MINERAÇÃO DE DADOS PARA A IDENTIFICAÇÃO DE FATORES QUE INFLUENCIAM A PRODUTIVIDADE DE CANA-DE-AÇÚCAR NO BRASIL

Elaine Conceição Venâncio Santos
Denise Fukumi Tsunoda
Celso Yoshikazu Ishida
Deborah Ribeiro Carvalho

RESUMO
O presente estudo tem como objetivo avaliar a capacidade de agregar conhecimento por meio da aplicação de métodos de mineração de dados, que é a principal etapa do processo de descoberta de conhecimento em bases de dados, em uma base de dados agrícola, analisando e descobrindo relações entre características de solos e fatores de produtividades da cultura de cana-de-açúcar em todo território nacional. Para atingir este objetivo foram utilizados dados referentes à produtividade da cultura por município, no período de 2008 a 2010, obtidos junto a base de Produção Agrícola Municipal (PAM) do IBGE, o mapa de classes de solos elaborado pela Embrapa e a malha digital contendo a divisão dos municípios do Brasil também do IBGE. O estudo foi proposto, em caráter exploratório e os dados analisados foram: área plantada, área colhida, valor, produtividade, classe de solo e localização. Para integração destes dados utilizou-se um banco de dados com suporte a informação espacial, sendo realizadas operações topológicas com a ferramenta PostGIS do PostgreSQL. Posteriormente, aplicaram-se dois métodos de mineração de dados, para encontrar regras que representem à base de dados e identificar como o processo de mineração agregou valor a informação que já era conhecida, possibilitando a descoberta de novos conhecimentos. Primeiramente aplicou-se o método de agrupamento a fim de identificar as similaridades entre os atributos, utilizando para isto o algoritmo *K-means*. Após a formação dos grupos foi utilizado um algoritmo para classificação, o C4.5, para identificar as relações em cada agrupamento. Como resultado, a pesquisa possibilitou complementar a informação dos tipos de solos mais produtivos, identificando quais *clusters* estão agrupados nestas áreas e quais os fatores de similaridades entre os grupos. Pretende-se no futuro expandir esta pesquisa utilizando uma série histórica de dados mais longa e um mapa de classe de solo com maior resolução bem como comparar outros métodos de mineração de dados, além dos dois aplicados neste estudo.

PALAVRAS-CHAVE
KDD, classificação, agricultura.

1 INTRODUÇÃO

Com o avanço de tecnologias como geoprocessamento, sensoriamento remoto e SIGs (Sistemas de Informações Geográficas), a agricultura passou a contar com ferramentas que auxiliam produtores, agrônomos e cooperativas em geral contribuindo para melhorar a eficiência da atividade agrícola. Estas ferramentas geram uma grande quantidade de dados

sobre a atividade agrícola local e, em nível nacional instituições como o IBGE (Instituto Brasileiro de Geografia e Estatística) produzem anualmente uma série de dados referente à atividade agrícola de acesso livre. Neste cenário, as bases de dados que contém informações agrícolas podem servir de fonte para o processo de KDD – *Knowledge Discovery in Database* (Descoberta de Conhecimento em Bases de Dados), uma vez que pela aplicação de métodos e técnicas em uma base de dados é possível buscar padrões que podem resultar em ações cujo objetivo pode ser, por exemplo, o aumento da produtividade, redução dos custos, auxílio às estratégias de plantios entre outros.

O armazenamento de grande quantidade de dados vem ocorrendo em diversas áreas do conhecimento, devido à redução de custos dos equipamentos tecnológicos e o próprio avanço tecnológico. Em contrapartida, o processo de identificar relações entre os dados é complexo e se torna cada vez mais difícil ao ser humano, pois, à medida que a quantidade de dados aumenta, torna-se inviável a capacidade de percepção e avaliação humana. Assim, este trabalho descreve a utilização do processo de KDD, em especial a etapa de *Data Mining* (mineração de dados) que conforme Fayaad *et a*l. (1996) é a principal etapa e consiste na aplicação de algoritmos que efetivamente busquem por padrões e relações em um determinado conjunto de dados. Estes padrões e relações descobertos podem ser relevantes para o gestor agrícola na identificação das melhores ações a serem tomadas. Para Han e Kamber (2006), as informações e os conhecimentos adquiridos por meio da mineração de dados podem ser usados em aplicações que variam desde análise de mercado, detecção de fraude e retenção de clientes até controle da produção.

Entretanto, sabe-se que nem sempre o processo de KDD pode ser viável a uma organização devido aos custos envolvidos e também pela incerteza nos resultados que podem ser alcançados. Neste contexto, este estudo tem como objetivo avaliar se a aplicação da técnica de mineração de dados poderá agregar conhecimento "novo" sobre a base obtida em comparação ao conhecimento já apresentado por métodos tradicionais de avaliação de dados.

2 ANÁLISE DO MERCADO DE CANA-DE-AÇÚCAR

Para UNICA (2012) a cana-de-açúcar ocupa cerca de 7 milhões de hectares (ha) ou seja, 2% de toda a terra arável do Brasil, que é o maior produtor mundial desta cultura, seguido por Índia, Tailândia e Austrália. A cana-de-açúcar é utilizada na produção de açúcar e etanol, sendo uma indústria importante na geração de emprego e renda para o país, já que ocupa o terceiro lugar em relação à área plantada, atrás apenas da soja e do milho (VIAN, 2012).

No que se refere ao etanol, atualmente este se apresenta como alternativa ecológica de combustível e, aliado ao cenário mundial de estímulo ao uso de fontes renováveis devido a crescente preocupação com o meio ambiente, torna-se um importante recurso energético. Com a decisão dos países desenvolvidos de substituir uma parte do petróleo por bicombustíveis, a procura mundial por etanol aumentou e o Brasil tem como vantagem a grande quantidade de cana-de-açúcar cultivada, além do fato do etanol ser mais produtivo por ser até sete vezes mais rentável em relação a energia do que o etanol de milho (BIOETANOL, 2012).

Mesmo com um grande mercado externo para o produto, o etanol também possui expressivo consumo interno, aquecido pela crise do petróleo em 1997 e a partir de 2000, devido ao aumento do ingresso de carros bicombustíveis. Conforme Fonseca, Paixão e Maria (2008), o governo brasileiro tem interesse no aumento da produção e exportação de etanol, pois este é uma alternativa limpa e barata e o país tem vantagem na produção dada à intensidade em recursos naturais e de mão-de-obra disponível.

Considerando a produção mundial de etanol, os maiores produtores são o Brasil, com o etanol proveniente da cana-de-açúcar e os Estados Unidos, com o etanol proveniente do milho, respondendo juntos por 70% da oferta global (VIAN, 2011).
O Brasil também é líder no mercado internacional na exportação de açúcar, já que é responsável por mais da metade do açúcar comercializado no mundo (MAPA, 2012b). O açúcar obtido a partir da cana é um produto antigo, sua produção foi incentivada pela Coroa Portuguesa, pois na época a produção era limitada a quantidades que não supriam a demanda do mercado, por esta razão o açúcar tinha valor tão alto quanto o do ouro em toda a Europa (UNICA, 2012b).
A produção de cana-de-açúcar se concentra nas regiões Centro-Sul e Nordeste do Brasil, com base no mapa da produção do setor sucroenergético apresentado na Figura 1. Em vermelho constam as áreas onde se concentram as plantações e usinas produtoras de açúcar, etanol e bioeletricidade, segundo dados oficiais elaborado pelo IBGE, NIPE-UNICAMP (Universidade Estadual de Campinas – SP) e CTC (Centro de Tecnologia Canavieira) (UNICA, 2012).

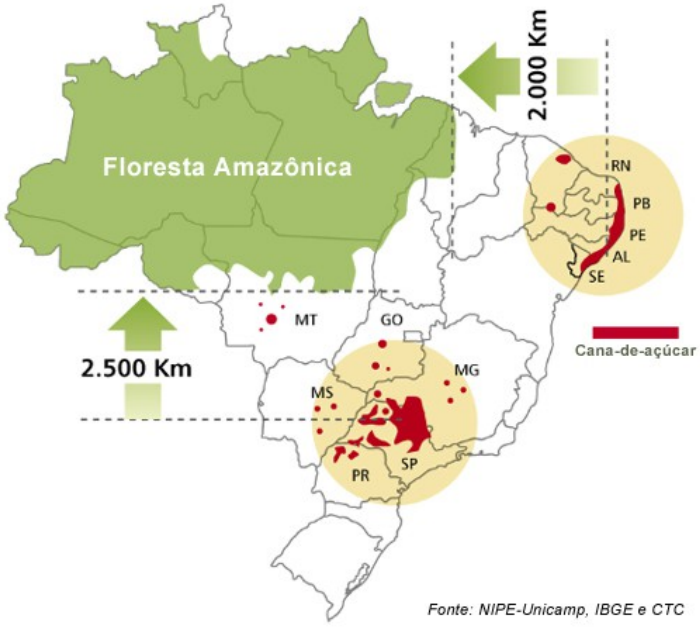

FIGURA 1 – SETOR SUCROENERGÉTICO - MAPA DA PRODUÇÃO
FONTE: UNICA (2012)

Conforme os dados da safra de 2010/2011 de cana-de-açúcar da MAPA (2012), os estados com maior produção são: São Paulo e Minas Gerais no Centro-Sul e Alagoas e Pernambuco na região Norte/Nordeste. Este fato motivou o trabalho para buscar identificar os motivos desta maior produção. Se há apenas relação com área plantada, clima, solo e variedades das culturas, que são as características principais apontadas por Marin (2011) como características agronômicas que regulam o desenvolvimento da cultura, ou se existem outras relações passíveis de descoberta por meio da mineração de dados que podem influenciar a produtividade.

3 MATERIAIS E MÉTODOS

3.1 BASE DE DADOS

A base de dados utilizada neste trabalho é composta por atributos com diferentes origens. Entre as origens encontram-se a base de dados PAM (Produção Agrícola Municipal), a malha digital da divisão política do Brasil em nível municipal referente ao ano de 2007 e a malha digital do mapa de solos do Brasil na escala 1:5.000.000, todos mantidos pelo IBGE. Todos os dados utilizados são públicos e estão disponíveis na página da internet do IBGE (IBGE, 2012b).
A base de dados original contém dezessete atributos agrupados em duas categorias. A primeira composta por dados relacionados à cultura da cana-de-açúcar, a segunda composta por dados relacionados à classificação de solos e municípios.

3.1.1 Dados relacionados à cultura da cana-de-açúcar

Os dados relacionados à cana-de-açúcar foram obtidos junto à base PAM utilizando a ferramenta de consulta disponível na própria página e são compostos por unidade da federação, município, quantidade de área plantada em hectares, área colhida em hectares, quantidade produzida em toneladas e valor da produção em mil reais referente aos anos de 2008, 2009 e 2010.
Segundo IBGE (2012b), a base PAM apresenta dados sobre área plantada, área destinada à colheita, área colhida, quantidade produzida, rendimento médio obtido e valor da produção dos produtos das culturas temporárias e permanentes, por grandes regiões, unidades da federação e municípios. Além disto, efetua uma análise sobre o desempenho das lavouras de maior relevância, tanto produtiva como comercial, destacando, entre outros aspectos, a distribuição espacial dos principais produtos agrícolas no território e sua participação relativa no valor total das produções regional e nacional, as colheitas obtidas nos principais municípios produtores, bem como os fatores de maior influência nos resultados e na produtividade dessas lavouras.

3.1.2 Dados relacionados à classificação de solos e malha municipal digital do Brasil

Segundo IBGE (2012), “O mapa de solos identifica os diferentes tipos de solos encontrados o Brasil e utiliza pela primeira vez a nomenclatura e as especificações recomendadas pelo Sistema Brasileiro de Classificação de Solos – SBCS da Embrapa (1999)”. Já a malha municipal do Brasil é um produto cartográfico do IBGE e retrata a situação vigente da DPA (Divisão Político-Administrativa) do Brasil, por meio da representação vetorial das linhas definidoras das divisas estaduais e municipais, referente ao ano base de 2007, contemplando 5.564 municípios.
As malhas digitais relacionadas à classificação de solos e divisão política dos municípios foram obtidas junto à página do IBGE (2012d) destinado a *downloads*. As malhas digitais são arquivos em formato *shapefile* (formato proprietário da empresa ESRI e que pode ser manipulado e visualizado por programas de computador como o ArcView, QGis e gvSIG entre outros) que armazenam a geometria e os atributos alfanuméricos correspondentes a cada município ou classe de solo no contexto deste trabalho.
A Figura 2 apresenta um mapa do território nacional dos tipos de solos com sua respectiva legenda.

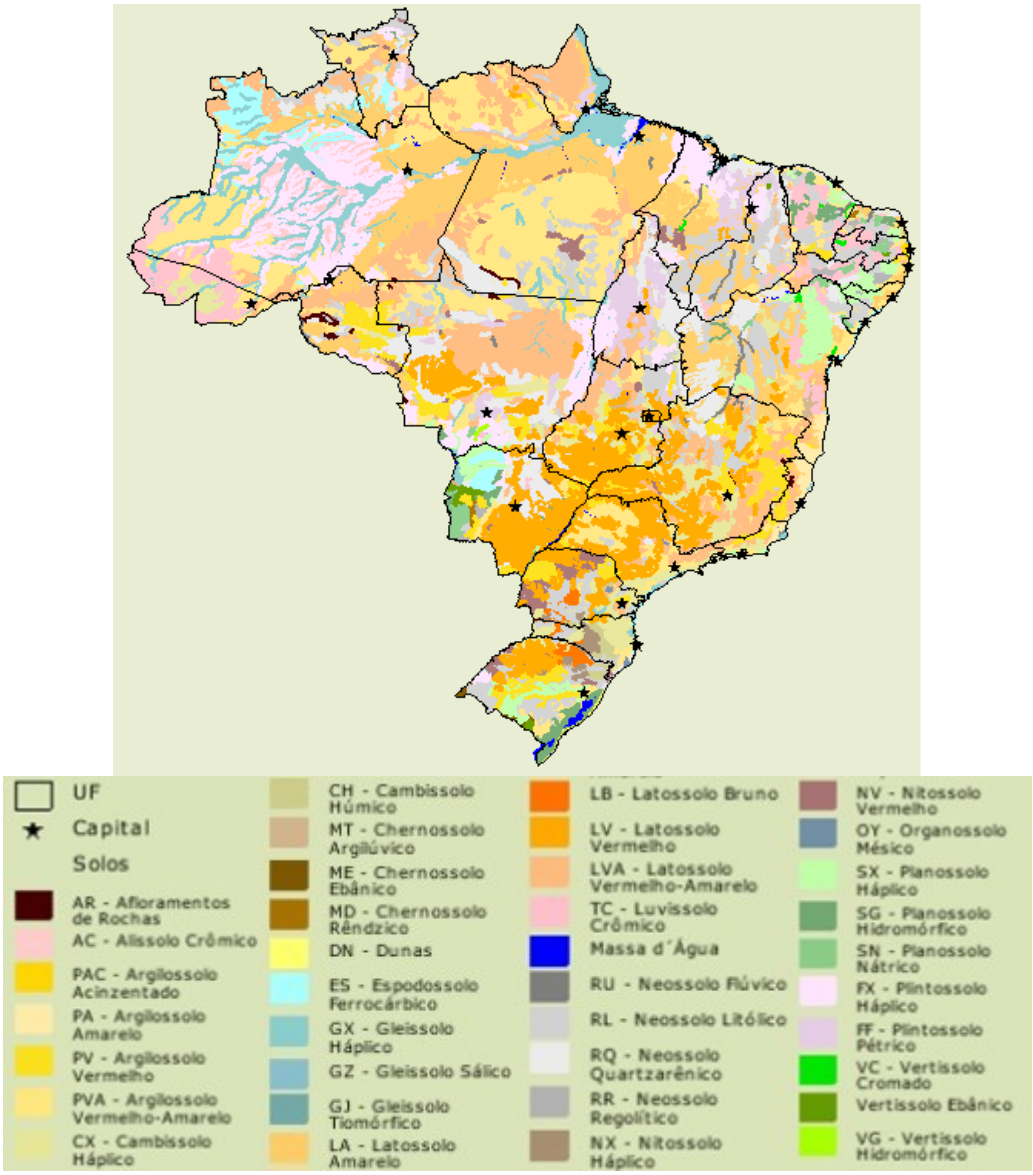

FIGURA 2 – MAPA DE SOLOS DO BRASIL
FONTE: IBGE (2012c)

Conforme a Figura 2, os nomes dos tipos de solo são compostos pelo 1º e 2º nível categórico. Segundo Embrapa (1999), o 1º nível categórico da classificação descreve as diversas classes que são identificadas pela presença ou ausência de atributos, horizontes diagnósticos ou propriedades que são características passíveis de serem identificadas no campo, mostrando diferenças no tipo e grau de desenvolvimento de um conjunto de processos que atuam na formação do solo. Como exemplo, no solo Argilossolo acinzentado, tem-se Argilossolo como o 1º nível categórico e acinzentado como 2º nível categórico.

3.2 PRÉ-PROCESSAMENTO

Os dados alfanuméricos referente ao PAM e as malhas digitais foram importados para o banco de dados relacional PostgreSQL (programa de banco de dados de código aberto) configurado com suporte a dados geográficos utilizando a extensão PostGIS (extensão do banco de dados

PostgreSQL utilizada para habilitar o suporte a dados geográficos. Os dados obtidos junto ao IBGE foram produzidos utilizando coordenadas geográficas.
O suporte a dados geográficos foi utilizado para calcular o centróide de cada município da malha digital. O centróide de cada município é representado por um ponto no espaço e cada ponto possui uma longitude e latitude. Em seguida, utilizou-se as ferramentas disponíveis no PostGIS (função *st_contains*) para calcular em qual polígono de classe de solo o centróide de cada município está localizado. Desta forma, associou-se a cada município da malha digital uma classe de solo. Existindo neste ponto duas tabelas no banco dados, a primeira contendo os dados do PAM associados a cada município e a segunda contendo os dados de classe de solo, longitude e latitude associados a cada município. Em seguida, utilizando os dados de município disponível nas duas tabelas foi possível unir os dados, representando todas as informações disponíveis em uma única tabela.
A tabela contendo o resultando das operações descritas anteriormente possui 5.564 registros/instancias, pois foram considerados todos os municípios do Brasil. Entretanto, muitos deles possuem atributos com valores zerados em todos os anos ou em pelo menos um dos anos selecionados, sendo a causa principal o fato que muitos municípios não serem produtores de cana-de-açúcar. Neste caso optou-se por retirar estes dados da base, que passou a ter 3.426 registros.
O atributo município também foi retirado da base, pois é um atributo específico e desta forma não há repetição deste atributo em nenhuma linha, portanto não contribui para a mineração de dados.
O atributo numérico produtividade foi acrescentado, cujo valor é a razão entre quantidade produzida por área colhida, para cada ano analisado, ou seja, de 2008 a 2010, sendo criado mais um atributo com a média da produtividade que representa a média destes três anos analisados.
No que se refere ao mapa de solos original, o atributo solo é composto pelo 1º e 2º nível categórico da classificação da Embrapa, conforme foi abordado na seção 3.1.2. Com o objetivo de diminuir a quantidade de categorias associadas a este atributo, optou-se por utilizar somente o 1º nível categórico da classificação. Após esta primeira simplificação das categorias, foram excluídos treze registros em que o centróide da divisão política do município constava como Massa d'água ou Afloramento Rochoso, devido ao fato de serem áreas inviáveis para a cultura da cana-de-açúcar, passando a base para 3.410 registros.
Portanto, os atributos utilizados foram: quantidade de área plantada em hectares, área colhida em hectares, quantidade produzida em tonelada por hectare, valor da produção em mil reais, produtividade tonelada por hectare, cada um destes dados referente aos anos de 2008, 2009 e 2010. Além da produtividade tonelada por hectare média destes três anos, longitude, latitude e solo, totalizando 19 atributos.

3.3 O MÉTODO

Foram aplicados dois algoritmos de mineração de dados para este experimento, utilizando o software WEKA, livremente distribuído. Primeiramente foi aplicado um método de agrupamento para buscar identificar similaridades entre os atributos, pois conforme Calil *et al.* (2008) o agrupamento visa identificar um conjunto finito de classes ou *clusters*, que consiste em agrupar um conjunto de objetos em função de sua proximidade e similaridade. O algoritmo de agrupamento utilizado foi o K-médias (*K-Means)*, onde foi dividido em 5 *clusters* e utilizado 66% para treinamento, passando a base para 1.160 registros.
O *K-Means* conforme Ghosh e Liu (2009) é um algoritmo de agrupamento iterativo simples nas quais as partições do conjunto de dados são definidos pelo número de *clusters*

especificado pelo usuário. É também um algoritmo muito utilizado em *Data Mining*. Segundo Han e Kamber (2006) o *K-Means* seleciona k elementos para formação inicial dos centróides do grupo. Após a seleção dos centróides, é calculada a distância (neste experimento foi utilizada a distância euclidiana) dos elementos restantes de cada elemento em relação aos centróides, sendo considerada a menor distância encontrada para identificar a similaridade, em seguida calcula a média para cada novo grupo. O processo termina somente quando todos os elementos estiverem agrupados.

Após a formação dos *clusters* foi utilizado o algoritmo de classificação C4.5 buscando identificar as relações entre as similaridades obtidas em cada agrupamento. O C4.5 é um sistema de aprendizado de máquina que constrói a árvore de decisão a partir da raiz, selecionando o melhor atributo classificador (maior ganho de informação) dentre todos os atributos do conjunto de dados. Após a escolha, os dados são separados de acordo com as classes do atributo escolhido, gerando uma subdivisão dos dados para cada descendente na árvore. O algoritmo é aplicado recursivamente a cada descendente, até que algum critério de parada seja atingido – por exemplo, a exaustão da análise de todos os atributos previsores.

O J48 é uma implementação em Java do algoritmo C4.5 release 8 – última versão pública – de Quinlan (1993). Ele foi desenvolvido para ser incorporado ao WEKA. Este método apresenta como resultado uma árvore de decisão com poda que pode ser transformada em um conjunto de regras de decisão.

Uma das opções de teste do algoritmo de classificação C4.5 é a validação cruzada com k-partições (k-*folds*), que consiste em dividir a base de dados aleatoriamente, em k partições. Na seqüência, uma das partições é selecionada para compor o conjunto de teste e os k-1 restantes são usados para compor o conjunto de treinamento. Este procedimento é repetido até que todas as k partições sejam selecionadas como conjunto de teste. O resultado final é fornecido pela média dos resultados obtidos em cada partição (KOHAVI, 1995).

4 RESULTADOS

4.1 ANÁLISE PRÉVIA

Ao analisar a base, descrita na seção 3, sem aplicação de qualquer técnica de mineração de dados, foi possível identificar duas informações importantes, que estão representadas nas Figuras 3 e 4.

Na Figura 3 pode-se identificar que os cinco tipos de solos com maior produtividade são: N – Nitossolo, L – Latossolo, P – Argilossolo, O – Organossolo, para esta análise foi considerada a produtividade média dos últimos três anos que esta sendo considerado no estudo. Esta constatação levou a pergunta sobre se a área plantada no Brasil também esta concentrada nestes solos mais produtivos. Para responder a este questionamento foram comparadas a média da área de plantio dos mesmos três anos com o tipo de solo, o resultado pode ser verificado na Figura 4.

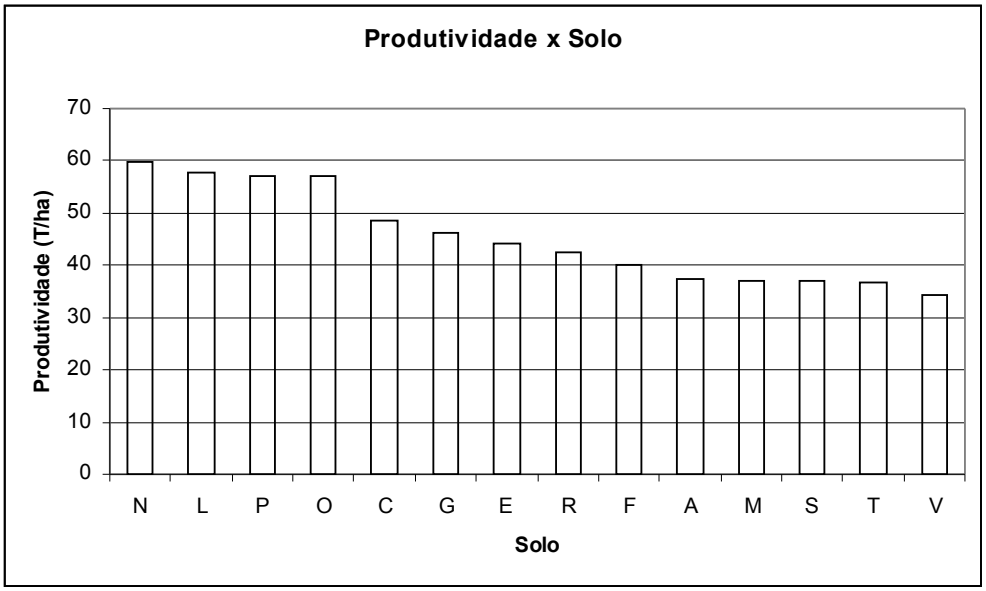

FIGURA 3 – PRODUTIVIDADE MÉDIA DOS ANOS DE 2008 A 2010 COMPARADA AO TIPO DE SOLO.
FONTE: Os Autores (2012).

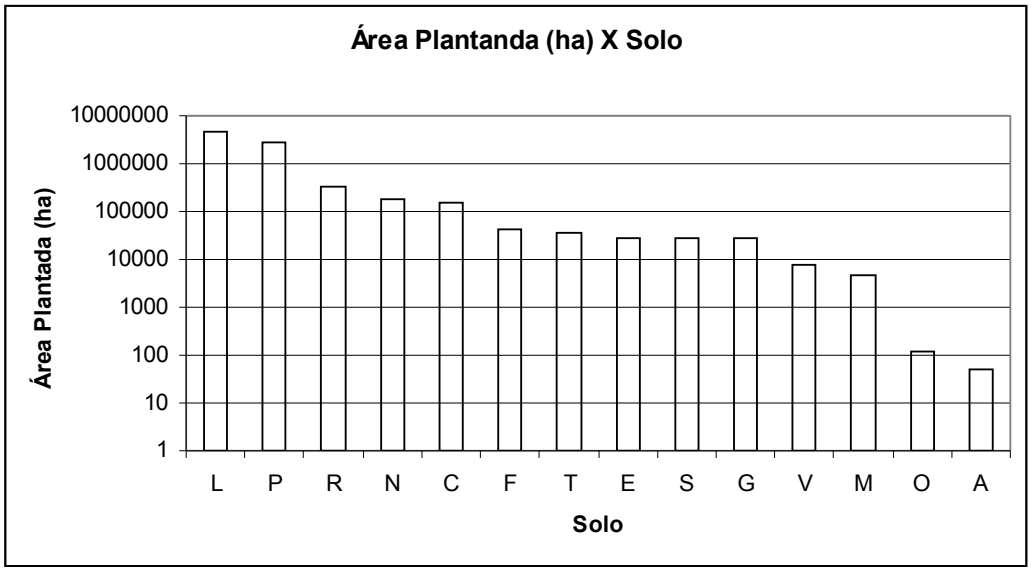

FIGURA 4 – MÉDIA DA ÁREA PLANTADA (ha) DOS ANOS DE 2008 A 2010 COMPARADO AO TIPO DE SOLO.
FONTE: Os Autores (2012).

O gráfico da Figura 4 está em escala logarítmica, pois há um crescimento exponencial da área plantada entre os diferentes tipos de solo. Pode-se identificar que os solos: L – Latossolo com área plantada de 4.785.576ha e P – Argilossolo com área plantada de 3.042.735ha são solos que representam 90% da área plantada de cana-de-açúcar no Brasil, ou seja, a produtividade maior que está vinculada ao solo da classe N – Nitossolo representa somente 2% da área plantada.

4.2 ANÁLISE DOS RESULTADOS OBTIDOS COM A MINERAÇÃO DE DADOS

Após as análises prévias, foram aplicados algoritmos de mineração de dados para avaliar a descoberta de conhecimento "novo" referente à base. Primeiramente foi aplicado o algoritmo

K-Means, sendo agrupados em 5 *clusters*, o resultado obtido foi: no "cluster0" foram agrupados 171 registros (15%), "cluster1" - 438 registros (38%), "cluster2" - 238 registros (21%), "cluster3" - 136 registros (12%) e o "cluster4" - 177 registros (15%).
Da aplicação do algoritmo C4.5 sobre a base resultante do algoritmo *K-Means* considerando como atributo meta o *cluster* e utilizando-se validação cruzada de 10 partições, foi induzida uma árvore com 50 folhas e tamanho 87.
O atributo que teve maior ganho de informação foi o Solo dividindo a área do experimento em quatorze classes principais, que se referem aos tipos de solos. O solo que teve o maior número de instâncias classificadas foi o Latossolo (L) onde 423 instâncias do "cluster1", ou seja, 50,54% do total de instâncias foram classificadas na regra.
A estrutura da árvore gerada pelo algoritmo C4.5 para a base citada foi convertida para um conjunto de regras. Este conjunto é composto por 47 regras para classificar os *cluster* nos quais as instâncias foram agrupadas, buscando assim identificar qual a similaridade ou proximidade entre os atributos que levaram as instâncias a ficarem em um mesmo grupo. Na Tabela 1 é possível visualizar as regras encontradas que obtiveram o maior número de classificações para cada *cluster*.

TABELA 1 – PRINCIPAIS REGRAS DE CLASSIFICAÇÃO PARA O AGRUPAMENTO DE CANA-DE-AÇÚCAR NO BRASIL

SE	E	E	E	E	ENTÃO
SOLO = C					cluster0
SOLO = L					cluster1
SOLO = P	PRODUT MEDIA <= 58	LONGITUDE <= 38.47	AREA PLANTADA <= 2300	LATITUDE > 3.86	cluster2
SOLO = P	PRODUT MEDIA <= 58	LONGITUDE > 38.47	LATITUDE <= 10.08		cluster3
SOLO = P	PRODUT MEDIA > 58	LATITUDE > 6.02	LONGITUDE <= 47.37		cluster4

FONTE: Os Autores (2012).

A avaliação das regras pela validação cruzada de 10 partições resultou em uma taxa de acerto de 93,3621% e taxa de erro de 6,6379%. A matriz de confusão que apresenta os acertos por classes consta na Tabela 2.

TABELA 2 – MATRIZ DE CONFUSÃO DO CLASSIFICADOR OBTIDO PELO ALGORITMO C4.5

Classificado como ➔	A	B	C	D	E
A = cluster0	161	3	5	2	0
B = cluster1	11	424	1	1	1
C = cluster2	4	1	222	11	0
D = cluster3	2	3	15	110	6
E = cluster4	3	2	0	6	166

FONTE: Os Autores (2012).

É possível identificar na matriz de confusão que houve um grande número de acertos em todas as classes. No entanto, a classe do cluster3 teve o maior percentual de erros, dos 136 exemplos 26 foram incorretamente classificados, representando 19,12% dos registros desta classe. Este resultado é explicado pelo conjunto de dados, pois como os dados dos demais *clusters* são mais significativos (com mais número de registros em cada *cluster*) o algoritmo tende a ter maior ganho de informação para as classes mais significativas, por isso, os demais *clusters* são melhores classificados do que o cluster3 que representa apenas 12% dos registros, como apresentado anteriormente.

5 CONCLUSÃO

Por meio da análise realizada foi possível identificar que com o processo de KDD houve uma importante complementação à informação que já se possuía antes da mineração, havendo uma contribuição para a descoberta de conhecimento considerando os dados analisados. Ao identificar os tipos de solo mais produtivos a mineração de dados possibilitou complementar esta informação identificando quais *clusters* estão agrupados nas áreas de solo mais produtivos e identificar os fatores de similaridades entre os grupos.

Na agricultura de precisão que visa tornar o campo mais produtivo sem que para isto seja necessário aumentar a área plantada a mineração de dados mostrou-se ser uma aliada, pois pode trazer ao gestor informações complementares de todos os atributos que influenciam na produtividade, como já se sabia que o solo e o clima são os fatores mais influentes, o agrupamento realizado permitiu identificar as similaridades entre os atributos, confirmando que o tipo de solo é a característica fundamental para os grupos, seguido pela produtividade e localização.

Na árvore de decisão gerada pelo algoritmo C4.5 pode-se identificar as proximidades entre os atributos que fizeram pertencer a uma mesma classe, o cluster1 por exemplo, que continha o maior número de registros (38% da base) teve 99,76% de seus registros classificado na mesma regra do solo L - Latossolo, sabe-se que este é um dos solos com maior área plantada no Brasil, no entanto não é o mais produtivo.

Finalmente, pretende-se ainda expandir esta pesquisa utilizando uma série histórica de dados mais longa e um mapa de classe de solo com maior resolução bem como comparar outros métodos de mineração de dados, além dos dois aplicados neste estudo.

REFERÊNCIAS

BIOETANOL – Bioetanol de Cana-de-açúcar. **BNDES publica livro com mapeamento inédito do setor de etanol.** Disponível em: <http://www.bioetanoldecana.org/pt/download/release.pdf> Acesso em: 28 de maio de 2012;

CALIL, L. A. A.; CARVALHO, D. R.; SANTOS, C. B.; VAZ, M. S. M. G. **Mineração de dados e pós-processamento em padrões descobertos**. Publicação UEPG Ci. Exatas Terra, Ci. Agr. Eng., Ponta Grossa, 14 (3): p. 207-215, dez. 2008;

EMBRAPA- Empresa Brasileira de Pesquisa Agropecuária. **SBCS - Sistema Brasileiro de Classificação de Solos**. Brasília, 1999;

FAYYAD, U. M.; PIATESTKY SHAPIRO, G.; SMYTH, P. From data mining to knowledge discovery: an overview. In: FAYYAD, U. M. et al. (Ed.). **Advances knowledge discovery and data mining**. Menlo Park: AAAI, p. 1-36, 1996;

FONSECA, M. B; PAIXÃO, M, C; MAIA, S, F. **Previsão de produção do etanol brasileiro para exportação: uma aplicação de vetores autoregressivos (VAR).** In: XLVI Congresso da Sociedade Brasileira de Economia, Administração e Sociologia Rural. 2008. Rio Branco – AC;

GHOSH, J.; LIU, A. The top ten algorithms in data mining. **K-means**. Cap. 2, p. 21-35. 2009;

HAN, J; KAMBER,M. **Data mining: concepts and techniques.** 2. ed. 2006;

IBGE – Instituto Brasileiro de Geografia e Estatística. Disponível em: <http://www.ibge.gov.br> Acesso em: 25 de maio de 2012;

IBGE – Instituto Brasileiro de Geografia e Estatística. **Produção agrícola municipal**. Disponível em: <http://www.ibge.gov.br/home/estatistica/economia/pam/2010/default.shtm> Acesso em: 25 de maio de 2012b;

IBGE - Instituto Brasileiro de Geografia e Estatística. **Mapa de solos**. Disponível em: <http://mapas.ibge.gov.br/solos/viewer.htm> Acesso em: 25 de maio de 2012c;

IBGE - Instituto Brasileiro de Geografia e Estatística. **Mapas interativos e Malhas digitais:** relacionadas à classificação de solos e divisão política dos municípios. Disponível em: <ftp://geoftp.ibge.gov.br/malhas_digitais/> e <http://mapas.ibge.gov.br/> Acesso em: 25 de maio de 2012d;

KOHAVI, R. A Study of CrossValidation and Bootstrap for Accuracy Estimation and Model Selection. In: **Appears in the International Joint Conference on Artificial Intelligence**. 1995. Disponível em: <http://robotics.stanford.edu/~ronnyk/accEst.pdf> Acesso em: 20 de abr. de 2012.

MAPA - MINISTÉRIO DA AGRICULTURA, PECUÁRIA E ABASTECIMENTO. **Evolução da produtividade e da produção de cana-de-açúcar no Brasil por safra**. 07 out. 2011. Disponível em: <http://www.agricultura.gov.br/arq_editor/file/Desenvolvimento_Sustentavel/Agroenergia/estatisticas/producao/SETEMBRO_2011/08_%20area_prod_brasil(1).pdf> Acesso em: 23 de maio de 2012;

MAPA - MINISTÉRIO DA AGRICULTURA, PECUÁRIA E ABASTECIMENTO. **Cana de açúcar.** Disponível em: <http://www.agricultura.gov.br/vegetal/culturas/cana-de-acucar> Acesso em: 24 de maio de 2012;

MARIN, F. R. EMBRAPA- Empresa Brasileira de Pesquisa Agropecuária. **Características**. Disponível em: <http://www.agencia.cnptia.embrapa.br/gestor/cana-de-acucar/arvore/CONTAG01_20_3112006152934.html> Acesso em: 03 de dezembro de 2011;

MARTINS, P. F. ; LAUGENI, F. P. **Administração da produção**. São Paulo: Saraiva, 2006;

POSTGIS. Disponível em: <http://www.postgis.org> Acesso em: 28 de maio de 2012.

POSTGRESQL Disponível em: <http://www.postgresql.org> Acesso em: 28 de maio de 2012.

UNICA- União da Indústria de Cana-de-Açúcar**, Cultivo da cana hoje**. Disponível em: < http://www.unica.com.br/content/show.asp?cntCode=9E97665F-3A81-46F2-BF69-26E00C323988> Acesso em: 26 de maio de 2012;

UNICA - União da Indústria de Cana-de-Açúcar. **Setor sucroenergético - mapa da produção**. Disponível em: < http://www.unica.com.br/content/show.asp?cntCode={D6C39D36-69BA-458D-A95C-815C87E4404D}> Acesso em: 26 de maio de 2012b;

UNICA- União da Indústria de Cana-de-Açúcar. **Setor sucroenergético – histórico**. Disponível em: < http://www.unica.com.br/content/show.asp?cntCode=8875C0EE-34FA-4649-A2E6-80160F1A4782> Acesso em: 26 de maio de 2012;

VIAN, C. E. F. Agência de informação Embrapa - Cana-de-açúcar. **Mercado**. Disponível em: <http://www.agencia.cnptia.embrapa.br/gestor/cana-de-acucar/arvore/CONTAG01_68_711200516719.html> Acesso em 02 de dezembro de 2011;

VIAN, C. E. F. Agência de informação Embrapa - Cana-de-açúcar. **Estatísticas**. Disponível em: <http://www.agencia.cnptia.embrapa.br/gestor/cana-de-acucar/arvore/CONTAG01_66_711200516719.html> Acesso em: 25 maio de 2012;

WEKA: Data mining software. Disponível em <http://www.cs.waikato.ac.nz >. Acessado em: 02 de dezembro de 2011.