DOI:10.5748/9788599693131-14CONTECSI/PS-4782

DATA MINING APPLICATION TO IDENTIFY ATTRIBUTES THAT INFLUENCE THE SUGARCANE YIELD

Maria das Graças J. M. Tomazela (Universidade Metodista de Piracicaba, Faculdade de Tecnologia de Indaiatuba, São Paulo, Brasil) – gtomazela@fatecindaiatuba.edu.br
Fernando Celso de Campos (Universidade Metodista de Piracicaba, São Paulo, Brasil) – fccampos@unimep.br
Luiz Antonio Daniel (Faculdade de Tecnologia de Indaiatuba, São Paulo, Brasil) – daniel51@terra.com.br

Brazil is one of the global leaders in producing and exporting lots of agricultural products. In order to keep this prominent position, it's important to apply computational methods that give support to managers in their planning and decision making activities. Therefore, the main goal of this work is to study the influence of various factors that controlthe sugarcane yield in a significant area of production in Brazil. Firstly, the clustering algorithm named K-means was used to identify the productivity groups. After that, the J48 classification algorithm was used to identify the most influent attributes in each of these groups. The soil, fertility and the texture were considerably distinctive in each of the clusters, nevertheless, the fertilization levels were similar on the 3 clusters. The attribute soil fertility was the most influent to define the groups. The soil texture and the sugarcane variety placed in the second hierarchic level of the decision tree. The overall accuracy of the model was 99.66%.

Keywords: Yield; Sugarcane; Data Mining; K-means; Decision Tree

APLICAÇÃO DE MINERAÇÃO DE DADOS PARA IDENTIFICAR ATRIBUTOS QUE INFLUENCIAM A PRODUTIVIDADE DA CANA-DE-AÇÚCAR

O Brasil é um dos líderes globais na produção e exportação de produtos agrícolas. Para manter esta posição de destaque, é importante aplicar métodos computacionais que dão apoio aos gestores em suas atividades de planejamento e tomada de decisão. Assim, o objetivo deste trabalho é estudar a influência dos vários fatores que controlam a produtividade da cana-de-açúcar em uma área de produção significativa no Brasil. Em primeiro lugar, o algoritmo de agrupamento denominado K-means foi utilizado para identificar os grupos de produtividade. Em seguida, o algoritmo de classificação J48 foi utilizado para identificar os atributos mais influentes em cada um desses grupos. O solo, a fertilidade ea textura foram consideravelmente distintos em cada um dos grupos, no entanto, os níveis de fertilização foram semelhantes nos três grupos. O atributo fertilidade do solo foi o mais influente para definir os grupos. A textura do solo ea variedade de cana-de-açúcar foram colocadas no segundo nível hierárquico da árvore de decisão. A acurácia geral do modelo foi de 99,66%.

Palavras-chave: Produtividade; Cana-de-açúcar; Mineração de dados,K-means; Árvore de Decisão

1. Introdução

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

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

As regiões produtoras de cana-de-açúcar concentram-se nos subsistemas regionais Centro-Sul e Norte-Nordeste. No Centro-Sul, destaca-se o Estado de São Paulo, que concentra mais de 50% da produção do país (Conab, 2015). As projeções do agronegócio para a safra 2023/2024 indicam que a produção de cana-de-açúcar do Estado de São Paulo deve ter um aumento de cerca de 24,6% na próxima década. As projeções indicam ainda que apenas em Minas Gerais o aumento da produção se dará pelos ganhos em produtividade. Nos demais estados, o crescimento previsto da produção se fará, principalmente, pelo aumento de área plantada (Brasil, 2014).

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

Ressalta-se ainda que a cana-de-açúcar possui uma relevante função estratégica na economia do país, pois o aquecimento global e a busca por alternativas que substituam a queima de combustíveis fósseis faz do etanol uma importante fonte de energia renovável, uma vez que é uma das melhores opções para reduzir a emissão de gases causadores do efeito estufa, haja vista que a sua queima como combustível reduz em 70% a emissão de CO_2 na atmosfera em relação à gasolina (Conab, 2015).

Diante do cenário exposto, verifica-se que o setor sucroenergético tem grande relevância para geração de saldo positivo na balança comercial brasileira, e sua contínua modernização e adequação à realidade do mercado impactam favoravelmente o desenvolvimento econômico do país. Destaca-se também que esse setor é uma *commodity*, dessa maneira o preço dos produtos é definido pelo mercado. Assim, aumentar os níveis de produtividade da cana-de-açúcar, tanto pelo aumento de produção como pela redução de custos, é uma atividade imprescindível para a manutenção do país em sua posição de destaque nesse mercado.

Entretanto, salienta-se que sistemas agrícolas são suscetíveis à variabilidade climática e biofísica (pragas, doenças, etc.), e isso aumenta muito a complexidade do planejamento e das tomadas de decisão subjacentes, desta forma, o uso das tecnologias de informação pode contribuir para melhorar a eficiência da gestão desses sistemas e, em consequência, obter melhores níveis de produtividade.

Técnicas de mineração de dados são projetadas para identificar relacionamentos implícitos em grandes bancos de dados que envolvem um grande

número de variáveis, além disso são capazes de identificar novos padrões, dar maior precisão em padrões conhecidos e modelar fenômenos do mundo real. De acordo com Tsai (2012) essa tecnologiafornecediversas metodologias para a tomada dedecisão, resolução de problemas, análise, planejamento, diagnóstico, detecção, integração, prevenção, aprendizagem e inovação.

A mineração de dadosé um campo interdisciplinarque combinainteligência artificial, gerenciamento de banco de dados,visualização de dados,aprendizagem de máquina, algoritmos matemáticose técnicasestatísticas (Han & Kamber, 2006; Tsai, 2012).Faz parte de um processo denominado "Descoberta de Conhecimento em Bases de Dados", conhecido como KDD (*Knowledge Discovery in Databases*).

Dessa forma, o objetivo deste trabalho foi estudar a influência dos vários fatores que controlam a produtividade da cana em uma importante área de produção de cana no Brasil, por meio da mineração de dados e, dessa forma, propiciar maior precisão no gerenciamento dessa cultura.

2. Mineração de dados da cana-de-açúcar

Um dos grandes desafios enfrentados pela agricultura brasileira é o desenvolvimento de técnicas e tecnologias que possam elevar os patamares de produtividade de cultivares como soja, café, cana-de-açúcar entre outros. O intuito é manter-se competitiva em um mercado cada vez mais acirrado e exigente.

Por essa razão existem diversos modelos de produtividade de cana-de-açúcar, propostos na literatura, utilizando as mais diferentes técnicas. Alguns trabalhos sugerem o uso de modelos matemáticos como os trabalhos de Rodriges Jr (2012), Marin e Carvalho (2012) e Silva, Bergamasco, Rodrigues, Godoy e Trivelin(2006). Já outros propõem a aplicação de modelos estatísticos baseados em regressão não linear, a exemplo de Bajpai, Prya eMalik (2012), e regressão linear múltipla, exemplificados com os estudos de Simões, Rocha eLamparelli (2005).

A característica interdisciplinar das técnicas de mineração de dados, bem como sua capacidade para trabalhar com um grande volume de dados, tem despertado a atenção dos pesquisadores para o uso dessa técnica na área agrícola. Vários desses trabalhos utilizam mineração de dados para análise de dados espectrais. Por exemplo, Everingham, Lowe, Donald, Coomans e Markley (2007) utilizaram dados espectrais para determinar a variedade da cana-de-açúcar e o estágio da plantação, enquanto Goltz, Arcoverde,Aguiar, Rudorff e Maeda(2009) utilizaram dados espectrais para classificar os tipos de colheita de cana sob diferentes tipos de solo. Fang, Li e Chen (2009) fizeram uso de uma biblioteca de dados espectrais em um sistema especialista para identificação de culturas, entre elas a cana-de-açúcar. Vieira *et al.* (2012) apresentaram um modelo de conhecimento para mapear as áreas de plantação de cana-de-açúcar prontas para a colheita. O trabalho de Nonato e Oliveira (2013) utilizou dados de satélite para a identificação de áreas de cultivo de cana.

Mineração de dados espectrais obtidos por meio de imagens NDVI (*Normalized Difference Vegetation Index*) foram utilizados ainda no trabalho de Gonçalves *et al.* (2011) que avaliou a produtividade da cana-de-açúcar em uma escala regional. Em Romani*et al.*(2011) foram mineradas imagens NDVI com o objetivo de monitorar a expansão das culturas de cana-de-açúcar. O uso de imagens NDVI foi combinado com dados meteorológicos para inferir sobre dados de produtividade de municípios e safras previamente selecionadas no trabalho de Fernandes, Rocha e Lamparelli (2011). Imagens NDVI e séries temporais de dados climáticos foram mineradas em um sistema de informações utilizado em Romani *et al*. (2013). Imagens NDVI do uso do solo foram utilizadas também no trabalho de Vintrou, Ienco, Béguée Teisseire (2013) para mapeamento da terra cultivada na África ocidental.

A mineração de dados agrícolas tem sido realizada por meio de diferentes tarefas e técnicas. O trabalho de Everingham *et al.* (2007) utilizou as técnicas de classificação *Support Vector Machines* (SVM) e *Random Forest.* Vintrou *et al.*(2013) apresentaram um algoritmo de classificação original baseado em padrões sequenciais. As pesquisas de Nonato e Oliveira (2013), Vieira *et al.* (2012), Ferraro, Ghersa e Rivero (2012),Fernandes*et al.* (2011), Souza *et al.* (2010), Fang*et al.* (2009), Goltz *et al.(*2009) e Ferraro, Rivero e Ghersa (2009) também utilizaram a tarefa de classificação, mas a técnica utilizada nesses trabalhos foi a indução por árvore de decisão, que é uma das principais técnicas de mineração de dados pela sua expressividade simbólica.

A tarefa de *clusterização*, com os algoritimos *K-means* e *K-medoids*, foi utilizada no trabalho de Romani*et al.* (2011). Gonçalves *et al.* (2011) utilizaram a técnica de redução de dimensão denominada Análise de Componentes Principais (PCA) e também fizeram uso do algoritmo de *clusterização K-means*. Um algoritmo de clusterização, baseado no comportamento das abelhas (*bee hive*), denominado CRY foi apresentado e comparado com outros algoritmos no trabalho de Ananthara, Arunkumar e Hemavathy (2013).

A aplicação de um novo algoritmo baseado em regras de associação foi verificada no trabalho de Romani *et al.* (2013). O algoritmo, denominado CLEARMiner foi incorporado em um sistema de informações de sensoriamento remoto desenvolvido para melhorar o acompanhamento dos campos de cana-de-açúcar. As regras de associação também foram utilizadas no desenvolvimento de um sistema de recomendação para conteúdos relacionados à cultura da cana-de-açúcar em Barros, Oliveira e Oliveira (2013), que utilizaram dados de navegação do usuário em páginas *Web* para a aplicação das regras de associação.Vale ressaltar que esse é um uso típico dessa técnica de mineração de dados.

3. Materiais e métodos

O método de pesquisa adotado, segundo Nakano (2010), é categorizado como Modelagem (e método analítico), porque a partir dos dados coletados há um tratamento estatístico cujo objetivo é caracterizar grupos de produtividade.

Inicialmente é realizada uma breve apresentação da usina que forneceu as informações, na sequência, são apresentadas as características dos dados disponibilizados.Este trabalho utiliza uma técnica de *clusterização* para caracterizar grupos de produtividade da cana-de-açúcar em seguida utiliza uma técnica de classificação para identificar os fatores que mais impactam na produtividade da cana-de-açúcar.

3.1 A usina

Neste trabalho, são utilizados os dados do censo varietal qualitativo referentes à cana-de-açúcar – 3 safras, 2006/2007 a 2008/2009, cedidos por um dos maiores grupos sucroenergéticos do Brasil, segundo a UNICA (União da Indústria de Cana-de-Açúcar) (UNICA, 2016), sediado no interior do Estado de São Paulo. O Grupo possui quatro usinas em operação, duas delas produzem açúcar e etanol, uma é dedicada à produção exclusiva de etanol e outra à produção de derivados de levedura. As usinas geram também energia elétrica a partir da queima do bagaço da cana (cogeração), garantindo autossuficiência e venda do excedente.

Segundo informações do site da empresa, o índice médio de mecanização da colheita do grupo é de 94%, chegando a 100% em uma das usinas, índices considerados referência no setor. A companhia compra, cultiva, colhe e processa a principal matéria prima usada na produção de açúcar e álcool. Na

safra 2016/2017, foram processadas um total de 19.281 milhões de toneladas de cana que resultaram em 1.301 toneladas de açúcar e 667 mil m³ de etanol.

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

3.2. Os dados

A planilha com os dados do censo varietal contém em cada instância os seguintes atributos: código da fazenda, código da gleba, código do talhão, tipo de solo, variedade da cana, datas (divididas em: plantio, corte 1, corte anterior e corte atual), estágio de corte, tipo de corte, condição de corte, fórmula do adubo, adubação, fertilidade, textura e produtividade.

Os atributos código da fazenda, código da gleba e código do talhãoidentificam cada instânciae, na condição de identificadores,não foram necessários neste processo de mineração de dados, as datas também não foram usadas porque o atributo estágio de corte resume estas informações. Desta forma o conjunto de dados resultante para este estudo é composto por 10 atributos referentes às características de 21.078 instâncias.

O ***atributo solo*** contém o código referente à classificação do tipo do solo, de acordo com a classificação brasileira de tipos de solo. Traz informação do solo em vários níveis: o primeiro nível diz respeito à classe do solo, de acordo com a morfologia (latossolo, argissolo, etc); o segundo nível considera as cores no horizonte B (horizontes são camadas mais ou menos paralelas à superfície do terreno, diferenciadas pela cor, textura e estrutura); o terceiro nível considera as condições químicas do horizonte subsuperficial (eutrófico, distrófico, etc).Detalhes dessatipificação podem ser encontrados em Prado *et al.*(2008). A base de dados estudada tem 39 tipos de solos distintos. O ***atributo variedade*** diz respeito à cultivar da cana-de-açúcar, são plantadas 76 diferentes cultivares. O ***estágio de corte*** é representado por um número que registra duas informações, a primeira representa o total de vezes que a cana foi cortada, e a segunda informa se a cana foi colhida em 12 ou 18 meses (cana de "ano" ou de "ano e meio"), por exemplo, o valor 3.12 indica terceiro corte de uma cana colhida em 12 meses.

O ***atributo tipo de corte*** informa se o corte da cana foi manual ou mecanizado e a ***condição de corte*** se a cana foi colhida após queima ou crua. A ***formulação do adubo*** informa resumidamente a fórmula do adubo utilizado no talhão, com 9 diferentes tipos de fórmulas, e a ***adubação*** diz respeito à quantidade desse adubo que foi aplicado, expresso em kg por hectare.

A***fertilidade do solo*** é representada por5 diferentes códigos: 1–Alta;2 – Média Alta; 3 –Média; 4 – Média baixa e 5 – Baixa. A***textura***refere-se à proporção de argila, silte e areia do solo, são utilizados os seguintes códigos: 1 – solo argiloso; 2 – solo arenoso e 3- solo argiloso/arenoso.

O ***atributo produtividade*** informa a quantidade de cana colhida no talhão em toneladas por hectare.

3.3*Clusterização*

A tarefa de *clusterização*consiste em particionar os registros da base de dados em subconjuntos (ou*clusters*) de maneira que elementos presentes em um *cluster* compartilhem um conjunto de propriedades comuns e que os diferenciem dos elementos de outros *cluster*s. Optou-se pela utilização do *k-means*, pela sua complexidade linear. O algoritmo *k-means* requer que seja informado o número de *cluster*s desejados. Como não se sabia a priori o número de *clusters* ideal, foram realizados testes com 3, 4 e 5 *clusters*. Após análise dos valores estatísticos fornecidos pela ferramenta Weka utilizada no processo de mineração, e por entender que essa opção é a mais prática para decisões gerenciais, decidiu-se pela divisão em 3 *cluster*s.

De acordo com as médias da produtividade apresentadas na Tabela 1, o *cluster* 0 foi designado como de produtividade alta, o *cluster* 1 como de produtividade baixa e o *cluster* 2 como de produtividade média. O total de instâncias em cada *cluster* também é apresentado na Tabela 1, salienta-se que o *cluster* 1 que apresentou menor produtividade média é o menor entre os 3 *cluster*s.

Tabela 1- Características dos *cluster*s

	Cluster 0	**Cluster 1**	**Cluster 2**
Média da produtividade	91.91988	79.30534	87.52384
Erro padrão	0.297736	0.424962	0.251816
Desvio padrão	27.0858	28.54215	22.9277
Assimetria	2.169151	1.049838	1.160361
Total de Instâncias	8276	4511	8291
%de instâncias	39%	21%	39%

3.4 Classificação

Após a *clusterização*, foi acrescentada uma coluna ao conjunto de dados discriminando a que *cluster* pertencia cada uma das instâncias. Esta coluna foi utilizada como atributo classe para a realização da tarefa de classificação. A Classificaçãoconsiste na busca por uma função que permita associar corretamente cada instância do banco de dados a uma classe. Para isso é necessário encontrar um modelo para o atributo alvo, utilizando uma função aplicada nos valores de outros atributos (Han &Kamber, 2006).

Os algoritmos de classificação utilizam uma parte do conjunto de dados para treinamento e uma parte para validação do modelo. Utilizou-se a abordagem, denominada na ferramenta Weka como *Percentage Split*, que divide os dados em dois grupos, um conjunto de treinamento e um conjunto de teste. Foi utilizada a divisão tradicional, dois terços para o conjunto de treinamento e um terço para o conjunto de teste.

Optou-se pela utilização da técnica de árvore de decisão, pela sua expressividade simbólica, e pelo algoritmo J48, que é a implementação da ferramenta

Weka do popular algoritmo denominado C4.5. A estrutura de uma árvore de decisão permite identificar quais atributos são mais relevantes para a determinação da classe. O atributo colocado na raiz da árvore é aquele com maior ganho de informação, ou seja, o que melhor discrimina as classes. Cada nó da árvore consiste em uma partição da base de dados que é recursivamente dividida até que se obtenha um nó folha com instâncias predominantemente de uma determinada classe.

A utilização de métodos de classificação permite também verificar a capacidade preditiva do modelo estudado. Para esse fim, várias medidas de desempenho são disponibilizadas pelos classificadores, essas medidas avaliam o desempenho do modelo geral, bem como para cada classe.

As seguintes medidas de desempenho geral são utilizadas pela ferramenta Weka:

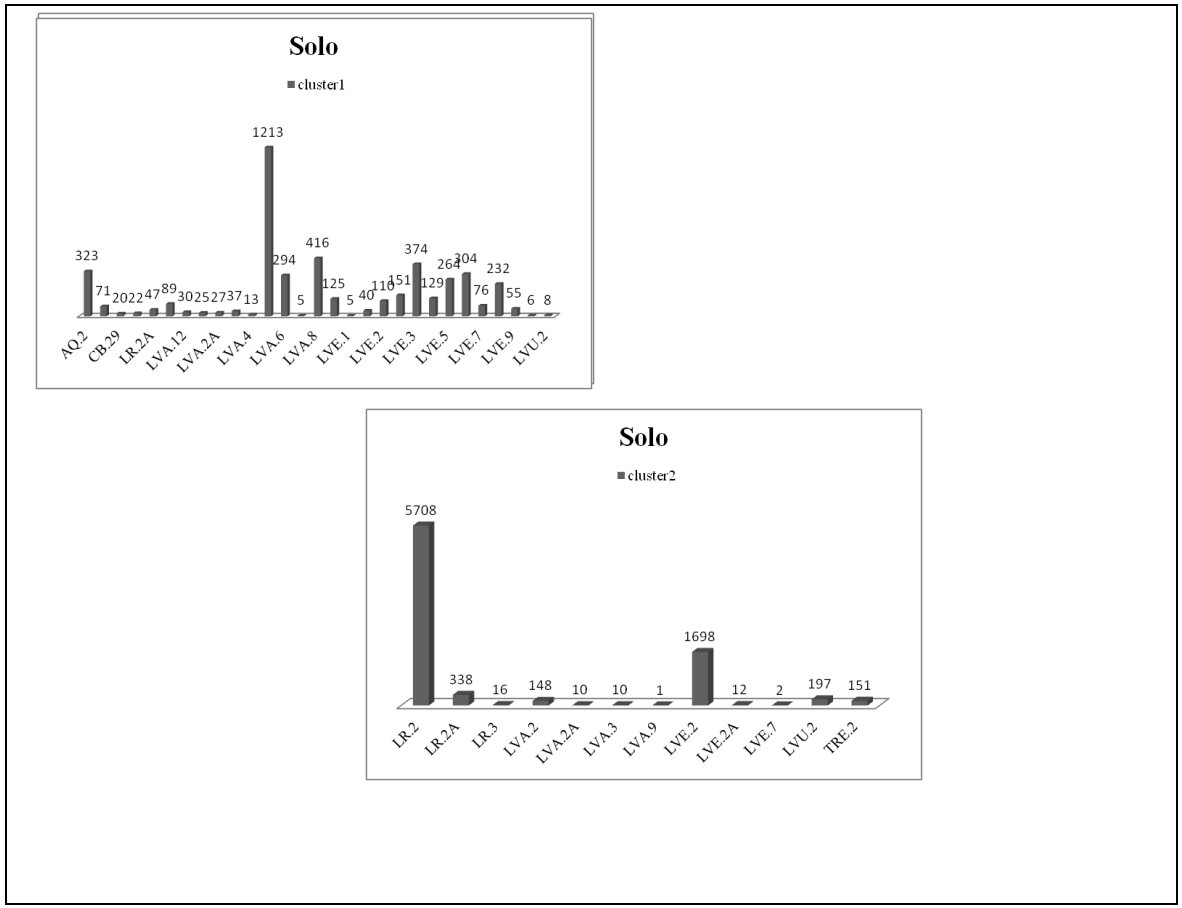
- Estatística*Kappa* - Índice que mede a concordância entre dois métodos de classificação. É uma medida de concordância e mede o grau de acurácia, além do que seria esperado tão somente pelo acaso. Seus valores variam de zero a um. Quanto menor o valor de *Kappa*, menor a confiança de observação, o valor um implica a correlação perfeita.
- *Erro médio absoluto* - média da diferença entre os valores reais e os preditos em todos os casos, é a média do erro da predição.
- Acurácia: proporção do número total de predições que foram corretas.

Algumas das medidas disponibilizadas para cada classe na ferramenta Weka são:

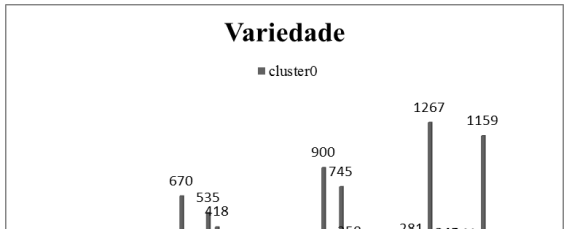
- Sensitividade ou Taxa de Verdadeiro Positivo (TP) - proporção de casos de uma classe que foram identificados corretamente;
- Taxa de Falso Positivo (FP) - proporção de casos de uma classe que foram classificados incorretamente como de outra classe;
- Precisão - proporção de casos positivos preditos que foram corretos.

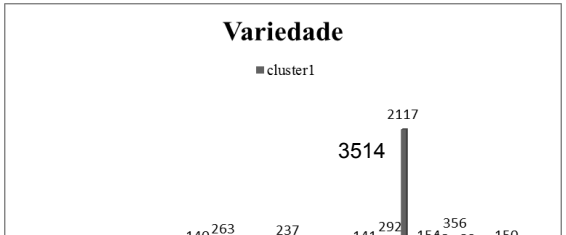
4. Resultados e discussões

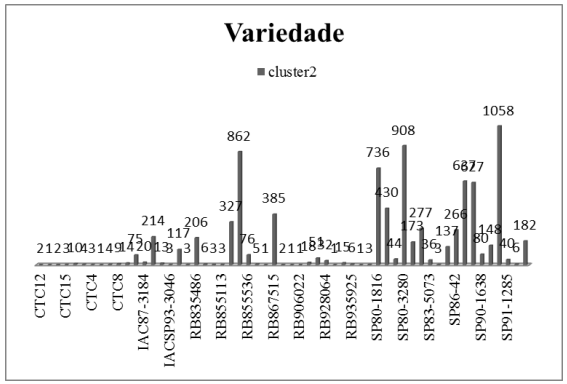
Cada um dos 3*cluster*s teve seu tipo de solo identificado com bastante clareza, conforme apresentado nas Figuras 1a, 1b e 1c. O tipo de solo mais encontrado no *cluster* 0 é o LR.1 – Latossolo Roxo (Texturas Finas,Eutróficos ou Endoeutróficos), 66% das instâncias possuem esse tipo de solo. No *cluster*1 destaca-se o solo LVA.5 – Latossolo vermelho amarelo(Texturas Médias, Distróficos ou Epieutróficos), com 28% de instâncias. No *cluster* 2 tem-se 69% de instâncias com solo LR.2– Latossolo Roxo (Texturas Finas,Distróficos ou Epieutróficos), o principal valor encontrado, mas também não se pode desconsiderar os 20% de solo do tipo LVE.2 - latossolo vermelho escuro.



As Figuras 2.a, 2.b e 2.c apresentam os resultados do atributo variedade. Percebe-se que houve concentração em mais de uma variedade para os *cluster* 0 e 2, enquanto no*cluster* 1prevaleceu a cultivar SP83-2847, presente em mais de 50% das instâncias. No *cluster* 0 as 3 principais cultivares foram: SP89-1115, SP91-1049 e SP80-1816, e no *cluster*2 as cultivares SP91-1049, SP80-3280 e RB85, 5453.É possível que a utilização das diversas cultivares nos *cluster*s 0 e 2, que possuem, predominantemente, o solo latossolo roxo, aconteça em decorrência da alta fertilidade desse tipo de solo, permitindo, portanto, experiências com muitos tipos de cultivares. Por outro lado, para o *cluster* 1, que possui solo menos fértil, investe-se no cultivar mais adaptada ao tipo de solo.

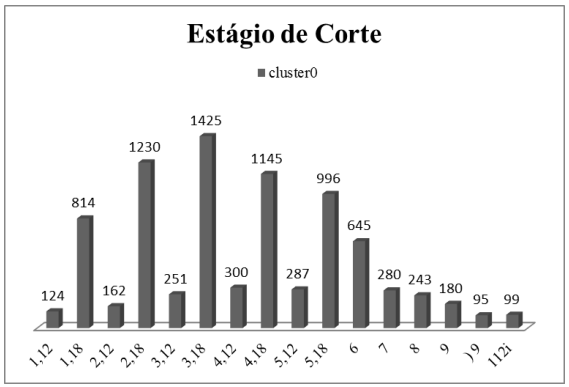


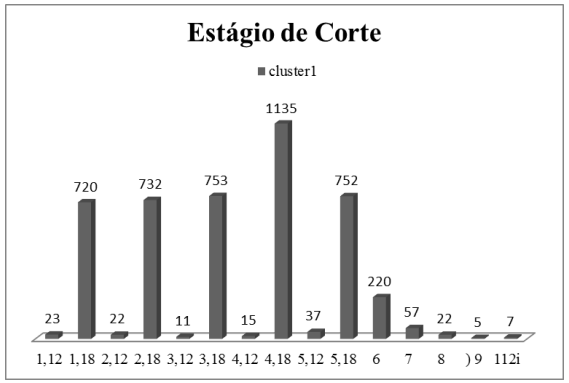


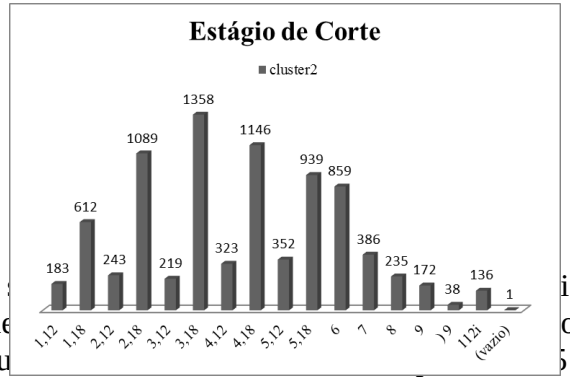


Em relação ao estágio de corte da cana-de-açúcar não houve um estágio específico que se destacasse em qualquer um dos *cluster*s. O gráfico da Figura 3.b assemelha-se ao que em Estatística se conhece por "distribuição uniforme"; nele observam-se frequências muito próximas para os resultados 1,18; 2,18; 3,18 e 5,18, mas com destaque para a maior frequência ao resultado 4,18.

Já nas Figuras 3.a e 3.c, que mostram os estágios de corte dos *cluster*s 0 e 2, respectivamente, esses mesmos 5 estágios tiveram as maiores frequências, mas não de maneira uniforme; em ambos os *cluster*s o estágio 3,18 possui frequência maior, e as frequências vão se tornando menores nos estágios adjacentes.







O atributo tipo de corte, s ição conforme a Tabela 2. Nota-se que o *cluster* 0, que tem a maior média de produtividade, possui a me o os *cluster*s 1 e 2 tem mecanização semelhante. O corte manual é realizado mais rente ao solo, enquanto qu 5 a 20 cm no comprimento do colmo cortado, em função do preparo da qualidade do solo, o que poderia explicar a produtividade maior no *cluster* com menor taxa de mecanização. A condição de corte, na maioria das instâncias, é queima se o corte é manual e "crua" se o corte é mecanizado, conforme pode ser visto também na Tabela 2. A porcentagem de corte manual sem queima é utilizada por razões agronômicas no sentido de obter "semente" para a "planta mãe" (colmos de cana utilizados para o plantio de primeiro ano).

Tabela 2– Dados do tipo e condição de corte

	Cluster 0	**Cluster 1**	**Cluster 2**
Corte Manual	44%	16%	19%
Corte Mecanizado	56%	84%	81%
Cana Crua	61%	88%	83%

Cana Queimada	39%	12%	17%

Na Tabela 3 são apresentadas as formulações de adubo mais encontradas nos 3 *cluster*s e a quantidade média de adubo utilizada. As 3 principais fórmulas são as mesmas em todos os *cluster*s, mudando ligeiramente a porcentagem encontrada em cada *cluster* para cada fórmula. Ressalta-se apenas que a percentagem de instâncias que usam a formulação com ureia agrícola é bem menor no *cluster* 1 e que a porcentagem de instâncias sem adubação é bem maior nesse mesmo *cluster*, que é o de menor produtividade.

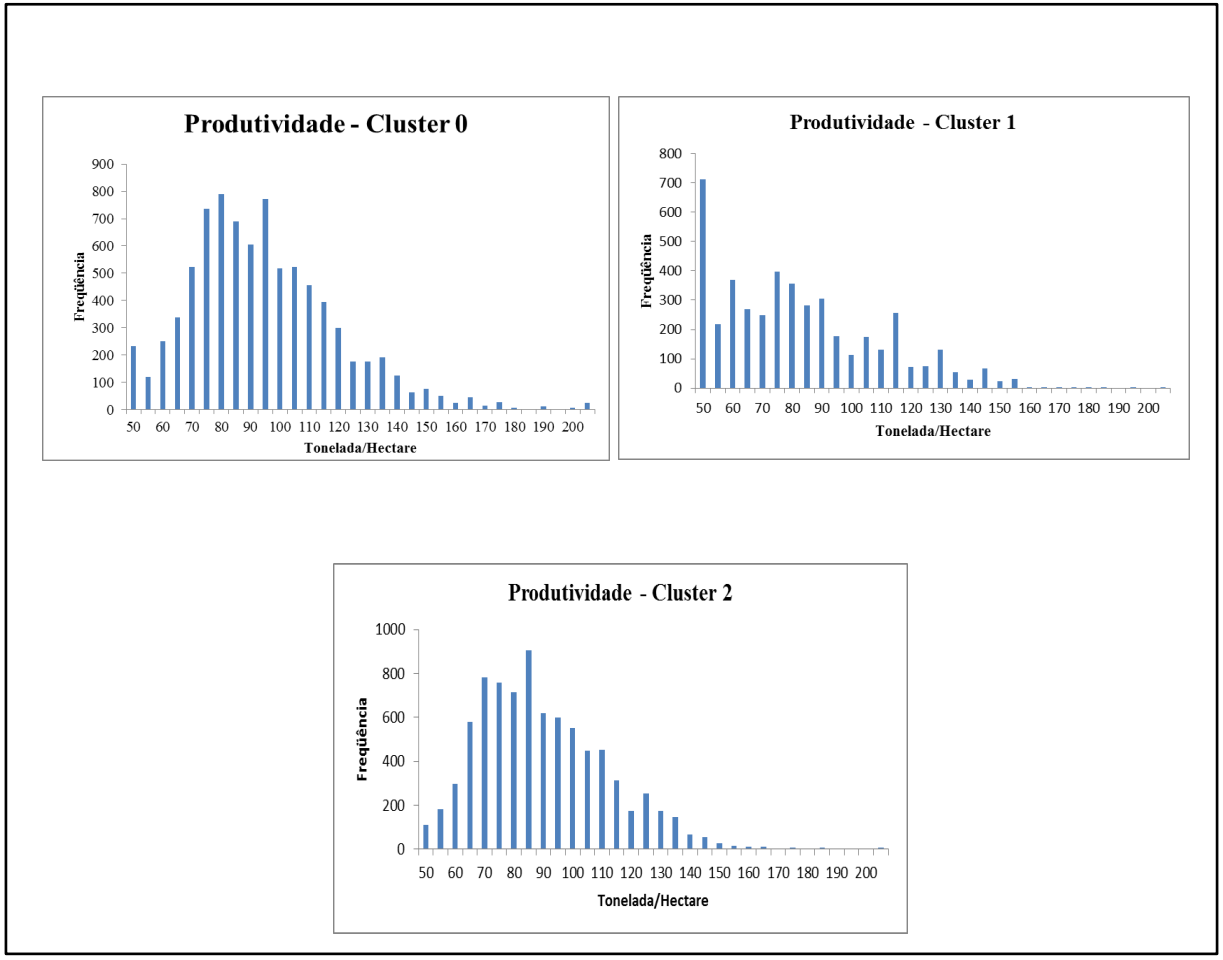
Tabela 3 – Características da adubação

	Cluster 0		**Cluster 1**		**Cluster 2**	
	% de instâncias	Média (kg/Ha)	% de instâncias	Média (kg/Ha)	% de instâncias	Média (kg/Ha)
Adubo 27-00-24	39%	442,76	38%	434,31	30%	441,45
Ureia Agricola (46-00-00)	23%	186,87	12%	267,98	28%	267,21
Adubo GR 21-00-21	16%	519,51	18%	517,70	15%	518,85
(Vazio)	16%	-	28%	-	17%	-

A fertilidade dos solos de cada grupo foi bastante característica, principalmente nos *cluster*s 0 e 2. O *cluster* 0 possui 94% das instâncias com fertilidadealta. O*cluster*1 possui 64% das instâncias com fertilidade média baixa, 19% com fertilidade média e 13% com fertilidade baixa. O*cluster* 2 possui 95% das instâncias com fertilidade média alta. Em relação à textura do solo, tem-se 86% de instâncias com textura argilosa no *cluster* 0, no *cluster* 1 tem-se 90% com textura argilosa/arenosa e o *cluster* 2 possui 75% de instâncias com textura argilosa e 25% com textura argilosa/arenosa. A fertilidade e textura são relacionadas ao tipo de solo, portanto esses resultados estão de acordo com os tipos de solo predominantes em cada um dos *cluster*s.

Nas Figuras 4.a, 4.b e 4.c são apresentados os histogramas de produtividade dos *cluster*s 0,1 e 2, respectivamente.Para os *cluster*s 0 e 2 nota-se a distribuição normal da produtividade. No *cluster* 0 a frequência maior de instâncias encontra-se no intervalo de 75 a 80 Kg/ha seguida pela frequência dos intervalos de 90 a 95 e 70 a 75 kg/ha. No *cluster* 2 a frequência mais elevada foi obtida no intervalo de 80 a 85kg/ha seguida pela frequência dos intervalos 65 a 70 e 70 a 75g kg/ha. Vale lembrar que o solo predominante no *cluster* 0 é um solo mais fértil que os tipos de solo que predominam no *cluster* 2 e como apresentado na Tabela 3 os níveis de adubação estão muito próximos nos 3 *cluster*s. Como a adubação, segundo Rossetto, Dias e Vitti (2008), é um importante fator de produtividade, mas também um elemento da planilha de custo, responsável por 17 a 25% dos custos do plantio da cana, sugere-se um estudo para a verificação e possível adequação nos níveis de adubação dos talhões do *cluster* 2.

Em relação à produtividade do *cluster* 1, o que surpreende é o alto número de instâncias com produtividade inferior a 50 ton/ha, 15% das instâncias desse *cluster* estão nessa faixa, ou seja, um número grande de talhões com produtividade muito baixa, que podem, até mesmo, estar gerando prejuízo.



A árvore de decisão gerada a partir do modelo de *clusters* é apresentada no Anexo A, possui 101 nós, com 51 nós folhas. O caminho mais longo, em profundidade, possui 10 nós e o mais curto possui 3. O atributo fertilidade está na raiz da árvore, indicando ser esse o atributo que melhor divide as instâncias da base de dados, ou, dito de outra forma, o atributo que traz maior ganho de informação para a classificação dessas instâncias de acordo com o modelo de *clusters* criado.

Se a fertilidade tiver valor igual a 1, indicando solo com fertilidade alta, e a textura tiver valor igual a 1, indicando solo argiloso, as instâncias serão do

cluster 0, que é o *cluster* com maior média de produtividade. Se a textura não for igual a 1 verifica-se o estágio de corte, a variedade e a adubação, para a previsão da classe. Destaca-se que para texturas diferentes de 1, com estágio de corte igual a 4 (cana de ano e meio) e variedade SP91-1049, as instâncias serão associadas ao *cluster* 1, o *cluster* com menor média de produtividade. Percebe-se por este resultado a influência do número de cortes na produtividade da cana. Também estão associadas ao *cluster* 1 as instâncias de variedade RB-855156, com textura diferente de 1, estágio de corte igual a 4 e adubação <= 314,42 kg/Ha.

Se a fertilidade for igual a 2, significando solo com fertilidade média alta, a definição do *cluster* se inicia pela variedade, seguida pelo tipo de solo e estágio de corte. Para esse tipo de fertilidade, as instâncias foram associadas apenas aos *clusters* 1 e 2. Para a variedade SP 83-2847, que é uma variedade frequente no *cluster* 1, se o solo for do tipo LR2 ou se o estágio de corte for diferente de 1 e 4 cortes as instâncias serão associadas ao *cluster* 2, caso contrário serão associadas ao *cluster* 1.

Para fertilidade igual a 2 e variedades diferentes de SP 83-2847, a associação aos clusters se inicia pelo estágio de corte, se for diferente de 4 cortes ou de textura igual a 1, as instâncias deverão pertencer ao *cluster* 2, senão a definição do *cluster* se dará pela quantidade de adubação e pela variedade da cana.

A sub árvore que se inicia com o predicado fertilidade diferente de 2 e textura igual a 1 é a que tem maior profundidade. As condições dessa sub árvore dizem respeito principalmente à adubação e à variedade da cana, entretanto envolveram também a fórmula do adubo e o tipo de corte. Ressalta-se que a condição sobre o tipo de corte se deu no último nível da árvore, antes do nó folha, para discriminar instâncias dos *clusters* 0 e 2, produtividade alta e média respectivamente, se o estágio de corte for igual a 8, a variedade for SP 80 1816 e o tipo de corte for manual a instância pertencerá ao cluster 0, se o corte for mecanizado pertencerá ao cluster 2. O corte manual é realizado mais rente ao solo, enquanto que no corte mecanizado há uma perda de 15 a 20 cm no comprimento do colmo cortado, em função do preparo da qualidade do solo, o que poderia explicar a associação ao cluster mais produtivo para o corte manual. Outro fato que chama a atenção é o estágio de corte da cana ser tão avançado e mesmo assim as instâncias poderem ser associadas ao *cluster* mais produtivo.

Se a fertilidade for diferente de 2 (nesse caso também diferente de 1) e a textura diferente de 1, verifica-se se a variedade é SP 91-1049, se essa condição for verdadeira será verificado se a fertilidade é igual ou diferente de 3. Se for diferente de 3 (nesse caso será 4 ou 5) as instâncias serão associadas ao *cluster* 1, caso contrário verifica-se o estágio de corte e o tipo de solo para discriminar as instâncias entre os *clusters* 1 e 2. Para variedades diferentes de SP 91-1049, verifica-se inicialmente se o atributo textura é igual a 2, a partir daí as verificações são a respeito do estágio de corte, da variedade, da adubação e sobre a fórmula do adubo, nessa sub árvore as instâncias são atribuídas aos *clusters* 0 e 1.

Em uma análise geral da árvore, pode-se afirmar que a fertilidade do solo é fator preponderante para discriminar a classe de uma determinada instância. O atributo textura e variedade aparecem no segundo nível da árvore. Embora a textura esteja relacionada à qualidade do solo, há variedades que se adaptam bem a solos menos férteis ou a menores quantidades de água, por exemplo. Assim é bastante plausível o fato do atributo variedade estar presente em diversos níveis da árvore de decisão.

Outros atributos frequentes são a adubação (quantidade de adubo) e o estágio de corte. A adubação do solo aumenta sua fertilidade e assim propicia melhores níveis de produtividade, razão pela qual este atributo aparece em diversos níveis da árvore. Também o estágio de corte é fator determinante de produtividade, os primeiros estágios de corte são os mais produtivos, a partir do sexto corte a produtividade dos talhões tende a declinar.

Salienta-se que o atributo tipo de solo apareceu somente em dois nós da árvore, indicando que na maioria das vezes esse atributo não é determinante na definição da classe. Isso implica que, em geral, a produtividade está associada a um conjunto de solos da mesma fertilidade e não a um tipo de solo específico.

A avaliação geral de desempenho do modelo de produtividade é apresentada na Tabela 4.

Tabela4 – Avaliação Geral do Modelo

	Resultado
Estatística Kappa	0,9948
Erro absoluto médio	0,0032
Acurácia	99,6651%

Os resultados obtidos com a classificação foram bastante satisfatórios, o índice *Kappa* está próximo ao valor máximo, indicando que a correlação entre os valores reais e preditos é perfeita. Por outro lado, o indicador referente ao erro médio está bem próximo de 0. A acurácia, que indica a proporção de acertos, também está próxima do valor máximo.

Na Tabela 5 são apresentadas as medidas de desempenho de cada classe.

Tabela5 – Avaliação das classes

	Cluster 0	**Cluster 1**	**Cluster 2**
TP	0.996	0.997	0.997
FP	0.002	0.003	0
Precisão	0.997	0.991	0.999

Os valores referentes à avaliação das classes também são muito próximos dos valores máximos no caso dos valores de TP e da precisão e quase nulos no caso de FP que indica uma porcentagem de erro na classificação.

Os níveis aceitáveis para os indicadores de desempenhosão dependentes da área de aplicação. Nonato e Oliveira (2013) obtiveram valores de acurácia entre 94,98% e 97,21% e coeficientes *Kappa* variando de 0,93 a 0,96. Resultadosinferiores foram obtidos em Vintrou *et al.* (2013), em que o modelo teve acurácia geral de 57,8%. No modelo de produtividade proposto por Fernandes*et al.* (2011), os valores de acurácia geral variaram entre 66,75 a 86,5%. No trabalho de Fang *et al.* (2009) a acurácia obtida foi de 93.5% com coeficiente *Kappa* de 0,85. No trabalho de Goltz *et al.*foram obtidos valores decoeficiente*Kappa* entre 0,69 e 0,84. No trabalho de Everingham *et al.* (2007) os modelos que classificaram a variedade da cana tiveram acurácia variando entre 71,3% a 92,3%, os modelos que classificaram o ciclo da cana (número de cortes) obtiveram acurácia entre 72,5% a 89,5%.

Assim, em comparação com os modelos de classificação da cultura da cana-de-açúcar encontrados na literatura,entende-se que o modelo gerado neste trabalho obteve desempenho bastante adequadoe pode ser usado como apoio em processos de tomada de decisão para a gestão da cultura da cana-de-açúcar.

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

5. Consideração Finais

Com a utilização da mineração de dados foi possível identificar 3 grupos distintos, caracterizados em produtividade baixa, média e alta, denominados, respectivamente, *cluster* 1,2 e 0.

Ao *cluster* 0, foram corretamente associados os talhões com fertilidade alta, predominantemente, solo latosssolo roxo, LR.1. Ao *cluster* 1, foram associados, em sua maior parte, talhões de fertilidademédia baixa, com maior frequência do solo latossolo vermelho amarelo (LVA.5). No *cluster* 2ficaram os talhões com fertilidade média alta, o solo mais frequente foi o latossolo roxo, LR.2, seguido de alguns tipos de solos vermelho escuro.

No*cluster* 1 houve grande incidência do cultivar SP83-2847, enquanto nos outros 2 *cluster*s em torno de 10 cultivares diferentes, em cada *cluster*, tiveram frequência relevante. O maior número de variedades abrange cultivares precoces e tardios e, dessa forma, permite-se antecipar ou prolongar períodos de cortes e isso poderia explicar a maior frequência de terceiro corte nos clusters 0 e 2,enquanto no cluster 1 há predomínio de quarto corte.

Foi possível identificar que os níveis de adubação estão muito semelhantes nos 3 *cluster*s, como o solo dos 3 *cluster*s se diferenciam em fertilidade sugere-se um estudo de viabilidade econômica para alteração dos níveis de adubação dos *cluster*s menos produtivos.Outro resultado importante, que requer atenção, é a porcentagem de instâncias do *cluster* 1 com produtividade abaixo de 50 ton/ha, valor esse, suscetível a gerarprejuízo.

A partir dos *clusters* gerados, foi utilizado um algoritmo de indução de árvore de decisão para identificar os atributos com maior ganho de informação para determinar a classe (*cluster* 0, 1 ou 2) de uma instância específica. Identificou-se que a fertilidade do solo foi o atributo que mais colaborou na determinação a classe, seguido pelo atributo textura e variedade, ressalta-se que dos 76 diferentes cultivares pertencentes à base de dados, apenas 9 estavam presentes nos predicados da árvore obtida. Uma análise agronômica poderia indicar as características dessas cultivares e assim explicar melhor sua participação nessa árvore.

Destaca-se ainda a importância dos atributos adubação e estágio de corte que apareceram em vários níveis da árvore, bem como o fato que, na maioria das vezes, o tipo de solo não foi fator determinante de identificação das classes.

Na avaliação do modelo os resultados foram bastante satisfatórios, tanto por classes, como na classificação geral, com acurácia de 99,66% e coeficiente *Kappa*de 0,99.

Com os resultados acima considera-se que os objetivos desta pesquisa foram atingidos, uma vez que foi possível caracterizar grupos de produtividade e identificar os atributos mais influentes na determinação da produtividade da cana-de-açúcar da usina em estudo. Esses resultados dão maior precisão aos padrões conhecidos e dessa forma podem auxiliar em processos de tomada de decisão referentes à cultura da cana.

Referências

Ananthara, M., Arunkumar, T., & Hemavathy, R. (2013). CRY—An improved crop yield prediction model using bee hive clustering approach for agricultural data sets. In Proceedings of the 2013 International Conference on Pattern Recognition, Informatics and Mobile Engineering. Retrieved from http://ieeexplore.ieee.org/xpls/abs_all.jsp?arnumber=6496717

Bajpai, P. K., Priya, K., & Malik, M. (2012). Selection of Appropriate Growth Model for Prediction of Sugarcane Area, Production and Productivity of India. *Sugar Tech*, *14*(2), 188–191. https://doi.org/10.1007/s12355-012-0142-4

Barros, F. M. M. de, Oliveira, S. R. de M., & Oliveira, L. H. M. de. (2013). Desenvolvimento e validação de um sistema de recomendação de informações tecnológicas sobre cana-de-açúcar. *Bragantia*, 387–395. Retrieved from http://www.scielo.br/pdf/brag/v72n4/aop_bragncea2061.pdf

Brasil (2014). Ministério da Agricultura, Pecuária e Abastecimento. *Sapcana*: Sistema de Acompanhamento de Produção Canavieira. Retrieved from http://www.agricultura.gov.br/comunicacao/noticias/2014/09/mapa-publica-projecoes-do-agronegocio-para-a-safra-20232024.

Brasil (2016). Ministério da Agricultura, Pecuária e Abastecimento. **Sapcana**: Sistema de Acompanhamento de Produção Canavieira. 2016. Retrieved from http://www.agricultura.gov.br/vegetal/culturas/cana-de-acucar.

Conab (2015).Acompanhamento da safra brasileira:cana-de-açúcar: monitoramento agrícola. Brasília: CONAB, v. 2, n. 1, p. 1-28. Retrieved from http://www.conab.gov.br/OlalaCMS/uploads/arquivos/15_04_13_08_49_33_boletim_cana_portugues_-_1o_lev_-_15-16.pdf.

Everingham, Y. L., Lowe, K. H., Donald, D. A., Coomans, D. H., & Markley, J. (2007). Advanced satellite imagery to classify sugarcane crop characteristics. *Agronomy Sustain. Dev.*, *27*, 111–117. Retrieved from http://link.springer.com/article/10.1051/agro:2006034

Fang, L., Li, H., & Chen, S. (2009). The design of intelligent expert classifier for featured crop mapping combining spectral library. In *2009 17th International Conference on Geoinformatics, Geoinformatics 2009*. Retrieved from http://ieeexplore.ieee.org/xpls/abs_all.jsp?arnumber=5293533

Fernandes, J. L., Rocha, J. V., & Lamparelli, R. A. C. (2011). Sugarcane yield estimates using time series analysis of spot vegetation images temporais de imagens spot vegetation. *Sci. Agric. (Piracicaba, Braz.)*, *v.68*, n.*2*(April), 139–146.

Ferraro, D. O., Ghersa, C. M., & Rivero, D. E. (2012). Weed Vegetation of Sugarcane Cropping Systems of Northern Argentina: Data-Mining Methods for Assessing the Environmental and Management Effects on Species Composition. *Weed Science*, *60*(1), 27–33. https://doi.org/10.1614/WS-D-11-00023.1

Ferraro, D. O., Rivero, D. E., & Ghersa, C. M. (2009). An analysis of the factors that influence sugarcane yield in Northern Argentina using classification and regression trees. *Field Crops Research*, *112*(2–3), 149–157. https://doi.org/10.1016/j.fcr.2009.02.014

Goltz, E., Arcoverde, G. F. B., Aguiar, D. A., Rudorff, B. F. T., & Maeda, E. E. (2009). Data mining by decision tree for object oriented classification

of the sugar cane cut kinds. In *2009 IEEE International Geoscience and Remote Sensing Symposium, IGARSS 2009* (pp. 405–408). Retrieved from http://ieeexplore.ieee.org/xpls/abs_all.jsp?arnumber=5417646

Gonçalves, R. R. V, Zullo Jr, J., Ferraresso, C. S., Sousa, E. P. M., Romani, L. A. S., & Traina, A. J. M. (2011). Analysis of NOAA / AVHRR multitemporal images, climate conditions and cultivated land of sugarcane fields applied to agricultural monitoring. In *2011 6th International Workshop on the Analysis of Multi-Temporal Remote Sensing Images, Multi-Temp 2011 - Proceedings* (pp. 229–232).

Han, J.& Kamber, M. (2006). *Data mining: concepts and techniques*.2. ed. São Francisco: Morgan Kaufmann, 2006. 770 p.

Marin, F., & Carvalho, G. (2012). Spatio-temporal variability of sugarcane yield efficiency in the state of São Paulo, Brazil. *Pesquisa Agropecuária Brasileira*, (1), 149–156. Retrieved from http://www.scielo.br/scielo.php?pid=S0100-204X2012000200001&script=sci_arttext

Nakano, B. (2010) *Metodologia da pesquisa em engenharia de produção e gestão de operações***.** Org.: Paulo Cauchick Miguel. In: Capítulo 4. Rio de Janeiro: Elsevier.

Nonato, R. T., & Oliveira, S. R. D. E. M. (2013). Data Mining Techniques for Identification of Sugarcane Crop Areas in Images Landsat 5. *Engenharia Agrícola*, *33*(6), 1268–1280.

Prado, H.; Pádua Jr., A.L.; Garcia, J.C.; Moraes, J.F.L.;Carvalho, J.P. &Donzeli,P.L. (2008); Ambientes de produção. In: Dinardo-Miranda, L. L.; Vasconcelos, A. C. M. de; Landell, M. G. de A. (eds.). *Cana-de-açúcar*.671–698, Campinas: Instituto Agronômico. Parte 4, 179-205.

Rodrigues Jr., F. A. (2012). *Análise e modelagem da influência de atributos do solo e planta na produtividade e qualidade da cana-de-açúcar*. Retrieved from http://www.bibliotecadigital.unicamp.br/document/?code=000862768

Romani, L. A. S., Goncalves, R. . R. V., Amaral, B. F., Chino, D. Y. T., Zullo Jr, J., Traina Jr, C. & Traina, A. J. . M. (2011). Clustering analysis applied to NDVI/NOAA multitemporal images to improve the monitoring process of sugarcane crops. In *Analysis of Multi- ...* (pp. 33–36). Retrieved from http://ieeexplore.ieee.org/xpls/abs_all.jsp?arnumber=6005040

Romani, L., Avila, A. De, Chino, D. Y. T., Zullo Jr, J., Chbeir, R., Traina Jr, C., & Traina, A. J. . M. (2013). A new time series mining approach applied to multitemporal remote sensing imagery. *EEE TRANSACTIONS ON GEOSCIENCE AND REMOTE SENSING*, *51*(1), 140–150. Retrieved from http://ieeexplore.ieee.org/xpls/abs_all.jsp?arnumber=6215038

Rossetto, R.;Dias, F.L.F.; Vitti,A.C.(2008); Nutrição e adubação. In: Dinardo-Miranda, L. L.; Vasconcelos, A. C. M. de; Landell, M. G. De A. (eds.). *Cana-de-açúcar*. 671- 698.Campinas: Instituto Agronômico. Parte 5, 221-337.

São Paulo (Estado) (2014). Investe São Paulo: Agência paulista de promoção de investimentos e competitividade.Retrieved fromhttp://www.investe.sp.gov.br/setores-de-negocios/agronegocios/cana-deacucar/

Silva, F. C.; Bergamasco, A. F. ; Rodrigues L. H. ; Godoy, A. P. & Trivelin, P. C. O. (2006). Manejo de N Fertilizantes para a cana-de-açúcar com Colheita crua, no Contexto Ecológico, por um Modelo de Simulação. In: Environmental And Health World Congress, Anais..., Santos, Brazil, 16 -19 july, 249-253.

Simões, M. dos S., Rocha, J.V., & Lamparelli, R. A. C. (2005). Growth indices and productivity in sugarcane. *Sci. Agric. (Piracicaba, Braz.)*, *62*(1), 23–30.

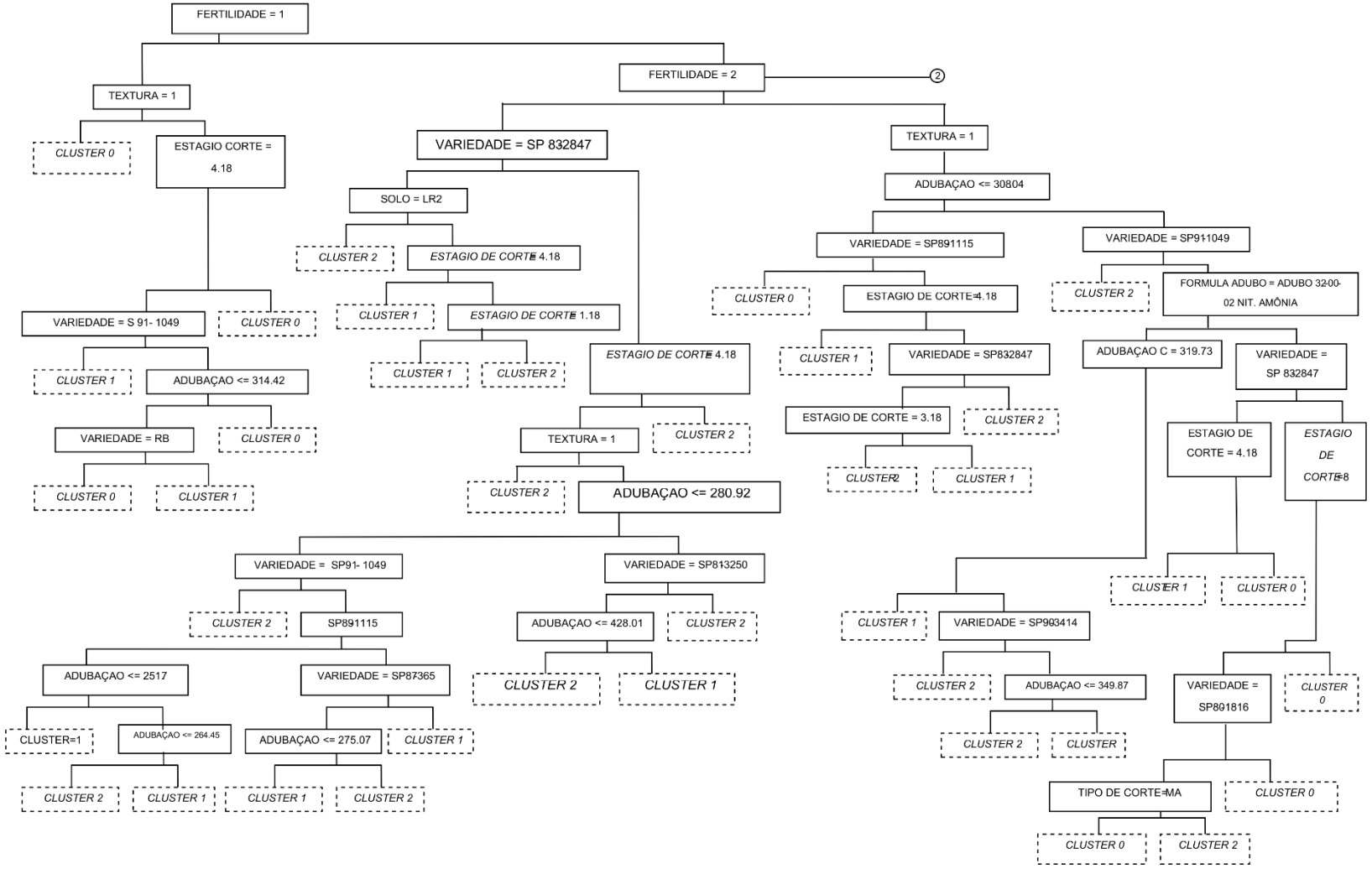
Souza, Z. M. De, Cerri, D. G. P., Colet, M. J., Rodrigues, L. H. A., Magalhães, P. S. G., & Mandoni, R. J. A. (2010). Análise dos atributos do solo e da produtividade da cultura de cana-de-açúcar com o uso da geoestatística e árvore de decisão. *Ciência Rural*, *40*(4), 840–847. https://doi.org/10.1590/S0103-84782010005000048

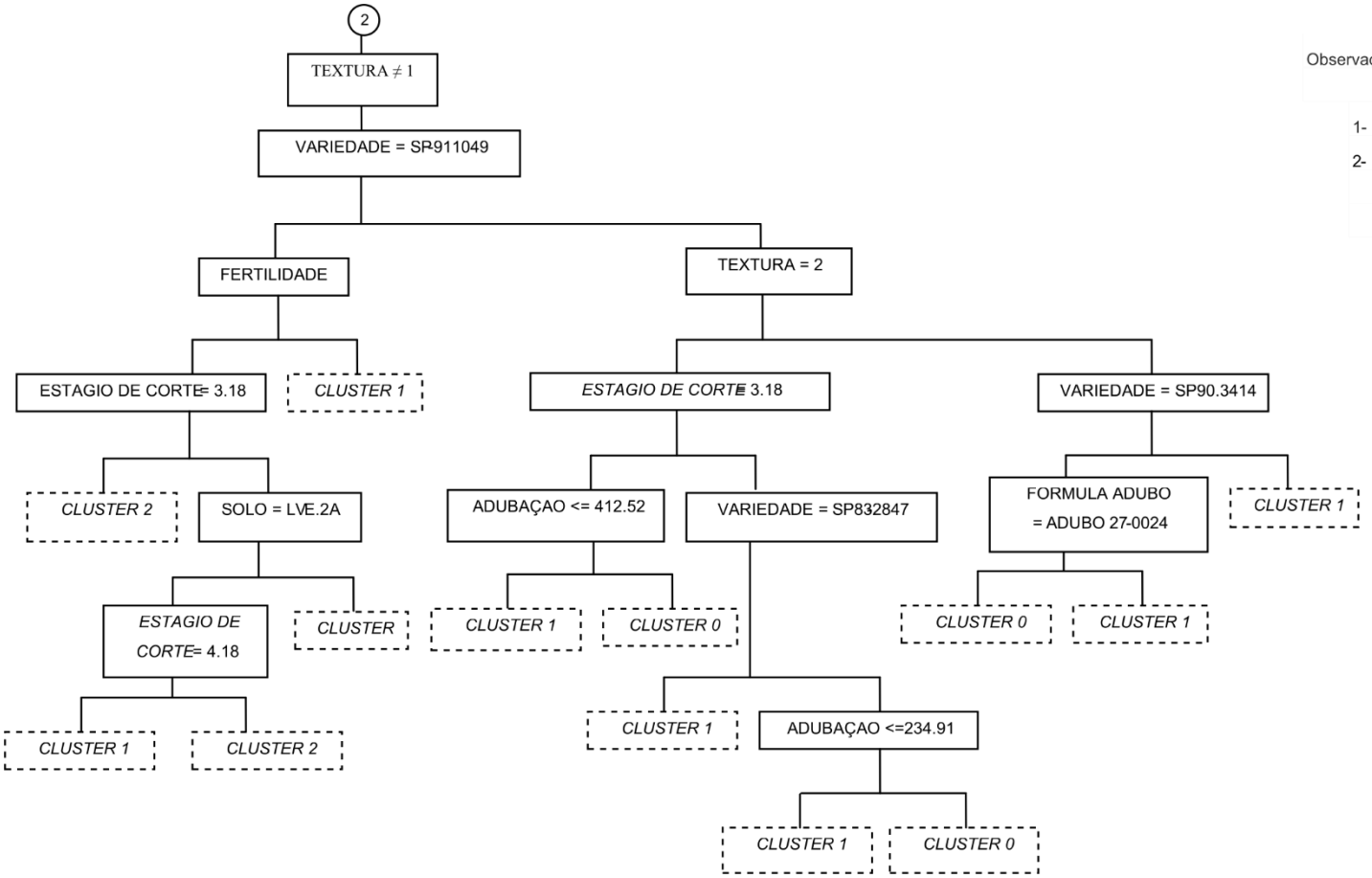
Tsai, H.-H. (2012). Global data mining: An empirical study of current trends, future forecasts and technology diffusions. *Expert Systems with Applications*, *39*(9), 8172–8181. https://doi.org/10.1016/j.eswa.2012.01.150

Vieira, M. A., Formaggio, A. R., Rennó, C. D., Atzberger, C., Aguiar, D. A., & Mello, M. P. (2012). Object Based Image Analysis and Data Mining applied to a remotely sensed Landsat time-series to map sugarcane over large areas. *Remote Sensing of Environment*, *123*, 553–562. https://doi.org/10.1016/j.rse.2012.04.011

Vintrou, E., Ienco, D., Bégué, A., & Teisseire, M. (2013). Data mining, a promising tool for large-area cropland mapping. *IEEE journal of selected topics in applied earth observations and remote sensing*, *6*(5), 2132–2138. Retrieved from http://publications.cirad.fr/une_notice.php?dk=571503

APÊNDICE A – ÁRVORE DE DECISÃO





Observações

1- As linhas tracejadas representam nós folha

2- 2- Todos os ramos à esquerda representam que a condição é verdadeira e os ramos à direita representam que a condição é falsa