ANÁLISE DE ACESSOS REPÚBLICA BRASILEIRA UTILIZANDO TÉCNICAS DE MINERAÇÃO DE DADOS

Isabela Cristina Teles Terra
Instituto Federal Fluminense
isabelacristinaterra@gmail.com

Tamys Luiz Fernandes
Universidade Federal Rural do Rio de Janeiro
tamyslf@gmail.com

Milton Erthal Junior
Universidade Candido Mendes
miltonerthal@hotmail.com

Henrique Rego Monteiro da Hora
Instituto Federal Fluminense
henrique.dahora@iff.edu.br

RESUMO

O objetivo desse trabalho foi utilizar técnicas de mineração de dados para verificar a relação das características da coleção de germoplasma da cana-de-açúcar com siglas República Brasileira (RB) do Banco Ativo de Germoplasma (BAG) 2018 da Rede Interuniversitária para o Desenvolvimento do Setor Sucroenergético (RIDESA). Foram utilizados registros de 414 acessos de sigla RB e 10 atributos. Para análise da tarefa de classificação, utilizando o método de árvore de decisão binária, o potencial do modelo de árvore de decisão foi avaliado em relação ao BRIX, TONELADA/HECTARE E BROTAÇÃO DE SOCARIA de cada genótipo. Dos resultados obtidos: Em relação ao Brix apresentou um acerto em 60,6% das instâncias com a árvore de decisão indicando influência direta da Maturação; em relação a Tonelada/Hectare apresentou um acerto em 56,2% das instâncias com a árvore de decisão apontando influência do Perfilhamento; e em relação a Brotação de Socaria apresentou um acerto em 74,6% das instâncias com a árvore de decisão indicando influência do Perfilhamento seguido da Tonelada/Hectare.

Palavras chaves: Mineração de Dados, Cana-de-açúcar, Germoplasma, Genótipo.

ABSTRACT

The objective of this work was to use data mining techniques to verify the relation of the characteristics of the sugarcane germplasm collection with the Brazilian Republic (RB) of the Germplasm Asset Bank (BAG) 2018 of the Rede Interuniversitária para o Desenvolvimento do Setor Sucroenergético (RIDESA). Used records of 414 accesses of RB and 10 attributes. For the analysis of the classification task, using the binary decision tree method, the potential of the decision tree model was evaluated in relation to the BRIX, TONNED/HECTARE AND BUDDING of each genotype. From the results obtained, Brix presented a 60.6% accuracy of the instances with the decision tree indicating direct influence of the Maturation; in relation to Tonned/Hectare presented a hit in 56.2% of the instances with the decision tree pointing to the influence of the tillering; and in relation to Budding showed a hit in 74.6% of the instances with the decision tree indicating influence of the tillering followed by the Tonned/Hectare.

Keywords: Data Mining, Sugarcane, Germplasm, Genotype.

Como Citar:
TERRA, Isabela Cristina Teles et al. Análise de variedades RB da cana-de-açúcar utilizando técnicas de mineração de dados. *In*: SIMPÓSIO DE PESQUISA OPERACIONAL E LOGÍSTICA DA MARINHA, 19., 2019, Rio de Janeiro, RJ. **Anais** […]. Rio de Janeiro: Centro de Análises de Sistemas Navais, 2019.

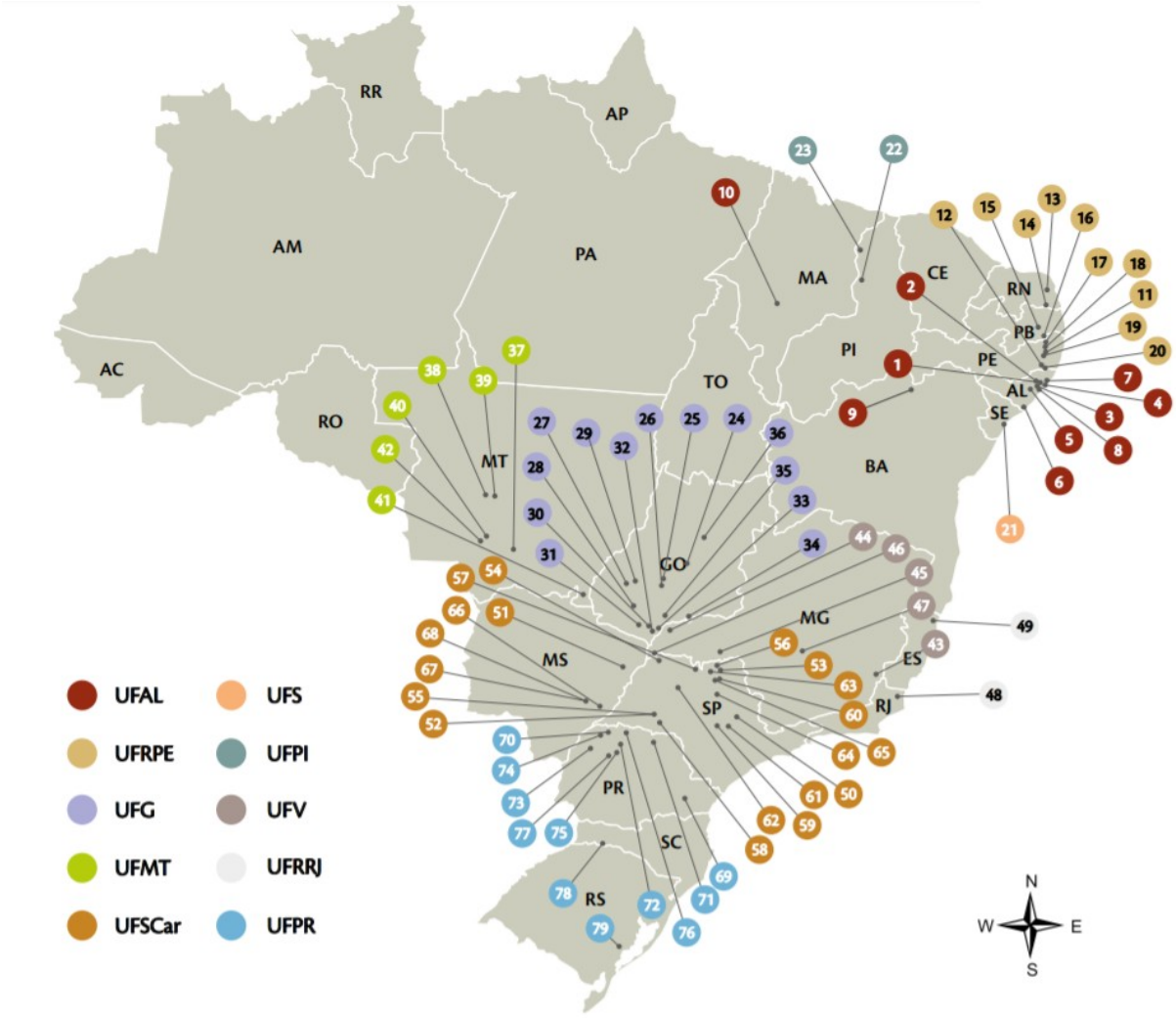
1. INTRODUÇÃO

O Brasil é atualmente o maior produtor e exportador de cana-de-açúcar do mundo. A cana-de-açúcar é a primeira fonte de energia renovável do país e responsável por 17,5% da matriz nacional (UNICA, 2018). Devido à relevância dessa cultura para a agricultura nacional é importante que haja investimento em pesquisas e desenvolvimento de novas tecnologias de manejo e produção.

Com a extinção do Instituto do Açúcar e do Álcool em 1990 ocorreu uma medida acertada, com a transferência dos recursos humanos, das estruturas físicas e tecnológicas do PLANALSUCAR para as Universidades Federais de Alagoas (UFAL), Rural de Pernambuco (UFRPE), Viçosa-MG (UFV), São Carlos-SP (UFSCar), Rural do Rio de Janeiro (UFRRJ), Paraná (UFPR) e Sergipe (UFS), que a partir de então criaram a Rede

Interuniversitária para o Desenvolvimento do Setor Sucroenergético (RIDESA). As atividades de pesquisa da RIDESA são desenvolvidas e partilhadas entre todas as Universidades, estimulando-se o intercâmbio de informações, de conhecimento e de resultados. Atualmente a RIDESA conta também com as Universidades Federais de Goiás (UFG), Mato Grosso (UFMT) e Piauí (UFPI), e atuam conjuntamente através de um Acordo de Parceria. Somam-se ao todo 79 bases de pesquisa, que englobam Laboratórios das Universidades, Estações de Cruzamento, Estações Experimentais e Bases de Seleção, localizadas nos estados onde a cultura da cana-de-açúcar apresenta maior expressão (Figura 1). Após a criação da RIDESA houve o lançamento de 59 variedades. Hoje já são mais de 78 materiais de cana distinguidos com a sigla RB, República Brasileira, que tem alto nível de adoção nos estados canavieiros e ocupam cerca de 70% dos canaviais do país, uma contribuição de cerca de 12,3% na matriz energética do Brasil (RIDESA, 2018).

Figura 1 – Estações experimentais e bases de pesquisa do PMGCA/RIDESA.



Fonte: RIDESA (2018)

Desde o seu início, em 1990, até os dias de hoje, o Banco de Germoplasma de Serra do Ouro vem sendo alimentado com milhares de genótipos potencialmente úteis às estratégias do melhoramento genético da cana-de-açúcar. Contudo, o fato das instâncias das espécies não estarem relacionadas, apenas tabeladas em uma planilha eletrônica, tornam-se vagas, sem informações e valores concretos para agregar na escolha dos genótipos a serem cruzados.

Considerando a importância da cultura da cana-de-açúcar para o Brasil e a diante deste cenário, este trabalho aplicará o método de mineração de dados, para o tratamento dos dados do Banco Ativo de Germoplasma (BAG) 2018 da RIDESA. O intuito é verificar a relação das características da coleção do germoplasma da cana-de-açúcar com siglas República Brasileira (RB), de modo a auxiliar e facilitar nas tomadas de decisões e estratégias dos programas de melhoramento genético da rede.

A exposição deste trabalho prossegue como segue. Na seção 2 descreve-se a base de dados utilizada no estudo e o método aplicado. Na Seção 3 apresentam-se os resultados do estudo. Por fim, a Seção 4 apresenta um resumo, as conclusões do estudo juntamente com ideias potenciais para trabalhos futuros.

2. MATERIAL E MÉTODOS

Hand *et al*. (2001) define mineração de dados como a análise de grandes conjuntos de dados com objetivo de encontrar relacionamentos inesperados e de resumir os dados de uma forma que eles sejam tanto úteis quanto compreensíveis. É classificada pela capacidade em realizar diversas tarefas e o presente trabalho apresenta análises com a utilização das tarefas de classificação.

A tarefa de classificação consiste em examinar uma determinada característica nos dados e atribuir uma classe previamente definida. A tarefa de classificação pode ser usada, por exemplo, para determinar quando uma transação de cartão de crédito pode ser uma fraude ou identificar em uma escola, qual a turma mais indicada para um determinado aluno.

Dentre as técnicas de classificação, destacam-se as árvores de decisão. Uma árvore de decisão é um fluxograma semelhante a uma estrutura de árvore, onde cada nodo interno denota um teste em um atributo, cada arco proveniente destes nodos são ramas que representam o resultado do teste e cada folha representa a distribuição dos registros (CAMILO; SILVA, 2009; HAND; MANNILA; SMYTH, 2001).

Para execução da Tarefa de Classificação dos dados, foi utilizado o método de árvore de decisão binária com algoritmo de indução J48, uma modificação do amplamente conhecido algoritmo C4.5, aplicado para o software Weka. Software este utilizado para aplicação do trabalho (GHOLAP, 2012).

Foram verificadas configurações diferentes, em que cada uma resultou em taxas diferentes de acerto do modelo, sempre igual ou superior a 50%. O que se refere a valores relativamente altos quando se trata de cana-de-açúcar, devido à compreensão do seu genoma ser limitada. Pois, diferente dos humanos, que tem duas cópias de cada um de seus pares de cromossomos, a cana-de-açúcar possui um arranjo genético muito mais complexo, com várias cópias de cada cromossomo e numerosas variantes de cada gene (SOUZA; SLUYS, 2010).

3. RESULTADOS E DISCUSSÃO

O conjunto de dados inicial contava com 30 atributos. Entretanto, devido à falta de informação de alguns atributos das espécies constantes no BAG 2018, foi necessária a adequação dos dados para serem processados e o número de atributos foi reduzido para 10, conforme Tabela 1. Foram selecionados somente atributos que continham dados de registro, sem prejuízo a quantidade de instâncias inicialmente disponíveis.

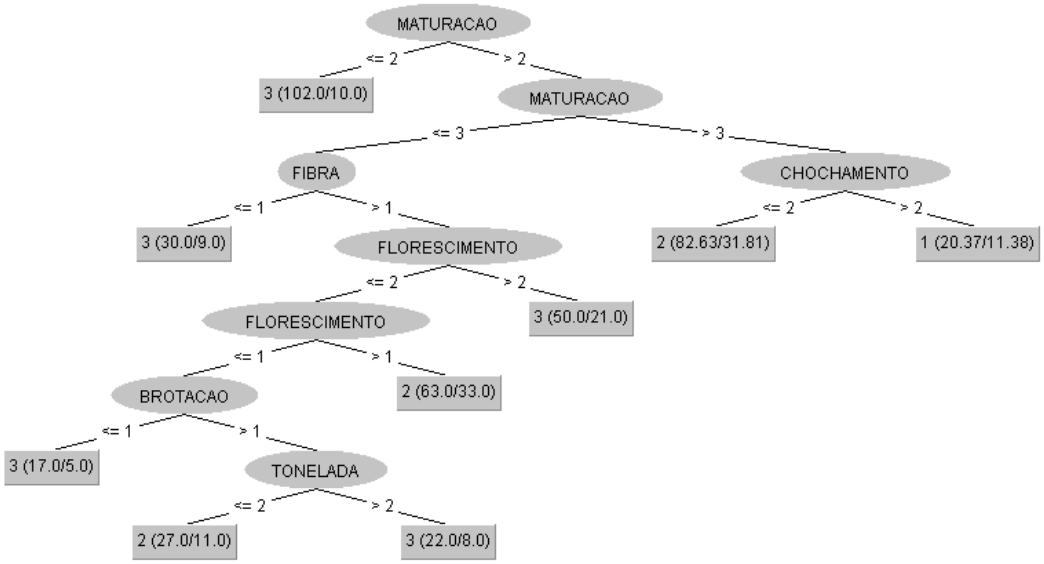
Foram utilizadas instâncias de 414 acessos RB da cana-de-açúcar disponíveis no BAG 2018. Para análise da tarefa de classificação, o potencial do modelo de árvore de decisão foi avaliado em relação ao BRIX, TONELADA/HECTARE E BROTAÇÃO DE SOCARIA de cada genótipo. Suas análises são de extrema importância para levantar a qualidade da matéria prima.

Tabela 1. Descrição e exemplos dos atributos contidos na base; os exemplos apresentados são de instâncias reais presentes no banco de dados.

Atributo	Descrição	Escala
brix	porcentagem em massa de sólidos solúveis contidos em uma solução de sacarose quimicamente pura	Baixo/Médio/Alto
maturação	processo fisiológico que envolve a formação de açúcares nas folhas e seu deslocamento e armazenamento no colmo	Hiper Precoce/ Precoce/Média/Tardia
tonelada/hectare	medida de produtividade	Baixa/Média/Alta
fibra	porcentagem de fibra na cana	Baixa/Média/Alta
brotação de socaria	tempo para crescimento dos brotos	Regular/Bom/ Médio/Ótimo
perfilhamento	processo de emissão de perfilhos por planta	Regular/Bom/Ótimo
desenvolvimento	velocidade de crescimento vegetativo	Regular/Bom/Ótimo
diâmetro	diâmetro do colmo da cana-de-açúcar	Fino/Médio/Grosso
florescimento	probabilidade de florescer	Raro/Baixo/Médio/Alto
chochamento	probabilidade de secamento do interior do colmo	Raro/Baixo/Médio/Alto

A avaliação em relação ao Brix apresentou um acerto em 60,6% das instâncias. A árvore de decisão demonstrou influência da Maturação, conforme Fig. 2, indicando que quando a Maturação é Precoce ou Hiper Precoce, o nível do Brix é Alto. A Maturação é muito específica de cada cana, porém, de acordo Marques e da Silva (2008), conforme citado por Delgado & César (1977); Paranhos (1987); Lopes & Parazzi (1992), “a maturação da cana-de-açúcar pode ser determinada pelos parâmetros tecnológicos (Brix, Pol, Pureza e Açúcares Redutores) e I.M. (índice de maturação).”.

Figura 2 - Árvore de decisão relacionada ao Brix a partir do conjunto de dados da Tabela 1.

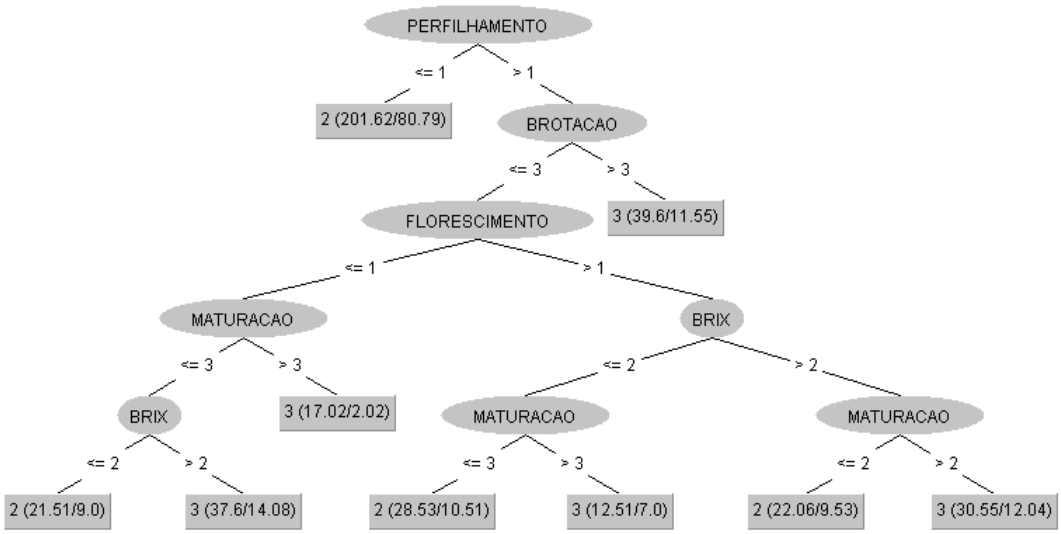


Fonte: Própria (2019).

Em seguida a avaliação em relação a Tonelada/Hectare apresentou um acerto em 56,2% das instâncias, onde a árvore de decisão demonstrou influência do Perfilhamento, conforme Fig. 3. A elevada Tonelada/Hectare está diretamente ligada a um Bom Perfilhamento. Silva *et al*. (2008) afirmam que os genótipos da cana-de-açúcar respondem à época de colheita em relação à produtividade, e respondem à altura de corte e à época de colheita dos colmos quanto ao perfilhamento.

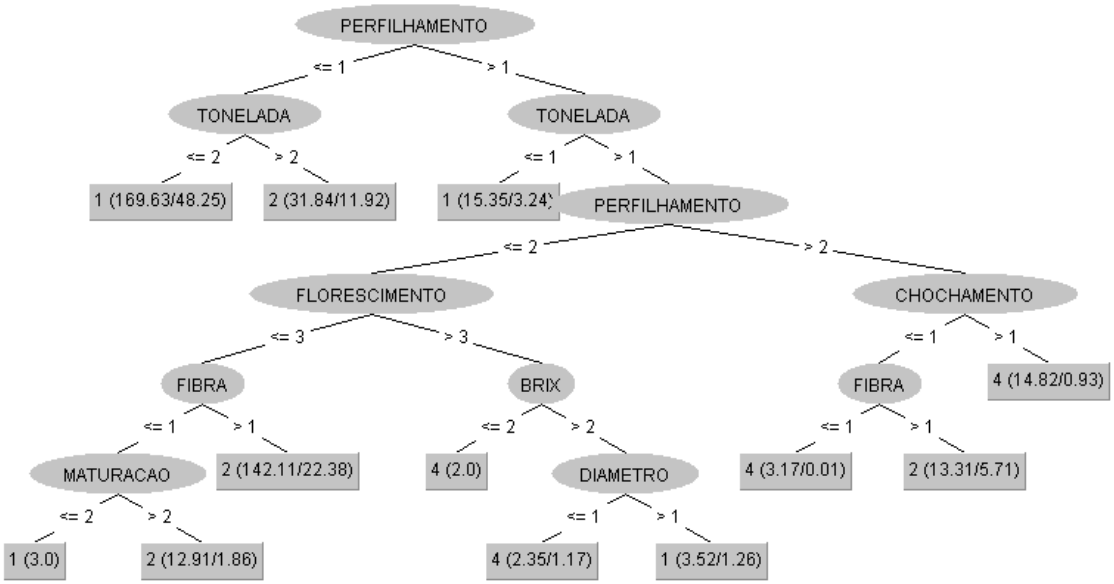
Já a avaliação em relação a Brotação de Socaria apresentou um acerto em 74,6% das instâncias e a árvore de decisão demonstrou influência do Perfilhamento seguido da Tonelada/Hectare, Fig. 4. O perfilhamento é uma variável importante para a cultura da cana-de-açúcar, e que se relaciona diretamente com a Tonelada/Hectare. Além disso, uma boa brotação aumenta a chance de ter maior número de colmos impactando em uma alta tonelada/hectare (SANTOS, 2015).

Figura 3 - Árvore de decisão relacionada a Tonelada/Hectare a partir do conjunto de dados da Tabela 1.



Fonte: Própria (2019)

Figura 4 - Árvore de decisão relacionada a Brotação de Socaria a partir do conjunto de dados da Tabela 1.



Fonte: Própria (2019)

4. CONCLUSÕES E TRABALHOS FUTUROS

Considerados satisfatórios, os resultados obtidos com a aplicação da Mineração de Dados no BAG 2018 apontam para uma possível viabilidade de se realizar inferências relativas ao melhoramento genético da cana-de-açúcar, descobrindo regras ou padrões interessantes, de forma rápida e fácil.

Se os genótipos estivessem catalogados por completo no BAG 2018, seria possível a mineração de todo o banco, não somente das siglas RB, o que influenciaria diretamente nos

resultados aqui obtidos. Futuros trabalhos poderiam ainda ser desenvolvidos com a utilização do conjunto de atributos original, assim como com o uso de todos os genótipos disponíveis.

5. REFERÊNCIAS BIBLIOGRÁFICAS

CAMILO, C. O.; SILVA, J. C. DA. Mineração de Dados: Conceitos, Tarefas, Métodos e Ferramentas. **Instituto de Informática - UFG**. 29, ago. 2009.

GHOLAP, J. Performance Tuning Of J48 Algorithm For Prediction Of Soil Fertility. **Published in Asian Journal of Computer Science and Information Technology**,Vol 2,No. 8. 20 ago. 2012.

HAND, D. J.; MANNILA, H.; SMYTH, P. **Principles of data mining**. Cambridge, Mass: MIT Press, 2001.

MARQUES, T. A.; DA SILVA, W. H. Crescimento vegetativo e maturação em três cultivares de cana-de-açúcar. **Revista de Biologia e Ciências da Terra**, v. 8, n. 1, 2008.

RIDESA, R. **Rede Interuniversitária para o Desenvolvimento do Setor Sucroenergético**. Disponível em: <https://www.ridesa.com.br/>. Acesso em: 23 ago. 2018.

SANTOS, M. A. L. DOS [UNESP. Balanço hídrico, crescimento e produtividade de genótipos RB de cana-de-açúcar em cultivo de sequeiro na região de Rio Largo-AL. **Aleph**, p. x, 60 f. : il. color., grafs., tabs, 27 fev. 2015.

SILVA, M. DE A.; JERONIMO, E. M.; LÚCIO, A. D. Height of cut and harvest period effects on tillering and yield of sugarcane. **Pesquisa Agropecuária Brasileira**, v. 43, n. 8, p. 979–986, ago. 2008.

SOUZA, G. M.; SLUYS, M.-A. V. Genômica e biotecnologia da cana-de-açúcar: estado da arte, desafios e ações. **Bioetanol de cana-de-açúcar: P&D para produtividade e sustentabilidade**, v. 1, p. 325–332, 2010.

UNICA, U. **União da Indústria de Cana-de-açúcar**. Disponível em: <https://www.unica.com.br/>. Acesso em: 23 ago. 2018.