

**UNIVERSIDADE ESTADUAL PAULISTA – UNESP
CÂMPUS DE JABOTICABAL**

Gabriela Mourão de Almeida

**APRENDIZAGEM DE MÁQUINA NA DETERMINAÇÃO DE AMBIENTES DE
PRODUÇÃO DE CANA-DE-AÇÚCAR**

2019

**UNIVERSIDADE ESTADUAL PAULISTA – UNESP
CÂMPUS DE JABOTICABAL**

Gabriela Mourão de Almeida

**APRENDIZAGEM DE MÁQUINA NA PREDIÇÃO DE AMBIENTES DE PRODUÇÃO
DE CANA-DE-AÇÚCAR**

**Orientador: Prof. Dr. Gener Tadeu Pereira
Coorientador: Prof^a. Dr^a. Angélica Santos Rabelo de Souza Bahía**

Dissertação apresentada à Faculdade de Ciências Agrárias e Veterinárias – Unesp, Campus de Jaboticabal, como parte das exigências para a obtenção do título de Mestre em Agronomia (Ciência do Solo)

2019

A447a Almeida, Gabriela Mourão de
Aprendizagem de máquina na predição de ambientes de produção
de cana-de-açúcar. / Gabriela Mourão de Almeida. -- Jaboticabal,
2019
49 p. : il., tabs., mapas + 1 CD-ROM

Dissertação (mestrado) - Universidade Estadual Paulista (Unesp),
Faculdade de Ciências Agrárias e Veterinárias, Jaboticabal
Orientador: Gener Tadeu Pereira
Coorientadora: Angélica Santos Rabelo de Souza Bahía

1. Agricultura de precisão. 2. Manejo específico. 3. Mapeamento do
solo. 4. Sustentabilidade produtiva. I. Título.

Sistema de geração automática de fichas catalográficas da Unesp. Biblioteca da Faculdade de
Ciências Agrárias e Veterinárias, Jaboticabal. Dados fornecidos pelo autor(a).

Essa ficha não pode ser modificada.

CERTIFICADO DE APROVAÇÃO

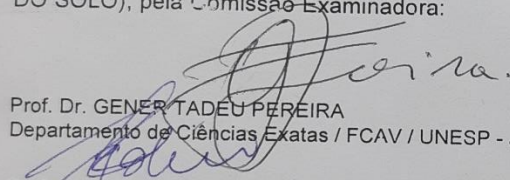
TÍTULO DA DISSERTAÇÃO: APRENDIZAGEM DE MÁQUINA NA DETERMINAÇÃO DE AMBIENTES DE PRODUÇÃO DE CANA-DE-AÇÚCAR

AUTORA: GABRIELA MOURÃO DE ALMEIDA

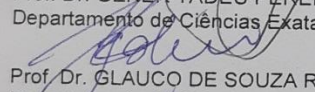
ORIENTADOR: GENER TADEU PEREIRA

COORDINADORA: ANGÉLICA SANTOS RABELO DE SOUZA BAHIA

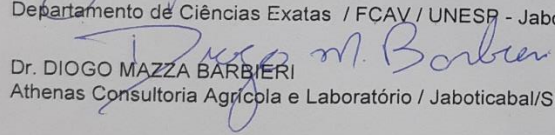
Aprovada como parte das exigências para obtenção do Título de Mestre em AGRONOMIA (CIÊNCIA DO SOLO), pela Comissão Examinadora:



Prof. Dr. GENER TADEU PEREIRA
Departamento de Ciências Exatas / FCAV / UNESP - Jaboticabal



Prof. Dr. GLAUCO DE SOUZA ROLIM
Departamento de Ciências Exatas / FCAV / UNESP - Jaboticabal



Dr. DIOGO MAZZA BARBIERI
Athenas Consultoria Agrícola e Laboratório / Jaboticabal/SP

Jaboticabal, 31 de julho de 2019

DADOS CURRICULARES DO AUTOR

Gabriela Mourão de Almeida, nascida em Belém do Pará, no dia 15/11/1994, formada em Engenharia Agrônômica pela Universidade Federal Rural da Amazônia, campus de Capitão Poço. Atualmente é mestranda do programa de pós-graduação em Agronomia (Ciência do Solo) da FCAV Unesp, campus de Jaboticabal.

“Nunca deixe ninguém te dizer que não pode fazer alguma coisa. Se você tem um sonho tem que correr atrás dele. As pessoas não conseguem vencer e dizem que você também não vai vencer. Se você quer uma coisa, corre atrás.”

A Procura da Felicidade

Dedico

A meus pais, Luiz Carlos Carneiro de Almeida e Katia do Socorro da Cunha Mourão de Almeida, grandes incentivadores e os maiores exemplos de amor na minha vida, esse trabalho é dedicado a vocês. A minha vitória também é de vocês!

AGRADECIMENTOS

Agradeço essa pesquisa primeiramente a Deus, autor do meu destino, companheiro de todos os momentos. Ele alimentou a minha alma com calma e esperança durante toda a jornada.

Aos meus pais Luiz Carlos Carneiro de Almeida e Katia do Socorro da Cunha Mourão de Almeida, meus heróis, que apesar de todas as dificuldades me fortaleceram, me apoiaram, incentivaram nas horas difíceis, de desânimo e cansaço. Vocês são o grande amor da minha vida e tudo que faço é por vocês, para poder proporcionar o mínimo que merecem.

A minha irmã Sabrina Mourão de Almeida que nos momentos de minha ausência dedicados ao estudo superior, sempre fez entender que o futuro é feito a partir da constante dedicação no presente.

Ao orientador Dr. Gener Tadeu Pereira, agradeço por todos os ensinamentos, conselhos, amizade, pela disponibilidade que sempre manifestou e a empatia com que recebeu as minhas ideias, foram o estímulo que me permitiu vencer as inseguranças deste processo. Obrigada por tudo.

A minha coorientadora Angélica Santos Rabelo de Souza Bahía por todos ensinamentos dados ao longo deste processo. Obrigada por tudo.

In memoriam ao meu avô Nelson de Azevedo Mourão e a minha avó e madrinha Maria da Cunha Mourão que acompanhou o início da minha vida acadêmica em matéria e está acompanhando a conclusão em espírito, pois, a existência material é apenas uma etapa de testes da vida eterna e quando esta se concluiu não quer dizer o fim de tudo e sim o início da verdadeira vida, você me faz muito feliz na infância. As minhas tias Sandra Maria Mourão, Sandra Cristina Mourão e Ana Madalena Mourão por todo o apoio e auxílio dados ao longo da vida desde a ajuda nas compras do material escolar da educação infantil a companhia para a viagem rumo ao mestrado.

Aos meus avós paternos Maria José Carneiro e Moacyr Almeida por terem dado a vida ao meu pai, Luiz Carlos que é uma pessoa maravilhosa e sou o que sou hoje graças a ele, bem como a todos da família Almeida.

A Vinicius Marchioro por toda a ajuda, companheirismo, paciência ao longo das etapas. Sem você e toda a sua visão prática, este trabalho não teria sido possível.

Ao professor Glauco de Souza Rolim e Alan Rodrigo Panosso por toda a paciência e conhecimento repassado ao longo da execução deste trabalho.

Aos companheiros do departamento de ciências exatas por todo o carinho e companheirismo ao longo da caminhada.

O presente trabalho foi realizado com apoio da Coordenação de Aperfeiçoamento de Pessoal de Nível Superior - Brasil (CAPES) - Código de Financiamento 001

Ao grupo de pesquisa CSME pela disponibilidade do banco de dados e as agências de pesquisa CAPES (Proc. nº149940); PROPE / UNESP (Pró-reitora de Pesquisa da Universidade Estadual Paulista, Edital Nº 15/2014); e CNPq (Conselho Nacional de Desenvolvimento Científico e Tecnológico, Edital Universal 01/2016 – Faixa C, Proc.nº 402796/2016-0).

SUMÁRIO

RESUMO.....	x
ABSTRACT	xi
LISTA DE FIGURAS	xii
LISTA DE TABELAS	xiii
1 Introdução.....	14
2 Revisão Bibliográfica	16
2.1 Cana de Açúcar	16
2.2 Ambientes de Produção	18
2.3 Aprendizagem de Máquina	19
2.4 Árvore de Decisão (AD)	21
3 Material e Métodos	21
3.1 Descrição da área	21
3.2 Análises Estatísticas.	25
3.2.1 Estatística Descritiva.....	25
3.2.2 Seleção pela Regressão “Stepwise”.	26
3.2.3 Multicolinearidade	26
3.2.4 Árvore de Decisão	27
3.2.5 Matriz de Confusão.....	27
3.2.6 Krigagem Indicadora.....	28
4 Resultados e Discussão	29
5 Conclusões	43
6 Referencial bibliográfico.....	44

APRENDIZAGEM DE MÁQUINA NA PREDIÇÃO DE AMBIENTES DE PRODUÇÃO DE CANA-DE-AÇÚCAR

RESUMO - A cana-de-açúcar é uma das culturas mais expressivas do mercado agrícola nacional. Visando um aumento de produtividade e qualidade da matéria prima, técnicas como a de manejo localizado, que já vem sendo adotada há muitos anos pelas usinas, porém, ainda de forma manual. O objetivo desse trabalho é determinar ambientes de manejo de cana-de-açúcar utilizando quantidade reduzida de variáveis de baixo custo, por meio de técnica de aprendizagem de máquina. Para atingir a máxima eficiência na predição, os dados foram submetidos à estatística descritiva, em seguida, à seleção de regressão “stepwise” para determinar quais variáveis seriam úteis ao modelo. Em seguida foi aplicado teste de multicolinearidade e, por fim, a árvore de decisão classificatória. Para avaliar a eficiência do modelo foi preparada uma matriz de confusão. Foi detectado que as variáveis ligadas às características de formação do solo foram as escolhidas para determinar os ambientes de produção, dando destaque a variável areia. A técnica de regressão “stepwise” mostrou-se eficiente na seleção de variáveis e a árvore de decisão mostrou eficiência na determinação dos ambientes, obtendo a satisfatória acurácia de 75%, além de ter gerado ambientes de manejo mais contínuos na área de cultivo.

Palavras-chave: Agricultura de precisão, Manejo específico, Mapeamento do solo, Sustentabilidade produtiva.

MACHINE LEARNING IN PREDICTING SUGARCANE PRODUCTION ENVIRONMENTS

ABSTRACT - Sugar cane is one of the most significant crops in the national agricultural market. Aiming to increase the quality and quantity of the raw material, techniques such as localized management, which has been adopted for many years by the plants, but still manually. The objective of this work is to determine the sugarcane management environments, using the reduced number of low-cost variables, through the machine learning technique. To achieve maximum prediction efficiency, the data were subjected to descriptive statistics, followed by stepwise regression selection to determine useful variable variables useful in the model. Then, the multicollinearity test was applied and, finally, a classification decision tree. To evaluate the efficiency of the model, a confusion matrix was prepared. It was detected that the variables selected to the soil characteristics were chosen to determine the production environments, highlighting a sand variable. A stepwise regression technique was efficient in the selection of variables and a reduced decision tree in the determination of environments, obtaining a satisfactory satisfaction of 75%, besides showing more continuous management environments in the cultivation area.

Keywords: Precision agriculture, Specific management, Soil mapping, Productive sustainability.

LISTA DE FIGURAS

Figura 1- Mapa de localização da área de estudo, a) mapa 3D demonstrativo da altimetria e grid dos pontos e b) Mapa de distribuição dos ambientes de produção desenvolvido pela CTC.	22
Figura 2- Mapa de distribuição dos atributos: a) mapa de ocorrência de areia na área de estudo. b) Mapa de classes de solo primeiro nível presentes na área.....	23
Figura 3- Mapa de distribuição dos atributos, a) mapa dos teores de susceptibilidade magnética ocorrida na área, d) Mapa de ocorrência do teor fósforo na área de estudo.	24
Figura 4-Mapa de distribuição dos atributos, a) mapa dos valores de pH presentes na área, b) mapa de variação de CTC potencial da área.	25
Figura 5- Fluxograma organizacional das análises realizadas.	25
Figura 6- – Esquema resultado explicativo da intersecção dos cinco mapas probabilísticos dos ambientes de produção resultando no mapa final.	29
Figura 7 - Gráficos violino representando a variabilidade das variáveis estudadas em função dos ambientes de produção. a) Teor de Fósforo nos ambientes; b) Valor da Susceptibilidade Magnética em cada ambiente, c) Valor da CTC potencial nos ambientes, d) Valor de pH em cada ambiente, e) Porcentagem de Areia nos ambientes.....	30
Figura 8- Dendrograma modelo da árvore de decisão com base no modelo 03.	35
Figura 9-Semivariogramas Indicadores dos ambientes de produção observados e dos preditos pela arvore de decisão (a) Ambiente A predito/observado; (b) Ambiente B predito/observado; (c) Ambiente C predito/observado; (d) Ambiente D predito/observado e (e) Ambiente E predito/observado. * Modelo: exp. = exponencial; esf. = esférico; efeito pepita (C0) ; alcance (a); patamar (C); Soma dos quadrados dos resíduos.	40
Figura 10- Mapas de probabilidade dos ambientes de produção observados e dos ambientes preditos pela arvore de decisão interpolados com a krigagem indicatriz (a) Ambiente A observado/predito; (b) Ambiente B observado/predito; (c) Ambiente C observado/predito; (d) Ambiente D observado/predito e (e) Ambiente E observado/predito.....	41
Figura 11- Mapas resultantes da intersecção das probabilidades dos ambientes de produção, a) Ambientes observados e b) Ambientes preditos pela AD.....	42

LISTA DE TABELAS

Tabela 1 - Potencial produtivo de cana-de-açúcar em toneladas de colmo por hectare em cada ambiente de produção.	19
Tabela 2 -Estatística descritiva associada aos gráficos box-plot.	31
Tabela 3- Métricas gerada com base em cada modelo desenvolvido pela seleção de stepwise sendo as variáveis representadas por: ambiente de produção (y), areia (X1), classe de solo (X2) e SM (X3).	33
Tabela 4-Valores de multicolinearidade por variável preditora.....	34
Tabela 5- Matriz de confusão do modelo utilizando dados de teste.	36
Tabela 6- – Métricas qualitativas desenvolvidas com base na matriz de confusão.	37

1 Introdução

A cana-de-açúcar (*Saccharum* sp.) é um dos cultivos tradicionais de maior importância no mundo, devido à versatilidade de produtos e subprodutos que, a partir dela, podem ser gerados. Nos últimos anos ocorreu aumento gradativo de interesse, a nível global, nesta cultura, devido à sua relevância econômica na produção de energia sustentável (NATARAJAN; SUBRAMANIAN; PAPAGEORGIOU, 2016).

Esta cultura é a principal matéria prima para a segunda maior indústria de agro-base do mundo, ficando atrás apenas do mercado têxtil. De acordo com dados da FAO (2019), atualmente, o Brasil é o maior produtor de cana-de-açúcar do mundo, produzindo 60% a mais que a Índia, segundo colocado. O estado de São Paulo atualmente é o maior produtor nacional da cultura, produzindo em média 80% a mais que o segundo colocado, o estado de Minas Gerais (IBGE, 2019).

Goldemberg et al. (2014) afirmam que no Brasil mais de 50% do que é produzido de cana-de-açúcar é voltado para a produção de etanol. Dada a importância desta cultura para o país, o governo brasileiro, na COP21, estabeleceu que a produção de etanol no ano de 2030 deverá atingir o total de 54 bilhões de litros, valor este que representa o dobro da atual produção. Já para a produção de açúcar, objetiva-se aumento de 38,7 para 46,4 milhões de toneladas e, para que essas metas sejam alcançadas, será necessário produzir 942 milhões de toneladas de cana por semestre em 2030 (SANCHES; MAGALHÃES; FRANCO, 2019).

Assim, o aumento da produtividade agrícola da cana-de-açúcar oferece alternativa sustentável, porém, para atingir estes volumes de produtividade será necessário investimento em tecnologia e pesquisa que possam contribuir significativamente para alcançar tais ganhos de produção.

Nos últimos anos, a agricultura de precisão (AP) vem se mostrando bastante promissora, pois envolve a aplicação de várias técnicas, com a finalidade de facilitar a gestão da variabilidade espacial e temporal de cultivos, visando assim, o aumento da produtividade agrícola e redução dos custos de produção. Segundo Moral; Terrón e Silva (2010) esses benefícios só são alcançados por meio do manejo específico local, que considera a variabilidade espacial e temporal dos campos.

Usinas canavieiras brasileiras já realizam o manejo de suas áreas levando em consideração a variabilidade local, os chamados “ambientes de cultivo” ou “ambientes de produção”, que nada mais são que, os fatores que compõem a formação de um determinado solo e suas interações, através de suas condições físicas, químicas, morfológicas e hídricas (PRADO, 2016).

Segundo Prado (2011), ambiente de produção consiste no conjunto de características físicas, hídricas, morfológicas, químicas e mineralógicas, de superfície e subsuperfície dos solos e nas condições climáticas (pluviometria, temperatura, radiação solar e evaporação) de um local.

É essencial ter conhecimento desses ambientes, pois é com base neles que as tecnologias de produção são adotadas, ao longo do cultivo, para que seja obtido o máximo rendimento de produção. Apesar de sua constante adoção e relevância para o sistema brasileiro de produção de cana-de-açúcar, a delimitação espacial destes ambientes de cultivo são atualmente baseadas na experiência dos profissionais, além destes se basearem em poucas amostras de solo, o que as tornam imprecisas, gerando condições locais irreais de variabilidade espacial do solo (SANCHES et al., 2018).

Nos últimos anos, alguns conceitos muito utilizados na tecnologia da informação (TI), vêm se mostrando ferramentas eficientes na extração de informações agrônômicas a partir de banco de dados. Segundo Bekker et al, (2019), os métodos estatísticos tradicionais lineares (por exemplo, modelos lineares generalizados) não são muito eficazes em revelar as complexas relações e padrões associados aos conjuntos de dados agrônômicos. Em contraste, está o aprendizado de máquina (AM), que detém a capacidade de modelar relações não-lineares em dados, sem a obrigação de satisfazer as premissas restritivas associadas às abordagens paramétricas.

A AM, que se origina dos campos da ciência da computação e da estatística, consiste em algoritmos matemáticos que aprendem por regressão ou classificação, à medida que são submetidos a treinamentos (CANIZO et al., 2019). Na ciência do solo, técnicas de AM têm sido mais comumente usadas na pedometria, para o desenvolvimento de mapas preditivos ou digitais do solo (HEUNG et al., 2016)

Modelos de classificação vêm sendo cada dia mais estudados no campo da AM. Os algoritmos mais comuns usados com sucesso para os problemas de classificação e predição, em estudos de pesquisa, são algoritmos baseados em lógica, como árvores de decisão (MASTROGIANNIS; BOUTSINAS; GIANNIKOS, 2009). Dentre estes algoritmos podemos citar ID3- Iterative Dichotomiser 3 (QUINLAN, 1986), C4.5 (QUINLAN, 1986) e o Ctree - Conditional Inference Trees (TORSTEN HOTHORN; KURT HORNIK; ACHIM ZEILEIS., 2009)

Portanto, pode-se questionar: De olho na já realidade “Agricultura digital”, quais são as vantagens de se utilizar AM na classificação de ambientes de produção de cana-de-açúcar?

A hipótese deste trabalho é que, a AM é capaz de classificar ambientes de produção, em áreas de cultivo de cana-de-açúcar, utilizando quantidade reduzida de variáveis. Sendo assim, o objetivo deste trabalho é selecionar variáveis pela técnica de regressão “stepwise” para serem utilizadas no algoritmo de Arvore de Decisão na determinação dos ambientes de produção da cana-de-açúcar.

2 Revisão Bibliográfica

2.1 Cana de Açúcar

A Cana-de-Açúcar apresenta como provável centro de origem a Polinésia, porém os primeiros registros da cultura foram feitos na Nova Guiné em tempos remotos, por volta de 7000 AC, onde a planta foi observada nas mais diversas pinturas e ilustrações. Junto a migração humana a cana se espalhou por todo sudeste asiático (AMARAL; SANTOS; BARBOSA, 2015)

Já no Brasil esta cultura chegou por volta do ano 1500 ou seja logo após o descobrimento do país, a introdução desta cultura objetivou acabar com o monopólio francês em relação ao suprimento mundial de açúcar que era produzido nas colônias caribenhas. Na segunda metade do século XVI engenhos começaram a espalhar pelo nordeste brasileiro em seguida expandiram por toda a região norte porem nesta os engenhos focavam a produção de cachaça. Esta expansão ocasionou intensa colonização do novo país bem como uma excelente fonte lucrativa foi neste período

que os estados de São Paulo e Rio de Janeiro firmaram-se como polos fornecedores para as regiões Sul e Sudeste (MARIN, 2014).

Atualmente a cana-de-açúcar é uma das commodities mais relevantes na economia mundial, sendo a terceira mais produzida. O continente americano detém 56% da produção mundial sendo o Brasil responsável por mais de 45% deste montante (FAO, 2019). De acordo com dados do IBGE, (2019) dentre as grandes regiões brasileiras o Sudeste por 85% na produção nacional da cultura, dando ainda destaque ao estado de São Paulo sendo responsável por 90% da produção da região ganhando assim o título de maior produtor nacional da cultura.

De acordo com (MARAFAON, 2012) a cultura apresenta seu desenvolvimento em formato de touceira podendo estas serem compostas por colmos eretos, semi-erectos e decumbentes, esta variabilidade é influenciada pelo material genético em uso. O colmo é a principal composição da parte aérea e este é em nós e entrenós nestes são localizados de forma alternada a inserção foliar, as folhas apresentam bainha, colar, lâmina foliar o que as torna completas e apresentam lamina foliar alongada (SCARPARI; BEAUCLAIR, 2008).

Além do açúcar e do etanol o Brasil detém a maior capacidade instalada para geração de eletricidade a partir de biomassa, sendo que o bagaço de cana representa 80% da matéria prima utilizada nesse setor (TROMBETA; FILHO, 2017)

Segundo Aude, (1993) a cana-de-açúcar é uma cultura semi-perene e pode ser colhida múltiplas vezes, sendo em média colhida de 5 a 8 vezes antes de ser realizado o replantio. O plantio é comumente realizado de forma mecanizada, são utilizados pedaços de colmos tendo aproximadamente 20cm de comprimento sendo conhecidos como toletes.

Cada tolete deve apresentar em sua formação uma gema para que estas possam produzir perfilhos e a partir desses possam crescer e formar novos colmos. Após cada colheita, as partes restantes da planta, chamadas soqueiras e produzem novos colmos, para que facilite a distinção entre os ciclos distinção, o primeiro ciclo da cultura é chamado de cana planta enquanto os demais são chamadas de cana-soca (BOCCA, 2018).

O conhecimento do ciclo da cultura da cana-de-açúcar e dos padrões de crescimento e desenvolvimento é de extrema importância para que seja executado

com eficiência o manejo, pois se sabe que toda e qualquer produção vegetal que tenha em vista a máxima produtividade econômica, fundamenta-se na interação de três fatores: a planta, o ambiente de produção e o manejo (ALMEIDA et al., 2008). Neste sentido, de acordo com Marafon, (2012), o processo canavieiro visa três objetivos básicos: (1) produtividade; (2) qualidade da matéria prima e (3) longevidade do canavial.

2.2 Ambientes de Produção

O Brasil é um dos países que mais detém tradição quando se trata de produção de cana-de-açúcar e devido a sua grande extensão territorial, a cultura encontra-se em vários tipos de solos e em diferentes condições climáticas o que resulta nos mais diferentes ambientes para produção (DIAS et al., 2014). Esta cultura pode ser influenciada por diversos fatores que interferem diretamente no seu cultivo e maturação, pode-se citar interação edafoclimática, o manejo da cultura e a cultivar escolhida (MAULE et al., 2001).

Em levantamentos pedológicos desenvolvidos para produtores rurais até o ano de 1993, os pesquisadores buscavam classificavam os tipos de solos das propriedades e a partir desta informação determinar seu potencial produtivo e recomenda o manejo adequado para cada tipo de solo (BARBOSA, 2017).

Nos anos subsequentes, pesquisadores divulgaram no seminário de tecnologia Agrônômica da Copersucar o conceito de ambientes de produção para a cultura da cana-de-açúcar. Estes pesquisadores foram os pioneiros na associação dos conhecimentos de pedologia e produtividade da cana-de-açúcar. O qual é embasado nas condições químicas da subsuperfície dos solos e nas diferentes texturas, e a partir destas informações os ambientes de produção foram classificados em A, B, C, D e E (JOAQUIM et al., 1997)

Os pesquisadores verificaram que as produtividades de cana-de-açúcar decresceram dos solos de maior para os de menor potencial de fertilidade, na seguinte ordem: eutróficos > distróficos > ácricos > álicos (SORIA, 2014). Com a definição das classes de potencial de produção, denominados ambientes de produção (A, B, C, D e E) foi também possível determinar as faixas de produtividade de cana-de-açúcar para cada ambiente (Tabela 01).

Devido a relevância deste conceito, estas classes são constantemente revisadas em sucessivos trabalhos como os de PRADO, (2011) e HÉLIO DO PRADO, (2016), cruzando estas informações com os níveis de produtividade da cana-de-açúcar para cada classe de solo cruzando as informações do banco de dados agrônômicos do Centro de Tecnologia Canavieira (CTC), composto por uma série histórica de dez anos de informações, este cruzamento possibilitou relacionar diferentes solos com as respectivas produtividades da cana-de-açúcar em todo o Estado de São Paulo (SILVA; RODRIGUES ALVES; FREITAS, 2015).

Tabela 1 - Potencial produtivo de cana-de-açúcar em toneladas de colmo por hectare em cada ambiente de produção.

Ambiente de produção	Potencial de produtivo	Tonelada de cana por ha
A	Alto	≥ 95
B	Médio/Alto	90 - 95
C	Médio	85 - 90
D	Médio/Baixo	80 - 85
E	Baixo	≤ 80

(JOAQUIM et al., 1997)

Após este estudo inicial, diversos outros foram desenvolvidos podendo citar (DEMATTE; DEMATTÊ, 2009), (MAZZA, 2011) e (MONTEIRO, 2012). O pesquisador (HÉLIO DO PRADO, 2016) de forma inédita, associou as condições químicas e pedológicas do solo a condições climáticas locais nos estudos de ambiente de produção e a partir do ano de 2015, o Centro de Tecnologia Canavieira (CTC) passou a considerar a influência do clima nos ambientes de produção, empregando o conceito de ambientes edafoclimáticos.

De acordo com a (EMBRAPA, 2018) os ambientes de produção são fundamentais para o planejamento da cultura da cana-de-açúcar. Assim, os ambientes de produção são definidos em função dos atributos dos solos relacionando-os com os seus respectivos potenciais produtivos. Portanto, a aplicação de tecnologia associada à variabilidade espacial e temporal faz-se necessária para os ambientes de produção, sobretudo nesta cultura para correlacionar o tipo de solo e a sua capacidade produtiva.

2.3 Aprendizagem de Máquina

A aprendizagem de máquina é uma vertente contida dentro da área de inteligência artificial, esta é focada principalmente em construir sistemas que sejam capazes de aprender de forma automatizada a partir do banco de dados. Este aprendizado é baseado em diversos treinos realizados ao longo da construção do modelo, tal capacidade de aprender se baseando em experiências passadas é algo desejado desde os primórdios da computação (BATISTA, 2019).

O aprendizado de máquina, para Stephen Marsland, (2015) , consiste na capacidade de computadores de modificar e ou adaptar suas ações para assim aumentar a precisão de suas atividades, já os pesquisadores Luxburg; Schölkopf, (2011) afirmam que a aprendizagem de máquina nada mais é que um processo de extração de regras generalizadas a cada banco de dado ou seja através da observação de exemplos um máquina ou computador aprende tarefas específicas seguindo algoritmos de aprendizagem, à ela são revelados exemplos particulares desta tarefa e o objetivo é inferir uma regra geral que pode tanto explicar os exemplos já vistos quanto generalizar para novos exemplos nunca antes vistos (BATISTA, 2019)

Os benefícios que uma máquina pudesse aprender como um humano seria imensurável: sistemas capazes de realizar diagnósticos com base em históricos médicos, sistemas que otimizassem o uso de energia elétrica em uma residência, de acordo com o modo em que seus ocupantes usam, e muitas outras.

Porém por hora a máquina ainda não possui a capacidade de aprendizado humano mesmo está evoluindo a passos significativos. Diversos algoritmos funcionam de forma eficiente em certas tarefas de aprendizado e a agronomia é uma das áreas que mais se beneficiam com a utilização do aprendizado de máquina (BRINK; RICHARDS; FETHEROLF, 2016).

De forma ampla pode-se dividir a AM em duas categorias aprendizado supervisionado e aprendizado não-supervisionado. Enquanto o aprendizado supervisionado, consiste em utilizar uma serie de exemplos já classificados, para induzir um modelo que seja capaz de classificar novas instancias de forma precisa, com base no aprendizado obtido com o conjunto de dados de treinamento variável de interesse.

O aprendizado não-supervisionado objetiva identificar o comportamento dos padrões e estruturas do banco de dados estudado. No aprendizado supervisionado,

quando a variável de interesse é discreta, ou seja é formada categorias ou classes o algoritmo a ser utilizado é o de classificação e se a variável de interesse for continua é utilizada a regressão (STEPHEN MARSLAND, 2015).

2.4 Árvore de Decisão (AD)

A árvore de decisão é um dos algoritmos de aprendizagem de máquina mais estudados e aplicados, o que justifica isso é o fato deste suportar diversos formatos de banco de dados, do binário ao categórico. Este algoritmo pode apresentar dois tipos de modelos sendo estes: classificação e regressão, sua estrutura muito se assemelha a uma árvore onde cada nó interno é responsável por um teste na característica que está sendo classificada, os ramos são responsáveis por representar o resultado do teste realizado anteriormente e os nós terminais representam as classes geradas no resultado final (WITTEN; FRANK; HALL, 2011).

Para realizar uma determinada classificação em uma determinada instância, a Árvore de Decisão (AD) é percorrida de cima para baixo, seguindo pelos ramos onde as características das instâncias satisfazem os testes realizados em cada nó até alcançar um nó terminal, que contém a nova classificação da instância (FÜRNKRANZ; GAMBERGER; LAVRAČ, 2012).

3 Material e Métodos

3.1 Descrição da área

A área objeto da pesquisa é localizada no nordeste do estado de São Paulo no município de Guataporá (figura 01) presente nas coordenadas de 21°28'45"S e 48°01'01"O e altitude média de 590 m acima do nível do mar, com extensão de 870 ha de lavoura comercial de cana-de-açúcar. O clima local de acordo com a classificação de Thornthwaite, (1948), é do tipo B1rB'4a', apresentando as seguintes características: Mesotérmico Úmido, com déficit hídrico em determinadas épocas do ano e evapotranspiração de verão menor que 48% da evapotranspiração anual.

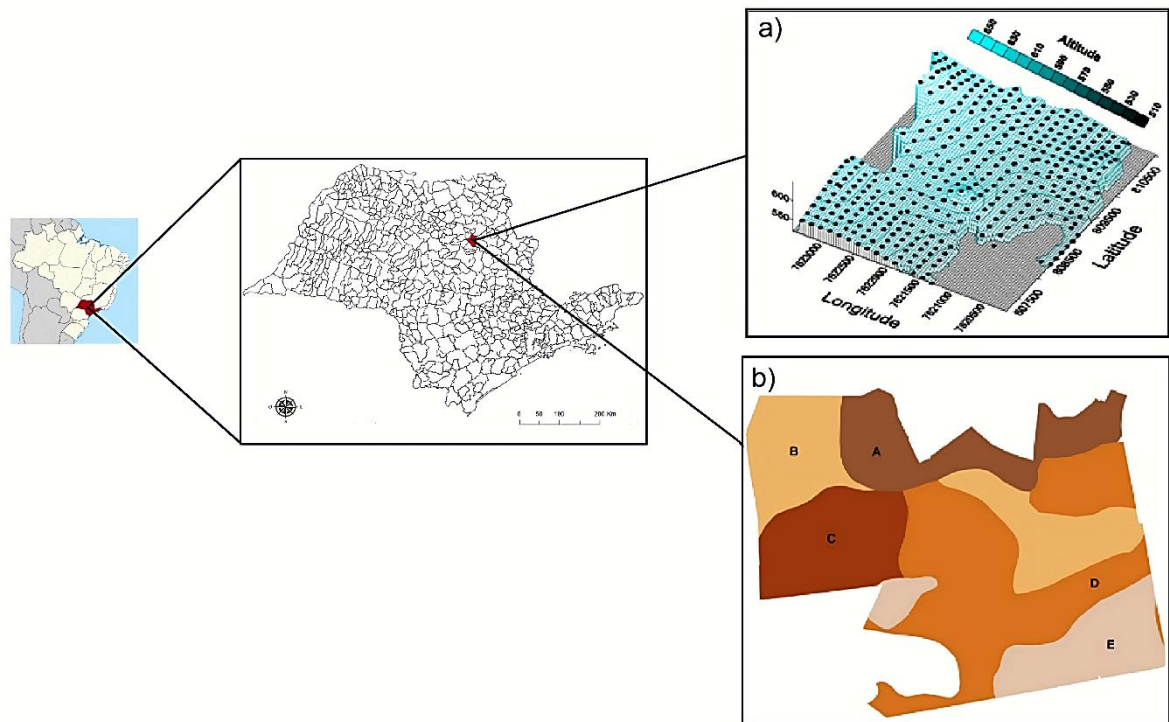


Figura 1- Mapa de localização da área de estudo, a) mapa 3D demonstrativo da altimetria e grid dos pontos e b) Mapa de distribuição dos ambientes de produção desenvolvido pela CTC.

Esta área foi selecionada pelo grupo de pesquisa CSME onde já foi realizado diversos estudos, ela pertence a província geomorfológica do Planalto Ocidental Paulista na proximidade do limite das “Cuestas Basálticas”, que nada mais são do que formações de relevo compostas por remanescentes de rochas vulcânicas como o basalto que se sobrepõe a um empilhamento de rochas de origem sedimentar como o arenito no estado de São Paulo este acontecimento está diretamente ligado ao surgimentos do relevo conhecido como Formações Serra Geral (CARNEIRO, 2018).

Este fenômeno é facilmente observado na área de estudo pois ao consultar a geologia local na plataforma Webmap (USP, 2019), nota-se que as áreas de maior altitude são as que apresentam maior quantidade de basalto e essa quantidade reduz a medida que a altitude diminui.

Para que fosse realizada a coleta de solo no local foi determinado previamente uma malha amostral de um ponto a cada 2,6ha, perfazendo a área total de 870 há, a cada ponto foram coletadas amostras de solos correspondentes a profundidade 0 –

0,25 m. A fazenda foco do estudo já apresenta um mapa de delimitação de ambiente de produção desenvolvido pela CTC- Centro de Tecnologia Canavieira.

O banco de dados desenvolvido com base nas amostras coletadas na área é composto pelo total de trinta variáveis, com base em critérios agronômicos foi realizada a seleção de seis deste total, três destas estão diretamente relacionadas a fatores de formação do solo o que as torna pouco mutáveis ao longo do tempo as três restantes estão ligadas ao manejo do solo o que a torna muito mutáveis em um pequeno espaço de tempo.

A primeira variável ligada a fatores de formação selecionada foi a porcentagem de areia. A figura 02 que corresponde a distribuição da areia pelo local nota-se que existem dois extremos, solos muito arenosos nas áreas de menor altitude e solos pouco arenosos em maiores altitudes, esta variável foi obtida pelo método estabelecido pela EMBRAPA, (2011).

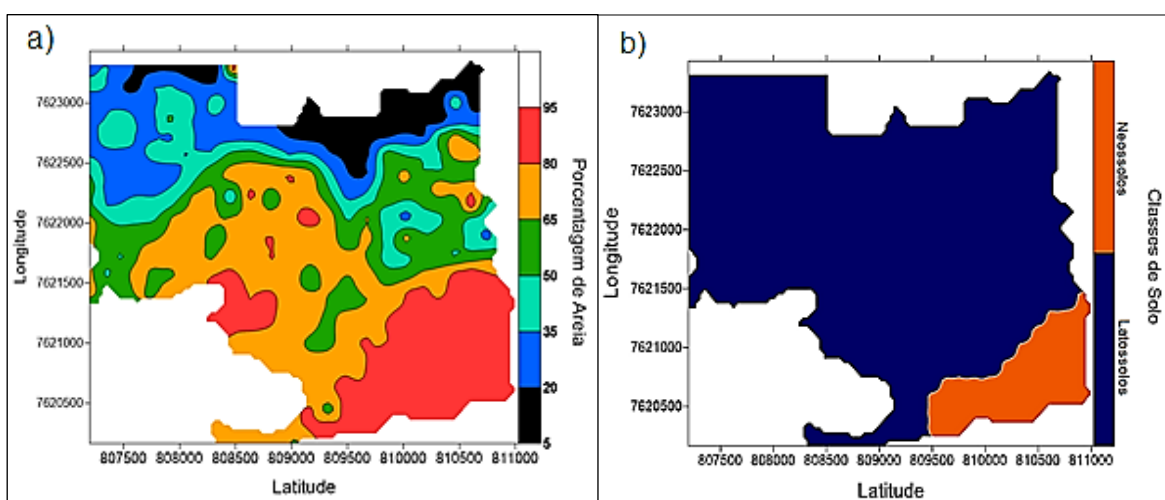


Figura 2- Mapa de distribuição dos atributos: a) mapa de ocorrência de areia na área de estudo. b) Mapa de classes de solo primeiro nível presentes na área.

A segunda variável selecionada foi a classe de solo. Foi registrada a ocorrência de quatro tipos diferentes de solos sendo estes : Latossolo Vermelho Amarelo distrófico textura média (LVAd3); Latossolo Vermelho distrófico textura média (LVd3); Latossolo Vermelho distroférico textura argilosa (LVdf1); Latossolo Vermelho eutroférico textura argilosa (LVef1); e Neossolo Quartzarênico órtico distrófico textura

arenosa (RQod4), o mapa pedológico da usina foi desenvolvido pelo Centro de Tecnologia Canavieira -CTC.

Para esse estudo foi utilizado apenas o primeiro nível de classificação do solo logo estes solos foram agrupados em Latossolos e Neossolos, tal medida foi tomada para reduzir a quantidade de informações necessárias para a estimativa dos ambientes.

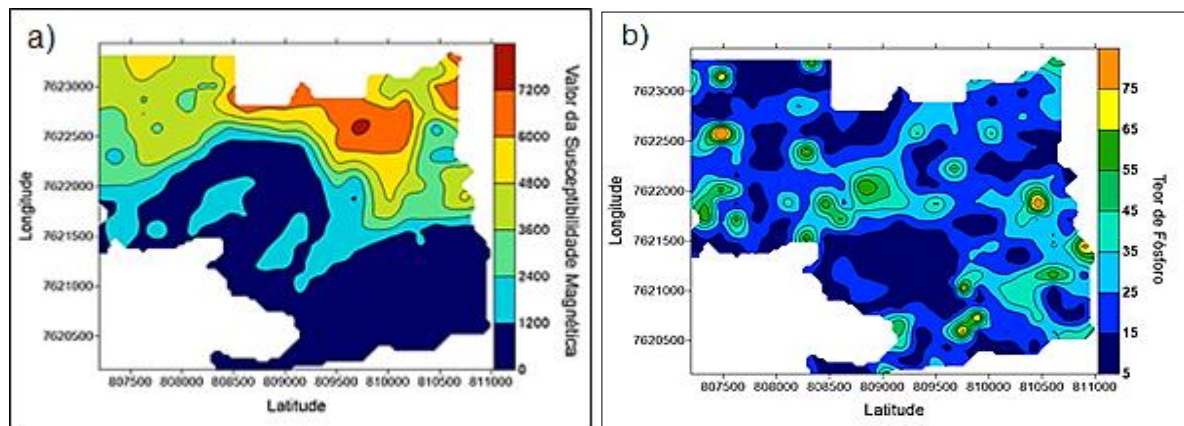


Figura 3- Mapa de distribuição dos atributos, a) mapa dos teores de susceptibilidade magnética ocorrida na área, d) Mapa de ocorrência do teor fósforo na área de estudo.

A terceira variável ligada a formação do solo selecionada foi a susceptibilidade magnética, na figura 04 nota-se que está também se mostra muito mutável na área, maior em altitudes elevadas e menor em altitudes reduzidas, essa variável foi determinada em 10 g de terra fina seca ao ar utilizando o equipamento Bartington MS2, acoplado ao sensor Bartington MS2B em baixa frequência (0,47 kHz) (Dearing, 1994).

Dentre as variáveis químicas, a primeira selecionada foi o fósforo (P). Na figura 03 nota-se que este é muito variável na área porem não apresenta nenhum padrão de variabilidade. A segunda foi o pH está se mostrou bastante uniforme no local e a última variável selecionada foi a CTC potencial (T) que mostrou variabilidade em função da altitude, ou seja, nas áreas de menor altitude menor foi a CTC e esta cresceu em função do aumento da altitude, estas três variáveis foram determinadas com base na metodologia de (RAIJ et al., 2001).

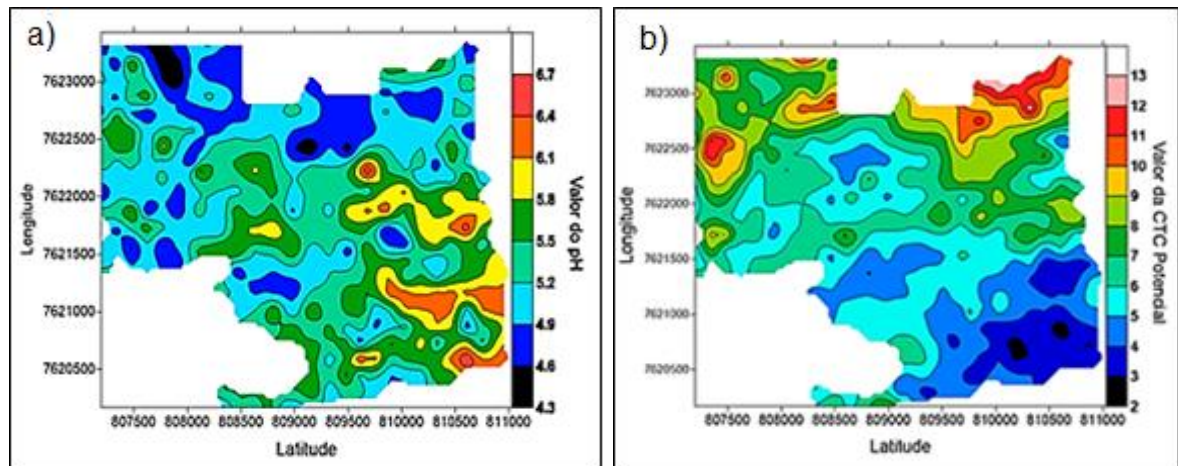


Figura 4-Mapa de distribuição dos atributos, a) mapa dos valores de pH presentes na área, b) mapa de variação de CTC potencial da área.

3.2 Análises Estatísticas.

3.2.1 Estatística Descritiva.

Para cada variável quantitativa foi calculada as estatísticas descritivas, elaborando media, mediana, quartis e foi elaborado um gráfico “violino” para cada variável que foi alocada em função do ambiente de produção para facilitar a compreensão dos dados. Em seguida, os dados foram submetidos à diversas etapas de processamento para melhorar a qualidade da predição (Figura 5).

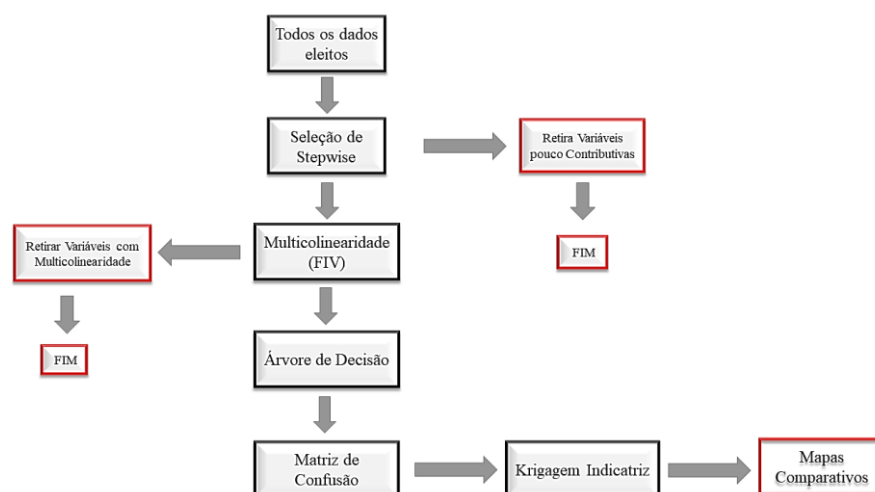


Figura 5- Fluxograma organizacional das análises realizadas.

3.2.2 Seleção pela Regressão “Stepwise”.

Na segunda etapa da análise os dados selecionados foram convertidos em variáveis dummy que nada mais é que uma transformação binária. Esta foi aplicada devido a presença de variáveis categóricas e em seguida submetidas à regressão stepwise(ZARE; FALLAH SHAMSI; ABTAHI, 2019) com o intuito de concluir quais dados são mais relevantes para a análise, o método foi feito de forma gradual e interativa, adicionando (forward) e removendo variáveis (backward), a partir do critério de seleção utilizado que foi o Akaike (AIC) (AKAIKE, 1998). Todos os resultados foram obtidos utilizando o pacote o “MASS” (VENABLES E RIPLEY, 2002) no Rstudio (RSTUDIO TEAM, 2016).

Quando se trabalha com variáveis independentes pode acontecer de alguma delas pouco influenciar no conjunto de variáveis dependentes, que são os dados de saída. A regressão de stepwise tem o intuito de determinar quais variáveis mais influenciam no conjunto de saída, reduzindo assim a quantidade de variáveis que irão compor o modelo, tornando-o mais eficiente.

3.2.3 Multicolinearidade

Quando se trabalha com muitas variáveis preditoras é de extrema importância verificar se estas apresentam relação entre si, isso é chamado de colinearidade. Porém, quando esta se mostra muito alta ou envolvendo três ou mais variáveis constitui redundância entre as variáveis preditoras.

A presença de multicolinearidade (DEMIRHAN, 2014) é prejudicial ao modelo pois pode ocasionar instabilidade, gerando dados errôneos e pouco confiáveis. A presença desta pode ser avaliada pelo cálculo de uma pontuação chamada fator de inflação de variância (FIV), que mede quanto a variância de um coeficiente de regressão é inflada devido à multicolinearidade no modelo. Foi utilizado o RStudio (RSTUDIO TEAM, 2016) para a realização dos cálculos e o pacote utilizado foi o “fmsb”(MINATO NAKAZAWA, 2007).

3.2.4 Árvore de Decisão

Na presente análise o algoritmo utilizado foi o Ctree de classificação, utilizando a interferência condicional como método de partição recursiva dos nós (HOTHORN; HORNIK; ZEILEIS, 2006). Foi separado 70% do banco de dados para treino e os 30% restantes para teste. Como nas demais análises, foi utilizado o RStudio (RSTUDIO TEAM, 2016) e o pacote utilizado para realizar a árvore foi o “party”(STROBL; MALLEY; TUTZ, 2009).

3.2.5 Matriz de Confusão

Para a validação do modelo foi realizado uma matriz de confusão (CAELEN, 2017), esta análise oferece uma medida efetiva do modelo de classificação, ao mostrar o número de classificações corretas versus as classificações preditas para cada classe. Para a elaboração foi utilizado o pacote “caret”(RSTUDIO TEAM, 2016), a partir da matriz gerada foram extraídos os valores verdadeiros positivo (VP) que ocorre quando o modelo prevê um caso positivo corretamente, falsos positivos (FP) são aqueles casos que o modelo classifica como um determinado ambiente porém na verdade não é, falso negativo (FN) quando o modelo indica que não é determinado ambiente mas na verdade é e por fim o verdadeiro negativo (VN) que é quando o modelo diz que não é determinado ambiente e ele acerta (HOTHORN et al., 2005). Após a determinação destas métricas iniciais foram calculadas outras seis com o objetivo de constatar a eficiência do modelo.

$$a) \text{ Valor preditivo positivo} = \frac{VP}{(VP + FP)} \quad (\text{eq.1})$$

$$b) \text{ Valor preditivo negativo} = \frac{VN}{(VN + FN)} \quad (\text{eq.2})$$

$$c) \text{ Acuracia} = \frac{(VN+VP)}{\text{Total de dados do conjunto}} \quad (\text{eq.3})$$

$$d) \text{ Sensibilidade} = \frac{VP}{(VP + FN)} \quad (\text{eq.4})$$

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

$$f) \text{Coeficiente } (\phi) = \frac{(VP*VN - FP*FN)}{\sqrt{((VP + FP)*(VP + FN)*(VN + FP)*(VN + FN))}} \quad (\text{eq.6})$$

3.2.6 Krigagem Indicadora

A krigagem indicadora (KI) (JOURNEL, 1980) foi o método geoestatístico utilizado para a análise da variabilidade espacial dos dados, esta foi selecionada devido a sua premissa de interpolação de variáveis categóricas. Estas variáveis são submetidas a uma transformação não linear e são convertidas em variáveis binárias (eq.7), ou seja, para cada ambiente foi dado uma coluna nesta foi incluído o valor 1 quando este ambiente era igual ao da coluna e 0 quando este era diferente em relação a coluna.

$$i) I_A(x) = \begin{cases} 1, & \text{se } x \in A, \\ 0, & \text{se } x \notin A, \end{cases} \quad (\text{eq.7})$$

Este tipo de aplicação da KI exige um variograma para cada tipo de variável categórica, a função do variograma (eq.8) é calculado pela média da diferença entre pontos separados por uma distância h , esta pode ser utilizada para derivar a função de distribuição acumulativa condicional e assim estimar a probabilidade do teor do ponto estimado ser menor que o valor de corte. Por fim foi utilizado a KI (eq.9) como método de interpolação aplicado para a estimativa dos valores de locais não amostrados.

$$h) \gamma(h, v_c) = \frac{1}{2N(h)} \sum_{i=1}^{N(h)} \{ [Z(x_i + h, v_c) - Z(x_i, v_c)]^2 \} \quad (\text{eq.8})$$

$$i) I^*(Z) = \sum_{i=1}^n \lambda_i 1_Z(x_i) < z \quad (\text{eq.9})$$

Foram gerados o total de dez mapas cinco para dados observados e cinco para dados preditos sendo um para cada ambiente, por fim foi feito uma intersecção, onde foi plotado no mapa final o pixel referente ao ambiente de maior probabilidade de ocorrência no local em questão, unindo os cinco mapas compostos pelos dados observados e os cinco compostos pelos dados preditos compondo um único mapa para cada categoria. Para a condução das análises foi utilizado a RStudio (RSTUDIO TEAM, 2016) para a realização dos cálculos e o pacote utilizado foi o “gstat”(PEBESMA; GRAELER, 2017).

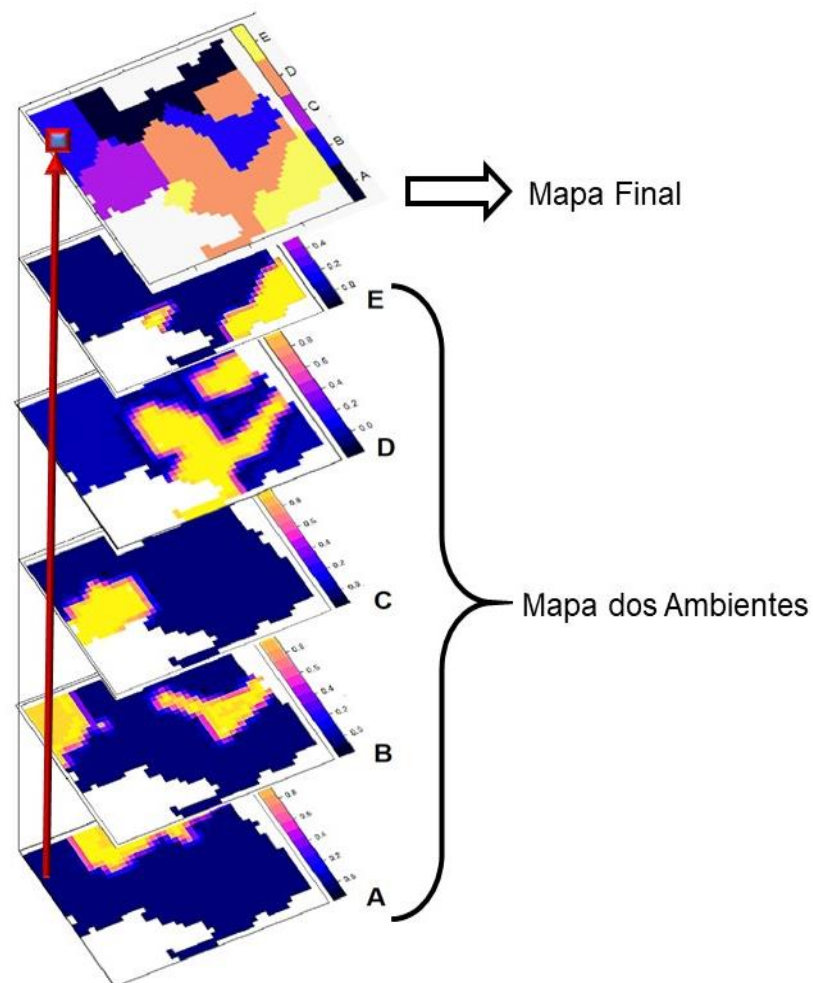


Figura 6- – Esquema resultado explicativo da intersecção dos cinco mapas probabilísticos dos ambientes de produção resultando no mapa final.

4 Resultados e Discussão

Análise exploratória dos dados (Tabela 02) associada a uma ferramenta gráfica (Tabela 07) se mostraram muito eficientes para sumarizar e simplificar o comportamento dos dados em cada ambiente de produção (Figura 07). A primeira variável analisada foi o fósforo lábil (P) e é possível notar que nos cinco ambientes avaliados, o fósforo apresentou comportamento semelhante nos ambientes o que mais diferiu foi o E, apresentando mediana maior que nos demais ambientes.

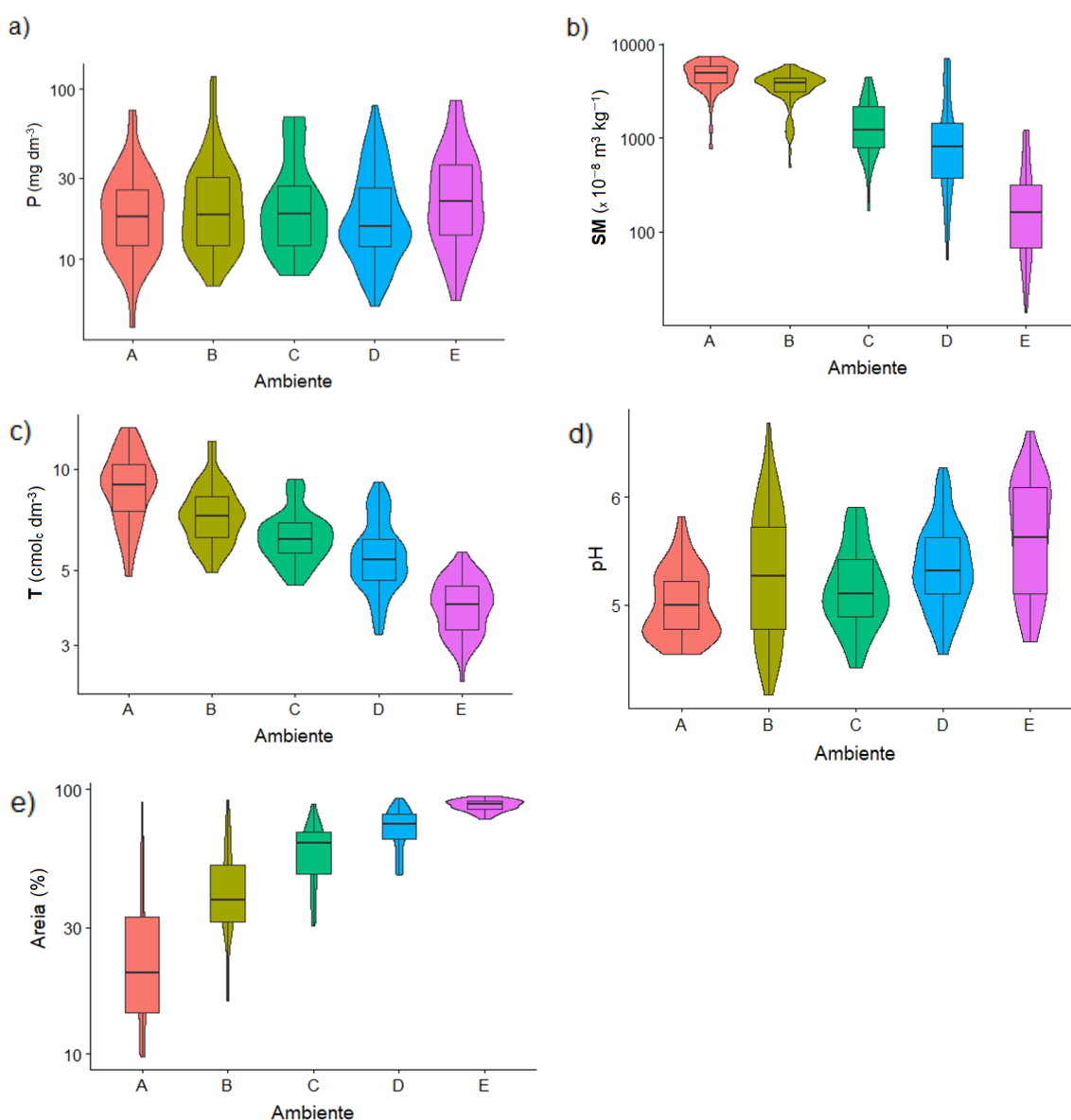


Figura 7 - Gráficos violino representando a variabilidade das variáveis estudadas em função dos ambientes de produção. a) Teor de Fósforo nos ambientes; b) Valor da Susceptibilidade Magnética em cada ambiente, c) Valor da CTC potencial nos ambientes, d) Valor de pH em cada ambiente, e) Porcentagem de Areia nos ambientes.

Tabela 2 -Estatística descritiva associada aos gráficos box-plot.

		AMB.A	AMB.B	AMB.C	AMB.D	AMB.E
AREIA	Máximo	89.71	90.82	88.03	92.70	94.51
	Q3	32.97	52.00	68.99	80.84	90.69
	Mediana	20.21	38.23	62.61	73.91	87.67
	Q1	14.30	31.65	47.98	63.73	83.09
	Mínimo	9.78	12.31	30.52	13.62	14.91
SM	Máximo	7458.00	6151.00	4471.00	7092.00	1217.00
	Q3	5864.25	4413.00	2170.25	1430.75	315.00
	Mediana	4891.00	3920.00	1224.00	809.00	160.00
	Q1	3864.25	3090.00	788.00	373.25	67.75
	Mínimo	780.00	492.00	168.00	51.00	14.00
T	Máximo	13.29	15.26	9.90	9.14	19.94
	Q3	10.31	8.30	6.90	6.18	4.59
	Mediana	8.98	7.27	6.18	5.39	4.01
	Q1	7.50	6.25	5.64	4.69	3.34
	Mínimo	4.80	4.64	3.74	3.23	2.35
PH	Máximo	6.00	6.80	6.40	6.60	6.70
	Q3	5.20	5.70	5.40	5.60	6.10
	Mediana	5.00	5.25	5.10	5.30	5.60
	Q1	4.80	4.80	4.90	5.10	5.10
	Mínimo	4.20	4.30	4.50	4.30	4.70
P	Máximo	97.00	119.00	120.00	80.74	85.68
	Q3	25.69	30.17	27.00	26.24	35.83
	Mediana	17.91	18.24	18.50	15.53	21.80
	Q1	12.00	12.00	12.00	11.87	13.82
	Mínimo	4.00	7.00	8.00	5.28	5.75

O mesmo comportamento foi observado nos valores máximos e mínimos, este aspecto pode ser associado à textura arenosa predominante no solo classificado como ambiente E, devido à menor quantidade de minerais de argila promovendo fixação do P (ZIN et al., 2015).

Na segunda variável analisada, susceptibilidade magnética (SM), nota-se que ocorre um decréscimo nos valores em função dos ambientes, apresentando a diferença de 96% entre a mediana dos ambientes A e E. Além disso, é possível observar que o intervalo interquartil das classes aumenta do ambiente A até o E, mostrando que a medida que a qualidade do ambiente de produção cai a dispersão dos dados de susceptibilidade aumenta.

A variável T e a SM mostrou-se decrescente do ambiente A para o E, estando estes resultados relacionados à variação textural presente nos diferentes

ambientes (MARQUES et al., 2014; TEIXEIRA et al., 2018). A dispersão dos dados se mostrou semelhante em todos os boxplots.

Nota-se que os valores de pH aumentam à medida que a qualidade do ambiente reduz, podendo este resultado estar relacionado às práticas de calagem feitas de forma homogênea na área como um todo, sem levar em consideração a variabilidade da necessidade de calagem nos diferentes ambientes. A quinta e última variável analisada foi a areia, tendo seus valores aumentado do ambiente A para o E. Resultado esperado, uma vez que a textura do solo está diretamente relacionado a fatores como fertilidade e armazenamento de água no solo (CARDENAS; KANAREK, 2014; LI et al., 2019). Outro aspecto analisado foi o intervalo interquartilico (Tabela 2) que reduziu à medida que a quantidade de areia aumentava bem como as caudas.

Após a análise descritiva dos dados foi realizado a regressão de “stepwise”, na tabela 03 é possível observar os modelos desenvolvidos nota-se que a análise discerniu com eficiência atributos atrelados às características do solo ou seja os atributos que são inerentes aos fatores de formação do solo uma espécie de DNA, este se mostraram mais relevantes que os atributos ligados às propriedades do solo ou seja aquele que são manejados anualmente.

As três variáveis mais contributivas ao modelo foram a areia, classe de solo e susceptibilidade magnética todas se mostraram significativas a 5% pelo teste F. Este fato pode ser associado a inerência destes atributos a fatores de formação do solo e do processo de diferenciação de horizontes o que faz com que estes sejam estáveis perante as práticas de manejo dos solos. Resultados semelhantes foram obtidos por Beucher, Møller e GREVE, (2017). Em sua pesquisa, Häring et al, (2012) afirmam que existem várias unidades complexas na elaboração de mapas de solo porem as variáveis geológicas são as mais relevantes e levam a indução de outras.

Já a variável T não foi significativa pelo teste F ($p > 0,05$), contribuindo muito pouco ao modelo. As demais variáveis não apresentaram diferença significativa e por consequência não foram incluídas isso pode ser associado à alta variabilidade no tempo destes atributos pois a cada novo ano estes são alterados devido o manejo. Para auxiliar na determinação de qual modelo é mais eficiente dentre os seis na predição dos ambientes de produção de cana foram determinados quatro métricas para cada modelo.

A primeira foi o coeficiente de determinação (R^2) e esta é a medida mais usual na seleção de modelos, indicando o quão próximo os dados estimados estão dos observados. A segunda medida avaliada foi o erro absoluto médio (MAE), está determina a amplitude média do erro dos dados preditos e todas as diferenças individuais têm o mesmo peso. A terceira medida determinada foi a raiz quadrada do erro médio (RMSE), esta consegue atribuir pesos em função do tamanho do erro (CHAI; DRAXLER, 2014).

Os modelos estimados pela técnica de regressão de “stepwise” com suas respectivas métricas de qualidade do modelo são apresentados na tabela 03. Com base nesta seleção a variável areia se mostra a mais importante a qual sozinha, explica 68,5% da variabilidade dos dados e menor coeficiente de $AIC = -1193.7$ (1º Modelo).

A segunda variável selecionada, dado que a areia já havia sido selecionada, foi a classe de solo aumentando o coeficiente de determinação para $R^2 = 72,4\%$ e a terceira variável selecionada dado que a areia e a classe de solo já foram selecionadas foi a SM elevando o coeficiente $R^2 = 72,8\%$. Nota-se que o incremento no coeficiente R^2 foi pequeno, porém este acréscimo foi significativo.

Tabela 3- Métricas gerada com base em cada modelo desenvolvido pela seleção de stepwise sendo as variáveis representadas por: ambiente de produção (y), areia (X_1), classe de solo (X_2) e SM (X_3).

Modelos	$R^2(\%)$	MAE	RMSE	AIC
1 $y = 0,15 + 0,0115 (X_1)$	68,5	0.14	0.18	-1193.7
2 $y = 0,126 + 0,00993 (X_1) - 0,217 (X_2)$	72,4	0.12	0.17	-1227.0
3 $y = 0,25 + 0,008 (X_1) - 0,215 (X_2) - 0,000019 (X_3)$	72,8	0.12	0.17	-1267.2

Na tabela 4 são expressos os valores gerados a partir do cálculo de fatores de inflação de variância (FIV) e é possível notar que todas as variáveis apresentaram colinearidade moderada. Esse aspecto pode ser concluído também com base na baixa variação do R^2 dos modelos determinados. Hengl; Macmillan, (2009) afirmam que ao trabalhar com dados ambientais como covariáveis para modelagem estatística sempre deve-se atentar a multicolinearidade. As variáveis que se mostraram mais correlacionadas foram a Areia, SM e T e este aspecto pode ser justificado pelo fato

da textura do solo ser diretamente ligado às propriedades magnéticas e à fertilidade natural do solo (MARQUES et al., 2014).

Tabela 4-Valores de multicolinearidade por variável preditora.

Areia	Classe de Solo	SM
4.452	1.470	3.517

Após a avaliação das medidas o 3º modelo foi o selecionado. Em seguida estas variáveis foram submetidas ao algoritmo de AM. O modelo de árvore classificou aleatoriamente o conjunto de dados em 70% dos dados para teste e os 30% restante para validação do modelo.

A árvore criada apresentou 5 nós terminais, que classificaram classificou corretamente 72% dos dados de treinamento. Além disso, mostrou que dentre as três variáveis utilizadas a areia foi a que apresentou maior influência no modelo, de 48%, passando assim a ser a variável mais relevante também na técnica da AM. Isto se deve ao fato de que nessas duas classes de solo (LV e RQ), no Brasil, a textura esteja diretamente correlacionada com a fertilidade e capacidade de retenção de água no solo (NOVAIS et al., 2007). A segunda variável de maior influência foi a susceptibilidade magnética (SM), com 36%, e por fim, a classe de solo com 22% de influência, esta alternância na ordem de importância pode ser associada ao fato de que na análise da AM a classe de solo difere apenas um grupo dos demais que é o ambiente E já a SM possibilita diferenciar três ambientes B, C e D.

Na figura 8 é possível observar o dendograma da árvore. Nota-se que o nó 1 (nó raiz) foi composto pela classe de solo e responsável pela divisão baseado nas duas classes presentes na área, sendo estas, Neossolo quartzarênico (RQ) e Latossolo vermelho (LV). A primeira divisão gerou o segundo nó, terminal, determinando que o RQ foi classificado como ambiente E, esta classe representa solos de menor potencial produtivo.

Resultado semelhante foi encontrado por Beucher, Møller e Greve (2017) onde em seu estudo para avaliar a drenagem do solo observou que sua árvore de decisão separou com exatidão solo com maiores quantidades de areia e solos com maiores quantidade de argila. Isso se dá devido a diferença entre as concentrações de areia nas duas classes, o que faz que elas se diferenciem com facilidade. A classificação

do ambiente E foi atribuída à alta quantidade de areia desses solos o que os torna pobres quimicamente e com baixa capacidade de retenção de água.

A segunda divisão da classe de solos levou os solos do tipo LV ao 3º nó, subdividindo a variável areia, onde a partir deste foram geradas duas novas divisões, sendo que uma delas terminou no 4º nó que determinou que os solos classificados como LV com teor de areia menor ou igual a 26 % devem ser classificados como ambiente de produção A sendo este de maior potencial produtivo. Este aspecto pode ser associado ao fato de que este ambiente detém os solos de melhor qualidade, logo, a baixa porcentagem de areia indica elevada concentração de argila e por consequência solos de melhor fertilidade e maior retenção de umidade.

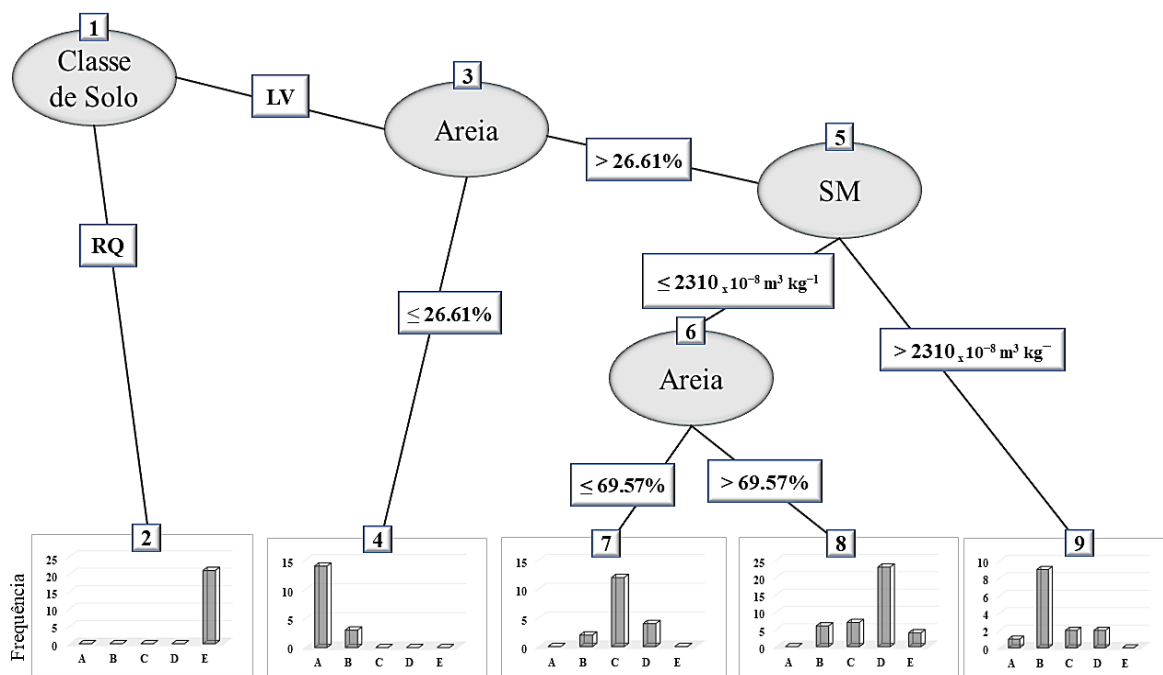


Figura 8- Dendrograma modelo da árvore de decisão com base no modelo 03.

Os minerais ferromagnéticos são a principal fonte de suscetibilidade magnética do solo. A suscetibilidade magnética do solo é afetada por uma ampla gama de fatores. O comportamento magnético do solo é afetado por fatores e processos de formação do solo, estando a textura diretamente relacionada a esta variável (SARMAST; FARPOOR; ESFANDIARPOUR BOROUJENI, 2017).

O 7º nó classifica os dados de classe LV, valores de SM inferiores ou iguais a $2310 \times 10^{-8} \text{ m}^3 \text{ kg}^{-1}$ e com teor de areia superior a 26% e igual ou menor que 69 % de

areia em ambientes C. O nó terminal 8 apresenta solos proveniente do 6º nó com valor de areia maior que 69 %, classificando-o como ambiente D. Voltando ao sub 5º nó, a segunda divisão deste ocasionou em um nó terminal que classificou solos LV, com teores de areia superiores a 26% e valores de SM maiores que $2310 \times 10^{-8} \text{ m}^3 \text{ kg}^{-1}$ em ambientes de produção B.

Nos nós terminais é possível observar a presença de gráficos que demonstram em qual ambiente a maior quantidade de dados foi agrupada, além de possibilitar a visualização dos erros. Nota-se que na classificação do ambiente E não ocorreu erros, ou seja, este ambiente foi determinado com maior precisão. Já o ambiente que apresentou a maior quantidade de erros de predição foi o C, o que faz com que este seja o mais difícil de prever. Isso se dá devido ao fato do conjunto de dados ser desequilibrado, sendo a ocorrência de ambientes C menor, o que dificulta a sua predição (TAGHIZADEH-MEHRJARDI et al., 2014).

Tabela 5- Matriz de confusão do modelo utilizando dados de teste.

		Dados preditos					1
Dados observados		Amb.A	Amb.B	Amb.C	Amb.D	Amb.E	
	Amb.A	14	3	0	0	0	3
	Amb.B	1	9	2	2	0	5
	Amb.C	0	2	12	4	0	6
	Amb.D	0	6	7	23	4	17
	Amb.E	0	0	0	0	21	0
FP		1	11	9	6	4	31

¹FP corresponde a quantidade de falsos positivos,² FN- corresponde a quantidade de falsos negativos

Para comprovar o desempenho do modelo foi desenvolvido uma matriz de confusão com 30% do banco de dados, separado antes do início da análise. Observando a tabela 5, onde em coluna são os dados preditos e na linha são os dados observados, nota-se que dentre os cinco ambientes o que apresentou mais acertos, ou seja, mais positivos verdadeiros (PV) foi o D com 23 predições corretas. Em relação aos falsos positivos (FP), os quais são aqueles que são determinado ambiente porém foram classificados em outro ambiente, neste caso é possível constatar que o ambiente B foi o que obteve a maior quantidade deste erro, já o ambiente D

apresentou uma maior quantidade de falsos negativos (FN) que são aqueles que são D porem foram estimados como outro ambiente.

Foram desenvolvidos seis métricas para determinar a eficiência do modelo (Tabela 6). A primeira foi a sensibilidade, na qual o ambiente A foi o que mais se destacou, apresentando a maior sensibilidade dentre os ambientes, ou seja, o modelo teve capacidade de prever corretamente a condição para os casos que realmente são este ambiente. O ambiente B foi o que apresentou menor valor de sensibilidade ambas as observações podem ser associadas aos valores de falso positivos que foram menores no ambiente A e maior no ambiente B.

A especificidade mede a capacidade do sistema em prever corretamente a ausência da condição para casos que realmente não a tem. O ambiente de maior especificidade foi o E, e o que apresentou menor foi o D, o que é associado a maior presença de falsos negativos no ambiente D e a ausência destes no ambiente E.

Tabela 6- – Métricas qualitativas desenvolvidas com base na matriz de confusão.

	Sensibilidade	Especificidade	Eficiência	PP ¹	PN ²	Coef ϕ ³
Amb.A	93%	96%	94%	82%	98%	0.850
Amb.B	45%	93%	69%	64%	86%	0.573
Amb.C	57%	92%	74%	67%	88%	0.441
Amb.D	79%	77%	78%	58%	90%	0.518
Amb.E	84%	100%	92%	100%	94%	0.886

¹ PP - Proporção de verdadeiros positivos em relação à todas as predições positivas; ²PN - Proporção de verdadeiros negativos em relação à todas as predições negativas; ³Coef ϕ - coeficiente de correlação de Matthews, que resume à qualidade da matriz em um único valor numérico.

A eficiência foi a terceira métrica testada e ela nada mais é do que a média aritmética da sensibilidade e da especificidade, que é utilizada para avaliar o equilíbrio entre essas métricas, pois, se o modelo se mostrar muito sensível a positivos, tende a gerar muitos falso-positivos e vice-versa, fazendo com que elas apresentem direções opostas.

O ambiente de maior eficiência foi o A e o que se mostrou mais propenso a desequilíbrio foi o B, podendo ser associado ao fato deste apresentar mais valores em falsos positivos que em verdadeiros positivos. A proporção de verdadeiros positivos (VP) em relação a todas as predições positivas foi observada com maior frequência no ambiente E, já na proporção de verdadeiros negativos (PN) em relação a todas as predições negativas o ambiente A se mostrou mais assertivo.

A sexta e última métrica avaliada foi o coeficiente de correlação de Matthews, que resume a qualidade da matriz em um único valor numérico, passível de ser comparado, sendo que quanto mais próximo de 1 melhor foi a predição. O ambiente E apresentou melhor porcentagem nesta métrica já o ambiente C foi o que apresentou menor desempenho aspecto este que pode ser associado a baixa quantidade de observações de tal ambiente em relação aos demais. No total, o modelo classificou de forma correta 75% dos dados, valor de acurácia condizente ao encontrado na literatura (JAFARI et al., 2012; SCULL, FRANKLIN e CHADWICK, 2005; TAGHIZADEH-MEHRJARDI et al., 2014)

O desempenho superior na estimativa do ambiente E deve-se à diferença entre as classes de solo RQ e LV, pois neste ambiente entram apenas dados classificados como solo RQ. Nos demais ambientes todos os dados são classificados como LV, então a diferenciação entre eles ocorre somente utilizando as demais variáveis preditoras.

A matriz de confusão fornece uma avaliação global e não espacial dos dados, utilizando uma única estatística, sendo assim a elaboração de mapas interpolados com o auxílio da geoestatísticas que possibilitem realizar um comparativo entre dados preditos e observados é essencial (FOODY, 2005). Sendo assim, foram desenvolvidos o total de dez semivariogramas indicadores dois para cada ambiente um com dados observados outro com dados preditos pela árvore de decisão (Figura 9).

Dos semivariogramas indicadores estimados, todos apresentaram dependência espacial, demonstrando patamares bem definidos. Segundo Souza et al., (2014) semivariogramas com estrutura de dependência espacial de moderada a forte geram mapas de krigagem mais precisos que aqueles gerados com fraca dependência espacial devido à menor contribuição do componente aleatório na variabilidade dos dados.

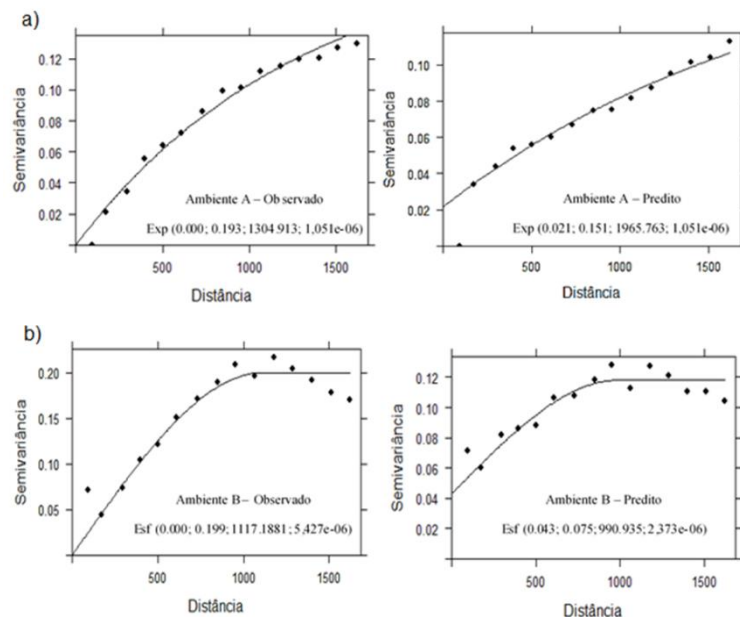
Os dados dos ambientes preditos e observados apresentaram comportamento esférico nos semivariogramas, com exceção dos dados referentes ao ambiente A, que apresentaram o comportamento exponencial. O modelo exponencial é similar ao esférico pelo fato de que ambos atingem o patamar de forma gradual, o que os difere é taxa por meio da qual o patamar é alcançado e pelo fato de que o modelo e o

patamar nunca convergem no modelo exponencial (MOTOMIYA; CORÁ; PEREIRA, 2006).

Além destes aspectos é possível observar que os semivariogramas indicadores dos dados preditos e dos observados apresentaram comportamentos semelhantes. Os que apresentaram maior diferenciação foram os do ambiente C o que pode ser associado ao baixo número de observações deste ambiente o que ocasiona instabilidade nos dados.

Nos semivariogramas gerados com os dados observados nota-se que em todos o valor de C_0 (efeito pepita) equivale a 0 mostrando a ausência de erro, nos semivariogramas dos dados preditos nota-se que assim como observado na matriz de confusão o ambiente E foi o que apresentou menor C_0 ou seja foi o que menos errou e o ambiente D apresentou o maior efeito pepita. Os semivariogramas dos dados preditos dos ambientes B, C e E apresentaram mais instabilidade quando comparado com os dos dados observados.

Por fim para estimar os valores não amostrados utilizando os parâmetros dos semivariogramas foi aplicada a técnica de interpolação krigagem indicatriz que indica a probabilidade de ocorrência dos ambientes ao longo da área. Com resultado desta análise foram elaborados mapas de probabilidade de ocorrência para cada ambiente.



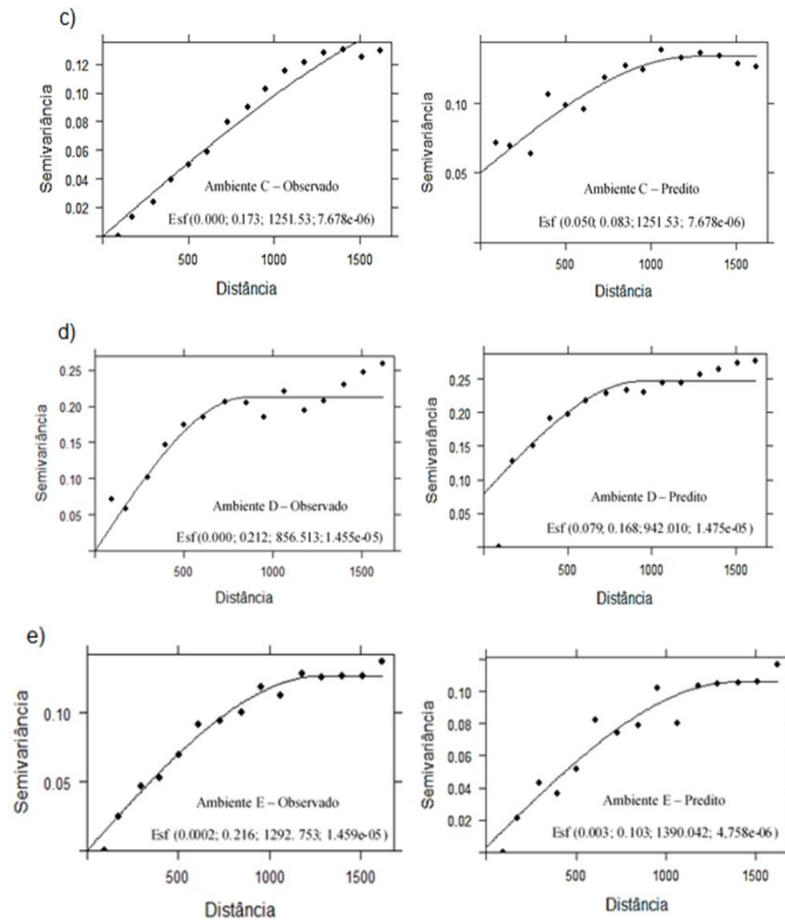


Figura 9-Semivariogramas Indicadores dos ambientes de produção observados e dos preditos pela arvore de decisão (a) Ambiente A predito/observado; (b) Ambiente B predito/observado; (c) Ambiente C predito/observado; (d) Ambiente D predito/observado e (e) Ambiente E predito/observado. * Modelo: exp. = exponencial; esf. = esférico; efeito pepita (C0); alcance (a); patamar (C); Soma dos quadrados dos resíduos.

Foram elaborados dez mapas probabilísticos dos ambientes de produção (figura 10). Os mapas da esquerda nesta (figura 10.a) foram obtidos dos dados observados, enquanto que os da direita (figura 10.b) foram obtidos após a classificação da AD. O ambiente E foi o que apresentou maior semelhança entre os mapas indicando que existe uma maior probabilidade deste ambiente se localiza na área de menor altitude no mapa, o ambiente A também apresentou muita semelhança entre os mapas porem o predito apresentou um ligeiro alongamento a esquerda porém ambos indicaram que existe uma maior probabilidade de ocorrer ambiente A em áreas de maior altitude.

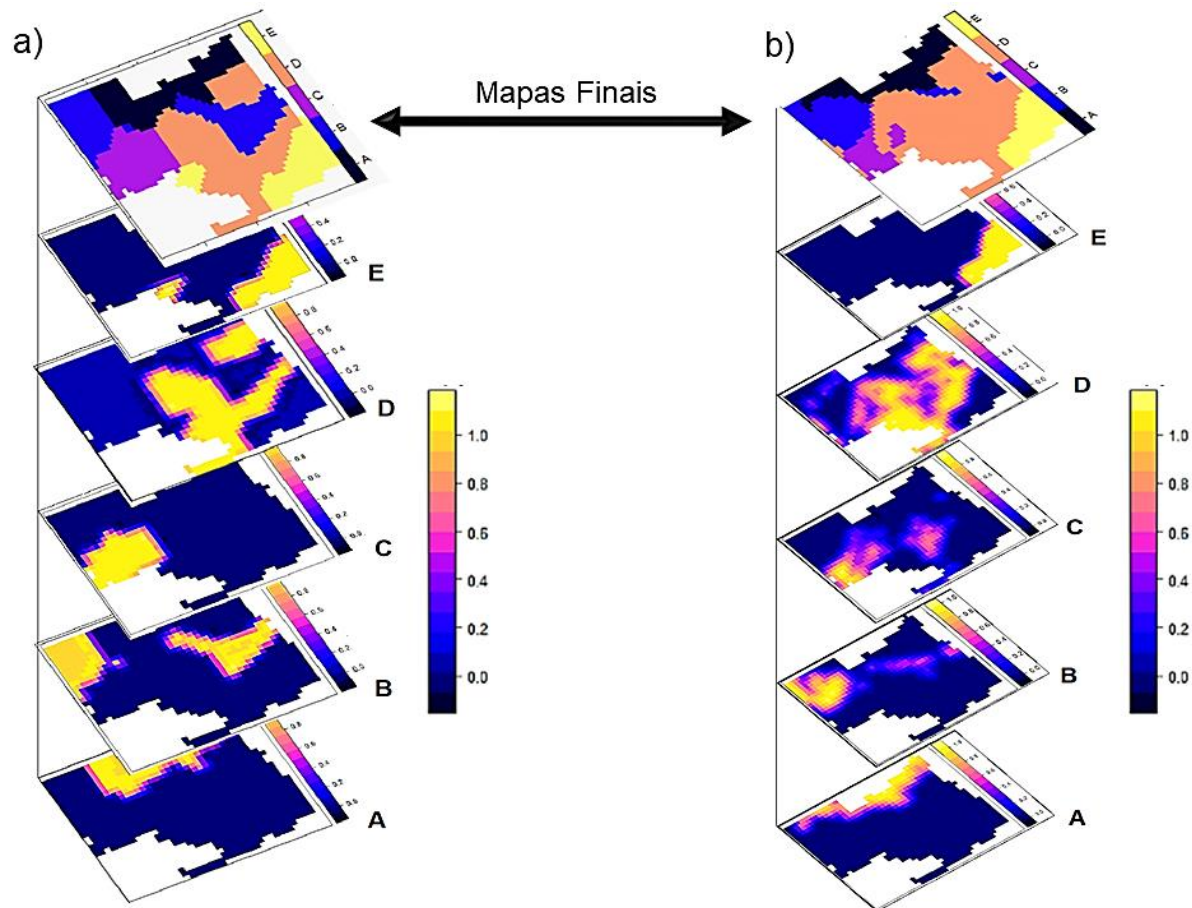


Figura 10- Mapas de probabilidade dos ambientes de produção observados e dos ambientes preditos pela arvore de decisão interpolados com a krigagem indicatriz (a) Ambiente A observado/predito; (b) Ambiente B observado/predito; (c) Ambiente C observado/predito; (d) Ambiente D observado/predito e (e) Ambiente E observado/predito.

Já os mapas que mais divergiram foram o do ambiente C, aspecto esse que já era esperado pois foi o ambiente mais problemático para estimar devido à baixa quantidade de observações no banco de dados. Ao analisar os mapas de uma forma geral nota-se que o que mais divergiu foram os valores de probabilidade dos pixels, ou seja, o mapa de dados observados apresentava um pixel de alta probabilidade de ocorrência do ambiente já no de dados preditos este passou a apresentar uma média probabilidade de ocorrência do ambiente.

Entretanto, nota-se que os mapas não apresentam erro de localização dos ambientes ou seja nos mapas com dados observados o ambiente A se localiza nas áreas de maior altitude e no mapa desenvolvido com dados preditos estes apresentam os ambientes A no mesmo local.

A elaboração destes mapas também possibilitou delinear com mais eficiência as zonas de transição entre ambientes, a redução da probabilidade de determinado ambiente em um local é indicio de uma possível zona de transição entre ambientes além de ter suavizado estas zonas de transição homogeneizando a área facilitando assim o trato cultural.

Por fim, foi elaborado os mapas finais (Figura 11), resultado da álgebra dos cinco mapas gerados, neste foram plotados apenas os pixels de maior probabilidade de ocorrência de cada mapa.

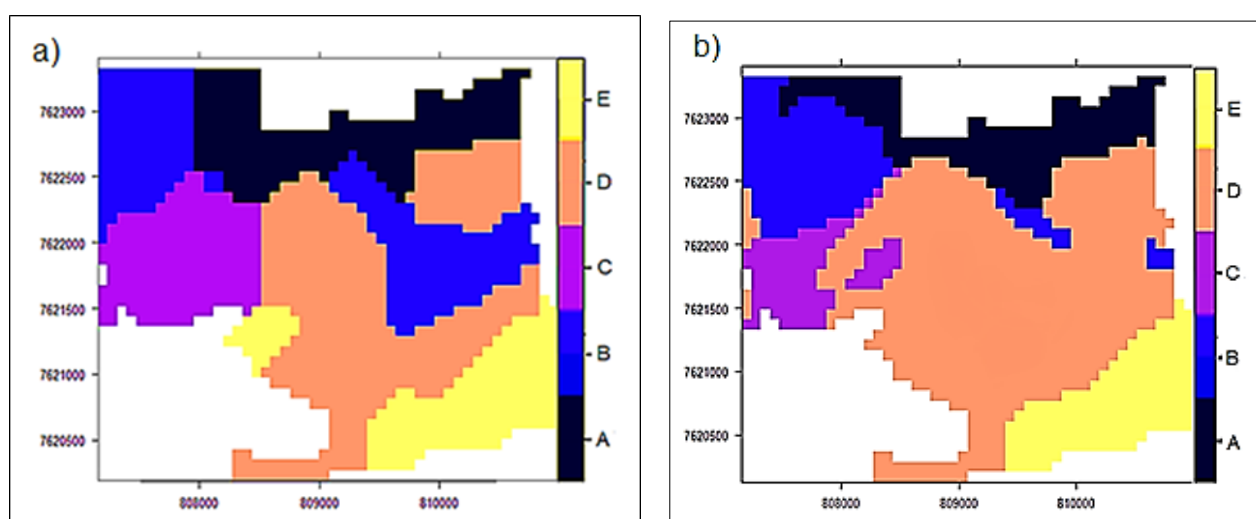


Figura 11- Mapas resultantes da intersecção das probabilidades dos ambientes de produção, a) Ambientes observados e b) Ambientes preditos pela AD.

Ao observar o mapa realizado com dados observados e com dados preditos nota-se que ambos são muito semelhantes, a pouca diferença observada é apresentada no ambiente D onde algumas áreas dos ambientes C e B foram reduzidas porem nota-se que o mapa predito deu mais continuidade a área o que otimiza o manejo, nesse sentido o desbalanço do banco de dados ocasionada pelo número reduzido observações é um fator prejudicial à predição dos ambientes pois os ambientes menos observados foram os mais prejudicados no mapa.

5 Conclusões

É possível prever ambientes de produção de cana-de-açúcar utilizando número reduzido de variáveis associado ao uso de algoritmos de aprendizado de máquina. Não é certo que tal abordagem funcionará em outras áreas, pois para isso seria necessário extrapolar o modelo. Entretanto, pode-se afirmar que somente se embasando em algumas variáveis de fácil acesso pode-se tirar orientações para avaliar cada situação de manejo.

A árvore de decisão, associada à geoestatística mostrou-se uma alternativa eficiente na determinação de ambientes de produção, indicando com a alta acurácia a localização de cada ambiente além de possibilitar a visualização das áreas de transição dos ambientes.

Agradecimentos

Este trabalho é parte das pesquisas do grupo de pesquisa CSME e pelas agências de pesquisa CAPES (Proc. nº 149940); PROPE / UNESP (Pró-reitora de Pesquisa da Universidade Estadual Paulista, Edital Nº 15/2014); e CNPq (Conselho Nacional de Desenvolvimento Científico e Tecnológico, Edital Universal 01/2016 – Faixa C, Proc.nº 402796/2016-0).

6 Referencial bibliográfico

AKAIKE, H. Information Theory and an Extension of the Maximum Likelihood Principle. In: Springer, New York, NY, 1998. p. 199–213.

ALMEIDA, A. C. S. et al. Desenvolvimento vegetativo e produção de variedades. **Ciência Agrotecnica**, v. 32, p. 1441–1448, 2008.

AMARAL, A. L. DO; SANTOS, J. M. DOS; BARBOSA, G. V. DE S. Denominação de Cultivares de Cana-deAçúcar como Suporte à Documentação e Utilização do Germoplasma Conservado. 2015.

AUDE, M. I. DA S. ESTÁDIOS DE DESENVOLVIMENTO DA CANA-DE-AÇÚCAR E SUAS RELAÇÕES COM A PRODUTIVIDADE. **Ciência Rural**, v. 23, n. 2, p. 241–248, ago. 1993.

BARBOSA, A. D. M. AMBIENTE DE PRODUÇÃO NA EFICIÊNCIA DA CONVERSÃO DE ENERGIA SOLAR EM CULTIVARES DE CANA-DE-AÇÚCAR. p. 20, 2017.

BATISTA, M. R. S. A utilização de algoritmos de aprendizado de máquina em problemas de classificação. 2019.

BEKKER, G. F. H. VAN G. et al. Using machine learning to identify the geographical drivers of *Ceratitis capitata* trap catch in an agricultural landscape. **Computers and Electronics in Agriculture**, v. 162, n. August 2018, p. 582–592, 2019.

BEUCHER, A.; MØLLER, A. B.; GREVE, M. H. Artificial neural networks and decision tree classification for predicting soil drainage classes in Denmark. **Geoderma**, n. November, p. 0–1, 2017.

BOCCA, F. F. **Estudo do efeito da adubação nitrogenada na produtividade de cana-de-açúcar com modelos de aprendizado de máquina produtividade de cana-de-açúcar com modelos de.** [s.l: s.n.].

BRINK, H.; RICHARDS, J. W.; FETHEROLF, M. **Real-world machine learning**, 2016.

CAELEN, O. A Bayesian interpretation of the confusion matrix. **Annals of Mathematics and Artificial Intelligence**, v. 81, n. 3–4, p. 429–450, 2017.

CANIZO, B. V. et al. Data mining approach based on chemical composition of grape skin for quality evaluation and traceability prediction of grapes. **Computers and Electronics in Agriculture**, v. 162, n. March, p. 514–522, 2019.

CARDENAS, M. B.; KANAREK, M. R. Soil moisture variation and dynamics across a wildfire burn boundary in a loblolly pine (*Pinus taeda*) forest. **Journal of Hydrology**, v. 519, n. PA, p. 490–502, 2014.

CARNEIRO, C. D. R. The “geological foundations of the relief of the State of São Paulo” nowadays. **Revista do Instituto Geológico**, v. 39, n. 3, p. 1–8, 2018.

CHAI, T.; DRAXLER, R. R. Root mean square error (RMSE) or mean absolute error (MAE)? -Arguments against avoiding RMSE in the literature. **Geoscientific Model Development**, v. 7, n. 3, p. 1247–1250, 2014.

DEMATTE, J. L. I.; DEMATTE, J. A. M. Ambientes de produção como estratégia de manejo na cultura da cana-de-açúcar. **Informações Agrômicas**, n. 127, p. 10–18, 2009.

DEMIRHAN, H. The problem of multicollinearity in horizontal solar radiation estimation models and a new model for Turkey. **Energy Conversion and Management**, v. 84, p. 334–345, 1 ago. 2014.

DIAS, F. L. F. et al. Produtividade da cana-de-açúcar em relação a clima e solos da região noroeste do estado de São Paulo. **Revista Brasileira de Ciência do Solo**, v. 23, n. 3, p. 627–634, 2014.

EMBRAPA. **Manual de Métodos de Análise de Solo**. 2. ed. Rio de Janeiro: Embrapa Solos, 2011.

EMBRAPA. **Ambientes de Produção**. Disponível em: <https://www.agencia.cnptia.embrapa.br/gestor/cana-de-acucar/arvore/CONTAG01_3_711200516715.html>. Acesso em: 30 jun. 2019.

FAO. **Organização das Nações Unidas para a Alimentação e a Agricultura**. Disponível em: <<http://www.fao.org/faostat/en/#data/QC/visualize>>. Acesso em: 27 jun. 2019.

FOODY, G. M. Local characterization of thematic classification accuracy through spatially constrained confusion matrices. **International Journal of Remote Sensing**, v. 26, n. 6, p. 1217–1228, 2005.

FÜRNKRANZ, J.; GAMBERGER, D.; LAVRAČ, N. Rule Learning in a Nutshell. p. 19–55; 2012.

GOLDEMBERG, J. et al. Meeting the global demand for biofuels in 2021 through sustainable land use change policy. **Energy Policy**, v. 69, p. 14–18, 2014.

HÄRING, T. et al. Spatial disaggregation of complex soil map units: A decision-tree based approach in Bavarian forest soils. **Geoderma**, v. 185–186, p. 37–47, 2012.

HÉLIO DO PRADO. **Pedologia Fácil: Aplicação em Solos Tropicais**. 5. ed. Piracicaba: [2016].

HENGL, T.; MACMILLAN, R. A. Geomorphometry-a key to landscape mapping and modelling. **Developments in Soil Science**, v. 33, n. C, p. 433–460, 2009.

HEUNG, B. et al. An overview and comparison of machine-learning techniques for classification purposes in digital soil mapping. **Geoderma**, v. 265, p. 62–77, 2016.

HOTHORN, T. et al. The design and analysis of benchmark experiments. **Journal of Computational and Graphical Statistics**, v. 14, n. 3, p. 675–699, set. 2005.

HOTHORN, T.; HORNIK, K.; ZEILEIS, A. Unbiased recursive partitioning: A conditional inference framework. **Journal of Computational and Graphical Statistics**, v. 15, n. 3, p. 651–674, 2006.

IBGE. **Instituto Brasileiro de Geografia e Estatística - SIDRA**. Disponível em: <<https://sidra.ibge.gov.br/home/lspa/brasil>>. Acesso em: 27 jun. 2019.

JAFARI, A. et al. Spatial prediction of USDA- great soil groups in the arid Zarand region, Iran: Comparing logistic regression approaches to predict diagnostic horizons and soil types. **European Journal of Soil Science**, v. 63, n. 2, p. 284–298, 2012.

JOAQUIM, A. C. . et al. **Potencial de produção de cana-de-açúcar.SEMINÁRIOS COPERSUCAR DE TECNOLOGIA AGRONÔMICA**. 1997.

JOURNEL, A. G. The lognormal approach to predicting local distributions of selective mining unit grades. **Journal of the International Association for Mathematical Geology**, v. 12, n. 4, p. 285–303, ago. 1980.

LI, N. et al. Digital soil mapping based site-specific nutrient management in a sugarcane field in Burdekin. **Geoderma**, v. 340, n. December 2018, p. 38–48, 2019.

LUXBURG, U. VON; SCHÖLKOPF, B. Statistical Learning Theory: Models, Concepts, and Results. **Handbook of the History of Logic**, v. 10, p. 651–706, 1 jan. 2011.

MARAFON, A. C. Análise quantitativa de crescimento em Cana-de-açúcar: Uma introducao ao procedimento práctico. **Embrapa Tabuleiros Costeiros**, v. 168, n. 1, p. 31, 2012.

MARIN, F. R. Universidade de São Paulo Escola Superior de A gricultura “ Luiz de Queiroz ” Eficiência de produção da cana-de-açúcar brasileira : estado atual e cenários futuros baseados em simulações multimodelos Fábio Ricardo Marin Engenheiro Agrônomo Piracicaba – SP. p. 0–262, 2014.

MARQUES, J. et al. Magnetic susceptibility and diffuse reflectance spectroscopy to characterize the spatial variability of soil properties in a brazilian haplustalf. **Geoderma**, v. 219–220, p. 63–71, 2014.

MASTROGIANNIS, N.; BOUTSINAS, B.; GIANNIKOS, I. A method for improving the accuracy of data mining classification algorithms. **Computers and Operations Research**, v. 36, n. 10, p. 2829–2839, 2009.

MAULE, R. F. et al. Produtividade Agrícola De Cultivares De Cana-De-Açúcar Em Diferentes Solos Productivity of Sugarcane Cultivars in Different. **Scientia**, p. 295–301, 2001.

MAZZA. **Manejo de Variedades de cana-de-açúcar nas unidades da Bunge**. Disponível em: <<https://docplayer.com.br/29521602-Manejo-de-variedades-de-cana-de-acucar-nas-unidades-da-bunge.html>>. Acesso em: 30 jun. 2019.

MINATO NAKAZAWA. **Pacote fmsb2018**, , 2007. Disponível em: <<https://cran.r-project.org/web/packages/fmsb/index.html>>. Acesso em: 30 set. 2019

MONTEIRO, L. A. **Modelagem agrometeorológica como base para a definição de ambientes de produção para a cultura da cana-de-açúcar no Estado de São Paulo**. Piracicaba: Biblioteca Digital de Teses e Dissertações da Universidade de São Paulo, 26 jan. 2012.

MORAL, F. J.; TERRÓN, J. M.; SILVA, J. R. M. DA. Delineation of management zones using mobile measurements of soil apparent electrical conductivity and multivariate geostatistical techniques. **Soil and Tillage Research**, v. 106, n. 2, p. 335–343, 2010.

MOTOMIYA, A. V. A.; CORÁ, J. E.; PEREIRA, G. T. Uso da krigagem indicatriz na avaliação de indicadores de fertilidade do solo. **Revista Brasileira de Ciencia do Solo**, v. 30, n. 3, p. 485–496, 2006.

NATARAJAN, R.; SUBRAMANIAN, J.; PAPAGEORGIOU, E. I. Hybrid learning of fuzzy cognitive maps for sugarcane yield classification. **Computers and Electronics in Agriculture**, v. 127, p. 147–157, 2016.

NOVAIS, R. F. et al. **Fertilidade do solo**. Viçosa: [2007].

PEBESMA, E.; GRAELER, B. **Spatial and Spatio-Temporal Geostatistical Modelling, Prediction and Simulation** Cran R Comprehensive R Archive Network (CRAN), , 2017. Disponível em: <<https://cran.r-project.org/web/packages/gstat/index.html>>. Acesso em: 30 set. 2019

PRADO, H. DO. **Pedologia fácil-aplicações na agricultura**. Piracicaba: 2016.

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

RAIJ, B. VAN et al. Análise química para avaliação da fertilidade de solos tropicais. **Campinas: Instituto Agrônômico**, 2001.

RSTUDIO TEAM. **RStudio: Integrated Development for R** RStudio, Inc., Boston, MA., 2016.

SANCHES, G. M. et al. Precision production environments for sugarcane fields. **Scientia Agricola**, v. 76, n. 1, p. 10–17, 2018.

SANCHES, G. M.; MAGALHÃES, P. S. G.; FRANCO, H. C. J. Site-specific assessment of spatial and temporal variability of sugarcane yield related to soil attributes. **Geoderma**, v. 334, n. August 2018, p. 90–98, 2019.

SARMAST, M.; FARPOOR, M. H.; ESFANDIARPOUR BOROUJENI, I. Magnetic susceptibility of soils along a lithotoposequence in southeast Iran. **Catena**, v. 156, p. 252–262, 1 set. 2017.

SCARPARI, M. S.; BEAUCLAIR, E. G. F. DE. Anatomia e Botânica. In: DINARDO-MIRANDA, L. L.; LANDELL, A. C. M. DE V.; ANDRADE, M. G. DE (Eds.). **Cana de Açúcar**. Campinas - SP: IAC, 2008. p. 47–56.

SCULL, P.; FRANKLIN, J.; CHADWICK, O. A. The application of classification tree analysis to soil type prediction in a desert landscape. **Ecological Modelling**, v. 181, n. 1, p. 1–15, 2005.

SILVA, F. C. DA; RODRIGUES ALVES, B. J.; FREITAS, P. L. DE. **Sistema de produção mecanizada da cana-de-açúcar integrada à produção de energia e alimentos**. 2015.

SORIA, J. E. **CORRELAÇÕES LINEARES E ESPACIAIS ENTRE ATRIBUTOS FÍSICO-QUÍMICOS DE SOLOS DE DIFERENTES AMBIENTES DE PRODUÇÃO DE CANA-DE-AÇÚCAR NO NOROESTE DO ESTADO DE SÃO PAULO**. [s.l.] UNIVERSIDADE ESTADUAL PAULISTA, 2014.

SOUZA, Z. M. et al. Número de amostras na análise geoestatística e na krigagem de mapas de atributos do solo. **Ciencia Rural**, v. 44, n. 2, p. 261–268, 2014.

STEPHEN MARS LAND. **MACHINE LEARNING An Algorithmic Perspective**. 2. ed. 2015.

STROBL, C.; MALLEY, J.; TUTZ, G. An Introduction to Recursive Partitioning: Rationale, Application, and Characteristics of Classification and Regression Trees, Bagging, and Random Forests. **Psychological Methods**, v. 14, n. 4, p. 323–348, 2009.

TAGHIZADEH-MEHRJARDI, R. et al. Digital Mapping of Soil Classes Using Decision Tree and Auxiliary Data in the Ardakan Region, Iran. **Arid Land Research and Management**, v. 28, n. 2, p. 147–168, 2014.

TEIXEIRA, D. D. B. et al. Mapping units based on spatial uncertainty of magnetic susceptibility and clay content. **Catena**, v. 164, n. December 2017, p. 79–87, 2018.

THORNTHWAITE, C. W. An Approach toward a Rational Classification of Climate Author (s): C . W . Thornthwaite Published by : American Geographical Society Stable URL : <http://www.jstor.org/stable/210739> Accessed : 11-08-2016 18 : 02 UTC. **Geographical Review**, v. 38, n. 1, p. 55–94, 1948.

TORSTEN HOTHORN; KURT HORNIK;; ACHIM ZEILEIS. ctree: Conditional Inference Trees. **R-cran**, 2009.

TROMBETA, N. DE C.; FILHO, J. V. C. Potencial e disponibilidade de biomassa de cana-de-açúcar na Região Centro-Sul do Brasil: Indicadores agroindustriais. **Revista de Economia e Sociologia Rural**, v. 55, n. 3, p. 479–496, set. 2017.

USP. **GeoHereditas | Webmap: geociências em mapas interativos**. Disponível em: <<http://webmap.igc.usp.br/#>>. Acesso em: 2 set. 2019.

WITTEN, I. H. (IAN H. .; FRANK, E.; HALL, M. A. (MARK A. **Data mining : practical machine learning tools and techniques**. [s.l.] Morgan Kaufmann, 2011.

ZARE, S.; FALLAH SHAMSI, S. R.; ABTAHI, S. A. Weakly-coupled geo-statistical mapping of soil salinity to Stepwise Multiple Linear Regression of MODIS spectral image products. **Journal of African Earth Sciences**, v. 152, p. 101–114, 1 abr. 2019.

ZIN, K. P. et al. Chemical properties and phosphorus fractions in profiles of acid sulfate soils of major rice growing areas in Brunei Darussalam. **Geoderma Regional**, v. 6, p. 22–30, 2015.