relações ocultas, padrões e, até mesmo, regras para predizer ou correlacionar dados.

Utilizada nas mais diversas áreas do conhecimento, a técnica de mineração de dados ganhou destaque, principalmente, nos problemas de marketing, detectando, por exemplo, perfis/hábitos de consumidores, fraudes, riscos e otimização de recursos. Neste cenário, nota-se que o uso desta técnica poderá agregar conhecimento, para que estratégias de decisões possam ser tomadas, com maior ênfase em informações dispostas no banco de dados.

Em se tratando da mineração de dados utilizada na pesquisa cafeeira, nota-se uma carência de resultados mencionados na literatura. Tal fato supostamente sugere que a organização de um banco de dados confiável, no sentido de proporcionar registros atualizados, que contemple questões regionais, ambientais, físico-químicas e sensoriais relacionados a diferentes tipos de café, seja inviável.

Especificamente a atribuição de um banco de dados confiável, em termos da qualidade do café, e a representatividade das variáveis sensoriais sabor e aroma estão diretamente associadas a diversos fatores. Dentre esses, características fenotípicas, como a cor do fruto, o ambiente de cultivo e o processamento pós-colheita do café são considerados fundamentais na obtenção de um produto final com qualidade (AVELINO et al., 2002; AVELINO et al., 2005; BORÉM, 2008; DECAZY et al., 2003, BARBOSA et al., 2012).

Convém ressaltar que a exigência por qualidade da bebida do café tornou-se critério consolidado para se atingir o mercado internacional que,

comparado ao nacional, é o que melhor remunera esse produto. Tamanho é este impacto que a demanda pelos grãos especiais cresce em torno de 15% ao ano, contra o crescimento de cerca de 2% do café *commodity*. O segmento representa, hoje, cerca de 12% do mercado internacional da bebida. O valor de venda atual, para alguns cafés diferenciados, tem sobrepreço que varia entre 30% e 40%, em relação ao café cultivado de modo convencional. Em alguns casos, pode ultrapassar a barreira dos 100% (BRAZIL SPECIALTY COFFEE ASSOCIATION, 2012).

Em diversos estudos, buscou-se estabelecer relações entre esses fatores, considerados formadores e definidores da qualidade sensorial da bebida do café, gerando, com isso, um número extensivo de informações. Porém, junto com os resultados encontrados em pesquisas com essa importância, surgem dificuldades na interpretação.

Em concordância com os objetivos principais destes estudos, uma alternativa aplicável se encontra no uso de técnicas estatísticas capazes de explorar ao máximo os conjuntos de dados e, ao mesmo tempo, garantir a melhor descrição do fenômeno. Neste contexto, a metodologia estatística utilizada é de fundamental importância para a descrição do perfil sensorial da qualidade do café arábica associada a fatores genéticos e ambientais, bem como tecnológicos, envolvidos no processo de produção.

Diante do exposto, este trabalho foi realizado com o objetivo de propor a utilização da técnica de *data mining Chi-square automatic iteration detection* (CHAID) no banco de dados do projeto intitulado “Protocolo de identidade, qualidade e rastreabilidade para embasamento da indicação geográfica dos cafés da Mantiqueira”, com o propósito de identificar características sensoriais da bebida do café arábica que, produzido em diferentes formas de processamento, associam-se com o ambiente.

2 REFERENCIAL TEÓRICO

No referencial teórico, aborda-se o tema *data mining*, em especial árvores de decisão, além de se realizar uma breve contextualização sobre a qualidade do café.

2.1 *Data mining*

O progresso na aquisição de dados digitais e a tecnologia de armazenamento resultaram no surgimento de bancos de dados com grande dimensão. Nos mais diversos segmentos, extrair informações a partir dessas bases não tem sido uma tarefa fácil.

A partir de um depósito de dados (*data warehouse*), as informações são analisadas por diversas agregações complexas e estatísticas, com o propósito de tentar descobrir regras e/ou padrões a partir dos dados (SILBERSCHATZ; KORTH; SUDARSHAN, 1999). Mediante esta problemática, surgiu a mineração de dados (*data mining*), definida como um conjunto de técnicas aplicadas a um conjunto de dados observacionais para encontrar possíveis relacionamentos e resumi-los, de modo compreensível e útil, para o usuário de interesse (HAND; MANILA, 2001).

Um problema especial pode ser respondido utilizando-se a estatística, bastando que um conjunto de dados pequeno e organizado seja amostrado de forma independente e identicamente distribuído (HAND, 1998). Essa situação não se aplica no contexto de mineração de dados, pois se trata, na maioria dos casos, de uma análise exploratória secundária por conta da origem dos dados (são coletados sem utilizar estratégias eficientes para responder a perguntas específicas).

Banco de dados com as dimensões gigabyte ou terabyte já são comuns. A varejista americana Walmart faz mais de 20 milhões de transações diárias (BABCOCK, 1994). Segundo Cortes e Pregibon (1997), a empresa de telecomunicações AT&T tem cerca de 100 milhões de clientes cadastrados e 200 milhões de chamadas de longa distância diárias. Números como esses colocam as técnicas estatísticas em um difícil contexto, por terem sido propostas quando a captação de informações era inferior, além da baixa capacidade dos computadores (HAND, 1998). A partir dessa problemática, o processo de mineração de dados foi proposto por Fayyad, Piatetsky e Smyth (1996), considerando as seguintes etapas:

a) seleção de dados: definir claramente o domínio e os objetivos da situação problema para que seja possível a seleção das bases de dados alvo e gerar o conhecimento requerido;
b) pré-processamento dos dados: efetuar um processo de limpeza, eliminando ruídos e registros duplicados; determinar soluções para problemas de campos com dados faltantes e campos com dados errados; correção de erros de digitação, dentre outros;
c) transformação dos dados: reduzir o espaço dimensional por meio de mecanismos de representação eficiente dos dados, redução da quantidade de atributos, redução do conjunto de dados utilizados para treinamento por amostragem, etc.;
d) técnicas de mineração de dados: executar técnicas de identificação e/ou reconhecimento de padrões. Podem-se integrar duas ou mais técnicas, a fim de aumentar a confiabilidade da metodologia utilizada;
e) interpretação dos resultados: a partir do resultado obtido, identificar sua satisfação ou retornar a etapas anteriores, para serem refeitas;

f) utilização: utilizar o conhecimento obtido para fins de tomada de decisões.

Todo o procedimento de *data mining* foi ilustrado utilizando conforme o fluxograma ilustrado na Figura 1.

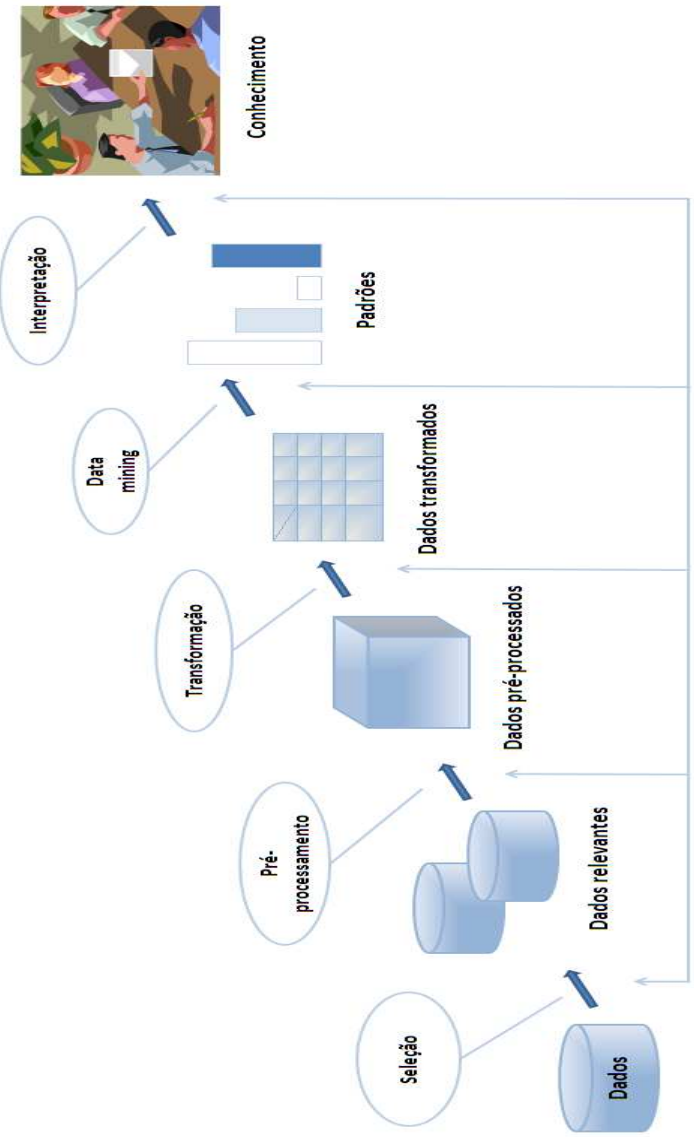

Figura 1 Processo de mineração de dados
Fonte: Fayyad, Piatetsky e Smyth (1996)

As principais metodologias de *data mining* se condicionam à situação problema e ao objetivo do estudo. Dentre as existentes, citam-se as árvores de decisão, que são comumente utilizadas para análise e extração de estruturas de decisão em um conjunto de dados multivariados com grande dimensão. A técnica particiona recursivamente o conjunto de dados, utilizando regras previamente definidas para gerar subconjuntos mutuamente exclusivos, exaustivos, que melhor descrevam a variável dependente de interesse (BIGGS; VILLE; SUEN, 1991).

O procedimento de partição é chamado de árvores de regressão, quando a variável resposta é numérica e árvores de classificação, quando a variável resposta é categórica (SICILIANO; MOLA, 2000).

Prezepiorski, Arns e Nievola (2005) apresentaram algumas vantagens nas árvores de decisão, como o fato de não assumir algum tipo de distribuição para os dados, as variáveis utilizadas podem possuir qualquer tipo de mensuração, além de ser possível construir modelos para qualquer função, dados uma amostra de treinamento suficiente e seu elevado grau de compreensão.

Algumas nomenclaturas são utilizadas para definir segmentos da árvore de decisão, como nós, ramos e folhas. Para Latorre et al. (2007), um nó raiz consiste no ponto inicial da análise, em que todos os dados encontram-se agrupados em um só conjunto; existe um conjunto hierárquico de nós internos (partições), responsáveis pela tomada de decisão a partir de critérios estabelecidos e que, posteriormente, define a próxima ramificação, denominada nó filho e, ao final da análise, dispõem-se dos nós terminais, chamados também de folhas. Na Figura 2 exibem-se as principais árvores de decisão existentes.

Figura 2 Principais árvores de decisão

2.2 Árvores de decisão baseadas em classificadores

Seja S um conjunto de dados contendo n casos $X_i(i = 1, \ldots, m)$ preditores e uma variável categórica dependente Y com k categorias. O algoritmo de árvore de decisão consiste em encontrar um modelo de classificação baseando-se nos valores de X que classifique casos em Y .

Suknovic et al. (2011) resumem uma árvore de decisão em nós, ramos e folhas. A predição de Y é definida pelas folhas, que também fornecem a quantidade de regras que podem ser extraídas do processo.

As árvores de decisões podem ser reproduzidas a partir de diferentes critérios. Para as árvores ID3, CART e CHAID, o algoritmo básico é resumido em três passos (SUKNOVIC et al., 2011), que são:

a) passo 1: para todas as categorias dos X_i preditores, definir todos os possíveis nós obtidos por meio da determinação de possíveis agrupamentos de categorias de atributos categóricos;

b) passo 2: avaliar as divisões e detectar a melhor. Ramificar a árvore, a partir de um critério pré-estabelecido, com a divisão selecionada. S é, então, dividido em subconjuntos disjuntos $S_j (j = 1, \ldots, l)$, em que l define o número de ramos da divisão escolhida, tendo $S = \cup S_j$;

c) passo 3: repetir, recursivamente, os passos anteriores para todos S_j, até que algum critério de parada pré-estabelecido seja atingido.

2.2.1 ID3

A árvore de decisão *Iterative Dichotomiser Tree*, ou ID3, é utilizada para dados categóricos e as ramificações acontecem para todas as categorias de um determinado preditor, produzindo árvores rasas, com muitos nós (QUINLAN, 1986). O algoritmo básico é executado em todos os nós e consiste em três passos, que são:

a) uma única divisão é gerada para cada categoria de um preditor, gerando ramificações da divisão proposta;
b) por meio da medida de entropia de Shannon (1948),

$$E(S) = -\sum_{r=1}^{k} (p_i log_2 p_i) \quad (1)$$

em que k representa o número de categorias da variável dependente e p_i a probabilidade associada à r-ésima categoria, calcula-se o de ganho de informação, dada pela equação 2,

$$I(X,S) = E(S) - E(X,S) \quad (2)$$

em que $E(X,S) = \sum_{u=1}^{t} \left(\frac{|S_u|}{|S|} . E(S_u) \right)$ é a entropia esperada de um preditor X com t categorias; $E(S_u)$ é a entropia da categoria de um preditor relacionada com a variável dependente e $|S_u|/|S|$ é a probabilidade da u -ésima categoria de um preditor. O ganho de informação é utilizado para selecionar a melhor divisão da árvore de decisão;

c) repetir os passos anteriores, recursivamente, para gerar novas ramificações, até que algum critério de parada pré-estabelecido seja atingido.

2.2.2 C4.5

Á árvore de decisão C4.5 é, basicamente, um aperfeiçoamento do algoritmo ID3 (QUINLAN, 1993). Pode ser utilizada quando os preditores são também numéricos, conforme os seguintes passos:

a) o critério de divisão é similar ao do ID3, no caso de o preditor ser categórico e, no caso numérico, é transformado em binário;
b) por meio da medida taxa de ganho

$$G(X,S) = \frac{I(X,S)}{SI(X,S)} \quad (3)$$

em que

$$SI(X,S) = -\sum_{j=1}^{k}\left(\frac{\frac{|S_j|}{|S|}\log|S_j|}{|S|}\right) \quad (4)$$

define-se a quantidade de informação das divisões para avaliar a melhor divisão da árvore gerada. Suknovic et al. (2011) relatam que essa medida é menos tendenciosa para a seleção de preditores com muitas categorias, além de auxiliar na construção da árvore com dados faltantes.

2.2.3 CART

A árvore de decisão *Classification and Regression Tree*, ou CART, produz classificadores ou regressores (BREIMAN et al., 1984). Cresce somente para divisões binárias, gerando árvores estreitas e com grande profundidade (SUKNOVIC et al., 2011). As etapas do algoritmo estão dispostas a seguir.

a) gerar todas as possíveis divisões. No caso numérico, o critério será similar ao utilizado em C4.5. Para um preditor categórico, gerar todos os possíveis grupos de atributos, dois a dois;
b) a avaliação da divisão é baseada na impureza do nó[1], descrita na equação 5,

$$G(X,S) = u(S) - p_E * u(S_E) - p_D * u(S_D) \quad (5)$$

[1] Silva (2007) define a impureza de um nó como a quantidade de informação necessária para chegar a um nó terminal.

em que $u(S)$ é a soma dos produtos de todas as combinações de pares de categorias, e p_E e p_D são a probabilidade de um caso ser alocado no ramo esquerdo e direito, respectivamente. As divisões são avaliadas por meio da medida de Gini, dada pela equação 6

$$u(S) = \sum_{u,l} p(u|S)p(l|S), u \neq l \quad (6)$$

em que $p(u|S)$ é a probabilidade da categoria u em S .

a) Os passos anteriores são repetidos até que algum critério de parada pré-estabelecido seja atingido.

2.2.4 CHAID

Introduzido por Kass (1980), o algoritmo C*hi-square Automatic Iteration Detection*, ou CHAID, foi criado a partir do processo *Automatic Iteration Detection*, ou AID (MORGAN; SONQUIST, 1963a; MORGAN; SONQUIST, 1963b). Sua raiz consiste em uma família de métodos de manuseio de dados quase livre dos pressupostos habituais necessários para processá-los (HAWKINS; KASS, 1982).

Utilizada no desenvolvimento de algumas das tradicionais árvores de decisão, a metodologia foi elaborada por conta da ausência de testes de significância na criação das ramificações (KASS, 1975), relacionamentos teoricamente falsos (EINHORN, 1972) e até tendência para má utilização nos casos de amostras pequenas (DOYLE; FENWICK, 1975). Baseando-se na estatística do qui-quadrado, mescla ou divide categorias de um preditor, identificando as divisões mais significativas (SUKNOVIC et al., 2011).

Os preditores utilizados nesse tipo de árvore de decisão têm caráter categórico e, no caso contínuo, Hawkins e Kass (1982) aconselham dividi-los em grupos. Existem cinco classes de preditores, sendo as três mais utilizadas apresentadas a seguir.

a) preditor monotônico: suas categorias se encontram em uma escala ordinal. Isto implica que apenas as categorias adjacentes podem ser agrupadas;
b) preditor livre: suas categorias são puramente nominais. Isto implica que quaisquer categorias podem ser agrupadas;
c) preditores flutuantes: as categorias se encontram em uma escala ordinal, à exceção de uma única categoria, que não pertence ao restante ou cuja posição na escala ordinal seja desconhecida.

Esta metodologia consiste, basicamente, em uma análise de agrupamentos adaptada, diferindo-se pelo fato de conter muitos casos para a redução de categorias de um preditor, gerando grupos heterogêneos, mas internamente homogêneos. A garantia de otimalidade ao final, anteriormente realizada a partir de uma programação dinâmica, foi substituída pelo método *stepwise*, não necessariamente ideal, porém, mais ágil e eficaz (KASS, 1980).

2.2.4.1 Tamanho amostral

Em cada fase do processo de criação da árvore de decisão, os dados são particionados em amostras e, a partir delas, outras análises são realizadas, resultando em subconjuntos da partição inicial. Consequentemente, é razoável constatar que os procedimentos de AID apresentem maior eficiência em bancos

de dados com grande dimensão, não sendo muito viáveis para pequenas amostras.

No caso CHAID, em que a variável dependente é de natureza categórica, o produto de divisão e o tamanho da amostra diminuem ao longo do processo, resultando em uma perda progressiva de poder (HAWKINS, 1982). Decorrente desse fato, recomenda-se incorporar um "controle" sobre o tamanho da amostra em qualquer subconjunto alcançado e encerrar as etapas do algoritmo nesse subconjunto, caso o tamanho da amostra não seja grande o suficiente para justificar as configurações obtidas.

Convém ressaltar que o tamanho da amostra real por subconjunto fica condicionado ao número de categorias, tanto nos preditores quanto na variável dependente, e, por isso, torna-se complexo especificar uma quantidade mínima por categoria. Hawkins (1982) utiliza a regra de que a frequência esperada na célula seja superior a 5, definindo que o tamanho mínimo do subconjunto seja $5kt$ observações, k e t denotando o número máximo de categorias da variável dependente e preditor, respectivamente. O tamanho mínimo da amostra global depende de um fator imponderável, que é o fato de definir em quantos grupos a mesma será particionada. Hawkins (1982) realizou diversas análises utilizando o algoritmo CHAID e obteve sucesso em amostras maiores do que 500, deixando de lado sua utilização no caso complementar.

2.2.4.2 Relação do teste qui-quadrado para independência e CHAID

Seja Y uma variável dependente com $k \geq 2$ categorias e um conjunto de preditores X_i $(i = 1, \ldots, m)$. Para duas variáveis, diga-se X_1 e X_2, realiza-se o teste qui-quadrado sob a hipótese nula de independência entre as duas variáveis.

O teste é realizado a partir da tabela de contingência $A_{k_1 x k_2}$, gerada pelas duas variáveis, em que as k_1 linhas e as t_1 colunas corresponderão às categorias de X_1 e X_2, respectivamente. O valor $a_{v,w}$, denotado por $O_{v,w}$, consiste no número observado, em que $X_1 = C(A_{v,})$ e $X_2 = C(A_{,w})$, em que $C(\square)$ é um operador que retorna a categoria representada nas linhas $(A_{v,})$ ou colunas $(A_{,w})$ da tabela A . A estatística teste χ^2 é obtida pela equação 7,

$$\chi^2 = \sum_{v,w} \frac{(O_{vw} - E_{vw})^2}{E_{vw}} \tag{7}$$

em que E_{vw} é o valor esperado para $a_{v,w}$ sob a hipótese nula. A estatística teste obtida mede o quão longe as observações atuais se encontram das observações esperadas (OUYANG; PATEL; SETHI, 2008) a partir da distribuição qui-quadrado com graus de liberdade $(k_1 - 1)(t_1 - 1)$. Para um nível de significância α , obtém-se a probabilidade p de os valores observados serem maiores ou iguais a um valor χ^2, disposto na equação 8,

$$p = 1 - \left(\chi^2_{(k_1-1)(t_1-1))}\right) \tag{8}$$

em que $\left(\chi^2_{(k_1-1)(t_1-1))}\right)$ é o valor da distribuição qui-quadrado acumulada com graus de liberdade $(k_1 - 1)(t_1 - 1)$. Rejeita-se a hipótese nula caso $p < \alpha$, e X_1 e X_2 são considerados dependentes.

2.2.4.2.1 Algoritmo CHAID

O processo de criação de uma árvore de decisão seguindo o método CHAID é resumido nos passos descritos a seguir (OUYANG; PATEL; SETHI, 2008; KASS, 1980).

Passo 1: criar uma tabela de contingência utilizando os preditores e a variável dependente e executar;

a) passo 1.1: para um nível de significância α, encontrar o par de categorias do preditor selecionado cuja subtabela $2 \times k$ seja o menos diferente, significativamente, a partir do teste qui-quadrado. Se a probabilidade p obtida supera o valor crítico α, as duas categorias, então, são mescladas, resultando em uma única categoria. Dessa forma, o passo é repetido até que pares significativos sejam encontrados;
b) passo 1.2: para cada categoria mesclada de três ou mais categorias, encontrar uma divisão binária, para gerar duas novas subcategorias mais significativas em que a mesclagem seja solucionada. Se a probabilidade p for menor do que o nível de significância α, implementar a divisão e retornar ao passo 1.1.

Passo 2: Obter a probabilidade p, por meio do teste qui-quadrado, para todas as possíveis divisões de categorias dos preditores. O preditor mais significativo para um nível de significância α pré-fixado será selecionado para criar uma nova ramificação.

Passo 3: S é particionado conforme os critérios de divisão e mesclagem de categorias. Os subconjuntos gerados são processados recursivamente pelo procedimento iniciado no passo 1.

Com estas especificações, o primeiro passo na execução do algoritmo, conforme Kass (1980), é criar as i tabelas de contingência reduzidas $w \times k$, $(w = 2, \ldots, t)$, e calcular a estatística $T_w^{(i)}$, dada na equação 9

$$T_w^{(i)} = \sum_{v}^{t} \sum_{w}^{k} \frac{(O_{vw} - E_{vw})^2}{E_{vw}} \sim \chi^2_{((v-1)(w-1))} \tag{9}$$

sendo O_{vw} a frequência observada na linha t, coluna k e E_{vw} a respectiva frequência esperada. Assumindo a estatística $T_w^{(*)} = max_i T_w^{(i)}$, obtém-se o $T_w^{(*)}$ mais significativo dado um α a partir do método de seleção de variáveis *stepwise* (KASS, 1980).

2.2.4.2.2 Seleção de variáveis a partir do método *stepwise*

Seja S um conjunto de dados formado por $X_i (i = 1, \ldots, m)$ preditores, com t categorias e Y uma variável dependente com k categorias. De forma similar ao método *stepwise* utilizado em modelos de regressão, a seleção de variáveis categóricas pelo método CHAID pode ser resumida no diagrama disposto na Figura 3.

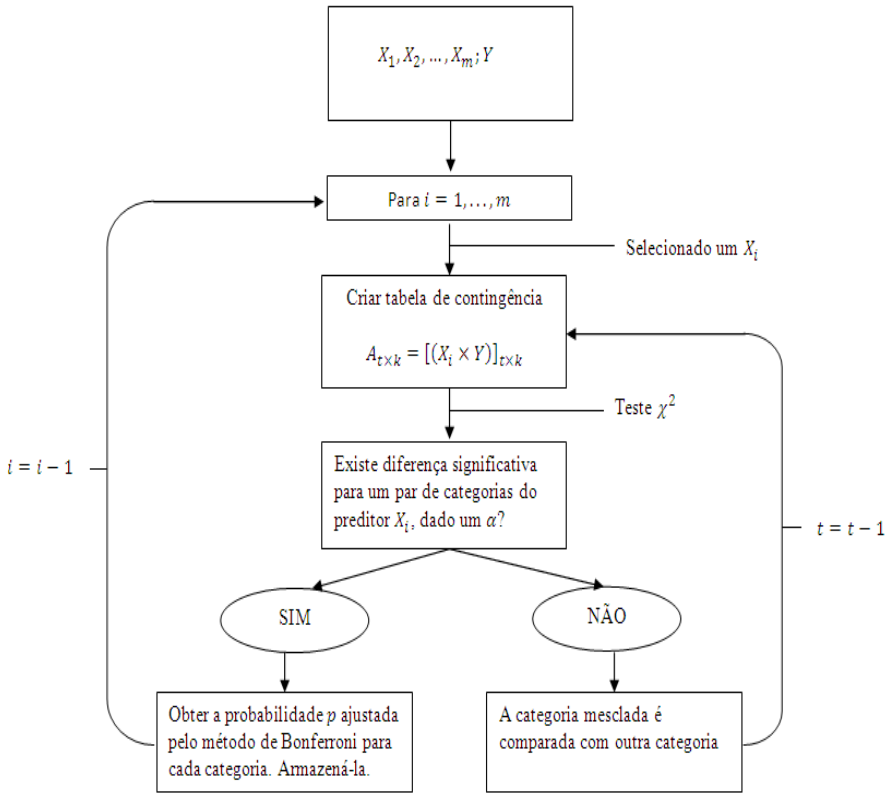

Figura 3 Diagrama da seleção de variáveis *stepwise* para o primeiro passo do método CHAID

No caso AID, a primeira etapa consiste em manter as categorias de cada preditor como distintas. As fases de mesclagem consistem, inicialmente, em separar o grupo em duas amostras para teste e comparar todas as categorias ou grupo de categorias mescladas que possam ser incluídas no modelo. Quando todos os testes forem realizados, as categorias menos significantemente diferentes são mescladas. Este processo continua enquanto existirem categorias ou grupos de categorias que possam ser incorporadas no modelo, e que não sejam significativamente diferentes (HAWKINS, 1982).

A segunda fase, ou fase de divisão, é executada quando surge um agrupamento de três ou mais categorias. Este agrupamento é investigado a partir de todas as possíveis divisões binárias do conjunto em dois subconjuntos de categorias e duas amostras estatísticas para criar, em um primeiro momento, uma identidade que separe as categorias. Para um nível de significância especificado, caso a divisão binária seja significativa, o composto de categorias é novamente dividido.

Os níveis de significância para divisão e mesclagem de categorias são especificados pelo usuário. Deve-se atribuir um valor mais representativo pra mesclagem de categorias, com o propósito de reduzir a ocorrência de *loop* infinito (HAWKINS, 1982).

2.2.4.2.3 Obtenção das probabilidades considerando o ajuste de Bonferroni

Kass (1980) propõe uma aproximação para a probabilidade p por meio do ajuste de Bonferroni, que é utilizado quando se têm comparações múltiplas em um processo estatístico. Esta metodologia altera p , a fim de minimizar a chance de ocorrência do erro tipo I, por conta das comparações múltiplas ocorridas no método CHAID (RITSCHARD, 2010).

Seja X um preditor com t categorias, define-se m conforme a natureza do preditor, como o número de possibilidades que todas as t categorias podem ser reduzidas em r grupos, sendo r o número final de categorias após mesclagem (KASS, 1980).

a) Preditores monotônicos:

$$m_{monotônico} = \binom{t-1}{r-1}$$

b) Preditores livres:

$$m_{livre} = \sum_{i=0}^{r-1} (-1)^i \frac{(r-i)^r}{i!\,(r-i)!}$$

c) Preditores flutuantes:

$$m_{flutuante} = \binom{t-2}{r-2} + r\binom{t-2}{r-1} = \frac{r-1+r(t-r)}{t-1} m_{monotônico}$$

Ao final do processo, caso as t categorias de um preditor sejam mescladas em r grupos $(2 \leq r \leq t)$, a probabilidade p será corrigida calculando m e utilizando-o como correção no ajuste de Bonferroni, conforme a equação 10,

$$p_0 \leq \frac{p}{m} \tag{10}$$

obtendo-se um limite p_0 para p (KASS, 1980; BIGGS; VILLE; SUEN, 1991).

Um exemplo de dados contendo 762 estudantes do primeiro ano, matriculados na Faculdade de Economia e Ciências Sociais da Universidade de Genebra, originado por Petroff, Bettex e Korry (2001), ilustra o algoritmo CHAID, utilizando as probabilidades obtidas corrigidas.

O estudo propunha identificar quais características pessoais apresentavam associação com a situação do aluno (eliminado, repetindo, passou), em outubro de 1999. Os preditores utilizados foram:

a) ano de nascimento: $(<= 76, (76-78], > 78)$;
b) ano de registro na universidade: $(<= 97, > 97)$;
c) curso (ciências sociais, administração/economia);
d) tipo de diploma secundário alcançado (clássico/latino, científico, economia, moderno, técnico, exterior, outro, sem informação);

e) local onde foi obtido o diploma secundário (Genebra, Suíça, exterior, outro);

f) idade de obtenção do diploma secundário $(<= 19, > 19)$;

g) nacionalidade (Genebra, Suíça, Europa, exterior);

h) lugar de residência da mãe (Genebra, Suíça, Exterior).

Para exemplificar uma primeira situação, na Tabela 1 observa-se a tabela de contingência entre a situação do aluno e o preditor tipo de diploma secundário.

Tabela 1 Tabela de contingência para a situação do aluno e o preditor tipo de diploma secundário

Preditor		Eliminado	Passou	Repetindo	Total
Clássico/Latino	1	23	16	87	126
Moderno	2	44	22	68	134
Científico	3	18	13	90	121
Economia	4	37	36	96	169
Técnico	5	5	0	4	9
Exterior	6	76	43	71	190
Outro	7	6	0	6	12
Sem informação	8	0	0	1	1
Total		209	130	423	762

O número de possibilidades para separar as categorias do preditor em questão, classificado como nominal, em r grupos, $(2 \leq r \leq 8)$, resulta em 4.140. Subtraiu-se em um a quantidade obtida pelo fato de todas as categorias agrupadas não serem consideradas uma divisão, totalizando 4.139 possibilidades (RITSCHARD, 2010).

Na Tabela 2 exibem-se as estatísticas qui-quadrado para cada par de categorias na parte superior à diagonal principal, e as probabilidades ajustadas p na parte inferior. Neste primeiro passo, a menor estatística qui-quadrado foi gerada no par $(5,7)$, identificando os alunos que tinham diploma técnico e os

que tinham outro tipo de diploma secundário. A estatística teste foi obtida considerando-se os graus de liberdade $(3-1)(2-1)=2$, resultando um $p=0,969$, permitindo concluir que não há diferenças significativas entre as categorias 5 e 7.

Tabela 2 Estatísticas qui-quadrado e probabilidades p para cada par de categorias da tabela de contingência completa 3×8

	1	2	3	4	5	6	7	8
1	0	9,62	0,87	5,25	7,53	30,64	7,37	0,45
2	0,008	0	15,66	4,79	2,81	5,88	2,92	0,96
3	0,647	0,000	0	9,88	9,84	40,80	9,65	0,34
4	0,073	0,091	0,007	0	6,25	16,65	6,37	0,76
5	0,023	0,245	0,007	0,044	0	2,66	0,06	1,11
6	0,000	0,053	0,000	0,000	0,264	0	3,47	1,66
7	0,025	0,232	0,008	0,041	0,969	0,177	0	0,93
8	0,800	0,618	0,842	0,685	0,574	0,436	0,629	0

Na Tabela 3 exibem-se os resultados da nova categoria criada, originada pela mesclagem das categorias técnico e outro tipo de diploma secundário.

Tabela 3 Tabela de contingência 3×2 para o primeiro caso de mesclagem, considerando a situação do aluno e o preditor tipo de diploma secundário

	Técnico	Outro	Mesclagem
	5	7	{5,7}
Eliminado	5	6	11
Passou	0	0	0
Repetindo	4	6	10
Total	9	12	21

O processo é repetido substituindo-se as colunas 5 e 7 da Tabela 2 pela coluna mesclada $\{5\times 7\}$, gerando uma tabela de contingência 3×7. Na Tabela

4 exibem-se as estatísticas qui-quadrado e probabilidades p para cada par de categorias, sendo o par $\{3,8\}$ menos significativo, com uma estatística qui-quadrado 0,34 e $p = 0,842$.

Tabela 4 Estatísticas qui-quadrado e probabilidades p para cada par de categorias da tabela de contingência reduzida 3×7

	1	2	3	4	$\{5,7\}$	6	8
1	0	9,62	0,87	5,25	7,53	30,64	0,45
2	0,008	0	15,66	4,79	2,81	5,88	0,96
3	0,647	0,000	0	9,88	9,84	40,80	0,34
4	0,073	0,091	0,007	0	6,25	16,65	0,76
$\{5,7\}$	0,002	0,066	0,000	0,003	0	5,97	1,05
6	0,000	0,053	0,000	0,000	0,264	0	1,66
8	0,800	0,618	0,842	0,685	0,574	0,436	0

O processo iterativo de mesclagem termina na 6ª iteração, com as estatísticas qui-quadrado e probabilidades p retratadas na Tabela 5. Verifica-se que o preditor em questão foi reduzido para cinco categorias e, em duas situações, ocorreu nova mesclagem.

Tabela 5 Primeira etapa do processo iterativo de mesclagem para o preditor tipo de diploma secundário

Iteração	Mesclagem	Qui-quadrado	p
1	$\{5,7\}$	0,06	0,967
2	$\{3,8\}$	0,34	0,846
3	$\{1,\{3,8\}\}$	0,95	0,623
4	$\{2,4\}$	4,79	0,091
5	$\{6,\{5,7\}\}$	5,97	0,051

A próxima etapa consistiu em averiguar se as categorias mescladas por mais de duas categorias, no caso $\{1,\{3,8\}\}$ e $\{6,\{5,7\}\}$, podem ser dicotomizadas

significativamente. Ainda no passo 1, a homogeneidade dos grupos foi avaliada pelo critério de selecionar o par com menor valor da estatística teste e maior probabilidade p , dado um α .

Pelos dados da Tabela 6, verifica-se que a proposta de mesclagem $\{1,3,8\}$, dividida em $\{\{1,3\},8\}$, resultou em não significância, assim como na mesclagem $\{5,6,7\}$, dividida em $\{\{5,6\},7\}$. Portanto, o preditor tipo de diploma secundário foi reduzido para três categorias.

O melhor preditor para iniciar partições no conjunto de dados, considerando a situação do aluno, foi tipo de diploma secundário, simplificado nas categorias {clássico/latino, científico, sem informação}, {moderno, economia} e {técnico, exterior, outro}.

Tabela 6 Estatística qui-quadrado e probabilidades p para a etapa final

Divisão		Qui-quadrado	p
{1,3} x {8}		0,39	0, 821
{5,6} x {7}		2,28	0, 194
	Clássico/latino Científico, sem informação {1,3,8}	Moderno, economia {2,4}	Técnico, exterior, outro {5,6,7}
{1,3,8}	0	18,04	52,94
{2,4}	0,000	0	14,56
{5,6,7}	0,000	0,001	0

No decorrer do processo, o sistema é análogo considerando os demais preditores. Ritschard (2010) descreve o processo de crescimento da árvore baseando-se na independência da tabela cruzada completa entre um preditor e a variável, a partir da probabilidade p , obtida pelo teste qui-quadrado.

O próximo passo é utilizar o ajuste de Bonferroni para corrigir p e encontrar árvores com ramificações diferentes. Na Tabela 7 resumem-se os resultados obtidos no nó raiz, considerando, no primeiro momento, a mesclagem dos oito preditores considerados. Em ordem decrescente, o preditor tipo de diploma secundário foi o mais significativo para avaliar a situação do aluno.

Tabela 7 Resumo para o primeiro nível de divisão, considerando o tipo de diploma secundário e a situação do aluno

Preditor	Categorias	Divisões	χ^2	p
Tipo de diploma secundário	8	3	54,8	0,000000000035
Ano de nascimento	25	3	53,0	0,000000000085
Local onde foi obtido o diploma secundário	3	3	42,7	0,0000000122
Local de residência da mãe	4	2	27,2	0,00000123
Nacionalidade	3	2	24,3	0,00000540
Ano de registro na faculdade	11+1	2	18,7	0,0000863
Idade quando obteve o segundo diploma	4	4	21,9	0,00128

Curso	2	2	1,4	0,499

Na Figura 4 exibe-se a árvore de decisão completa sem o ajuste de Bonferroni. O primeiro nó foi obtido pela tabela de contingência entre o preditor tipo de diploma secundário e a variável dependente situação do aluno, sendo o preditor em questão o primeiro fator discriminador para a situação do aluno. Neste caso, os graus de liberdade para esse caso são definidos como $(k-1)(r-1)$, sendo $k=3$ categorias da variável dependente e h , categorias após mesclagem do preditor.

As categorias do preditor mais significativo foram agrupadas em três grupos, gerando as primeiras divisões. Nos nós resultantes, o processo foi iterado com os demais preditores.

Para o grupo de 248 alunos com diploma secundário clássico/latim, científico ou sem informação, obteve-se significância com o ano de nascimento, não havendo mesclagem em suas categorias. Os 303 estudantes com diploma secundário moderno ou economia foram divididos conforme o curso escolhido e os 211 estudantes restantes foram distinguidos conforme o ano de registro.

O processo iterativo foi encerrado quando se obteve nós terminais, que obedeceram aos critérios de parada de parada pré-estabelecidos. A significância para mesclagem e divisão, considerando número de observações em nós secundários inferior, divisão de categorias em nós secundários com menos que 50 observações ou a quantidade de nós terminais são exemplos de critérios de parada (RITSCHARD, 2010).

Considerando uma significância de 5%, a situação do aluno foi classificada na categoria “eliminado” para os tipos de diploma secundário exterior, técnico ou outro e o seu ano de registro menor ou igual a 97.

Nos demais nós terminais, a situação do aluno foi classificada como “passou”, existindo diferenças nas porcentagens finais dessa classificação. Um

exemplo, 86,67% dos alunos cujo tipo de diploma secundário era clássico/latino, científico ou sem informação, o ano de nascimento na categoria $(76,78]$ e idade ≤ 19 , tiveram situação “passou”.

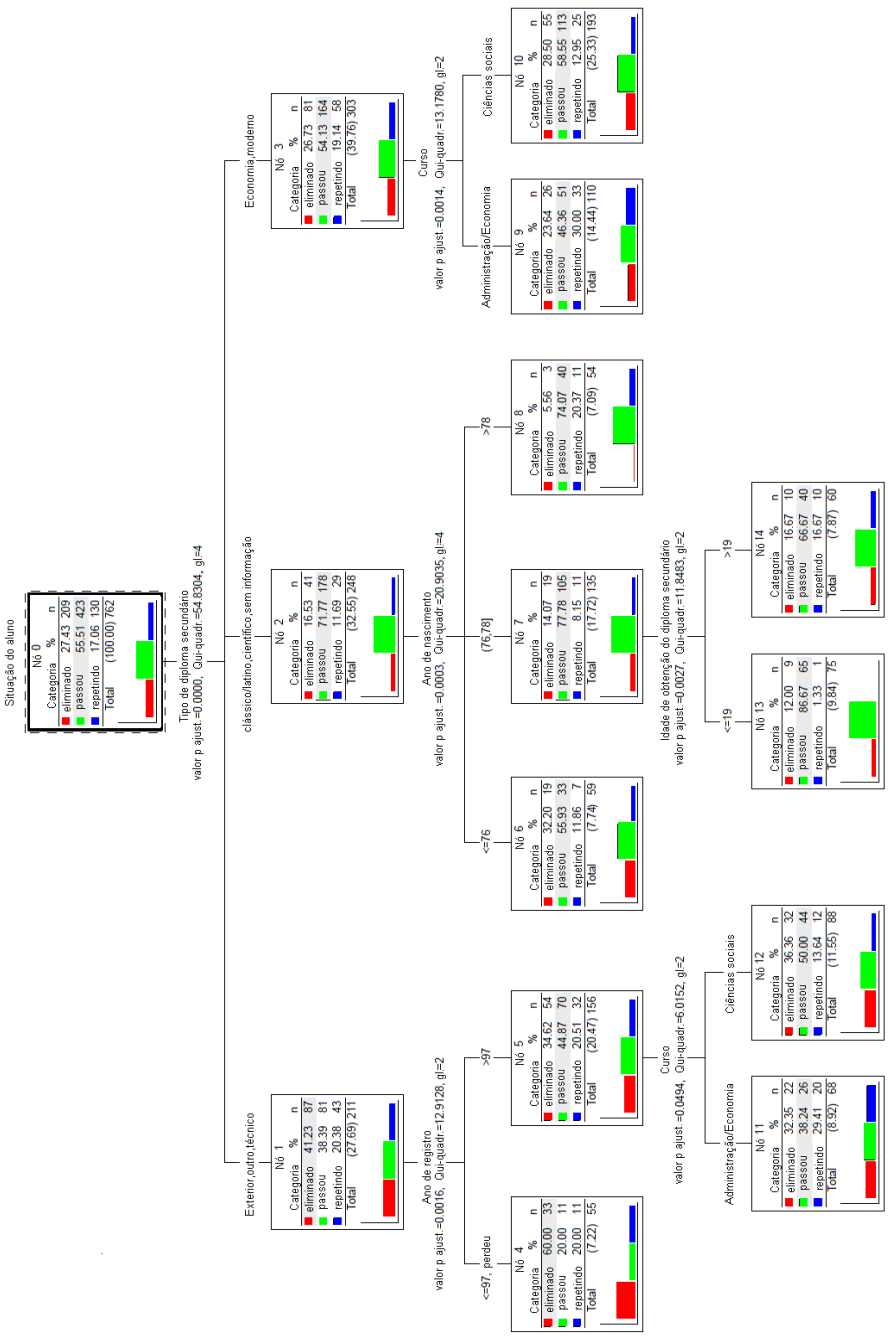

Figura 4 Método CHAID sem o ajuste de Bonferroni
Fonte: Petroff, Bettex e Korry (2001)

O ajuste de Bonferroni, quando utilizado, pode ser excessivo e inverter a tendência dos preditores (KASS, 1980). No exemplo, para todos os preditores, a Tabela 8 apresenta a probabilidade p corrigida dado o nó raiz.

A coluna de classificação (*rank*) permitiu identificar mudanças na prioridade para a classificação dos preditores. O preditor tipo de diploma secundário, que ocupava, anteriormente, a 1ª posição, caiu para a 3ª. Já o local do diploma secundário, com um número menor de categorias, ocupou o posto de melhor classificador. Esta situação exibe a penalidade levada a um preditor com muitas categorias, quando o ajuste de Bonferroni for utilizado (RITSCHARD, 2010).

Tabela 8 Resumo para o primeiro nível de divisão, considerando o ajuste de Bonferroni

Preditor	Categorias	Divisões	χ^2	p	p corrigida	Rank
Tipo de diploma secundário	8	3	54,8	3,5E-11	3,41E-08	3
Ano de nascimento	25	3	53,0	8,5E-11	2,34E-08	2
Local onde foi obtido o diploma secundário	3	3	42,7	1,22E-08	1,22E-08	1
Local de residência da mãe	4	2	27,2	1,23E-06	8,64E-06	4
Nacionalidade	3	2	24,3	5,40E-06	1,62E-05	5
Ano de registro na faculdade	11+1	2	18,8	8,63E-05	1,64E-03	7
Idade de obtenção do segundo diploma	4	4	21,9	1,28E-03	1,28E-03	6
Curso	2	2	1,4	4,99E-01	4,99E-01	8

Na Figura 5 é mostrado o resultado da árvore de decisão utilizando o ajuste de Bonferroni. A nova configuração gerou seis nós terminais com novos fatores classificadores, para uma significância de 5%.

A situação do aluno foi “eliminado” em dois nós terminais, cuja porcentagem final atingida alocou no máximo 45% dos casos. Quando o local de obtenção do diploma secundário foi exterior/outro, houve significância estatística com o preditor curso, mas, em ambos os nós, a classificação final foi “eliminado”.

Para o local de obtenção do diploma secundário Genebra ou Suíça, as ramificações subsequentes levaram a nós terminais cuja classificação final foi “passou”. Considerando que o local de obtenção do diploma foi Genebra, o tipo de diploma secundário clássico/latino ou científico e a nacionalidade Genebra, 79,55% dos casos deste nó terminal foram classificados na categoria “passou”.

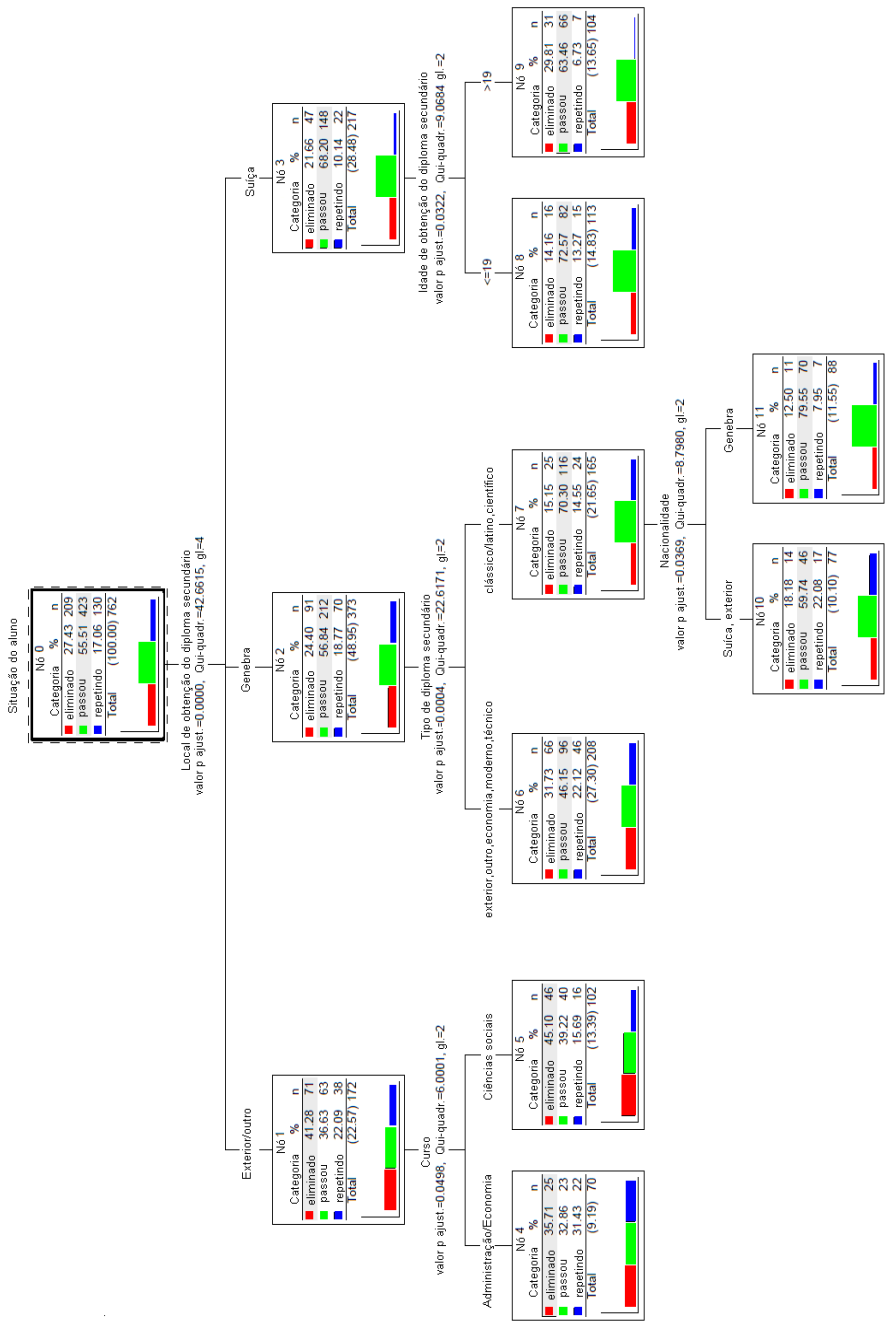

Figura 5 Método CHAID com o ajuste de Bonferroni
Fonte: Petroff, Bettex e Korry (2001)

2.2.4.3 Desempenho da árvore de decisão

Em problemas de classificação, em que a variável dependente atrelada ao estudo é não métrica, a avaliação preditiva do modelo acontece por meio da matriz de confusão (HAIR et al., 2009; KOHAVI; PROVOST, 1998).

Seja S um conjunto de dados com n observações, Y uma variável dependente com duas categorias, k_1 e k_2, e θ uma observação cuja classificação seja desconhecida. Para n_1 observações da categoria k_1 e n_2 observações da categoria k_2, a matriz de confusão (Tabela 9) apresenta a tabulação cruzada entre a classificação real de uma amostra e sua classificação predita pelo método considerado (JOHNSON; WICHERN, 2007).

Tabela 9 Matriz de confusão

		Classificação predita		
		k_1	k_2	Total
Classificação real	k_1	n_{1C}	$n_{1M} = n_1 - n_{1C}$	n_1
	k_2	$n_{2M} = n_2 - n_{2C}$	n_{2C}	n_2

sendo

n_{1c}: número de observações de π_1 classificados corretamente como π_1;

n_{1M}: número de observações de π_1 classificados incorretamente como π_2;

n_{2c}: número de observações de π_2 classificados corretamente como π_2;

n_{2M}: número de observações de π_2 classificados incorretamente como π_1.

Baseando-se na matriz de confusão, as seguintes medidas de desempenho podem ser calculadas (WEISS; KULIKOWSKY, 1991):

a) erro 1: probabilidade condicional de uma amostra ser predita na categoria k_2, dado que ela pertence à categoria k_1, dada pela equação 11:

$$S = \frac{n_{1M}}{n_{1M} + n_{1C}} \tag{11}$$

b) erro 2: probabilidade condicional de uma amostra ser predita na categoria k_1, dado que ela pertence à categoria k_2, dada pela equação 12:

$$E = \frac{n_{2M}}{n_{2M} + n_{2C}} \tag{12}$$

c) acurácia: denotada também por probabilidade de acerto global, é uma estimativa para o acerto do modelo na predição das classes da variável dependente de interesse no caso de novos exemplos, calculada pela Equação 13 (WITTEN; FRANK, 2005):

$$A = \frac{n_{1C} + n_{2C}}{n_1 + n_2} \tag{13}$$

d) erro global: consiste na estimativa para o erro do modelo na predição das classes da variável dependente de interesse no caso de novos exemplos. A probabilidade de erro global (Equação 14) pode ser também obtida pelo complementar da medida acurácia (WITTEN, FRANK, 2005).

$$Er = \frac{n_{1M} + n_{2M}}{n_1 + n_2} = 1 - A \tag{14}$$

Os métodos usuais para estimar os erros de classificação incorreta de uma observação amostral são: método da resubstituição, método da resubstituição com divisão amostral, método pseudojackknife ou validação cruzada (LACHENBRUCH; MICKEY, 1968) e método das probabilidades de classificação incorretas estimadas. Em todos os casos, a classificação apresentará melhores resultados quando os erros 1 e 2 forem mínimos, além de maior acurácia (MINGOTI, 2007).

2.3 O ambiente e o processamento na qualidade sensorial da bebida do café

Algumas cultivares, dentre os materiais de *Coffea arabica* L. mais estudados, destacam-se por apresentar elevado potencial de expressão da qualidade sensorial, sendo, por isso, altamente valorizadas no mercado. Diante disso, na tentativa de identificar genótipos de café arábica com qualidade superior da bebida, Ferreira et al. (2012) encontraram diferença entre materiais de frutos amarelos e vermelhos. Esses resultados permitem inferir que diferentes genótipos de cafeeiro podem apresentar qualidades distintas, sendo mais ou menos influenciados pelo ambiente.

Dentre os fatores ambientais considerados mais impactantes na qualidade sensorial do café, a altitude e a precipitação têm sido os mais estudados (AVELINO et al., 2005; DECAZY et al., 2003; GUYOT et al., 1996). Segundo Guyot et al. (1996), para determinadas variedades de café arábica, a altitude e o sombreamento são fatores que influenciam positivamente a qualidade sensorial da bebida. Sabe-se, ainda, que altitudes elevadas e precipitação anual inferior a 1.500 mm favorecem a produção com qualidade superior (DECAZY et al., 2003).

Entretanto, para atender às necessidades do mercado como produto pronto para consumo, o café colhido passa por uma série de processos.

Tradicionalmente, os principais métodos utilizados para o processamento do café são via seca e via úmida. Por definição, via seca consiste em secar os frutos na sua forma integral. Já o processamento por via úmida pode ser realizado de diferentes formas, como removendo-se a casca e parte da mucilagem mecanicamente, dando origem ao café descascado; removendo-se a casca mecanicamente e a mucilagem por fermentação biológica, originando o café despolpado ou removendo-se a casca e a mucilagem mecanicamente, dando origem ao café desmucilado (BORÉM, 2008).

Com relação à qualidade sensorial da bebida, é comumente aceito que os cafés obtidos a partir das diferentes formas de processamento apresentam características distintas (ILLY; VIANI, 1995). De modo geral, os cafés naturais apresentam mais corpo e os cafés despolpados, acidez mais desejável. No entanto, considerando que a qualidade sensorial do café se origina na interação planta e ambiente e se define no processamento, pode-se inferir a ocorrência de eventos anteriores à colheita capazes de, junto com as operações seguintes, como o processamento pós-colheita, conferir atributos distintos para a qualidade da bebida.

3 MATERIAL E MÉTODOS

3.1 Natureza dos dados

Os dados utilizados neste trabalho são pertencentes ao projeto intitulado "Protocolo de identidade, qualidade e rastreabilidade para embasamento da indicação geográfica dos cafés da Mantiqueira". Especificamente, com este projeto, buscou-se obter a indicação geográfica, na modalidade denominação de origem, dos cafés da microrregião da serra da Mantiqueira, no estado de Minas Gerais. Para tanto, considerando a grande extensão de abrangência do projeto e a complexidade da paisagem dessa região, selecionou-se uma área piloto para estudos detalhados, incluindo a coleta de amostras de café. Dentre os municípios envolvidos, Carmo de Minas (-22°6', 45°8') foi selecionado como área de estudo por representar satisfatoriamente o ambiente característico da região quanto à pluviosidade, à temperatura, à altitude, à declividade e à área de produção de café.

Dessa forma, foram coletadas amostras de café (*Coffea arabica* L.), ao longo de duas safras agrícolas (2009/10 e 2010/11), em lavouras comerciais de propriedades localizadas no município de Carmo de Minas, Minas Gerais, Brasil.

O delineamento experimental foi baseado no estudo da interação entre variáveis ambientais, características fenotípicas e de processamento. O ambiente de cultivo do café foi estratificado em três classes de altitude (inferior a 1.000 m, entre 1.000 e 1.200 m e superior a 1.200 m) e dois grupos de vertentes, sol (NE, N, NO e O) e sombra (L, SE, S e SO), resultando na combinação de seis variáveis ambientais. Para cada um dos ambientes, foram coletados frutos vermelhos e amarelos. Assim, para todas as combinações envolvendo o ambiente de produção e a cor do fruto, foram coletadas cinco repetições e

processadas nas duas formas distintas (via seca e úmida), totalizando, dessa maneira, 120 amostras por safra.

Todos os procedimentos de colheita, processamento e secagem foram realizados segundo Borém (2008). Por fim, as amostras foram armazenadas, beneficiadas e, após a torração dos grãos, analisadas sensorialmente por quatro degustadores treinados e qualificados como juízes certificados de cafés especiais, utilizando-se a metodologia proposta pela Associação Americana de Cafés Especiais – SCAA (LINGLE, 2001). Nessa avaliação foram atribuídas notas, no intervalo de 0 a 10 pontos, para cada um dos seguintes atributos: fragrância/aroma, uniformidade, ausência de defeitos, doçura, sabor, acidez, corpo, finalização, equilíbrio e impressão global.

3.2 Variáveis utilizadas

Para compor o banco de dados do presente estudo, foram consideradas as variáveis dispostas na Tabela 10 que, além da pontuação final (Nota) e altitude da amostra, descreveram de maneira qualitativa as características sensoriais da bebida do café produzido em diferentes faces de exposição ao sol (Vertente).

Tabela 10 Variáveis e respectivas codificações de categorias utilizadas no processo da árvore de decisão

Variável	Categorias		Codificação
Processamento	Via seca (natural)		VS
	Via úmida (desmucilado)		VU
Nota	Amarelo – $Ni(i: 78, 79, \ldots, 94{,}95)$		$N >= i$
			$N < i$
	Vermelho – $Ni(i: 76, 77, \ldots, 89{,}90)$		$N >= i$
			$N < i$
Altitude	$j(j = 950, 1000, \ldots, 1300, 1350)$		$>= j$
			$< j$
Tipo de sabor	Baunilha	Presença	BA
		Ausência	N
	Chocolate	Presença	CH
		Ausência	N
	Cítrico	Presença	CI
		Ausência	N
	Floral	Presença	FL
		Ausência	N
	Frutado	Presença	FR
		Ausência	N
Tipo de acidez	Cítrica		CI
	Málica		MA
	Indefinida		I
Tipo de corpo	Cremoso		CR
	Oleoso		OL
	Indefinido		I
Intensidade da acidez	Alta		
	Média		
	Baixa		
Intensidade do corpo	Alta		
	Média		
	Baixa		
Intensidade da doçura	Alta		
	Média		
	Baixa		

A partir da variável nota final, novas variáveis *dummies*, para cada cor de fruto, foram criadas, levando-se em consideração os valores mínimo e máximo deste atributo, gerando “pontos de corte”. Para a cor do fruto amarelo,

com notas mínima e máxima de 77 e 96,25, respectivamente, variáveis *dummies* geraram 18 novas classes $i(i = 78,79,\ldots,94{,}95)$. Foi atribuída a codificação $N >= i$ quando a nota final de uma amostra foi maior ou igual a i, e codificação $N < i$, caso contrário. Para a cor de fruto vermelha, com notas mínima e máxima de 75 e 91,75 respectivamente, foram criadas variáveis *dummies* que originaram 15 novas classes $i(i = 76,77,\ldots,89{,}90)$. Atribui-se a codificação $N >= i$ quando a nota final de uma amostra foi maior ou igual a i, e codificação $N < i$, caso contrário.

A variável altitude foi categorizada a partir de variáveis *dummies*, com o propósito de identificar faixas que delimitaram (discriminaram/separaram) as características sensoriais do estudo. Com altitude mínima de 932 m e máxima de 1.391 m, foram geradas variáveis *dummies* que originaram 9 classes $j(j = 950,1000,\ldots,1300,1350)$. Foi atribuída a codificação $>= j$ quando a altitude de uma amostra foi maior ou igual a j, e codificação $< j$, caso contrário.

Para as variáveis relacionadas ao tipo de sabor, utilizou-se a moda das descrições feitas na análise sensorial pelo grupo de degustadores, quanto à presença ou à ausência (Figura 6).

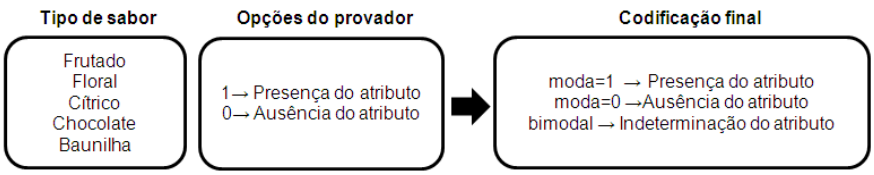

Figura 6 Codificação de variáveis relacionadas ao tipo de sabor

Para o tipo de corpo, utilizou-se a moda das descrições, obtidas na análise sensorial como critério final de classificação, conforme a Figura 7.

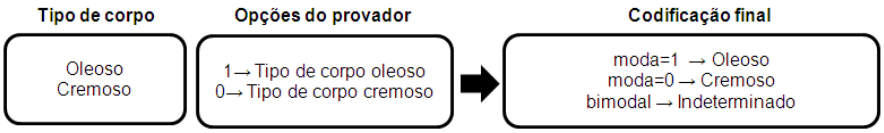

Figura 7 Codificação da variável relacionada ao tipo de corpo

Foi utilizada, também, a moda das descrições atribuídas pelos degustadores como critério final de classificação para a variável tipo de acidez (Figura 8).

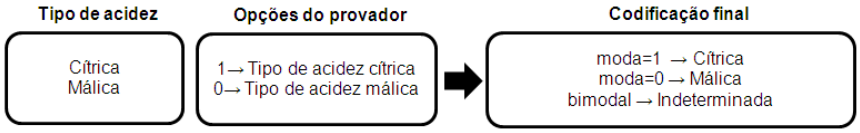

Figura 8 Codificação da variável relacionada ao tipo de acidez

As variáveis intensidade do corpo, intensidade da acidez e intensidade da doçura, classificadas inicialmente como qualitativas ordinais com três categorias (baixa, média e alta), foram recodificadas para os valores numéricos. Foram atribuídos os valores 0, 1 e 2, para a descrição das intensidades como baixa, média e alta, respectivamente. A amostra foi classificada com intensidade baixa quando a média aritmética obtida apresentou valor menor ou igual a 0,5; média, quando maior que 0,5 e menor que 1,5 e alta, quando maior ou igual a 1,5 e menor ou igual a 2 (Figura 9).

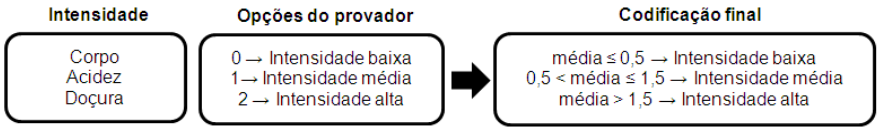

Figura 9 Codificação de variáveis relacionadas à intensidade do corpo, intensidade da acidez e intensidade da doçura

3.3 Critérios para a metodologia CHAID

Como variáveis dependentes, foram utilizadas as classes de altitude, tendo como evento de interesse a categoria $>= j$. Como preditores, foram consideradas as variáveis tipos de sabor (baunilha, chocolate, cítrico, floral, frutado), tipo de acidez, tipo de corpo, intensidade da acidez, do corpo e da doçura, bem como o processamento e a nota final codificada obtida pela análise sensorial.

Os critérios de calibração utilizados foram: nível de significância de 5% para mesclagem e/ou divisão de categorias de um preditor no passo 1.1 do algoritmo e nível de significância de 5%, para associação entre preditores e número mínimo de observações em nós seja igual a 10. O algoritmo foi encerrado ao se obter um nó puro ou quando as categorias de um preditor apresentaram valores idênticos.

Para avaliação do desempenho das árvores de decisão, acurácia, erro 1 e erro 2 foram considerados, sendo:

a) erro 1: probabilidade condicional de uma amostra ser predita na categoria $>= j$, dado que ela pertence à categoria $< j$;
b) erro 2: probabilidade condicional de uma amostra ser predita na categoria $< j$, dado que ela pertence à categoria $>= j$.

4 RESULTADOS

O estudo foi dividido em duas partes, para uma melhor compreensão dos resultados. Em cada momento, uma cor do fruto do café arábica (amarelo ou vermelho) foi abordada.

4.1 Árvores de decisão para o café arábica – cor do fruto amarelo

As árvores de decisão obtidas para o café arábica com cor de fruto amarela, indicaram melhora na acurácia (>0,95) em faixas de altitude mais baixas e mais altas (Anexo A). Embora, nessas situações, a quantidade de casos classificados corretamente tenha sido alta, as probabilidades de classificação incorreta foram máximas, justificável pelo fato de a amostra de estudo compor praticamente uma categoria da variável dependente (Figura 10). Nas faixas de altitude >=950 e >=1.350, por exemplo, os preditores considerados não apresentaram significância com a altitude, indicando que a classificação de uma amostra em maiores ou menores faixas de altitudes independe dos preditores considerados.

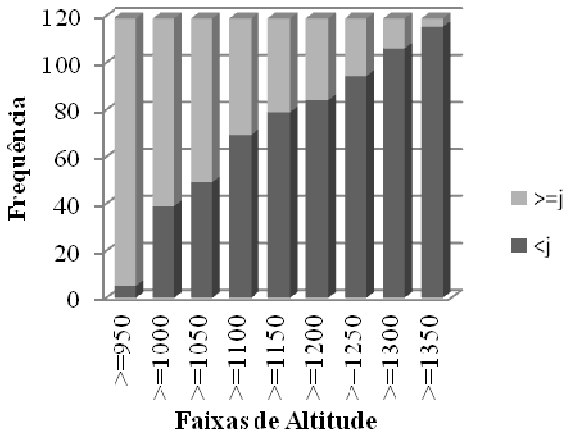

Figura 10 Frequência absoluta para as faixas de altitude do café com cor do fruto amarelo

Os modelos considerando faixas de altitude >=1.150 e >=1.200 obtiveram resultados similares (Anexo A). Foram abordados os modelos com a maior faixa de altitude (>=1.200), tendo em vista sua maior influência na qualidade do café (BARBOSA et al., 2012).

Para a nota corte >=84, obteve-se acurácia de 0,7983 e erros menores do que 0,25. Uma amostra foi classificada na altitude >=1.200, para o tipo de sabor cítrico ou indeterminado, seguido da intensidade do corpo baixa ou alta (Figura 11). Demais casos foram classificados na faixa <1.200, para a ocorrência do tipo de não cítrico, ou tipo de sabor cítrico e indeterminado seguido da intensidade do corpo ser média.

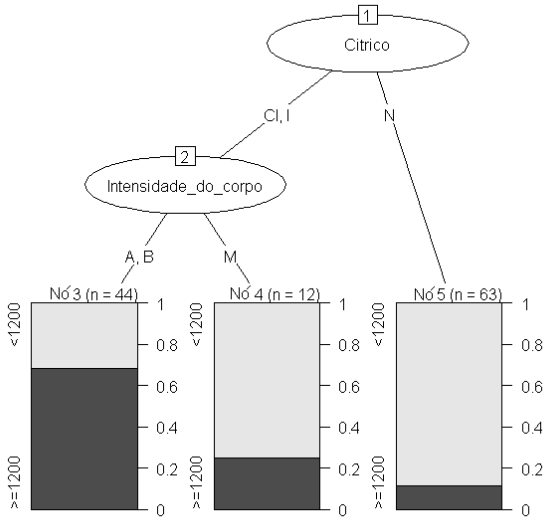

Figura 11 Árvore de decisão para a cor do fruto amarelo, considerando a faixa de altitude >=1.200 e nota corte >=84

Com classificação correta de 79,83% dos casos, 14 das 79 amostras pertencentes à faixa de altitude <1.200 (erro 1=0,1772) e 10 das 40 amostras pertencentes à faixa de altitude >=2.000 (erro 2=0,25) foram classificados incorretamente (Tabela 11).

Tabela 11 Matriz de confusão para a cor do fruto amarelo, considerando a nota corte >=84 e faixa de altitude >=1.200

	Altitude predita		
Altitude verdadeira	<1.200	>=1.200	Total
<1.200	65	14	79
>=1.200	10	30	40
Total	75	44	119

Para a nota corte >=87, obteve-se uma acurácia similar ao modelo anteriormente apresentado (0,7983), com erros inferiores a 0,22 (Anexo A). Uma amostra foi classificada na faixa de altitude >=1.200 para a nota corte >=87 e na faixa <1.200, caso contrário (Figura 12).

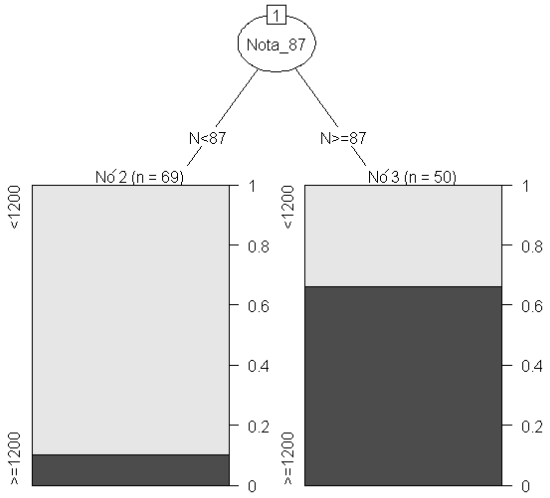

Figura 12 Árvore de decisão para a cor do fruto amarelo, considerando a faixa de altitude >=1.200 e nota corte >=87

O modelo considerando a nota corte >=87 penalizou os casos pertencentes à faixa de altitude <1.200 (erro 1=0,2152) e obteve melhor desempenho, considerando a faixa de altitude >=1.200 (erro 2=0,175) (Tabela 12).

Tabela 12 Matriz de confusão para a cor do fruto amarelo, considerando a nota corte >=87 e faixa de altitude >=1.200

	Altitude predita		
Altitude verdadeira	<1.200	>=1.200	Total
<1.200	62	17	79
>=1.200	7	33	40
Total	69	50	119

Considerando uma nota corte >=88, obteve-se maior acurácia perante os modelos apresentados (0,8487), com erros inferiores a 0,25 (Anexo A). Uma amostra foi classificada na faixa de altitude >=1.200 para a nota corte >=88, e na faixa <1.200, caso contrário (Figura 13).

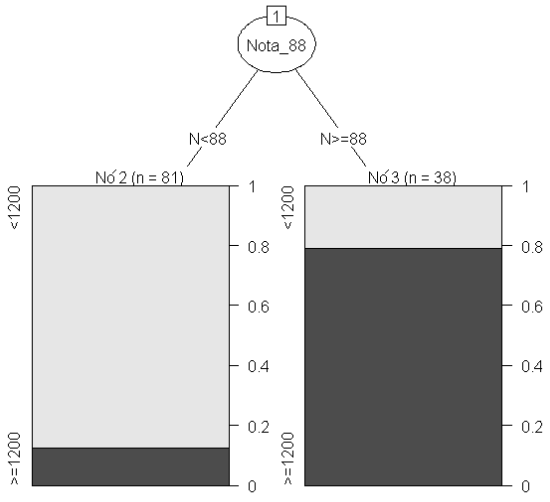

Figura 13 Árvore de decisão para a cor do fruto amarelo, considerando a faixa de altitude >=1.200 e nota corte >=88

O modelo considerando a nota corte >=88 obteve melhor desempenho, considerando a faixa de altitude <1.200 (erro 1=0,1012), por conta da

quantidade de casos classificados incorretamente ser menor, além da frequência de casos nesta categoria ser maior (Tabela 12).

Tabela 13 Matriz de confusão para a cor do fruto amarelo, considerando a nota corte >=88 e faixa de altitude >=1.200

	Altitude predita		
Altitude verdadeira	<1.200	>=1.200	Total
<1.200	71	8	79
>=1.200	10	30	40
Total	81	38	119

4.2 Árvores de decisão para o café arábica – cor do fruto vermelho

As árvores de decisão obtidas para o café arábica com cor de fruto vermelha indicaram melhora na acurácia (>0,90) em faixas de altitude mais baixas e mais altas (Anexo B), assim como na cor de fruto amarela. As probabilidades de classificação incorreta foram máximas, pelo fato de a amostra de estudo compor praticamente uma categoria da variável dependente (Figura 14). Para as faixas de altitudes inferiores ou iguais a 1.100 m ou maiores ou iguais a 1.300 m, os preditores considerados não apresentaram significância com a variável dependente, indicando que a classificação de uma amostra nessas faixas de altitudes independe dos atributos considerados.

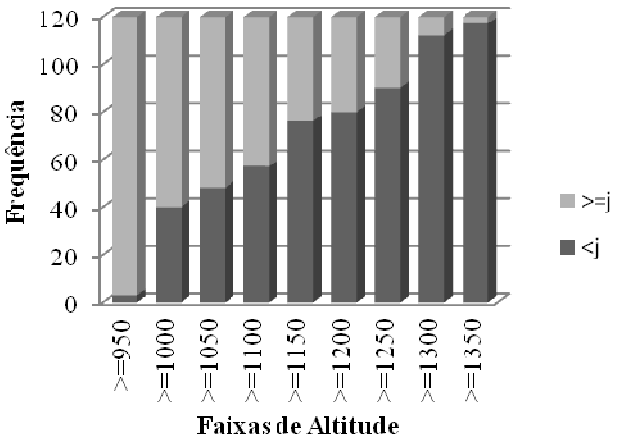

Figura 14 Frequência absoluta para as faixas de altitude do café com cor de fruto vermelha

De modo geral, os modelos apresentaram maior acurácia para maiores faixas de altitude, além de terem erros elevados, superiores a 0,5, em alguns casos (Anexo B). A árvore de decisão criada a partir da faixa de altitude >=1.250 e nota corte >=87 obteve acurácia de 0,80 e erros menores do que 0,5. Uma amostra foi classificada na altitude >=1.250 para a nota corte >=87 ou nota corte <87, seguida do tipo de sabor ser frutado ou indeterminado e tipo de sabor floral ou indeterminado (Figura 11). Nos casos restantes, as amostras foram classificadas na faixa de altitude <1.250.

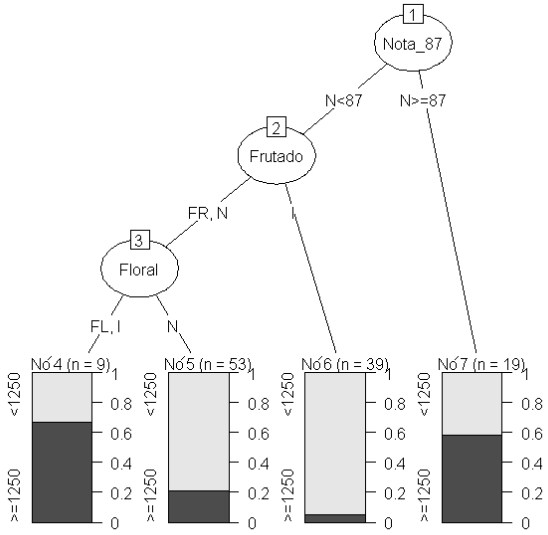

Figura 15 Árvore de decisão para a cor de fruto vermelha, considerando a faixa de altitude >=1.250 e nota corte >=87

O modelo considerando a nota corte >=87 classificou incorretamente 11 dos 90 casos pertencentes à faixa de altitude <1.250 (erro 1=0,1222) e 13 das 30 amostras pertencentes à faixa de altitude >=1.250 (erro 1=0,4333) (Tabela 14).

Tabela 14 Matriz de confusão para a cor de fruto vermelha, considerando a nota corte >=87 e faixa de altitude >=1.250

	Altitude predita		
Altitude verdadeira	<1.250	>=1.250	Total
<1.250	79	11	90
>=1.250	13	17	30
Total	92	28	119

5 CONCLUSÃO

A utilização da técnica de data mining CHAID propiciou respostas eficazes, permitindo seu emprego na pesquisa cafeeira. Os resultados obtidos permitiram concluir que as características sensoriais da bebida do café arábica estão associadas ao meio físico, dado por faixas de altitude, independente das formas de processamento consideradas.

De maneira geral, os dados referentes ao café com cor do fruto amarelo obtiveram melhores acurácias e combinações de erros, sugerindo a aplicabilidade da metodologia na descrição do perfil sensorial da qualidade deste café.

REFERÊNCIAS

ANDREICA, M. E. **Financial distress prediction of the romanian companies using chaid models.** Bucharest: Computers & applied sciences complete, 2012.

AVELINO, J. et al. Effects of slope exposure, altitude and yield on coffee quality in two altitude *terroirs* of Costa Rica, Orosi and Santa María de Dota. **Journal of Science Food and Agriculture,** Sussex, v. 85, n. 11, p. 1869-1876, Aug. 2005.

AVELINO, J. et al. **Ver une identification de cafés-terroir au Honduras.** Montpeller Cedex: Plantations Recherche Developpement, 2002.

BABCOCK, C. Computer world. In: PARALLEL MINES RETAIL DATA, 6., 1994, Cambridge. **Processing…** Cambridge: [s.n]: 1994.

BALEMBA, S.; BEAUREGARD, E. to resist or not to resist? The effect of context and crime characteristics on sex offenders' reaction to victim resistanc. **Proceedings...** Philadelphia: ASC Annual Meeting, PA, 2012. Disponível em <http://www.allacademic.com/meta/ p372117_index.html>. Acesso em: 29 jan. 2013.

BARBOSA, J. N. et al. Coffee quality and its interactions with environmental factors in Minas Gerais, Brazil. **Journal of Agricultural Science,** Cambridge, v. 4, n. 5, p. 182-190, May 2012.

BIGGS, D.;, VILLE, B. D.; SUEN, E. A method of choosing multiway partitions for classication and decision trees. **Journal of Applied Statistics**, Abingdon, v. 18, n. 1, p. 49-62, Jan. 1991.

BORÉM, F. M. **Pós-colheita do café**: volume 1. Lavras: UFLA, 2008.

BRAZIL SPECIALTY COFFEE ASSOCIATION. Mercado de cafés especiais no Brasil dobra em três anos. **Brazil Specialty**, Varginha, out. 2012. Disponível em: <http://bsca.com.br/ noticia.php?id=118>. Acesso em: 11 nov. 2012.

BREIMAN, L. et al. **Classification and regression trees.** Monterey: Wadsworth & Brooks, Wadsworth. 1984.

CHIA, F. C.; YUAN, L.; MOHAMAD, I. Flow diagram analysis of electrical fatalities in construction industry. **Safety Science**, Amsterdam, v. 50, n. 5, p. 1205-1214, 2012.

CORTES, C.; PREGIBON, D. **Mega-monitoring.** Washington: Microsoft Summer Research Institute on Data Mining, 1997.

DECAZY, F. et al. Quality of different Honduran coffees in relation to several environments. **Journal of Food Science**, Chicago, v. 68, n. 7, p. 2356–2361, 2003.

DOYLE, P.; FENWICK, I. The pitfalls of AID analysis. Journal of Marketing Research, v. 12, n. 4, 408-413, Chicago, Mar. 1975.

EINHORN, H. J. Alchemy in the behavioral sciences. **Public Opinion Quarterly**, Chicago, v. 36, n. 3, p. 367-378, 1972.

FAYYAD, U.; PIATETSKY, S. G.; SMYTH, P. The KDD process for extracting useful knowledge from volumes of data. **Communications of the ACM**, v. 59, n. 11, p. 27-34, Nov. 1996.

FERREIRA, D. A. et al. Análise sensorial de diferentes genótipos de cafeeiros Bourbon. **Interciencia**, Catanduva, v. 37, n. 5, p. 390– 394, May 2012.

FISHER, W. D. On grouping for maximum homogeneity. **American Statistical Association Journal,** New York, v. 53, p. 789-798, Dez. 1958.

GOODMAN, L. A. How to ransack social mobility tables and other kinds of cross-classification tables. Chicago**:** University of Chicago Press, 1969.

GOODMAN, L. A. Simple models for the analysis of association in cross-classifications having ordered categories. **American Statistical Association Journal,** New York, v. 74, p. 537-552, 1979.

GUYOT, B. et al. Influence de l'altitude et de l'ombrage sur la qualité des cafés arabica. **Plantation Recherche, Développement,** Versalhes, v. 3, n. 4, p. 272-280, 1996.

HAND, D. J. Data mining: statistics and more? **The American Statistician**, New York, v. 52, n. 2, 112-118, May 1998.

HAND, D. J.; MANILA, H. S. P. **Principles of data mining**. Cambridge: The MIT Press, 2001.

HAIR, F. J. et al. **Análise multivariada de dados.** 6. ed. Porto Alegre: Bookman, 2009.

HAWKINS, D. M.; KASS, G. V. Automatic interaction detection. In: HAWKINS, D. M. **Topics in applied multivariate analysis.** Cambridge: Cambridge University Press, 1982. p. 269–302.

ILLY, A.; VIANI, R. **Espresso coffee:** the chemistry of quality. London: Academic Press, 1995.

JOHNSON, R. A.; WICHERN, D. W. **Applied multivariate statistical analysis.** 6. ed. New Jersey: Pearson, 2007.

KAPIKIRAN, N. Idea-real self-concept and state-trait anxiety in Turkish university students according to CHAID analysis. **College Student Journal**, Chula Vista, v. 45, n. 4, p. 715, Dec. 2011.

KASS, G. V. An exploratory technique for investigating large quantities of categorical data. **Journal of Applied Statistics**, Cleveland, v. 29, n. 2, p. 119–127, 1980.

KASS, G. V. Significance testing in automatic interation detection. **Journal of the Royal Statistical Society. Series C. Applied Statistics**, London, v. 24, n. 2, p. 119-127, 1975.

KOHAVI, R.; PROVOST, F. Glossary of terms. **Machine Learning**, Boston, v. 30, n. 2-3, p. 271-274, 1998.

LACHENBRUCH, P. A.; MICKEY, M. R. Estimation of error rates in discriminant analysis. **Technometrics**: a journal of statistics for the physical, chemical and engineering sciences, Washington, v. 10, n. 1, 1-11, Jan. 1968.

LAHMANN, N. A.; KOTTNER, J. Relation between pressure, friction and pressure ulcer categories: A secondary data analysis of hospital patients using CHAID methods. **International Journal of Nursing Studies**, Elmsford, v. 48, n. 12, p. 1487-1494, 2011.

LATORRE, M. L. et al. Integração de dados de sensoriamento remoto multiresoluções para a representação da cobertura da terra utilizando campos contínuos de vegetação e classificação por árvores de decisão. In: SIMPÓSIO BRASILEIRO DE SENSORIAMENTO REMOTO, 13., 2007, Florianópolis. **Anais...** Florianópolis: INPE, 2007. p. 6771-6778.

LINGLE, T. R. **The coffee cupper's handbook**: a systematic guide to the sensory evaluation of coffee's flavor. 3. ed. Long Beach: Speciality Coffee Association of America, 2001.

MAGIDSON, J. The CHAID approach to segmentation modeling: chi-squared automatic interaction detection. In: BAGOZZI, R. P. (Ed.). **Advanced methods of marketing research**. Cambridge: Blackwell, 1994. P. 118–159.

MINGOTI, S. A. **Análise de dados através de métodos de estatística multivariada:** uma abordagem aplicada. Belo Horizonte: Editora UFMG, 2005.

MORA, B.; CASTELA, G. A CHAID contribution to the detection of criminal profiles: developments in trategic planning and tactical guidance of police resources in the algarve region. **Quantitative Methods Applied to Social Sciences**, Algarve, n. 3, p. 39-48, July 2010.

MORGAN, J. A.; SONQUIST, J. N. Problems in the analysis of survey data: and a proposal. **American Statistical Association Journal,** New York, v. 58, n. 302, p. 415-434, June 1963a.

MORGAN, J. A.; SONQUIST, J. N. Some results from a non-symmetrical branching process that looks for interaction effects. In: SOCIAL STATISTICS SECTION, AMERICAN, STATISTICAL ASSOCIATION. **Proceedings…** Amsterdam: ASA, 1963b. p. 40-53.

NILS, A. L.; JAN K. Relation between pressure, friction and pressure ulcer categories: A secondary data analysis of hospital patients using CHAID methods. **International Journal of Nursing Studies**, Elmsford, v. 48, n. 12, p. 1487-1494, 2011.

OUYANG, J.; PATEL, N.; SETHI, I. K. Chi-square test based decision trees induction in distributed environment. In: INTERNATIONAL CONFERENCE ON DATA MINING WORKSHOPS, 8., 2008, Washington. **Proceeding...** Washington: IEEE Computer Society, 2008. p. 477-485. Disponível em: <http://dl.acm.org/citation.cfm?id=1490299.1490729>. Acesso em: 20 nov. 2012.

PETROFF, C.; BETTEX, A. M.; KORRY, A. **Itineraires d'etudiants a la faculte es sciences economiques et sociales:** le premier cycle. Geneve: Faculte SES, 2001.

PREZEPIORSKI, E. L.; ARNS, M. T. S; NIEVOLA, J. C. Análise de crédito bancário por meio de redes neurais e árvores de decisão: uma aplicação simples de data mining. **Revista de Administração**, São Paulo, v. 40, n. 3, p. 225-234, jul./ago./set. 2005. Disponível em: <http://redalyc.uaemex.mx/src/inicio/ ArtPdfRed.jsp?iCve=223417392002>. Acesso em: 16 nov. 2012.

QUINLAN, J. R. Induction of decision trees. **Machine Learning**, Boston, v. 1, n. 1, p. 81-106, 1986.

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

R DEVELOPMENT CORE TEAM. **R:** a language and environment for statistical computing. Vienna: R Foundation for Statistical Computing, Austria. Disponível em: <URL http://www.R-project.org/>. Acesso em: 16 nov. 2012.

RITSCHARD, G. **CHAID and earlier supervised tree methods**. Geneve: Cahiers du Département d'économétrie, 2010. Disponível em: <http://www. unige.ch/ ses/metri/cahiers/2010_02.pdf>. Acesso em: 10 set. 2011.

SHANNON, C. E. A mathematical theory of communication. **The Bell System Technical Journal**, Oxford, v. 27, p. 379–423, Oct. 1948.

SICILIANO, R.; MOLA, F. Multivariate data analysis through classification and regression trees. **Computational Statistics & Data Analysis**, Amsterdam, v. 32, p. 285–301, 2000.

SILBERSCHATZ, A.; KORTH, H. F.; SUDARSHAN, S. **Sistema de banco de dados**. São Paulo: Pearson Education do Brasil, 1999.

SILVA, J. P. D. **Algoritmos de classificação baseados em análise formal de conceitos**. 2007. 96 f. Dissertação (Mestrado em Ciências da Computação) - Universidade Federal de Minas Gerais, Belo Horizonte.

SILVEIRA, M.; ECHEVESTE, M. E. S. Utilização de árvores de decisão (CHAID) para alinhamento de atributos no desenvolvimento de novo produto. IN: CONGRESSO BRASILEIRO DE GESTÃO DE DESENVOLVIMENTO DE PRODUTO, 8., 2011, Porto Alegre. **Anais…** Porto Alegre: ABEPRO, 2011.

STEVENSON J. C. et. al. Technical note: prediction of sex based on five skull traits using decision analysis (CHAID). **American Journal of Physical Anthropology,** Washington, v. 139, n. 3, p. 434–441, 2009.

SUKNOVIC, M. et al. Reusable components in decision trees induction algorithms. **Computational Statistics**, Heidelberg, v. 27, n. 1, p. 127-148, 2011.

TANHAN, F.; KAYRÍ, M. An examination of the factors affecting prospective teachers' perceptions of faculty members using chaid analysis. **Kuram ve Uygulamada Egitim Bilimleri**, Turkey, v. 12, n. 2, p. 816–821, 2012.

WEISS, S. M.; KULIKOWSKY, C. A. **Computer systems that learn**. San Mateo: Morgan Kaufmann, 1991.

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

ANEXOS

ANEXO A

Tabela 1A Medidas de desempenho para as árvores de decisão do café arábica com cor de fruto amarela

			Altitude								
		Desempenho	950	1.000	1.050	1.100	1.150	1.200	1.250	1.300	1.350
Nota	78	Acurácia	0,9579	0,7311	0,7143	0,7311	0,7983	0,7983	0,8067	0,8571	0,9664
		Erro 1	1,0000	0,5897	0,3061	0,2754	0,1772	0,1772	0,1596	0,0000	0,0000
		Erro 2	0,0000	0,1125	0,2714	0,2600	0,2500	0,2500	0,3200	1,0000	1,0000
	79	Acurácia	0,9579	0,7311	0,7143	0,7311	0,7983	0,7983	0,8067	0,8571	0,9664
		Erro 1	1,0000	0,5897	0,3061	0,2754	0,1772	0,1772	0,1596	0,0000	0,0000
		Erro 2	0,0000	0,1125	0,2714	0,2600	0,2500	0,2500	0,3200	1,0000	1,0000
	80	Acurácia	0,9579	0,7311	0,7143	0,7311	0,7983	0,7983	0,8067	0,8571	0,9664
		Erro 1	1,0000	0,5897	0,3061	0,2754	0,1772	0,1772	0,1596	0,0000	0,0000
		Erro 2	0,0000	0,1125	0,2714	0,2600	0,2500	0,2500	0,3200	1,0000	1,0000
	81	Acurácia	0,9579	0,7479	0,7143	0,7311	0,7983	0,7983	0,8067	0,8571	0,9664
		Erro 1	1,0000	0,6923	0,3061	0,2754	0,1772	0,1772	0,1596	0,0000	0,0000
		Erro 2	0,0000	0,0375	0,2714	0,2600	0,2500	0,2500	0,3200	1,0000	1,0000
	82	Acurácia	0,9579	0,7479	0,7143	0,7311	0,7983	0,7983	0,8067	0,8571	0,9664
		Erro 1	1,0000	0,5897	0,3061	0,2754	0,1772	0,1772	0,1596	0,0000	0,0000
		Erro 2	0,0000	0,0875	0,2714	0,2600	0,2500	0,2500	0,3200	1,0000	1,0000
	83	Acurácia	0,9579	0,7395	0,7143	0,7311	0,7983	0,7983	0,8067	0,8571	0,9664
		Erro 1	1,0000	0,5385	0,3061	0,2754	0,1772	0,1772	0,1596	0,0000	0,0000
		Erro 2	0,0000	0,1250	0,2714	0,2600	0,2500	0,2500	0,3200	1,0000	1,0000

Tabela 1A, continua

		Altitude								
	Desempenho	950	1.000	1.050	1.100	1.150	1.200	1.250	1.300	1.350
84	Acurácia	0,9579	0,7311	0,7311	0,7311	0,7983	0,7983	0,8067	0,8571	0,9664
	Erro 1	1,0000	0,4102	0,4286	0,2754	0,1772	0,1772	0,1596	0,0000	0,0000
	Erro 2	0,0000	0,2000	0,1571	0,2600	0,2500	0,2500	0,3200	1,0000	1,0000
85	Acurácia	0,9579	0,7059	0,7395	0,7227	0,7227	0,7227	0,7899	0,8571	0,9664
	Erro 1	1,0000	0,3077	0,3061	0,3768	0,3924	0,3924	0,0000	0,0000	0,0000
	Erro 2	0,0000	0,2875	0,2286	0,1400	0,0500	0,0500	1,0000	1,0000	1,0000
86	Acurácia	0,9579	0,7479	0,7479	0,7563	0,7899	0,7899	0,7899	0,8571	0,9664
	Erro 1	1,0000	0,6410	0,3061	0,2609	0,2658	0,2658	0,0000	0,0000	0,0000
	Erro 2	0,0000	0,0625	0,2143	0,2200	0,1000	0,1000	1,0000	1,0000	1,0000
87	Acurácia	0,9579	0,7395	0,7143	0,7479	0,7983	0,7983	0,8571	0,8571	0,9664
	Erro 1	1,0000	0,6410	0,3061	0,2174	0,2152	0,2152	0,1170	0,0000	0,0000
	Erro 2	0,0000	0,0750	0,2714	0,3000	0,1750	0,1750	0,2400	1,0000	1,0000
88	Acurácia	0,9579	0,7479	0,7143	0,7815	0,8487	0,8487	0,8235	0,8571	0,9664
	Erro 1	1,0000	0,6154	0,3061	0,1014	0,1012	0,1012	0,1809	0,0000	0,0000
	Erro 2	0,0000	0,0750	0,2714	0,3800	0,2500	0,2500	0,1600	1,0000	1,0000
89	Acurácia	0,9579	0,7311	0,7143	0,7395	0,8235	0,8235	0,8151	0,8571	0,9664
	Erro 1	1,0000	0,5897	0,3061	0,0580	0,0506	0,0506	0,1277	0,0000	0,0000
	Erro 2	0,0000	0,1125	0,2714	0,5400	0,4250	0,4250	0,4000	1,0000	1,0000
90	Acurácia	0,9579	0,7311	0,7143	0,7311	0,8151	0,8151	0,8403	0,8571	0,9664
	Erro 1	1,0000	0,5897	0,3061	0,0145	0,0127	0,0127	0,0745	0,0000	0,0000
	Erro 2	0,0000	0,1125	0,2714	0,6200	0,5250	0,5250	0,4800	1,0000	1,0000
91	Acurácia	0,9579	0,7311	0,7143	0,7311	0,7815	0,7815	0,8319	0,8571	0,9664
	Erro 1	1,0000	0,5897	0,3061	0,2754	0,0127	0,0127	0,0851	0,0000	0,0000
	Erro 2	0,0000	0,1125	0,2714	0,2600	0,6250	0,6250	0,4800	1,0000	1,0000

Tabela 1A, conclusão

		Altitude								
	Desempenho	950	1.000	1.050	1.100	1.150	1.200	1.250	1.300	1.350
92	Acurácia	0,9579	0,7311	0,7143	0,7311	0,7983	0,7983	0,8319	0,8655	0,9664
	Erro 1	1,0000	0,5897	0,3061	0,2754	0,1772	0,1772	0,0319	0,0490	0,0000
	Erro 2	0,0000	0,1125	0,2714	0,2600	0,2500	0,2500	0,6800	0,6471	1,0000
93	Acurácia	0,9579	0,7311	0,7143	0,7311	0,7983	0,7983	0,8235	0,8571	0,9664
	Erro 1	1,0000	0,5897	0,3061	0,2754	0,1772	0,1772	0,0000	0,0000	0,0000
	Erro 2	0,0000	0,1125	0,2714	0,2600	0,2500	0,2500	0,8400	1,0000	1,0000
94	Acurácia	0,9579	0,7311	0,7143	0,7311	0,7983	0,7983	0,8067	0,8739	0,9664
	Erro 1	1,0000	0,5897	0,3061	0,2754	0,1772	0,1772	0,1596	0,0000	0,0000
	Erro 2	0,0000	0,1125	0,2714	0,2600	0,2500	0,2500	0,3200	0,8824	1,0000
95	Acurácia	0,9579	0,7311	0,7143	0,7311	0,7983	0,7983	0,8067	0,8739	0,9664
	Erro 1	1,0000	0,5897	0,3061	0,2754	0,1772	0,1772	0,1596	0,0000	0,0000
	Erro 2	0,0000	0,1125	0,2714	0,2600	0,2500	0,2500	0,3200	0,8824	1,0000

ANEXO B

Tabela 1B Medidas de desempenho para as árvores de decisão do café arábica com cor de fruto vermelha

			Altitude								
		Desempenho	950	1.000	1.050	1.100	1.150	1.200	1.250	1.300	1.350
Nota	76	Acurácia	0,9917	0,6667	0,6833	0,6417	0,7083	0,7250	0,7583	0,9333	0,9750
		Erro 1	1,0000	1,0000	0,6458	0,6491	0,1200	0,1375	0,1556	0,0000	0,0000
		Erro 2	0,0000	0,0000	0,0972	0,0952	0,5778	0,5500	0,5000	1,0000	1,0000
	77	Acurácia	0,9917	0,6667	0,6000	0,6417	0,7083	0,7250	0,7583	0,9333	0,9750
		Erro 1	1,0000	1,0000	1,0000	0,6491	0,1200	0,1375	0,1556	0,0000	0,0000
		Erro 2	0,0000	0,0000	0,0000	0,0952	0,5778	0,5500	0,5000	1,0000	1,0000
	78	Acurácia	0,9917	0,6667	0,6833	0,6417	0,7083	0,7250	0,7583	0,9333	0,9750
		Erro 1	1,0000	1,0000	0,6458	0,6491	0,1200	0,1375	0,1556	0,0000	0,0000
		Erro 2	0,0000	0,0000	0,0972	0,0952	0,5778	0,5500	0,5000	1,0000	1,0000
	79	Acurácia	0,9917	0,6667	0,6833	0,6417	0,7083	0,7250	0,7583	0,9333	0,9750
		Erro 1	1,0000	1,0000	0,6458	0,6491	0,1200	0,1375	0,1556	0,0000	0,0000
		Erro 2	0,0000	0,0000	0,0972	0,0952	0,5778	0,5500	0,5000	1,0000	1,0000
	80	Acurácia	0,9917	0,6667	0,6833	0,6417	0,7083	0,7250	0,7583	0,9333	0,9750
		Erro 1	1,0000	1,0000	0,6458	0,6491	0,1200	0,1375	0,1556	0,0000	0,0000
		Erro 2	0,0000	0,0000	0,0972	0,0952	0,5778	0,5500	0,5000	1,0000	1,0000
	81	Acurácia	0,9917	0,6667	0,6833	0,6417	0,7083	0,7250	0,7583	0,9333	0,9750
		Erro 1	1,0000	1,0000	0,6458	0,6491	0,1200	0,1375	0,1556	0,0000	0,0000
		Erro 2	0,0000	0,0000	0,0972	0,0952	0,5778	0,5500	0,5000	1,0000	1,0000
	82	Acurácia	0,9917	0,6667	0,6833	0,6417	0,7083	0,7250	0,7583	0,9333	0,9750
		Erro 1	1,0000	1,0000	0,6458	0,6491	0,1200	0,1375	0,1556	0,0000	0,0000
		Erro 2	0,0000	0,0000	0,0972	0,0952	0,5778	0,5500	0,5000	1,0000	1,0000
	83	Acurácia	0,9917	0,9917	0,6833	0,6417	0,7083	0,7250	0,7583	0,9333	0,9750
		Erro 1	1,0000	1,0000	0,6458	0,6491	0,1200	0,1375	0,1556	0,0000	0,0000

	Erro 2	0,0000	0,0000	0,0972	0,0952	0,5778	0,5500	0,5000	1,0000	1,0000

Tabela 1B, conclusão

		Altitude								
	Desempenho	950	1.000	1.050	1.100	1.150	1.200	1.250	1.300	1.350
84	Acurácia	0,9917	0,6667	0,6417	0,6417	0,7083	0,7250	0,7583	0,9333	0,9750
	Erro 1	1,0000	1,0000	0,2708	0,6491	0,1200	0,1375	0,1556	0,0000	0,0000
	Erro 2	0,0000	0,0000	0,4167	0,0952	0,5778	0,5500	0,5000	1,0000	1,0000
85	Acurácia	0,9917	0,6667	0,6083	0,6583	0,7083	0,7250	0,8000	0,9333	0,9750
	Erro 1	1,0000	1,0000	0,1667	0,6491	0,1200	0,1375	0,0889	0,0000	0,0000
	Erro 2	0,0000	0,0000	0,5417	0,0635	0,5778	0,5500	0,5333	1,0000	1,0000
86	Acurácia	0,9917	0,6667	0,6000	0,6500	0,7083	0,7250	0,7583	0,9333	0,9750
	Erro 1	1,0000	1,0000	0,0625	0,6491	0,1333	0,1375	0,1556	0,0000	0,0000
	Erro 2	0,0000	0,0000	0,6250	0,0794	0,5556	0,5500	0,5000	1,0000	1,0000
87	Acurácia	0,9917	0,6667	0,6833	0,6417	0,7083	0,7333	0,8000	0,9333	0,9750
	Erro 1	1,0000	1,0000	0,6458	0,6491	0,1200	0,1750	0,1222	0,0000	0,0000
	Erro 2	0,0000	0,0000	0,0972	0,0952	0,5778	0,4500	0,4333	1,0000	1,0000
88	Acurácia	0,9917	0,6667	0,6000	0,6417	0,7083	0,7417	0,7583	0,9333	0,9750
	Erro 1	1,0000	1,0000	1,0000	0,6491	0,1200	0,1500	0,1556	0,0000	0,0000
	Erro 2	0,0000	0,0000	0,0000	0,0952	0,5778	0,4750	0,5000	1,0000	1,0000
89	Acurácia	0,9917	0,6667	0,6000	0,6417	0,7083	0,7417	0,7583	0,9333	0,9750
	Erro 1	1,0000	1,0000	1,0000	0,6491	0,1200	0,1375	0,1556	0,0000	0,0000
	Erro 2	0,0000	0,0000	0,0000	0,0952	0,5778	0,5000	0,5000	1,0000	1,0000
90	Acurácia	0,9917	0,6667	0,6833	0,6417	0,7083	0,7250	0,7583	0,9333	0,9750
	Erro 1	1,0000	1,0000	0,6458	0,6491	0,1200	0,1375	0,1556	0,0000	0,0000
	Erro 2	0,0000	0,0000	0,0972	0,0952	0,5778	0,5500	0,5000	1,0000	1,0000