UNIVERSIDADE ESTADUAL DE CAMPINAS

Faculdade de Engenharia Agrícola

Felipe Ferreira Bocca

Estudo do efeito da adubação nitrogenada na produtividade de cana-de-açúcar com modelos de aprendizado de máquina

Campinas

2018

Felipe Ferreira Bocca

Estudo do efeito da adubação nitrogenada na produtividade de cana-de-açúcar com modelos de aprendizado de máquina

Tese apresentada à Faculdade de Engenharia Agrícola da Universidade Estadual de Campinas como parte dos requisitos exigidos para a obtenção do título de Doutor em Engenharia Agrícola, na área de Gestão de Sistemas na Agricultura e Desenvolvimento Rural

Orientador: Prof. Dr. Luiz Henrique Antunes Rodrigues

Este exemplar corresponde à versão final da tese defendida pelo aluno Felipe Ferreira Bocca, e orientada pelo Prof. Dr. Luiz Henrique Antunes Rodrigues

Campinas

2018

Agência(s) de fomento e nº(s) de processo(s): CAPES; CNPq, 140615/2017-2
ORCID: https://orcid.org/0000-0003-2956-5286

Ficha catalográfica
Universidade Estadual de Campinas
Biblioteca da Área de Engenharia e Arquitetura
Luciana Pietrosanto Milla - CRB 8/8129

Bocca, Felipe Ferreira, 1988-
B63e Estudo do efeito da adubação nitrogenada na produtividade de cana-de-açúcar com modelos de aprendizado de máquina / Felipe Ferreira Bocca. – Campinas, SP : [s.n.], 2018.

Orientador: Luiz Henrique Antunes Rodrigues.
Tese (doutorado) – Universidade Estadual de Campinas, Faculdade de Engenharia Agrícola.

1. Aprendizado de máquina. 2. Cana-de-açúcar. 3. Análise de sensibilidade. I. Rodrigues, Luiz Henrique Antunes, 1959-. II. Universidade Estadual de Campinas. Faculdade de Engenharia Agrícola. III. Título.

Informações para Biblioteca Digital

Título em outro idioma: The effect of nitrogen fertilization in sugarcane yield evaluated with machine learning models
Palavras-chave em inglês:
Machine learning
Sugarcane
Sensitivity analysis
Área de concentração: Gestão de Sistemas na Agricultura e Desenvolvimento Rural
Titulação: Doutor em Engenharia Agrícola
Banca examinadora:
Luiz Henrique Antunes Rodrigues [Orientador]
Romis Ribeiro de Faissol Attux
Paulo César Sentelhas
Gustavo Enrique de Almeida Prado Alves Batista
Henrique Coutinho Junqueira Franco
Data de defesa: 14-08-2018
Programa de Pós-Graduação: Engenharia Agrícola

Este exemplar corresponde à redação final da **Tese de Doutorado** defendida por **Felipe Ferreira Bocca**, aprovada pela Comissão Julgadora em 14 de agosto de 2018, na Faculdade de Engenharia Agrícola da Universidade Estadual de Campinas.

_____

Prof. Dr. Luiz Henrique Antunes Rodrigues – Presidente e Orientador

_____

Prof. Dr. Romis Ribeiro de Faissol Attux – Membro Titular

_____

Prof. Dr. Paulo César Sentelhas – Membro Titular

_____

Prof. Dr. Gustavo Enrique de Almeida Prado Alves Batista – Membro Titular

_____

Prof. Dr. Henrique Coutinho Junqueira Franco – Membro Titular

A Ata da defesa com as respectivas assinaturas dos membros encontra-se no processo de vida acadêmica do discente.

Aos meus irmãos e irmã. Eu ainda faço lição de casa...

Agradecimentos

Lique, obrigado! Pela confiança, pela autonomia, pelos conselhos. Você é forte influência da pessoa e pesquisador que me tornei.

Monique, obrigado! Minha vida, dentro e fora da academia, foi, e é, absolutamente melhor por você estar ao meu lado. Não vai ser por meio deste agradecimento que vou conseguir detalhar tudo, mas a partir do momento que você entrou na minha vida, onde quer que seja, tenho me tornado uma pessoa melhor.

Aos meus pais e irmãos: Mesmo que de longe, mesmo sem entender muito do que eu faço, o apoio de vocês foi muito importante para chegar onde estou.

Victor, Diego e Walter: Foi um prazer estar aqui com os senhores. Muito da riqueza deste percurso se deve a vocês. Obrigado pela paciência, pelos ensinamentos e pela companhia.

Aos amigos César, Bruno, Victor e Guilherme, obrigado por estarem do meu lado nessa jornada. Não deve ter sido fácil escutar tanto sobre "a pós"...

Thiago, Matheus, Rafaella, Vinicius, Tobias, Yuri, Henrique, e todos os demais que passaram pelo SISDA. Vocês me ensinaram colaboração, vocês me ensinaram responsabilidade, vocês me ensinaram a ser um profissional melhor. Aprendemos juntos muitas coisas, e entre linhas de R e posters de IC, emergiram amizades que quero levar para vida.

Ritinha, obrigado! Se não fosse por você, esse árduo período talvez fosse impossível.

Aos colegas que me receberam tão bem na Academia dos Campeões.

A todos aqueles que não pude incluir nominalmente neste agradecimento, funcionários da Feagri, colegas da pós-graduação, colegas de trabalho, alunos que dividiram as trincheiras, alunos para os quais tive a honra de participar da formação.

Agradeço a Feagri e a Unicamp, por me proporcionarem esta oportunidade tão rica.

Agradeço a CAPES e ao CNPq pela concesssão de bolsas que permitiram minha dedicação a vida acadêmica em sua plenitude. Graças a isso, posso dizer que sei como a pós-graduação é mais que a pesquisa. O presente trabalho foi realizado com apoio da Coordenação de Aperfeiçoamento de Pessoal de Nível Superior - Brasil (CAPES) - Código de Financiamento 001. Bolsa CPNq processo: 140615/2017-2.

Resumo

Na produção de cana-de-açúcar, é frequentemente reportada a falta de resposta para adubação com nitrogênio (N), fazendo com que sua efetividade e necessidade sejam questionadas. Dado o potencial de uso ineficiente dos recursos e de potenciais efeitos ambientais negativos, a recomendação adequada da quantidade de N aplicado é essencial para sustentabilidade financeira e ambiental da produção. Neste estudo, modelos de aprendizado de máquina foram desenvolvidos e aplicados para avaliação do efeito da adubação nitrogenada na produtividade da cana-de-açúcar, assim como dos fatores que interagem com essa prática. Para isso, modelos de produtividade de cana-de-açúcar foram desenvolvidos com técnicas de aprendizado de máquina aplicados a dados de produção comercial de cana-de-açúcar. Foi conduzida a análise de sensibilidade de primeira ordem, que foi comparada com a importância das variáveis nos conjuntos de dados, e a análise de sensibilidade de segunda ordem para estudo das interações de outras variáveis com a fertilização com N. Os resultados foram analisados com base em gráficos de resposta parcial, priorizados pela importância ou sensibilidade das variáveis. Gráficos de resposta condicional foram utilizados para distinguir o padrão de resposta geral (resposta marginal) do padrão de resposta local (resposta condicional para cada condição encontrada nos dados). Foi constatado que o padrão de respostas individuais não apresenta respostas consistentes para a produção de cana-de-açúcar, embora as respostas gerais sejam mais coerentes. Considera-se então que não é recomendável utilizar a saída de modelos gerados utilizando as técnicas empregadas neste trabalho para análises de respostas individuais, o que seria por exemplo, necessário para recomendação de adubação para cada talhão de cana-de-açúcar. Usos pautados pela resposta geral parecem não ser afetados e devem ser avaliados em trabalhos futuros.

Palavras-chaves: Aprendizado de Máquina; Fertilização de Nitrogênio; Cana-de-açúcar; Análise de Sensibilidade; Análise visual de modelos.

Abstract

Lack of response to nitrogen fertilization is often reported for sugarcane production, leading to questions regarding its necessity and effectiveness. Given the potential for inefficient resource usage and potential negative environmental impacts, properly recommending the amount of N fertilizer is essential for a financial and environmental sustainable sugarcane production. In this thesis, machine learning models of sugarcane yield were developed and applied to evaluate the effects of Nitrogen fertilization in sugarcane yield, as well as factors that interacts with this practice. First order sensitivity analysis was performed and compared with feature importance measured in the datasets used for modeling, and second-order sensitivity analysis was performed to evaluate interactions with N fertilization in the model. Results were evaluated based on the partial response plots, prioritized by feature importance and variable sensitivity. Independent conditional expectancy graphics are also used to evaluate the individual response of plots (conditioned response in each condition modeled) and to evaluate the differences from the general response pattern (marginalized response). From the results of the visual analysis, it can be seen that individual responses are not consistent with common knowledge for sugarcane production, even though some of the general responses are more coherent. Based on these results, the use of such models for individual analysis and recommendations, such as needed for nitrogen fertilization recommendation, are not recommended. The use based on the general response may not be affected and could be further evaluated in future works.

Keywords: Machine learning; Nitrogen fertilization; Sugarcane; Sensitivity analysis; Visual model analysis.

"There is no real ending. It's just the place where you stop the story."
(Frank Herbert, 1969)

Lista de ilustrações

Lista de tabelas

Sumário

1 Introdução

1.1 Produção de cana-de-açúcar e adubação nitrogenada

A produção da cana-de-açúcar e seus derivados é de grande importância para o Brasil. O valor bruto da produção de cana-de-açúcar — R$ 59,22 bilhões — é o segundo maior valor produzido de um total de R$ 307,72 bilhões de valor bruto produzido pela agricultura em 2017 (MAPA, 2018). O país é o maior produtor de cana-de-açúcar do mundo, com uma produção de 768 milhões de toneladas em 2017, seguido da Índia e da China, que produziram 348 e 122 milhões de toneladas respectivamente, nesse mesmo ano (FAO, 2018). Além da produção de etanol e açúcar, derivados de cana-de-açúcar, a produção de etanol de segunda geração e eletricidade nas usinas de cana-de-açúcar tem grande potencial para geração de energia sustentável no Brasil (Goldemberg et al., 2008).

A cana-de-açúcar é uma cultura semi-perene que pode ser colhida múltiplas vezes, sendo tipicamente colhida de 5 a 8 vezes antes de ser replantada. O plantio é predominantemente mecanizado e utiliza pedaços dos colmos com aproximadamente 20 cm, chamados de toletes. Cada tolete precisa conter uma gema para produzir perfilhos que possam crescer e formar novos colmos. Após cada colheita, as partes restantes da planta, chamdas soqueiras, produzem novos colmos. Para distinção, o primeiro ciclo da cultura é chamado de cana planta enquanto os demais são chamadas de cana-soca.

Em uma descrição simplificada, o desenvolvimento da cultura pode ser dividido em brotação, perfilhamento, desenvolvimento dos colmos e maturação. A fase de brotação é bem caracterizada no primeiro ciclo da cultura, quando o primeiro perfilho brota e se estabelece. Para soqueiras, o primeiro perfilho que emergiu também pode ser considerado para caracterizar a botação, porém o desenvolvimento não é linear como ocorre para cana planta, pois diversos perfilhos podem emergir antes que o primeiro se estabeleça. Após o estabelecimento do primeiro perfilho, é iniciada a fase de perfilhamento e mais perfilhos serão emitidos pela planta, até que uma população máxima de perfilhos é atingida e começa a diminuir devido à competição intra-específica. Com o declínio da população de perfilhos, é encerrado o perfilhamento e os colmos começam a se desenvolver com uma fase de rápido alongamento, quando a maior parte da biomassa é acumulada. Após o alongamento do colmo, condições meteorológicas induzem a planta ao amadurecimento. Essa fase é estimulada por baixas temperaturas e baixa disponibilidade de água no solo.

Para o cultivo da cana-de-açúcar no Brasil, a fertilização com nitrogênio é recomendada em taxas [1] de 3 g·m^{-2} para cana planta e taxas de 6 a 12 g·m^{-2} para soqueiras (Raij et al., 1996). Extrapolando essas taxas de aplicação para área de expansão de 6,4 milhões de hectares (Goldemberg et al., 2014) [2], é possível estimar um aumento na demanda de nitrogênio próximo a 400 mil toneladas por ano.

A adubação nitrogenada é necessária para a produção de cana-de-açúcar, mas seu uso em excesso pode causar danos ambientais (Robertson e Vitousek, 2009). Esses danos estão relacionados à contaminação de corpos de água em caso de lixiviação e consequente eutrofização dos mesmos. Além disso, a volatilização do nitrogênio na forma de NO^3 pode contribuir para o agravamento do efeito estufa (Reay et al., 2012). Cabe destacar que o efeito residual da adubação em ciclos posteriores é muito baixo (Franco et al., 2015), fazendo com que excessos na adubação tenham apenas impactos negativos. Considerando o custo direto relacionado ao uso do insumo e as potenciais consequências negativas da aplicação, a correta recomendação da adubação nitrogenada é uma necessidade para a produção da cana-de-açúcar.

O nitrogênio é importante para as plantas por participar da composição dos ácidos nucleicos, proteínas e enzimas, em especial as ligadas à fotossíntese. Para cana-de-açúcar, o efeito do nitrogênio se destaca no perfilhamento e no crescimento da cultura (Kingston, 2013). Embora o nitrogênio seja abundante na forma de N_2 na atmosfera, os vegetais não conseguem aproveitá-lo nesta forma, sendo necessária sua fixação por bactérias, que convertem o nitrogênio gasoso em amônio e nitrato. Outra forma de disponibilização do nitrogênio aos vegetais é pela degradação e mineralização da matéria orgânica no solo. Uma discussão completa sobre o ciclo do nitrogênio pode ser vista em Tartowski e Howarth (2001).

A relação entre a adubação nitrogenada e a produtividade de cana-de-açúcar é assunto de extensa pesquisa no contexto produtivo do Brasil. Uma síntese das revisões realizadas por Otto et al. (2016), Penatti (2013), Vitti et al. (2010) e Franco e Trivelin (2010) aponta que [3]:

- A resposta da cana-de-açúcar à adubação nitrogenada é uma questão aberta e em diversas situações, não é observada.

[1] Usualmente, a adubação é reportada em kg· ha^{-1}, porém neste trabalho optou-se pela unidade no Sistema Internacional. Para conversão das medidas para unidade usual, basta multiplicar o valor por 10.

[2] Soma ponderada para cultivos de 1 a 6 anos, com peso 1/6 na aplicação de 3 g·m^{-2} para cana planta e 5/6 para 7 g·m^{-2} na média para soqueiras.

[3] Cabe destacar que a maior parte do material referenciado nessas revisões diz respeito a pesquisas conduzidas para cana queimada antes da colheita, o que está sendo substituída pela colheita mecanizada de cana crua. As mudanças de manejo podem impactar nas relações de nitrogênio e produtividade da cana-de-açúcar. No trabalho de Otto et al. (2016), consta uma seção dedicada a essa discussão.

  - Em solos arenosos, a cana-de-açúcar tem uma resposta maior e responde a doses maiores de adubo.
  - Recomendações de adubação em diversos experimentos vão de 2 a 9 $g \cdot m^{-2}$.

- A resposta da cana-de-açúcar à fertilização nitrogenada é mais frequente e intensa para soqueiras do que a resposta da cana planta.
  - A intensa movimentação do solo durante o plantio de cana-de-açúcar eleva a taxa de mineralização da matéria orgânica, suprindo a demanda da cana planta. Isso pode explicar porque a cana planta responde menos à adubação nitrogenada do que as soqueiras.
  - A quantidade de nitrogênio nos toletes plantados não é capaz de suprir o desenvolvimento da cana planta e nem explicar a menor resposta à adubação.

- A fixação de nitrogênio por vegetais e bactérias pode contribuir para suprir nitrogênio para a cana-de-açúcar, mas as taxas providas não são capazes de suprir completamente a demanda.

- A recuperação do nitrogênio aplicado vai de 0,2 a 63,0 %. Dentre os fatores que afetam a taxa de recuperação estão:
  - Regime hídrico
  - Granularidade, textura e estrutura do solo
  - Tipo de adubo utilizado (e.g. Ureia ou Amônio)
  - Como o fertilizante foi depositado (e.g. sobre a palha ou enterrado)

Duas grandes mudanças estão impactando a produção de cana-de-açúcar no Brasil: a intensa mecanização e a rápida expansão da área nos últimos anos. A cana-de-açúcar passou a ser plantada e colhida mecanicamente e a mecanização da colheita deve atingir 100 % nos próximos anos. Dado que a cana-de-açúcar não é mais queimada para colheita, a palha da cana-de-açúcar pode ser utilizada para produção de energia elétrica e/ou etanol de segunda geração. Além do uso comercial da palha, existem benefícios agronômicos e ambientais da palha deixada no campo (Leal et al., 2013). A palha deixada no campo protege o solo de erosão, diminui a brotação de plantas daninhas e permite a ciclagem de nutrientes a partir da matéria orgânica. Impactos negativos são a redução da temperatura do solo, o que pode impactar a brotação da cana-de-açúcar em algumas regiões e o aumento de pragas e doenças que eram eliminadas com a queima da palha. Um impacto altamente desejado é a redução da emissão de gases de efeito estufa associados à produção da cana-de-açúcar.

Quanto à expansão de áreas, ela é um efeito da grande concorrências pelas áreas nas regiões típicas de produção, levando o aumento da produção de cana-de-açúcar para o oeste do estado de São Paulo (Rudorff et al., 2010) e para as regiões de Cerrado em Minas Gerais no Sudeste brasileiro e nos estados do Mato Grosso, Mato Grosso do Sul e Goiás, no Centro-Oeste brasileiro (Koizumi, 2014).

Esse cenário de mudanças na produção de cana-de-açúcar no Brasil pode levar a um aumento no uso de nitrogênio para adubação necessário. Enquanto a deposição da palha no solo pode levar a uma redução na demanda de nitrogênio necessário para produção (Trivelin et al., 2013), isso não acontece logo no início da mudança. Devido à alta proporção de carbono para nitrogênio (razão C:N), o nitrogênio do solo é imobilizado (Meier et al., 2006), demandando uma aplicação maior de fertilizante para acomodar esse efeito. Outro aspecto é que solos da região de Cerrado são de baixa fertilidade natural e alta acidez, além de apresentar baixa capacidade de troca catiônica, baixo teor de matéria orgânica e alta disponibilidade de alumínio (Lopes e Cox, 1977). Em áreas onde a agricultura comercial está estabelecida, boa parte dessas características estão corrigidas, porém a expansão da cana-de-açúcar se dá primariamente em áreas de pastagens degradadas, onde o solo possui frequentemente as características negativas originais. Em função dessas características, a produção nessas regiões pode demandar alto uso de insumos. Nesse contexto, a expansão da cana-de-açúcar para áreas de baixa fertilidade pode levar a um fenômeno similar ao da expansão da cana-de-açúcar na Flórida, E.U.A. Na Flórida, a cana-de-açúcar era tipicamente produzida em áreas ricas em argila e matéria orgânica, condições características de baixa demanda de nitrogênio. A expansão para áreas com solos mais arenosos e de baixa matéria orgânica resultaram no aumento nas taxas de nitrogênio aplicadas (Zhao et al., 2014; McCray et al., 2010).

Ainda que deixar a palha da cana-de-açúcar no campo possa aumentar o conteúdo de matéria orgânica no solo, contribuindo na quantidade de nitrogênio disponível, isso não é suficiente para suprir a necessidade de nitrogênio da cultura da cana-de-açúcar. Ferreira et al. (2015) encontraram uma taxa de recuperação de nitrogênio da palhada de 0,76 $g \cdot m^{-2}$ para soqueiras em três anos. A baixa taxa de recuperação está ligada à liberação lenta promovida pela degradação da matéria orgânica no solo. Robertson e Thorburn (2007) conduziram um experimento e mediram uma taxa de recuperação do nitrogênio da palhada no solo variando de 0,1 a 0,5 $g \cdot m^{-2}$ no final da safra australiana, quando o clima é úmido e quente. No experimento, 1 $kg \cdot m^{-2}$ de palhada de cana-de-açúcar eram devolvidos ao solo após a colheita. Construindo um modelo a partir de diversos resultados publicados, Trivelin et al. (2013) encontraram um potencial de reduzir a fertilização no longo prazo (17 anos) se a palha for deixada no campo. Esses resultados destacam que

pode haver um benefício direto por meio da redução da adubação necessária, concorrendo com o uso comercial da palha.

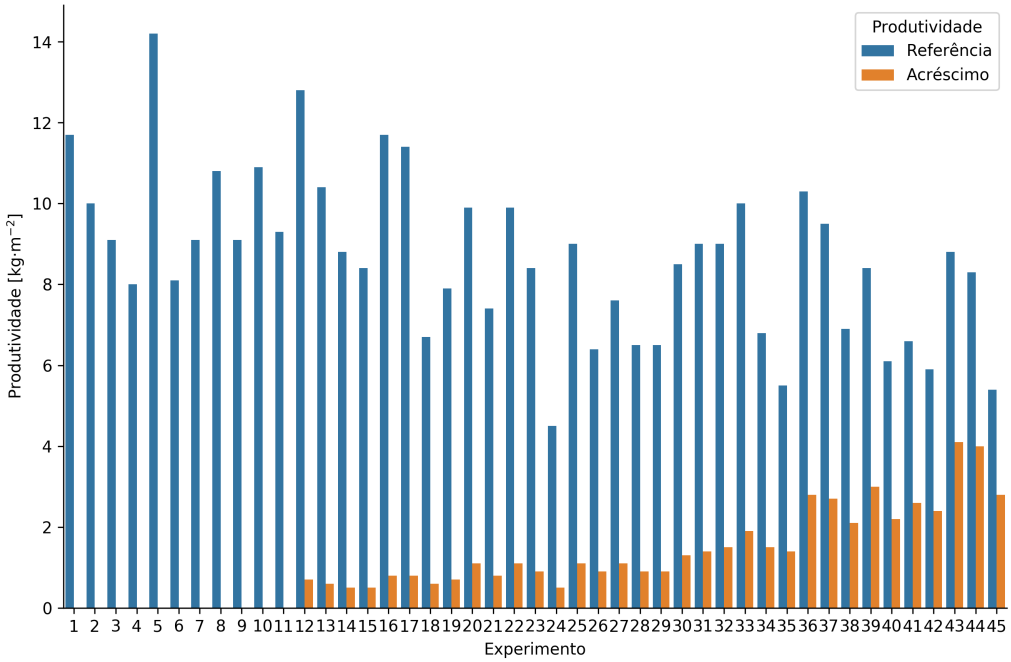
A baixa recuperação do nitrogênio aplicado [4] na cana colhida não implica um baixo uso do nitrogênio. Medindo a recuperação do nitrogênio proveniente da fertilização ao longo do tempo, e não apenas na colheita, Franco et al. (2011) identificaram que de 40 a 70 % do nutriente na planta era proveniente do adubo aplicado. Isso mostra que a importância da adubação pode ser maior nas etapas iniciais do que para a composição final. Zhao et al. (2014) mostraram que diferentemente de outros vegetais, a resposta da cana-de-açúcar à adubação nitrogenada não afeta o nível de clorofila das folhas ou a taxa de fotossíntese nas folhas. Para a cana-de-açúcar, o efeito do nitrogênio se deu por meio de um aumento da área foliar e no número de perfilhos, sendo o resultado dependente da variedade utilizada. Outra diferença entre a cana-de-açúcar e outros vegetais no que diz respeito ao nitrogênio é a preferência pelo amônio sobre o nitrato (Robinson et al., 2011), o inverso do observado para as demais plantas. Em situações em que ambos os nutrientes estão presentes, a cana-de-açúcar absorve o amônio em detrimento do nitrato disponível.

Ainda que existam diversos resultados sobre como diversos fatores afetam a resposta à adubação com nitrogênio, resultados de 45 experimentos compilados por Otto et al. (2016) mostram que em 75 % dos casos a cultura teve resposta baixa ou moderada. Para os 25 % restantes, nos quais a cultura foi responsiva, isso se deu em condições de baixa fertilidade natural. Os resultados dos 45 experimentos são apresentados na Figura 1. Além disso, diferentemente do que pode ser realizado para outros nutrientes, para os quais é possível recomendar a quantidade de fertilizantes com base na concentração do nutriente no solo, ainda não existe uma metodologia que correlacione medidas de campo com o estoque de N disponível para as plantas (Mariano et al., 2017). Esses resultados indicam que existe espaço para melhorar a recomendação da quantidade de N para adução de cana-de-açúcar. Idealmente, a recomendação consideraria o contexto de produção e ajustaria a quantidade recomendada em função da resposta esperada da cultura.

Considerando as limitações envolvidas em aplicar experimentos convencionais em larga escala para investigar todas as possíveis interações nas diversas condições de crescimento, uma alternativa é explorar as relações existentes nos dados de produção. Aspectos negativos dessas bases são a qualidade menor se comparadas com dados de experimentos e os vieses implícitos associados ao ambiente de produção. Por outro lado, esses dados estão prontamente disponíveis em larga escala, e refletem as condições reais de produção. Dado o grande volume de dados, Lawes e Lawn (2005) consideraram que os erros devem ser pequenos em estudos relacionados à variabilidade espaço-temporal da produtividade e dos fatores que afetam a produtividade. Quando esses dados são

[4] 0,2 a 63,0 % do N aplicado é encontrado na planta, pag. 16

Figura 1 – Resposta à adubação nitrogenada em diferentes experimentos compilados por Otto et al. (2016). Para os experimentos são apresentadas a produtividade da cana-de-açúcar sem adubação (referência) e o acréscimo obtido com a adubação nitrogenada. Os experimentos são ordenados em função da resposta relativa de produtividade (acréscimo de produtividade em relação a produtividade de referência). Adaptado de Otto et al. (2016).



enriquecidos com dados meteorológicos e de solo, análises podem ser feitas para investigar como esses fatores impactam a produtividade da cana-de-açúcar.

Considerando que, no Brasil, as usinas de cana-de-açúcar controlam grande parte da área de produção, dados de manejo estão disponíveis junto aos dados de produção. Diversos estudos recentes foram conduzidos com a aplicação de técnicas de mineração de dados e aprendizado de máquina para usufruir dos potenciais destes bancos de dados. Peloia e Rodrigues (2016) avaliaram como diferentes fatores estão relacionados às diferenças relativas na produtividade de cana-de-açúcar para áreas com diferentes números de colheitas. Peloia et al. (no prelo) avaliaram o uso de árvores de decisão para identificar condições nas quais padrões acionáveis têm potencial de recomendar intervenções para aumento de produtividade. Outros trabalhos buscaram o desenvolvimento de modelos com ênfase na performance final, sendo Bocca e Rodrigues (2016) e Hammer (2016) para produtividade da cana-de-açúcar e Oliveira et al. (2017) para o teor de açúcar da cana na colheita.

1.2 Aprendizado de máquina

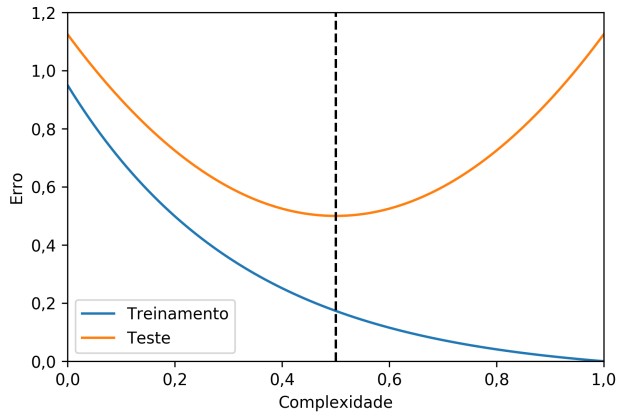
A área de aprendizado de máquina tem como objetivo desenvolver programas que são capazes de aprender com base em conjuntos de dados. Esses conjuntos de dados, embora geralmente sejam associados a alguma representação de tabela, ou conjunto de tabelas, podem assumir diversas formas, como imagens, sons, grafos, ou mesmo o resultado da interação de um agente em um ambiente, entre outras. Para atingir o objetivo de aprendizado, a área de aprendizado de máquina combina elementos de estatística e ciência de computação, utilizando diversos princípios para generalizar padrões com base em observações.

De forma geral, pode-se dizer que o aprendizado de máquina se divide entre aprendizado supervisionado e aprendizado não-supervisionado. Enquanto no aprendizado supervisionado, busca-se uma relação entre as observações disponíveis e alguma variável de interesse, no aprendizado não-supervisionado, busca-se identificar padrões e estruturas entre as observações. No aprendizado supervisionado, quando a variável de interesse é discreta, isto é, composta de diferentes categorias ou classes, temos uma tarefa de classificação. Se a quantidade de interesse é contínua, a tarefa é de regressão. Cabe destacar que diversas técnicas estão disponíveis para realizar ambas as tarefas, sendo que grande parte das técnicas podem ser utilizadas para classificação e para regressão com as devidas adaptações[5]. Um modelo é criado a partir da aplicação de uma técnica a um conjunto de dados, sendo que neste processo ocorre o chamado aprendizado. Esta etapa é comumente chamada de treinamento do modelo.

Na modelagem utilizando técnicas de aprendizado de máquina, geralmente é utilizada alguma forma de validação cruzada do desempenho do modelo, em que a performance de um modelo é avaliada em um conjunto de dados diferente dos utilizados para treinamento do modelo. O objetivo principal deste procedimento é obter uma boa estimativa da performance do modelo em novos dados. Isso se dá pois a grande capacidade de modelagem de técnicas de aprendizado de máquina permite que modelos sejam ajustados com grande precisão ao conjunto de treinamento sem que isso seja necessariamente um indicador da performance posterior. Essa situação caracteriza o *overfitting*, em que o aumento da complexidade leva o modelo a se ajustar especificamente ao conjunto de treino perdendo capacidade de generalizar em novos conjuntos de dados, conforme ilustrado na Figura 2. A complexidade de um modelo está ligada a sua flexibilidade e capacidade de representar hipóteses complexas. Para modelos baseados em árvores, por exemplo, o número de folhas está relacionado ao aumento da capacidade.

[5] Uma visão mais completa de técnicas não supervisionadas e as diferentes abordagens, assim como detalhes sobre as diferentes tarefas pode ser vista em James et al. (2013).

Figura 2 – Ilustração de *overfitting*. A taxa de erro do modelo (eixo y) avaliada nos dados de treinamento (curva azul) e em um conjunto de dados independente (curva laranja) diminuem com o aumento da complexidade do modelo, até que em certo ponto (indicado pela reta tracejada) a diminuição do erro nos dados de treinamento não é acompanhada por uma queda do erro em um conjunto independente, que passa aumentar.



Entre as diversas formas de validação cruzada existentes, três formas mais gerais se destacam, sendo elas a validação em uma amostra separada (tradução livre de *holdout*), a validação em K-partições (*K-fold*) e reamostragem. Na validação *holdout*, o conjunto de dados é dividido aleatoriamente em duas partes, sendo a primeira utilizada para treinamento do modelo, e a segunda reservada para estimar a performance do modelo. Na validação *K-fold*, o conjunto de dados é dividido em K partições. A avaliação é repetida K vezes, sendo que para cada iteração, uma parcela é reservada para teste e as demais são utilizadas para treinamento, de forma que cada parcela foi utilizada para teste uma vez. A avaliação de performance é então agregada, sendo utilizada geralmente a média dos K valores obtidos. Na avaliação por reamostragem, repetidas divisões em treino e teste são feitas e agregadas, podendo a amostra de treino ser obtida com reposição (*bootstrap*).

Para que o treinamento de um modelo seja efetivo, é necessário ajustar configurações das técnicas utilizadas. Esse ajuste é feito por meio da especificação de parâmetros da técnica, usualmente denominados hiper-parâmetros. Os hiper-parâmetros são responsáveis pela condução do processo de aprendizado, estando relacionados à função que a técnica otimiza para criar os modelos, ou ligados à complexidade do modelo que será criado. Um distinção entre os parâmetros de um modelo e os hiper-parâmetros de uma técnica é que os parâmetros são ajustados em função dos dados observados, o que reflete o aprendizado dos modelos. Já os hiper-parâmetros são definidos antes da etapa de aprendizado, e portanto, não fazem parte do aprendizado a partir dos dados. Embora as técnicas produzam modelos a partir dos dados, os hiper-parâmetros fazem com que diferentes modelos possam resultar dos mesmos dados, sendo necessária a avaliação empírica do efeito dos hiper-parâmetros na performance dos modelos. A otimização da performance

de um modelo em função dos diferentes hiper-parâmetros é comumente chamada de *tuning* (ajuste do modelo, em tradução livre). Considerando que a performance do conjunto de treino não é uma estimativa confiável dado que os modelos podem ser facilmente ajustados para ter erro desprezível no conjunto de dados de treinamento, isso faz com que usualmente os conjuntos de dados utilizados sejam divididos em três partes. Uma parcela para treinamento do modelo, uma parcela utilizada para avaliar a performance do modelo durante o *tuning* do modelo e uma última parcela de dados para estimar a performance do modelo. Essas parcelas são comumente denominadas de treino, validação e teste respectivamente. Dado que a separação de mais uma parcela de dados pode diminuir ainda mais os dados disponíveis para treino do modelo, usualmente é feita a validação cruzada *k-fold* para *tuning* do modelo. Para grandes conjuntos de dados, isso não se faz necessário, podendo até ser contra-produtivo, dada a multiplicação do esforço computacional para a estratégia *k-fold*.

Um procedimento frequentemente empregado na modelagem com técnicas de aprendizado de máquina é a seleção de variáveis. Para um conjunto de dados, o objetivo é encontrar um subconjunto de variáveis que otimiza a performance do modelo criado. Esta etapa pode simplificar o conjunto de dados e portanto reduzir o esforço computacional e a complexidade do modelo. Em alguns casos, a seleção de variáveis permite melhorias na performance por eliminar variáveis que podem introduzir ruídos na modelagem. Uma visão geral sobre seleção de atributos e sua relação com a modelagem é apresentada por Guyon e Elisseeff (2003), entre outros assuntos.

Entre as diversas estratégias de seleção de variáveis, têm-se a estratégia de filtro, onde métodos são utilizados para estimar o quanto uma variável independente informa sobre a variável dependente. Esta estratégia se dá sem o desenvolvimento do modelo e permite quantificar o efeito individual de cada variável. Essa quantificação nos dá a importância de uma variável presente no conjunto de dados e permite priorizar variáveis de um conjunto de dados. O uso da importância de variáveis para seleção de atributos tem mais sucesso se essa medida de importância de uma variável é capaz de capturar quanto uma técnica de modelagem será capaz de aproximar a relação entre a variável independente e a variável dependente. A título de ilustração, se tivermos uma relação linear aditiva entre as variáveis dependentes e independentes, avaliar a importância das variáveis utilizando a correlação individual será efetivo.

1.2.1 Técnicas de modelagem

Modelos são representações matemáticas das leis que governam um fenômeno natural, ganhando importância na medida que compõem sistemas de tomada de decisão, levam a novas descobertas ou se tornam uma forma de organizar o conhecimento científico

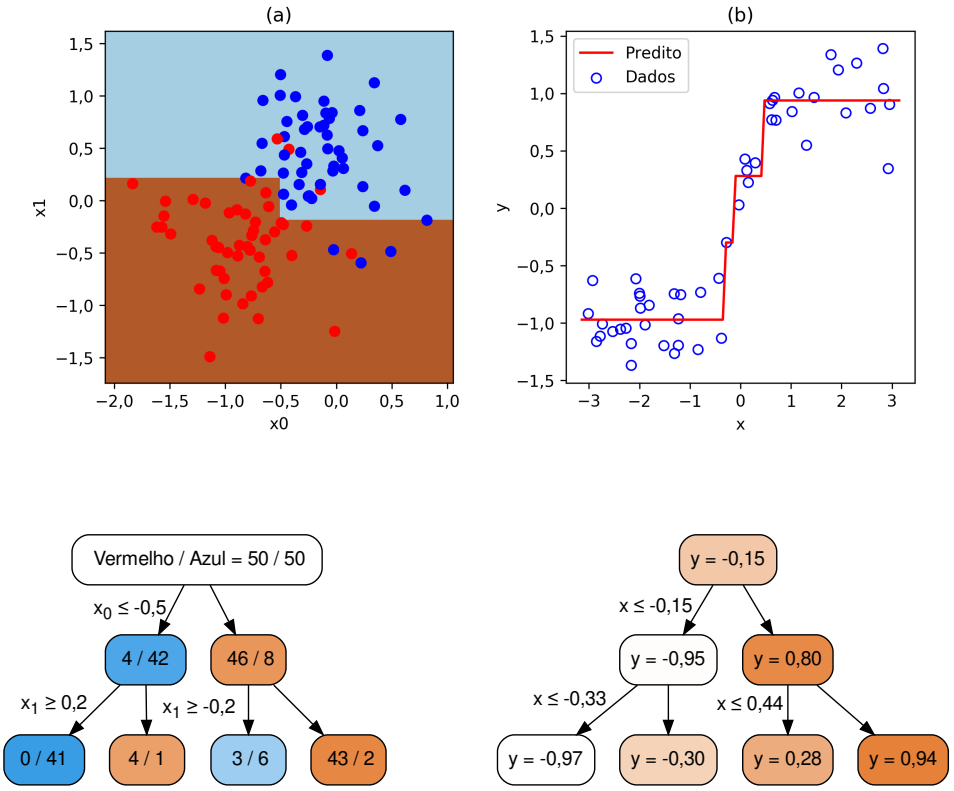
(Tedeschi, 2006). Enquanto modelos podem ser construídos a partir de leis conhecidas, é possível derivar relações entre quantidades a partir de conjuntos de dados, uma linha de modelagem chamda de orientada por dados (tradução livre de *data-driven*).

Diferentes técnicas de aprendizado de máquina podem ser utilizadas para criar modelos que relacionam "entradas"(variáveis independentes ou atributos de modelagem) a uma "saída"(variável dependente ou atributo alvo). Entre as diferentes técnicas de aprendizado de máquina, é amplamente admitido um *trade-off* entre a capacidade de modelagem de uma técnica e a facilidade em interpretar o modelo criado (James et al., 2013, p.25). Enquanto técnicas como árvores de decisão ou modelos de regressão podem ser facilmente interpretados em grande parte dos casos de uso, técnicas mais complexas como redes neurais artificiais multicamadas (Rumelhart et al., 1985) e máquinas de vetores de suporte (Cortes e Vapnik, 1995) geram modelos difíceis de serem interpretados. Outra fonte de complexidade são técnicas de combinação de diversos modelos simples para formar um modelo de maior capacidade ou *ensemble* (Dietterich, 2000) cuja interpretabilidade é comprometida. Em função da dificuldade de interpretar essas técnicas, elas são geralmente consideradas caixas-pretas.

Árvores de decisão são estruturas do tipo árvore onde a bifurcação é representada por testes lógicos. A representação gráfica desses modelos se assemelha a uma árvore invertida. Com base em um conjunto de dados, o algoritmo busca uma partição que maximize a pureza da variável de interesse nos subconjuntos (Breiman et al., 1984). No geral, os algoritmos de indução de árvores de decisão usam aproximações para quantificar a pureza, como o índice Gini e a entropia em tarefas de classificação, ou desvio padrão para tarefas de regressão. Assim, para um determinado conjunto de dados, o algoritmo varre todas as variáveis de entrada buscando maximizar a pureza da variável de saída em cada partição. Para cada partição, esse procedimento é repetido até que um critério de parada seja atingido. Na representação de árvore, o nó inicial representa todas as amostras do conjunto de dados utilizado para treinamento. Cada nó subsequente representa uma divisão do conjunto de dados com base em um teste lógico, até que seja atingido um critério de parada criando nós denominados "folhas". Para cada "folha" da árvore, a predição do modelo é feita considerando a "saída" das amostras que foram alocadas na folha. Em tarefas de classificação, usualmente é predita a classe majoritária das amostras na "folha", enquanto, para regressão, é utilizada a média das amostras (Figura 3). Uma árvore de decisão é facilmente interpretável quando possui poucas "folhas"ou pouca profundidade (número de divisões até atingir uma "folha").

Árvores de regressão incrementadas por gradiente (tradução livre de *Gradient boosted regression trees*, BRT) são um tipo de *ensemble* em que modelos são combinados incrementalmente de forma que cada modelo adicionado busca reduzir o erro cometido

Figura 3 – Ilustrações de uma árvore de decisão (a) e de uma árvore de regressão (b).



antes da sua inclusão (Friedman, 2001). Um modelo fraco é gerado inicialmente, sendo geralmente uma árvore com limitação na profundidade máxima. O segundo modelo é ajustado então no erro no primeiro modelo, de forma que a soma da saída dos dois modelos se aproxime dos dados de resposta observada. Para estabilizar esse processo, é utilizada uma taxa de aprendizado que multiplica a saída de cada modelo. Com o uso desse processo, cada modelo é responsável por aproximar uma fração da resposta final. A saída do modelo tende então a ficar cada vez mais próxima da resposta observada nos dados para cada modelo adicionado ao *ensemble*.

Florestas aleatórias (*Random Forest*, RF) são uma variação de *ensembles* do tipo *bagging*. Um *ensemble* do tipo *bagging* (Breiman, 1996) usa duas estratégias, *bootstrap* e *aggregating* (agregação). Uma coleção de modelos é obtida por meio do treinamento de modelos em múltiplas versões do conjunto resultantes da reamostragem com reposição (*bootstrap*) dos dados de treinamento. O resultado final do modelo é obtido pela agregação da saída dos modelos, sendo usual a média para modelos de regressão e a moda para modelos de classificação. Mesmo que os modelos criados para cada amostra do conjunto de dados seja diferente, em uma RF o algoritmo de treinamento da árvore é modificado para diminuir a similaridade entre os modelos gerados (Breiman, 2001) No treinamento de árvores para uma RF, um número de atributos menor do que o total de atributos disponíveis

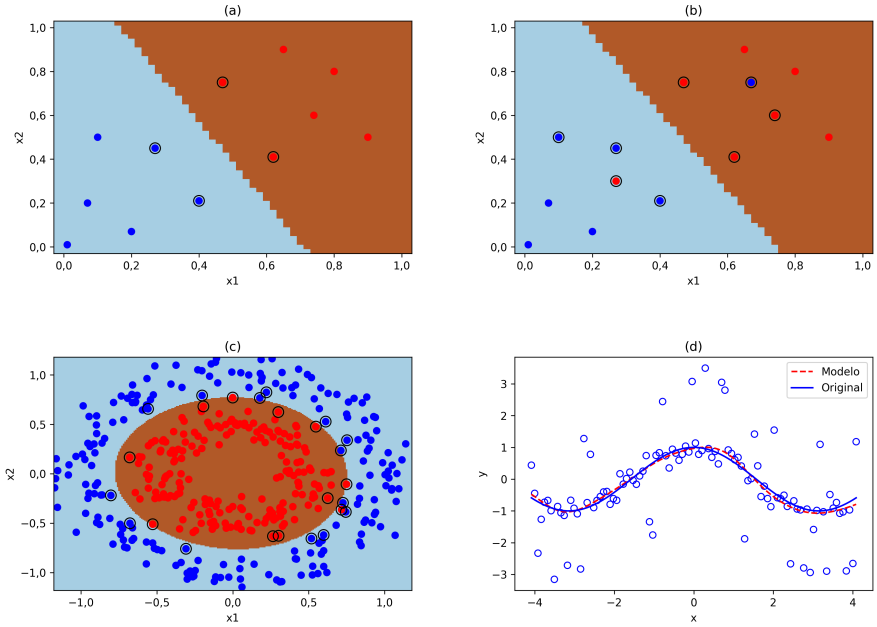
para avaliação das partições é especificado. Para cada iteração de treinamento da árvore, são sorteados então os possíveis candidatos para partição com base em um subconjunto dos atributos disponíveis. Isso diminui ainda mais a correlação entre os diferentes modelos criados e suas saídas, melhorando a performance do modelo final.

Regressão de vetores de suporte (*support vector regression*, SVR) é a adaptação (Smola e Schölkopf, 2004) das máquinas de vetores de suporte (Cortes e Vapnik, 1995) para a tarefa de regressão. Originalmente, as máquinas de vetores de suporte foram concebidas para classificação, buscando maximizar a margem de separação entre classes, como ilustrado na Figura 4 a. A margem é obtida utilizando programação quadrática, garantindo que pode ser encontrada uma solução ótima na solução que tem a maior margem. Duas importantes características da técnica incluídas após a formulação original são o uso de uma constante de custo, que caracteriza o *trade-off* entre maximizar a margem e tolerar erros (Figura 4 b) e o truque do *kernel* (Figura 4 c), que permite mapear os dados de entrada em um espaço de dimensão superior, permitindo a separação de dados não separáveis linearmente. Na figura, é ilustrada a fronteira de decisão resultante da aplicação do truque do *kernel*. De forma simplificada, o truque se dá pois a relação entre dois vetores estabelecida pelo kernel depende apenas do produto interno desses vetores, permitindo quantificar uma relação que se dá em um espaço de dimensão maior, sem que seja necessário transformar os dados. Quando adaptada para regressão, o problema torna-se buscar os vetores suporte tal que uma região do espaço por eles definida contenha o máximo possível de pontos no seu interior (Figura 4d). Diferentes tipos de *kernel* estão disponíveis, sendo que, para modelagem de dados numéricos, o usual é a utilização de *kernel* linear, polinomial ou de base radial. No caso do polinomial ou de base radial, é necessário ajustar parâmetros específicos de cada *kernel*.

1.2.2 Ajuste de parâmetros e validação de modelos

Dado o efeito dos hiper-parâmetros de cada técnica no modelo resultante, é usual que sejam avaliados diferentes valores para cada hiper-parâmetro. Em função da interação entre hiper-parâmetros, frequentemente são testadas as combinações de valores dos diferentes hiper-parâmetros. Uma estratégia para isso é realizar uma busca em *grid* (grade, em tradução livre). Para isso, são especificados valores mínimos, valores máximos e número de intervalos que serão avaliados para cada hiper-parâmetro. São avaliadas então todas as combinações entre os valores possíveis dos hiper-parâmetros. Para uma SVR, supondo que sejam avaliados os valores de custo 10, 100 e 1000, e do grau do polinômio utilizado no *kernel* com valores 2 e 3, haveria 6 combinações (10 e 2, 10 e 3, 100 e 2, ..., 1000 e 3). Essa abordagem também é chamada de força bruta.

Figura 4 – Ilustração dos princípios de *Support Vector Machines*. (a) Separação de um conjunto linearmente separável utilizando o hiper-plano de maior margem. Vetores suportes estão circulados em preto (idem nas sub-figuras b e c). (b) Diante de dois pontos atípicos, ainda é possível identificar uma fronteira de separação com grande margem. (c) Classificação utilizando o *kernel* rbf para separação não-linear. (d) Regressão SVR utilizando o kernel rbf de uma função co-seno em que é adicionado um ruído gaussiano e *outliers*.

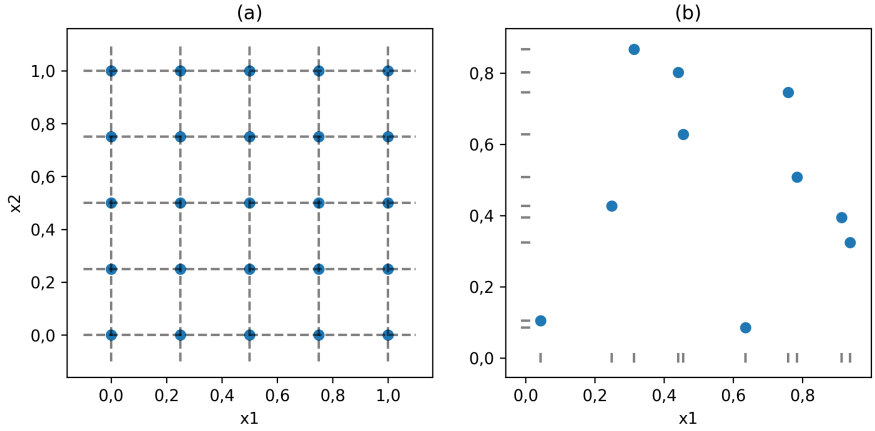


Dado o crescimento exponencial de combinações a serem testadas em função do número de parâmetros, uma alternativa prática é utilizar a busca aleatória de hiper-parâmetros. A busca aleatória permite que cada atributo seja explorado melhor do que no cenário de busca em *grid* (Bergstra e Bengio, 2012) e, em geral, atinge uma performance equivalente à busca em *grid* para um mesmo número de iterações. Isso se dá pois os hiper-parâmetros podem ser mais ou menos relevantes para cada caso, mas sem conhecimento prévio, diversas combinações de configurações serão geradas para avaliar diferentes valores de um hiper-parâmetro que pode não ser relevante, de modo que não há benefício. Embora a busca aleatória pareça contra-intuitiva, ela permite uma busca efetiva em espaços de grande dimensão[6], onde seria proibitiva a busca em *grid*. Na Figura 5 são ilustradas as buscas em grid e aleatória.

Um aspecto em comum entre a busca em *grid* e a busca aleatória é que as configurações são pré-fixadas antes do início do ajuste dos parâmetros, negligenciando as informações obtidas durante o ajuste dos parâmetros. Outras estratégias são capazes de incorporar essas informações parciais e escolher as próximas configurações em função dos

[6] Entendendo que N hiper-parâmetros formam um espaço de N dimensões onde é buscada a configuração do algoritmo.

Figura 5 – Ilustração dos pontos avaliados para busca em grid com 25 pontos (a) e busca aleatória com 10 pontos (b) em duas dimensões.

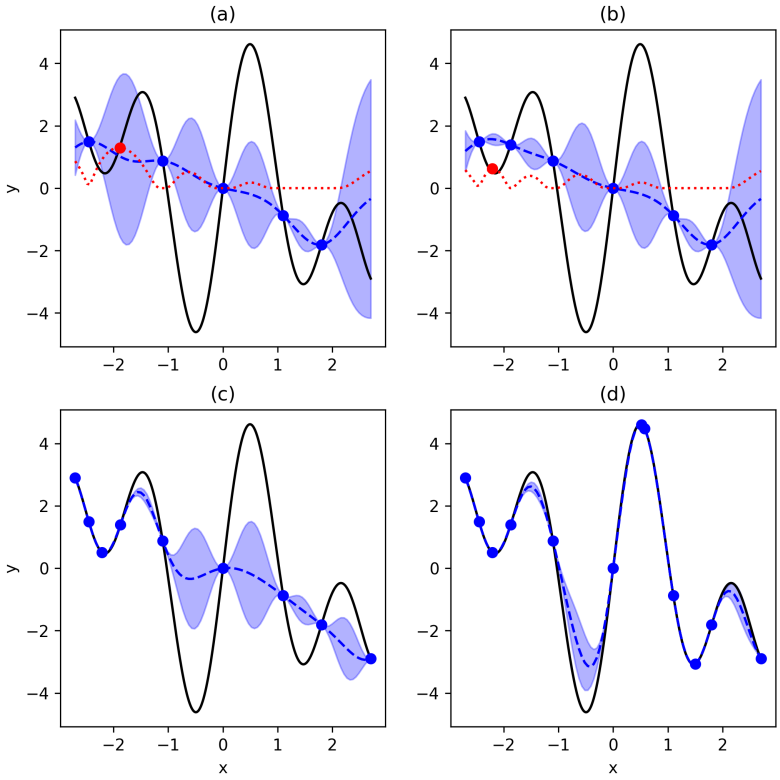


resultados obtidos. Para isso, são escolhidas estratégias que permitem otimizar uma função sem que seja conhecida sua forma ou seu gradiente (pode-se dizer que estamos otimizando uma função caixa-preta, que é a performance do modelo em função dos hiper-parâmetros da técnica utilizada). Uma proposta inicial é o uso de processos gaussianos ou o uso de estimadores Parzen (Bergstra et al., 2011). Um processo gaussiano é um processo estocástico tal que, para cada coleção finita de valores de y — $Y = y_1, y_2, ..., y_n$ — segue uma distribuição normal multivariada. Sendo $y = f(x)$ a relação entre os pontos $y_1, y_2, ..., y_n$ é estabelecida em função da proximidade dos pontos $x_1, x_2, ..., x_n$ representando a matriz de co-variância. Essa representação pode ser feita usando uma diversidade de funções de *kernel*. Uma abordagem completa do uso de processos gaussianos para modelagem pode ser vista em Williams e Rasmussen (2006). Para otimizar uma função usando um processo gaussiano, são utilizadas inicialmente um conjunto de configurações aleatórias. A partir destas posições, é estimado o possível ganho para cada ponto do espaço que será explorado, e o ponto com maior expectativa de melhora. O algoritmo então avalia essa posição, colhendo novas informações sobre a performance do modelo no espaço de hiper-parâmetros. São testadas então sequencialmente as posições nas quais existe uma maior expectativa de melhora, até que um certo número pré-fixado de configurações sejam testadas. Esse procedimento é ilustrado na Figura 6.

1.3 Análise de sensibilidade

O objetivo da análise de sensibilidade é quantificar como a variabilidade da saída de um modelo pode ser relacionada a variações nas suas entradas e parâmetros. Essas variações podem ser intencionais, potenciais incertezas ou mesmo a suscetibilidade a erros de medição e estimativa. De forma geral, o procedimento é utilizado para quantificar

Figura 6 – Processo gaussiano para buscar x que maximiza $y = f(x)$ (curva preta contínua). Na inicialização do algoritmo, a função é avaliada em quatro pontos e o modelo é ajustado (curva azul tracejada). O ganho esperado (curva vermelha pontuada) é calculado e o ponto de maior ganho esperado é avaliado na primeira iteração (b), quando então o ganho esperado é recalculado. Algoritmo após quatro iterações (c) e depois de mais duas iterações (d).



como variações das entradas ou parâmetros afetam a saída de um modelo. Será utilizada a denominação fatores para se referir a entradas ou parâmetros de um modelo de forma genérica. Embora, neste trabalho, seja enfatizado o uso da análise de sensibilidade para quantificar o efeito da variação das diferentes entradas na saída do modelo, o estudo do efeito dos parâmetros de modelos é análogo. Contextos típicos de uso e motivações possíveis para análise de sensibilidade são (Saltelli et al., 2008, Cap. 1):

- Priorização de fatores: Identificar fatores com maior influência na variabilidade da saída de um modelo.
- Fixação de fatores: Identificar fatores que podem ser fixados sem que o modelo perca sua capacidade de representação.
- Corroborar um modelo: Identificar quão robusta é a saída de um modelo ou quanto ela pode ser dependente de um valor específico de um fator.

- Priorização em pesquisa: Qual fator deve ser mais bem estudado ou medido?
- Simplificação de modelos: Indicar se algum comportamento de um modelo pode ser simplificado ou assumido constante.
- Identificar regiões críticas: Verificar quais regiões do espaço de entradas de um modelo levam a valores extremos.
- Antes de estimativas: Identificar a região do espaço de fatores na qual um modelo é mais sensível a um fator que se deseja estimar ou medir.

Uma primeira distinção entre formas de análise de sensibilidade diz respeito ao quanto o espaço de fatores é explorado, podendo ser local ou global. Métodos locais tendem a serem mais rápidos e simples. Exemplos de métodos numéricos são derivadas numéricas e a análise da variação de um fator por vez. No caso da derivada numérica, o modelo é perturbado com uma pequena variação de um fator e é quantificado o efeito na saída do modelo em relação à variação do fator $((f(X_i + \delta X) - f(X_i))/\delta X)$. Na variação de um fator por vez, um ponto é fixado (por exemplo, todas as entradas são configuradas para seus valores médios) e então cada fator é variado e é medida a variação da saída do modelo. Esses métodos possuem sérias limitações quando o modelo utilizado não é linear ou não é um modelo aditivo. Além disso, a caracterização será feita na vizinhança do ponto inicial de análise e pode não ser representativa da influência do fator ao longo do espaço de fatores. A realização deste tipo de análise para modelos complexos pode levar a resultados espúrios (Saltelli e Annoni, 2010).

Para suprir as limitações de métodos locais, uma série de métodos considerados globais foi proposta. No geral, grande parte dos métodos globais são baseados na medição do efeito de um fator na variância da saída de um modelo ou em princípios de teoria de informação (Borgonovo e Plischke, 2016). De acordo com (Saltelli et al., 2008, cap. 1), partindo de um modelo f de uma quantidade Y em função de k variáveis:

$$Y = f(X_1, X_2, ..., X_k) \tag{1.1}$$

é possível verificar a importância de uma variável X_i de entrada fixando-a em um valor x_i^* e analisando o efeito disso na variabilidade de f. Será assumido ao longo desta seção que estas variáveis tem distribuição normal, são mutuamente independentes e que cada uma pode ser amostrada da sua distribuição marginal. Para isso, seja $V_{X_{\sim i}}(Y|X_i = x_i^*)$ a variância de Y dado $X_{\sim i}$ (todas as variáveis exceto X_i). Essa variância é dita condicionada em $X_i = x_i^*$. Dado que uma fonte de incerteza (X_i) é fixada, é esperado que a variância de Y diminua. Caso a variância condicional seja pequena em relação à variância original (variância não condicionada de Y), o efeito de X_i em Y é grande. Caso a variância

condicional seja próxima da variância original, podemos concluir que o efeito de X_i em Y é pequeno. Embora essa métrica possa ser considerada uma medida da sensibilidade de Y em relação a X_i, ela ainda é dependente de x_i^* e sujeita a casos patológicos em que a variância condicionada pode ser maior que a original em função do ponto x_i^* escolhido (Saltelli et al., 2008, pag. 48). Tomando a média dessa quantidade para todos os valores de x_i possíveis[7], a dependência de x_i desaparece e temos $E_{X_i}(V_{X_{\sim i}}(Y|X_i))$, que é sempre menor ou igual a $V(Y)$, dado que[8]:

$$E_{X_i}(V_{X_{\sim i}}(Y|X_i)) + V_{X_i}(E_{X_{\sim i}}(Y|X_i)) = V(Y) \tag{1.2}$$

A variação condicional $V_{X_i}(E_{X_{\sim i}}(Y|X))$ é chamada de efeito de primeira ordem de X_i em Y e permite definir:

$$S_i = \frac{V_{X_i}(E_{X_{\sim i}}(Y|X_i))}{V(Y)} \tag{1.3}$$

como o índice de sensibilidade de primeira ordem de X_i em Y, que é um número limitado entre 0 e 1.

Considerando a possibilidade de modelos não-aditivos, é possível também estudar o efeito de uma variável e sua interação com as demais variáveis. Enquanto no caso aditivo a soma dos efeitos de primeira ordem é igual a 1, isso não é verdade na presença de interações. Nestes casos, uma variável passa a ter importância em função da sua interação com as demais e pode ser necessário analisar os termos de interação. Em um modelo com quatro variáveis, a importância da variável X_1 seria dada portanto por:

$$S_{T1} = S_1 + S_{12} + S_{13} + S_{14} + S_{123} + S_{124} + S_{134} + S_{1234} \tag{1.4}$$

sendo os índices S_{12}, ..., S_{1234} obtidos condicionando simultaneamente nas respectivas variáveis. Enquanto o cálculo do índice S_{T1}, chamado de efeito total de uma variável possa ser computacionalmente intensivo se forem calculadas as interações de muitas variáveis, é possível explorar a equação 1.2 para o cálculo direto. Re-arranjando os termos, temos:

$$E_{X_i}(V_{X_{\sim i}}(Y|X_i)) = V(Y) - V_{X_i}(E_{X_{\sim i}}(Y|X)) \tag{1.5}$$

a quantidade da variância de y que seria explicada, em média, caso fossem conhecidos todos os valores de $X_{\sim i}$. Dividindo por $V(y)$ e re-arranjando os termos, tem-se:

$$S_{Ti} = \frac{E_{X_i}(V_{X_{\sim i}}(Y|X_i))}{V(y)} = 1 - \frac{V_{X_i}(E_{X_{\sim i}}(Y|X_i))}{V(y)} \tag{1.6}$$

[7] Uma forma usual desta avaliação é fixar o valor de x_i e avaliar a função em diversos valores de $x_{\sim i}$. Esse procedimento é repetido então para um conjunto de valores x_i ao longo do domínio da variável.

[8] Esta equação decompõe a variância da saída de uma função na variação causada por uma variável X_i e a variação causada por todas as demais variáveis $X_{\sim i}$. Maiores detalhes podem ser vistos em Saltelli et al. (2008, pag. 20-22)

cujo cálculo numérico pode ser otimizado utilizando diversas estratégias (Saltelli et al., 1999; Sobol, 2001).

1.4 Estatística de Kolmogorov-Smirnov para análise de sensibilidade: o índice *PAWN*

Ao caracterizar a variabilidade da saída de um modelo a partir de sua variância, estamos caracterizando uma distribuição em função do seu segundo momento [9], o que pode ser errôneo na presença de assimetria ou multi-modalidade, entre outras condições (Pianosi e Wagener, 2015). Índices de sensibilidade baseados na função densidade de probabilidade [10](FDP) são alternativas mais adequadas no cenário de distribuição fortemente assimétrica, porém apresentam uma complexidade associada à representação adequada da FDP. Uma alternativa de fácil implementação é o índice PAWN, proposto por Pianosi e Wagener (2015). O índice utiliza a função de distribuição acumulada [11](FDA) para representar a distribuição da saída de um modelo. Para quantificar o efeito de uma variável X_i na saída de um modelo, é feita a comparação da FDA da saída do modelo ($F_y(y)$) com a FDA condicionada em algum valor de $X_i = x_i^*$ ($F_{y|X_i=x_i^*}(y)$) utilizando a estatística de Kolgomorov-Smirnov (KS), que mede a distância máxima entre duas FDA[12]:

$$KS_{x_i^*} = max|F_y(y) - F_{y|X_i}(y)| \tag{1.7}$$

Como a FDA condicionada caracteriza a variabilidade da saída do modelo fixando o valor da variável X_i e, portanto, removendo a incerteza associada, a distância entre a FDA condicionada e original caracteriza o efeito da variável na variabilidade da saída do modelo. O caso limite inferior é quando a variável não tem importância nenhuma e a FDA condicionada coincide com a FDA original. Neste caso, a distância entre a FDA original e condicionada é zero e pode-se concluir que X_i não tem influência sobre y. Conforme aumenta a sensibilidade de y em relação a X_i, maior será a distância observada. Na Figura 7 é ilustrada a diferença entre variáveis com diferentes médias. Como a FDA condicionada é dependente do valor de x_i^*, é tomado como índice de sensibilidade o valor agregado da distância KS para vários valores de X_i. Um estatística apresentada pelos autores é uso da mediana, embora o valor médio ou máximo sejam alternativas.

[9] Uma apresentação destes e outros conceitos estatísticos que serão utilizados nesta seção pode ser vista em Kerns (2010).

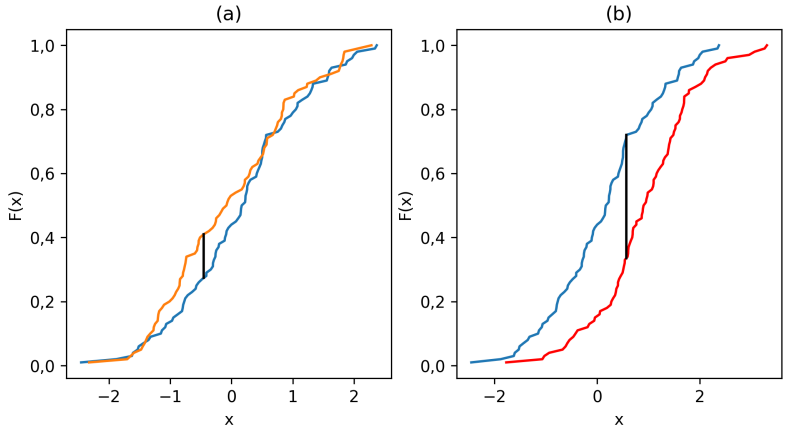
[10] A função de densidade de probabilidade de uma variável aleatória X associa cada valor de X a uma probabilidade de sua ocorrência.

[11] A função de distribuição acumulada de uma variável X associa para cada valor X^* a probabilidade de X assumir um valor igual ou menor que $X*$.

[12] Para este cálculo, a variável X_i é fixada em um valor x_i^* e as demais são variadas.

Dadas as dificuldades associadas ao cálculo analítico do índice, é apresentado o uso da FDA empírica para cálculo da análise de sensibilidade. O método também permite a análise de interações entre duas variáveis. Neste caso, é avaliada a FDA condicionada simultaneamente nas duas variáveis sendo analisadas.

Figura 7 – Ilustração da distância KS entre variáveis com pouca diferença (a), apresentando distância KS de 0,14, e grande diferença (b), apresentando distância KS de 0,39. Na Figura (a), 100 amostras de uma variável normal com média 0,1 (azul) são comparadas com outras 100 amostras de uma variável de média -0,1 (laranja). Na Figura (b) a comparação é entre variáveis com média 0,1 (azul) e 1,0 (vermelha).



1.5 Inspeção visual de modelos

Embora a apresentação da análise de sensibilidade tenha enfatizado a abordagem quantitativa, a inspeção de gráficos da saída do modelo em função das suas entradas é uma alternativa simples para entender o comportamento de um modelo em função de uma variável. Saltelli et al. (2008, Cap. 1) apresentam o uso de gráficos de dispersão da saída de um modelo em função de cada variável para inspeção. Variáveis com pouca importância tendem a apresentar uma dispersão quase uniforme de pontos ao longo do seu domínio. Conforme a importância de uma variável aumenta, passa a ser possível observar a formação de padrões ao longo do domínio de uma variável. Duas outras formas de apresentar o efeito de uma variável de forma visual são as curvas de resposta parcial e as curvas de esperança condicional individual.

Partindo de um modelo de regressão $y = f(X_1, X_2, ..., X_i, ..., X_k)$, uma alternativa de visualizar o efeito de uma variável X_i é fixar os valores das demais variáveis em um ponto de interesse $X_{\sim i} = x^*_{\sim i}$ e variar o valor de X_i observando sua resposta. Essa resposta caracterizaria a resposta local na vizinhança de $x^*_{\sim i}$, sujeita às mesmas críticas apresentadas para os métodos locais de análise de sensibilidade. De acordo com Hastie et al. (2009, p. 369), uma forma de representar a resposta global do modelo é marginalizar

a resposta em relação a $X_{\sim i}$, obtendo a esperança da resposta de f em função de X_i. A resposta parcial então é:

$$f_i(X_i) = E_{X_{\sim i}} f(X_i, X_{\sim i}) \tag{1.8}$$

que pode ser estimada por:

$$\hat{f}_i(X_i = x_i^*) = \frac{1}{N} \sum_{j=1}^{N} f(X_i = x_i^*, X_{j \sim i}) \tag{1.9}$$

em que $X_1 \sim i, X_2 \sim i, \ldots, X_N \sim i$ são os valores de $X_{\sim i}$ que ocorrem no conjunto de treinamento. Conforme destacado por Hastie et al. (2009), a resposta parcial como foi definida representa o efeito de X_i em y após considerar o efeito de $X_{\sim i}$ em y. Os autores destacam que as curvas de resposta parcial são de grande valia para entender modelos com interações de baixa ordem. A resposta parcial de um modelo SVR ajustado em um conjunto de dados sintéticos pode ser visto na Figura 8.

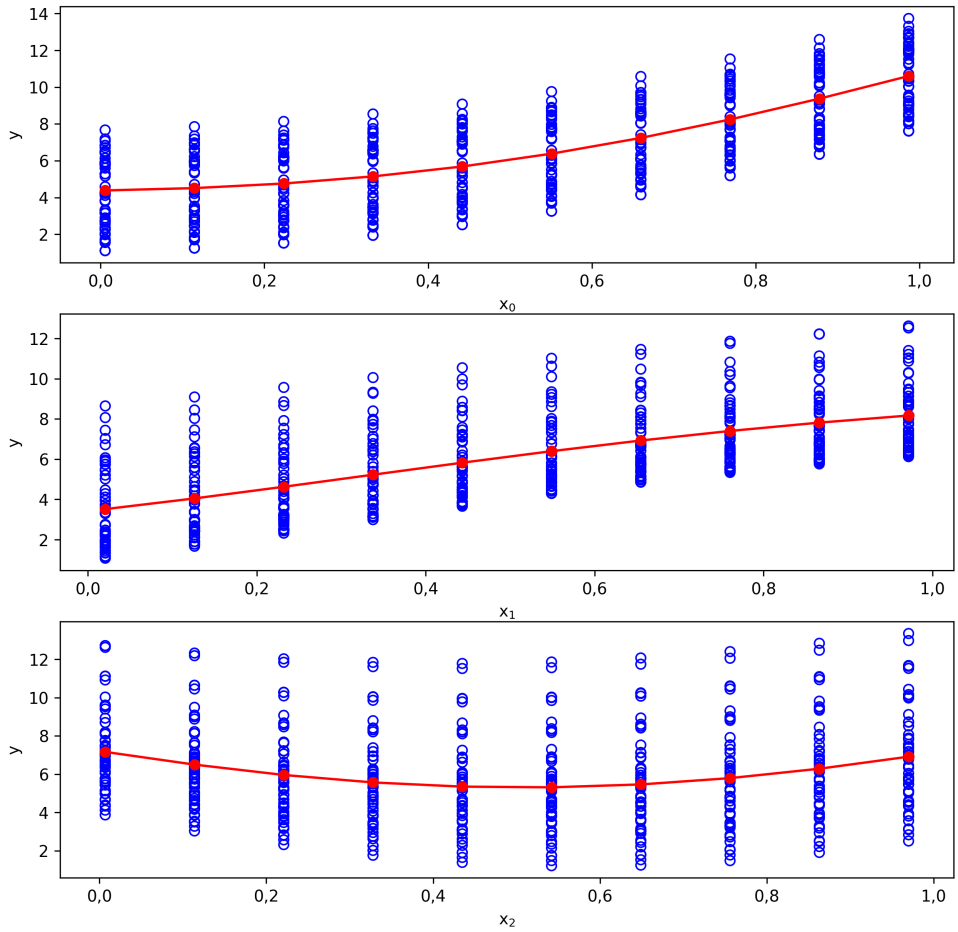
Na presença de interações, a agregação produzida pode ocultar efeitos gerados por interações. Como alternativa, existem as curvas de esperança condicional individual (Goldstein et al., 2015). Para produção do gráfico de curvas de resposta condicional, são produzidas as curvas de resposta ao longo de X_i para cada valor de $X_{\sim i}$. No geral, as curvas são enriquecidas com cores ou símbolos representando outras variáveis do conjunto de dados para caracterizar da interação com outras variáveis na saída do modelo. Um exemplo de curva de esperança condicional pode ser visto na Figura 9. Quando é utilizado um único valor $x_{\sim i}^*$, temos a curva condicional a $x_{\sim i}^*$, que é uma exploração local de f na vizinhança de $x_{\sim i}^*$ conforme inicialmente discutido.

1.6 Seleção de variáveis e sensibilidade de modelos

Considerando a construção de modelos baseados em dados, podemos estabelecer uma relação entre a importância de uma variável no conjunto de dados e a sensibilidade às diferentes variáveis de entrada de um modelo. A importância da variável quantifica a relação entre as grandezas conforme apresentado nos dados, enquanto a análise de sensibilidade quantifica a relação entre as grandezas conforme estabelecido pelo modelo analisado. Em um cenário em que essas medidas sejam comparáveis, podemos analisar a importância das variáveis como uma informação a *priori* e a sensibilidade à variável como uma informação a *posteriori* da modelagem.

Considerando que métodos baseados em regressão linear ou correlações não são indicados para análise de sensibilidade em cenários de modelos complexos, em especial na presença de respostas não lineares e interações (Saltelli e Annoni, 2010), o uso de um

Figura 8 – Ilustração da resposta parcial de um modelo SVR ajustado nos conjunto de dados gerado por $y = exp(2 \cdot x_0) + 5 \cdot x_1 + 10 \cdot (x_2 - 0.5)^2 + \epsilon, \epsilon \sim N(0,\ 0, 5)$. Os círculos azuis representam as combinações dos valores de $X_{\sim i}$ para os diferentes valores de X_i avaliados, enquanto o ponto vermelho representa as médias que formam a resposta parcial.

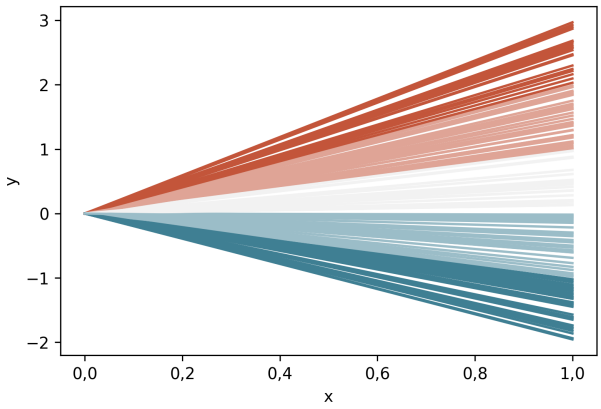


método comum entre análise de sensibilidade e seleção de variáveis deve ser direcionado para métodos de seleção de variáveis análogos aos métodos de análise de sensibilidade global.

1.6.1 O índice *PAWN* para seleção de variáveis

Partindo de um conjunto de dados existente, Pianosi e Wagener (2015) sugerem que o índice *PAWN* poderia ser adaptado com base em intervalos pré-especificados das variáveis independentes X_i para condicionar a variável dependente y. Nesse caso, $F_y(Y)$ é estimada para todas as amostras do conjunto de dados, enquanto $F_{y|x_i^*}(y)$ é estimada para os valores de y tal que os valores de X_i estão contidos em um intervalo centrado

Figura 9 – Ilustração da resposta condicional independente de um modelo $y = (b_0 + b_1) \cdot x, b_0 \in [-2, -1, 0, 1, 2], b_1 \sim U(0, 1)$ em função de x e colorida por b_0.



em x_i^*. Assim, X_i pode ser dividido em diversos intervalos e o valor da distância KS obtida entre $F_y(y)$ e $F_{y|x_i^*}(y)$ para os vários intervalos pode ser agregado. Os autores não especificam a metodologia para determinação dos intervalos, e sugerem o uso do valor médio ou mediano na maior parte dos casos para determinação do valor agregado. O valor agregado da distância KS é utilizado como medida da importância da variável. O algoritmo 1.1 apresenta o cálculo da importância de variáveis utilizando o índice *PAWN*. Como valores maiores representam uma maior relevância da variável, o índice permite então que as variáveis sejam ordenadas em relação à sua importância.

Pseudocódigo 1.1 Cálculo do índice PAWN para seleção de variáveis. É assumido que estão disponíveis as funções FDA(y) para calcular a função de distribuição acumulada de uma variável y, mediana(y) para calcular a mediana de uma variável y e KS(Y_1, Y_2) para calcular a estatística de Kolmogorov-Smirnov entre duas funções de distribuição acumuladas Y_1 e Y_2

Entrada: matriz X com dimensões $n \times m$, Y com comprimento n
  $FY \leftarrow FDA(Y)$
  $PAWN \leftarrow$ lista vazia
  para $i = 1$ até m **faça:**
    $listaKS \leftarrow$ lista vazia
    $X_i \leftarrow$ a coluna i de X
    $intervalos \leftarrow$ intervalos equiprováveis de X_i
    para $intervalo$ em $intervalos$ **faça:**
      $y|Xi \leftarrow$ listar todos y_j em Y se x_j em X_i está contido no $intervalo$
      $FY|X_i \leftarrow FDA(y|Xi)$
      $KS_{intervalo} \leftarrow KS(FY, FY|X_i)$
      acrescente $KS_{intervalo}$ em $listaKS$
    fim para
    acrescentar mediana($listaKS$) em $PAWN$
  fim para
  Retornar: PAWN

1.7 Síntese

É hipótese desta tese que as interações entre a adubação nitrogenada e características do solo, práticas de manejo e condições meteorológicas podem ser capturadas por técnicas de aprendizado de máquina aplicadas em conjuntos de dados de produção de cana-de-açúcar. É, então, objetivo desta tese desenvolver modelos produtividade de cana-deçaúcar utilizando técnicas de aprendizado de máquina, e avaliar a interação da adubação nitrogenada com as demais variáveis. O índice *PAWN* de análise de sensibilidade adaptado para avaliar a importância de variáveis em conjuntos de dados, conforme apresentado na seção 1.6, foi utilizado e comparado com os resultados de sensibilidade das variáveis de entrada do modelo. Para explorar as interações no modelo, foram utilizados gráficos de resposta condicional individual. Como objetivo secundário de mitigar o aspecto de "caixa-preta"dos modelos gerados, buscou-se interpretar os modelos gerados a partir da importância das variáveis no conjunto de dados, sensibilidade das variáveis e gráficos de resposta parcial. Analisando os resultados, constatou-se que o padrão de respostas individuais não apresenta respostas consistentes para a produção de cana-de-açúcar, embora as respostas gerais sejam mais coerentes. Considera-se então que não é recomendável utilizar a saída de modelos gerados utilizando as técnicas empregadas neste trabalho para análises de respostas individuais, o que seria por exemplo, necessário para recomendação de adubação para cada talhão de cana-de-açúcar. Usos pautados pela resposta geral parecem não ser afetados e devem ser avaliados em trabalhos futuros.

2 Metodologia

2.1 Dados

2.1.1 Dados de produção de cana-de-açúcar

Os dados utilizados foram fornecidos pela empresa Atvos[1] e são referentes ao plantio, histórico de produção, área, aplicação de insumos e contorno geo-referenciado das áreas de produção em quatro unidades da empresa para as safras de 2011 a 2015. Estão disponíveis dados referentes às unidades Usina Conquista do Pontal (UCP, Mirante do Paranapanema, SP), Usina Costa Rica (UCR, Costa Rica, MS), Usina Rio Claro (URC, Caçú, GO) e Usina Santa Luzia (USL, Nova Alvorada do Sul, MS). Dado que são necessários dados anteriores ao início do ciclo de crescimento da planta para que seja realizada a modelagem da produtividade, foi modelada a produtividade nos anos de 2012 a 2015. A representação das bases no conjunto de dados foi padronizada para representar a produtividade na colheita de cada talhão para cada safra. Cada produtividade é associada a um ciclo de crescimento que vai do plantio à primeira colheita para cana planta e da rebrota (colheita anterior) à colheita para soqueiras. Esse ciclo de crescimento é utilizado para descrever as condições meteorológicas durante o crescimento da cana-de-açúcar, assim como alocar os insumos que são referentes a esse ciclo de crescimento.

Os dados referentes ao histórico de produção, área, plantio e manejo aplicado foram fornecidos na forma de arquivos csv [2]. As informações de plantio foram utilizadas para atribuir o início do primeiro ciclo de crescimento e a variedade utilizada. Foram incluídos como insumos aplicados no primeiro ciclo os insumos aplicados desde a colheita do último cultivo na área antes do plantio corrente. Com isso, são consideradas as aplicações de insumo que são realizadas antes do plantio, durante a re-sistematização da área. Para os ciclos subsequentes, os insumos são atribuídos a um ciclo se ele é aplicado depois da colheita anterior e antes da colheita corrente. Foram considerados apenas os insumos de correção de solo, adubação, vinhaça, torta de filtro e irrigação.

O contorno georreferenciado das áreas (arquivo *shape*) foi fornecido para parte das áreas, sendo excluídas as áreas para as quais o contorno não estava disponível, dado que isso inviabilizaria a caracterização das condições meteorológicas, conforme será detalhado na seção 2.1.2. O arquivo continha também a classificação do solo predominante nos talhões, até o segundo nível hierárquico do Sistema Brasileiro de Classificação de Solos

[1] Odebrecht Agroindustrial até 12 de Dezembro de 2017

[2] *comma separated values*, valores separados por vírgulas em tradução livre

Tabela 1 – Distribuição da classificação (ordem e sub-ordem) do solo dos talhões conforme o Sistema Brasileiro de Classificação de Solos.

Sigla	Ocorrências	Descrição
LV ou LVA	6982	Latossolo Vermelho ou Vermelho Amarelo
RQ	6380	Neossolo Quartzarênico
LA	285	Latossolo Amarelo
PV	243	Argissolo Vermelho
NV	171	Nitossolo Vermelho
GX	166	Gleissolo Háplico
CX	114	Cambissolo Háplico
FF	20	Plintossolo Pétrico

(Santos et al., 2006). Devido à dispersão de diferentes classificações de solo nos talhões da usina, as classificações foram agrupadas em três grupos. Os Neossolos permaneceram como um grupo (categoria N), enquanto os Latossolos e Argissolos foram agrupados (categoria LP). As demais classificações de solo foram agrupadas como 'Outros' (categoria O). Um procedimento de agrupamento similar foi realizado para as variedades de cana-de-açúcar que ocorriam em menos de 200 talhões e para os ambientes de produção A e B, dado a baixa ocorrência destes ambientes no conjunto de dados.

Informações adicionais de solo continham a classificação do ambiente de produção conforme a metodologia proposta por Demattê e Demattê (2009), além da categorização da fertilidade e textura do solo, conforme apresentado no apêndice A.

Para parte dos talhões, estavam disponíveis dados de análise da química do solo. Foi considerado que a condição química do solo vai impactar o ciclo de crescimento subsequente à realização da análise. Dessa forma, os resultados da análise química foram adicionados como variáveis preditoras para o ciclo seguinte à análise. Para parte dos talhões, também estavam disponíveis dados da análise textural do solo dos talhões. Os dados de textura do solo foram utilizados como variáveis preditoras de um talhão para todos os anos safra, assumindo que não há mudança nessas propriedades ao longo do período utilizado para modelagem.

Dessa forma, quatro subconjuntos de variáveis estavam disponíveis, e considerando cada talhão-safra como um registro, o primeiro subconjunto tem 12057 registros sem nenhuma informação de solo (subconjunto 1), o segundo tem 10389 registros com a adição da classificação do solo no SiBCS, categorias de textura e fertilidade e ambiente de produção (subconjunto 2), o terceiro tem 1000 registros com adição da análise química do solo (subconjunto 3), e o quarto tem 542 com informações de textura (subconjunto 4). Os subconjuntos não são mutuamente exclusivos, e são entendidos como subconjuntos de subconjuntos. Os atributos adicionados complementam os atributos anteriores dispo-

Tabela 2 – Resumo das características dos subconjuntos de dados utilizados.

Subconjunto	Amostras	Variáveis
1	12057	Manejo e meteorologia
2	10389	Idem 1 + Classe de solo e variáveis categórias de solo
3	1000	Idem 2 + Variáveis de química do solo
4	542	Idem 3 + Variáveis de textura do solo

níveis, apresentando então um *trade-off* entre número de registros disponíveis e número de variáveis disponíveis. Um resumo das características dos subconjuntos de dados pode ser visto na Tabela 2.

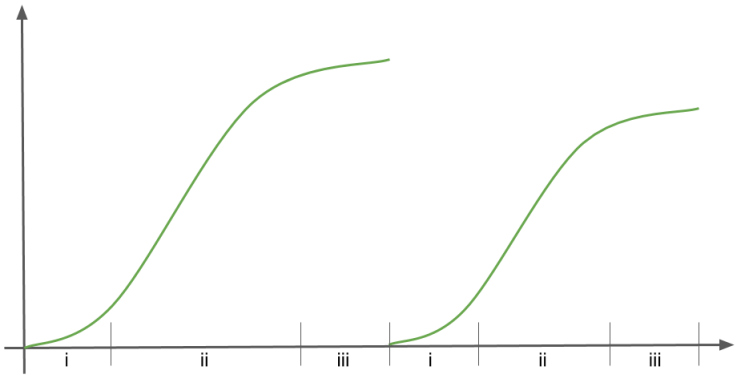
2.1.2 Dados meteorológicos

Mediante a indisponibilidade de dados de estações meteorológicas suficientemente próximas das áreas das usinas, optou-se pela caracterização das condições meteorológicas com base em dados obtidos por sensoriamento remoto. Foram utilizados dados de precipitação do sensor TRMM (Huffman et al., 2010) e dados de temperatura da superfície terrestre dos sensores MODIS Aqua (Wan et al., 2015a) e Terra (Wan et al., 2015b). Considerando que já existe boa concordância entre os valores de precipitação mensal no Brasil usando dados do TRMM (Melo et al., 2015), o uso dos dados agregados em períodos maiores, conforme será explicado na sequência, deve oferecer uma boa representação da precipitação acumulada. Os dados dos sensores Aqua e Terra foram combinados seguindo a metodologia proposta por Crosson et al. (2012), em que a diferença média de cada pixel durante um período de tempo é utilizada para compatibilizar a medição de temperatura dos dois sensores que é realizada em horários diferentes. Dados dos anos de 2007 a 2011 foram utilizados para estimar a média da diferença entre os sensores Aqua e Terra. A combinação de sensores foi empregada para minimizar a falta de dados em função da presença de nuvens. Embora o uso da temperatura do ar seja mais indicado para modelar respostas de plantas, a temperatura de superfície é uma aproximação efetiva para modelagem empírica de produtividade na ausência de dados de estações meteorológicas (Huang et al., 2015).

Os dados do sensor TRMM representam dados diários não agregados com tamanho de pixel de 0,25 graus, o que corresponde a aproximadamente 27,7 km na região da usina UCR (mais ao norte), sendo o valor dependente da latitude. Os dados de precipitação não apresentam valores faltantes. Os dados dos sensores MODIS utilizados correspondem a dados diários compostos a cada 8 dias, com tamanho de pixel de aproximadamente 1 km. Após a compatibilização, os dados dos sensores Aqua e Terra foram unificados. Quando um pixel apresentava valores para os dois sensores, o valor médio foi

utilizado. Quando apenas um dos sensores apresentava dados faltantes, era utilizado o dado do outro sensor. Essa estratégia é similar à empregada por Huang et al. (2015), exceto pela complementação com base em pixeis vizinhos para o caso de dados faltantes em ambos os sensores. Não houve estratégia de complementação de dados faltantes quando faltavam dados de ambos os sensores [3].

Figura 10 – Ilustração dos períodos utilizados para caracterizar a meteorologia durante o desenvolvimento da cultura. Os períodos *i* e *iii* no início e final do ciclo tem 90 dias de duração. O período *ii* tem duração variável em função dos diferentes ciclos de desenvolvimento. Para a cana planta, o primeiro ciclo se inicia no plantio, enquanto para a soqueira subsequente, o ciclo se inicia na colheita da cana planta.



Os dados meteorológicos foram agrupados em três variáveis representando três períodos de tempo distintos, conforme ilustrado na Figura 10. Um primeiro conjunto de variáveis foi criado para representar as condições meteorológicas nos primeiros 90 dias após o início do ciclo de crescimento. Dado que os ciclos de crescimento têm duração variável (entre 10 e 20 meses, com mediana de 11 meses), foi criado um conjunto de variáveis meteorológicas para representar as condições nos últimos 90 dias. O período compreendido entre os 90 dias iniciais e 90 dias finais é então caracterizado por um último conjunto de variáveis. Essa segmentação visa representar as condições meteorológicas no período de brotação e perfilhamento (primeiro período), crescimento (segundo período) e maturação (terceiro período). Entende-se que esta simplificação não corresponde ao desenvolvimento fenológico da cana-de-açúcar, porém representa uma simplificação do procedimento adotada em Bocca e Rodrigues (2016) onde os períodos possuiam comportamento variável em função da meteorologia esperada. Não houve perda de performance dos modelos em função da simplificação. Para cálculo dos atributos, foi assumido o intervalo aberto à direita para os intervalos de datas. Para representar o efeito da temperatura, foi utilizada a temperatura média no período ignorando os valores faltantes, enquanto, para precipitação, foi utilizada a soma da precipitação no período. Considerando que, para os dados de

[3] Para caracterização do efeito da temperatura na produtividade, foi utilizado o valor médio da temperatura em períodos determinados, ignorando os pontos faltantes da série.

temperatura de superfície, estão disponíveis os dados da temperatura diurna e noturna, em conjunto com os dados de precipitação para os três períodos foram gerados no total nove variáveis para representar as condições meteorológicas.

2.1.3 Variáveis disponíveis para modelagem

A Tabela 3 apresenta as variáveis disponíveis para modelagem, sua descrição e a identificação do código que aparecerá em tabelas no decorrer deste texto. Além das variáveis tabeladas, constam para o subconjunto 3 as variáveis soma de bases (sb), saturação de bases (v), capacidade de troca catiônica efetiva (ctce), capacidade de troca catiônica potencial (ctcp), e pH. Também estão disponíveis as concentrações de alumínio (al), magnésio (mg), potássio (k), acidez trocável (hal). No subconjunto 4 estão disponíveis as frações granulométricas de areia, silte e argila.

Tabela 3 – Apresentação das variáveis disponíveis no conjunto de dados e identificação utilizada.

Identificação	Descrição	Tipo
ambprod	Classificação das áreas de produção variando de AB (melhor) até E (pior)	categórico
fert	Classificação da fertilidade do solo de 1 (mais fértil) até 7 (menos fértil)	categórico
text	Classificação da textura do solo variando de 1 (mais arenoso) até 6 (mais argiloso) e 7 para siltoso	categórico
solo	Grupo de classificação do solo	categórico
espac	Indicador de espaçamento 1,5 m (padrão) ou 1,4 m	categórico
ncol	Número de colheitas desde o plantio	numérico
queima	Codificação indicando se a área foi queimada na colheita anterior	categórico
variedade	Variedade de cana de açúcar plantada	categórico
forman	Forma de nitrogênio aplicada (binaria para Nitrato e Ureia)	categórico
dfert	Número de dias entre início do ciclo de crescimento e a data da fertilização	numérico
qfert_{n,p,k}	Quantidade de nitrogênio, fósforo e potássio aplicados em $[kg \cdot m^{-2}]$	numérico
{q,d}agua	Lâmina (q) de irrigação aplicada [mm] e dias (d) desde início do ciclo para aplicação	numérico
{q,d}vin	Lâmina (q) de vinhaça aplicada [mm] e dias (d) desde início do ciclo para aplicação	numérico
qcalc	Quantidade de calcário aplicada $[kg \cdot m^{-2}]$	numérico
qgesso	Quantidade de gesso aplicada $[kg \cdot m^{-2}]$	numérico
qtorta	Quantidade de torta de filtro aplicada $[kg \cdot m^{-2}]$	numérico
ppt_{i,ii,iii}	Soma de precipitação [mm] nos primeiros 90 dias (i), últimos 90 dias (iii) e entre períodos i e iii (ii)	numérico
ts{d,n}_{i,ii,iii}	Média da temperatura diurna (d) e noturna (n) de superfície [K]. Períodos idênticos aos da precipitação	numérico

2.2 Modelagem

Em função da auto-correlação espacial inerente aos dados, optou-se por uma estratégia de validação cruzada com blocos espacializados (Roberts et al., 2017). O uso de uma estratégia convencional em que amostras são sorteadas aleatoriamente para partição dos conjuntos de dados distribui amostras próximas espacialmente (e, portanto, espacialmente correlacionadas) para os diferentes conjuntos, comprometendo a independência entre os conjuntos de treinamento e teste. Na modelagem de produtividade de cana-de-açúcar, negligenciar a auto-correlação faz com que as taxas de erro encontradas sejam substancialmente subestimadas (Ferraciolli et al., no prelo). As amostras foram

agrupadas seguindo os blocos de colheitas conforme indicado no histórico de produção [4]. O conjunto para treino do modelo foi composto por 75 % dos blocos, sendo os 25 % restantes alocados para o conjunto de teste. Para ajuste de parâmetros da modelagem, foi realizada a validação cruzada *k-fold* com 5 *folds* no conjunto de treino.

Foram utilizadas as implementações de BRT, RF e SVR com *kernel* rbf disponíveis no pacote *scikit-learn*, versão 0.19.1 (Pedregosa et al., 2011). Os atributos numéricos foram normalizados utilizando *Z-score* (média zero e variância unitária), usando a média e a variância do conjunto de treino. Os atributos categóricos foram binarizados na forma *one-hot-encoding*, em que uma variável binária é criada para cada valor da variável categórica codificada. Na binarização, foi omitida a variável binária que corresponde à primeira categoria da variável original, dado que a representação com todas as variáveis binárias é redundante. A seleção de variáveis utilizou o método proposto na seção 1.6, utilizando 10 intervalos equiprováveis e agregação utilizando a mediana para cálculo do índice *PAWN*. Os modelos foram avaliados utilizando o erro médio absoluto. Para referência, é reportado o erro ao predizer a produtividade do conjunto de teste utilizando-se a produtividade média do conjunto de treino. Essa estratégia será denominada como modelo nulo. Os hiper-parâmetros foram ajustados utilizando um processo gaussiano com o *kernel* Matérn (Williams e Rasmussen, 2006, Pag. 84). Foram utilizadas 15 avaliações aleatórias dos hiper-parâmetros seguidas de 5 buscas do processo gaussiano. A implementação de processo gaussiano do pacote scikit-optimize (Head et al., 2018) foi utilizada com as demais configurações com valores padrão. Além do ajuste de cada técnica, foi ajustado o número de atributos mais importantes, seguindo a importância dada pelo índice *PAWN* para seleção de variáveis utilizando uma distribuição discreta uniforme com no mínimo 10 atributos e máximo igual ao número de atributos total[5]. Detalhes para os parâmetros ajustados para cada técnica podem ser vistos na Tabela 4.

2.3 Avaliação dos modelos

Para análise de sensibilidade, o cálculo do índice PAWN foi implementado seguindo o algoritmo proposto por Pianosi e Wagener (2015). Para obtenção da FDA não-condicionada, foi feita a predição de todas as amostras do conjunto de treino. Para obtenção da FDA condicionada, foram determinados 10 valores de X_i igualmente espaçados entre os valores máximo e mínimo, e determinados os valores únicos de $X_{\sim i}$. Para cada valor de X_i foram feitas as predições com todos os valores únicos de $X_{\sim i}$. A mediana

[4] Um bloco de colheita é uma unidade gerencial que agrupa talhões próximos que foram colhidos sequencialmente. No geral, os talhões de um bloco são colhidos em um período inferior a uma semana e então as colhedoras são deslocadas para outra região onde será colhido outro bloco.

[5] Para máximo inferior ao número total, a exclusão de atributos seria forçada

Tabela 4 – Parâmetros ajustados para cada técnica (Téc.) e limites inferior (LI) e superior (LU) utilizados para sorteio das configurações aleatórias no ajuste dos modelos. Os parâmetros foram amostrados de uma distribuição contínua uniforme, exceto quando marcados por [1] para distribuição log-uniforme e [2] para valores discretos.

Téc.	Parâmetro	LI	LS
BRT	Número de árvores [2]	$1,0 \cdot 10^{2}$	$1,0 \cdot 10^{3}$
	Taxa de aprendizado	$1,0 \cdot 10^{-2}$	$2,5 \cdot 10^{-1}$
	Profundidade máxima[2]	$1,0 \cdot 10^{0}$	$6,0 \cdot 10^{0}$
	Mínimo de amostras para split	$1,0 \cdot 10^{-2}$	$1,0 \cdot 10^{-1}$
	Fração de amostragem	$1,0 \cdot 10^{-1}$	$1,0 \cdot 10^{0}$
	Fração de sorteio	$1,0 \cdot 10^{-1}$	$1,0 \cdot 10^{0}$
RF	Número de árvores[2]	$1,0 \cdot 10^{1}$	$1,0 \cdot 10^{3}$
	Fração de sorteio	$5,0 \cdot 10^{-3}$	$7,5 \cdot 10^{-1}$
	Mínimo de amostras para split	$1,0 \cdot 10^{-2}$	$5,0 \cdot 10^{-2}$
SVR	Custo [1]	$1,0 \cdot 10^{-1}$	$1,0 \cdot 10^{3}$
	Gamma[1]	$1,0 \cdot 10^{-4}$	$1,0 \cdot 10^{0}$
	Epsilon	$5,0 \cdot 10^{-2}$	$3,0 \cdot 10^{-1}$

dos 10 valores foi utilizada como importância da variável. Para avaliar a sensibilidade à interação entre as demais variáveis e a taxa de aplicação de nitrogênio, foi utilizada a combinação de 10 valores para o nitrogênio e 10 valores da variável analisada. Para os 100 valores resultantes da combinação, foram avaliados os resultados do modelo que então foram comparados com a FDA não condicionada, sendo utilizada a mediana destes valores como resposta. Isso não garante que a análise de sensibilidade tenha se limitado ao espaço observado no treinamento do modelo, pois é possível que combinações de X_i e $X_{\sim i}$ avaliadas não estejam presentes no conjunto de treino. As consequências ou potenciais impactos dessa condição não serão tratados neste trabalho.

Para inspeção visual dos modelos, foi implementada a metodologia proposta por Hastie et al. (2009, pag. 369) para resposta parcial dos modelos. Para inspeção da resposta condicional, para cada talhão-safra foram fixados os valores de todas as variáveis exceto o nitrogênio. Com os valores das outras variáveis fixados, a curva de resposta condicional independente é obtida com a variação da quantidade de adubação nitrogenada. Foram avaliados 10 valores igualmente espaçados dentro do intervalo observado nos dados de treinamento. Para explorar visualmente a interação, a análise é similar, porém são variados simultaneamente a taxa de adubação e a variável cuja interação será analisada. Na construção do gráfico, as curvas de adubação foram coloridas conforme a discretização da variável de interação em intervalos de igual frequência. Quando possível, foram utilizados cinco intervalos de igual frequência (divisão nos quintis), se não, foi utilizada a divisão nos tercis, ou ainda a divisão na mediana. A inspeção visual foi realizada para variáveis que se destacaram em função da análise de sensibilidade.

3 Resultados e discussão

3.1 Avaliação dos modelos

A configuração de hiper-parâmetros para cada técnica com o menor erro avaliado por validação cruzada *K-fold* no conjunto de treino foi utilizada para análise. Os hiper-parâmetros escolhidos para cada técnica são apresentados na Tabela 13 do apêndice B. Na Tabela 5, são apresentados os valores de MAE obtidos para os modelos nos diferentes subconjuntos de dados. Não houve grande diferença entre a performance dos modelos criados. Considerando o erro do modelo nulo de 1,64 kg· m^{-2} no subconjunto 1, 1,65 kg· m^{-2} no subconjunto 2, 1,54kg · m^{-2} no subconjunto 3 e 1,56 kg· m^{-2} no subconjunto 4, houve uma maior eficiência (relação entre o erro de teste e o erro nulo) de modelagem nos subconjuntos 1 e 2 do que nos subconjuntos 3 e 4 [1]. Com exceção do modelo gerado por RF no subconjunto 3, os erros dos modelos são próximos a 90 % do erro do modelo nulo, chegando a 100,15 % para o caso do modelo gerado pela SVR, indicando que este modelo é pior que a predição pela média. Para os subconjuntos de dados 1 e 2, o erro de validação cruzada no treinamento foi superior ao erro de teste, enquanto o contrário aconteceu para os subconjuntos 3 e 4. Nos subconjuntos 1 e 2, considerando a diferença entre os erros de validação cruzada e o de teste e o desvio padrão do erro da validação cruzada, pode-se admitir que os erros são suficientemente próximos para que a diferença possa ser desconsiderada, especialmente dado que os valores de desvio padrão são subestimados (Bengio e Grandvalet, 2004). Para os subconjuntos 3 e 4, todos os modelos tiveram um desvio padrão superior a 0,20, com exceção do modelo SVR para o subconjunto 4.

De forma geral, os valores do erro de teste foram pouco maiores que 1,30 para todas as técnicas nos diferentes subconjuntos, exceto o modelo SVR no subconjunto 4. O mesmo não ocorre para o erro no conjunto de treinamento, em que valores de erro menores foram obtidos para os conjuntos com menos registros. Podem explicar esse padrão um potencial *overfitting* no conjunto de treino, a despeito do procedimento de validação cruzada, ou o fato de que é mais fácil ajustar um modelo com menor erro de treinamento para um menor número de amostras (Abu-Mostafa et al., 2012, p. 66)[2].

Analisando o erro absoluto percentual médio (MAPE), também são observadas (Tabela 6) a tendência de diminuição do erro de validação cruzada no conjunto de

[1] Usualmente, a produtividade é reportada em toneladas por hectare, porém optou-se pelo uso da unidade compatível com o Sistema Internacional. Para obter os valores na escala usual, basta multiplicar os valores por 10.

[2] Não está sendo considerada a hipótese de que os modelos podem ter aprendido melhor, dado o aumento da diferença entre erro de treino e teste.

Tabela 5 – Erro absoluto médio (MAE [kg·m^{-2}]) para as diferentes técnicas nos diferentes subconjuntos de dados e relação entre erro de teste e erro do modelo nulo (Teste/Nulo, [%]). Erro médio do modelo no conjunto de treino (MAE-CV$_{\text{médio}}$) e o respectivo desvio padrão (MAE-CV$_{\text{desv.}}$) se referem à validação cruzada no conjunto de treino. O erro de teste (MAE-Teste) é o erro do modelo obtido no conjunto de treino e avaliado no conjunto de teste.

Modelo	Subconjunto	MAE-CV$_{\text{médio}}$	MAE-CV$_{\text{desv.}}$	MAE-Teste	Teste/Nulo
BRT	1	1,46	0,14	1,37	83,89
RF	1	1,44	0,16	1,35	82,50
SVR	1	1,45	0,15	1,36	83,35
BRT	2	1,46	0,17	1,37	83,65
RF	2	1,42	0,17	1,34	81,66
SVR	2	1,46	0,13	1,33	81,17
BRT	3	1,20	0,24	1,34	87,06
RF	3	1,17	0,27	1,27	82,55
SVR	3	1,24	0,21	1,37	89,48
BRT	4	1,16	0,21	1,32	90,74
RF	4	1,04	0,20	1,36	86,92
SVR	4	1,02	0,15	1,56	100,15

Tabela 6 – Erro absoluto percentual médio (MAPE [%]) para as diferentes técnicas nos diferentes subconjuntos de dados e relação entre erro de teste e erro do modelo nulo (Teste/Nulo, [%]). Erro médio do modelo no conjunto de treino (MAPE-CV$_{\text{médio}}$) e o respectivo desvio padrão (MAPE-CV$_{\text{desv.}}$) se referem à validação cruzada no conjunto de treino. O erro de teste (MAPE-Teste) é o erro do modelo obtido no conjunto de treino e avaliado no conjunto de teste.

Modelo	Subconjunto	MAPE-CV$_{\text{médio}}$	MAPE-CV$_{\text{desv.}}$	MAPE-Teste	Teste/Nulo
GBR	1	24,13	2,09	24,55	74,38
RF	1	22,83	2,34	21,42	64,92
SVR	1	24,19	2,60	21,26	64,43
GBR	2	23,72	2,10	22,05	68,91
RF	2	22,29	2,68	21,63	67,59
SVR	2	23,64	1,43	21,37	66,78
GBR	3	17,57	3,50	20,19	77,66
RF	3	17,52	4,62	19,15	73,64
SVR	3	16,79	2,87	20,72	79,71
GBR	4	18,35	4,78	20,16	77,55
RF	4	14,90	3,66	20,02	77,01
SVR	4	16,10	3,82	23,01	88,51

modelagem e a tendência de aumento no desvio padrão que foi observada para o MAE . Os modelos desenvolvidos com a RF obtiveram o melhor MAPE nos subconjuntos 2 a 4, apresentando pequena diferença para o modelo SVR que teve a melhor performance no subconjunto 1. Assim como ocorreu para o MAE, também não foi observado um decrés-

cimo dos erros no conjunto de teste em paralelo ao decréscimo no erro de modelagem. Os erros do modelo nulo para os subconjuntos 1 a 4 foram 32,73, 32,33, 26,47 e 26,33 % respectivamente. No caso do MAPE, a relação de performance dos modelos em comparação com o modelo nulo é mais favorável. Mesmo o modelo SVR no subconjunto quatro, que apresentava MAE maior que o modelo nulo, teve um desempenho melhor neste critério. Esse comportamento geral pode ser explicado pelo grande peso do erro percentual para talhões de baixa produtividade, que não está sendo compensando pelo menor erro percentual nos talhões de alta produtividade, um reflexo da assimetria da distribuição da produtividade.

A Figura 11 apresenta os valores reais e preditos no conjunto de teste para os diferentes subconjuntos. Ao observar a região delimitada pela banda construída a com o erro absoluto médio e a referência 1:1 e comparar com a distribuição dos pontos, nota-se que grande parte dos pontos está próximo da referência ou contida na região delimitada. Em alguns casos, porém, é possível observar valores de erros superiores a 7,0 $kg.m^{-2}$, o que é comparável com a amplitude observada na variação do atributo meta. Isso sugere uma distribuição de erros altamente assimétrica, com longa cauda à direita. Esse padrão de grandes erros ocorre principalmente para valores de produtividade real abaixo de 4,0 $kg.m^{-2}$ em um padrão de super-estimativa da produtividade. Para valores acima de 7,5 $kg.m^{-2}$, os modelos tendem a subestimar a produtividade. A tendência de superestimar valores baixos e subestimar valores altos é corroborada pela comparação entre a reta 1:1 e a linha de tendência obtida com a regressão linear entre valores reais e preditos. O comportamento de agrupar valores perto da média, isto é, a superestimativa de valores baixos e subestimativa de valores altos, sugere um modelo sub-ajustado ou incapaz de descrever a variabilidade da variável dependente.

Conforme apresentado na Tabela 7, as melhores configurações para modelagem não utilizaram todos os atributos disponíveis, sendo as exceções os modelos RF e SVR para o subconjunto de dados 3. No subconjunto 4, os modelos RF e SVR utilizaram todos os atributos exceto a variável binária da variedade SP803280. De forma geral, a seleção de atributos foi mais agressiva para os modelos BRT nos subconjuntos 3 e 4. Embora os algoritmos BRT e RF sejam *ensembles* de árvores, a seleção de atributos usando o índice PAWN se mostrou benéfica para a técnica BRT e não beneficiou a RF. Analisando a sensibilidade aos atributos nos subconjuntos 3 e 4, é possível notar que o modelo BRT tem uma sensibilidade maior às variáveis que os modelos RF, que foi ajustada em mais variáveis (Tabelas 16, no apêndice C).

Tabela 7 – Número de atributos excluídos pela seleção de atributos para cada técnica (BRT - *Boosted Regression Trees*, RF - *Random Forest*, SVR - *Support Vector Regression*) e número inicial de atributos disponíveis em cada subconjunto de dados (CD). [1]Com o menor número de registros, menos rótulos ocorrem nas variáveis categóricas, reduzindo o número total de atributos de modelagem em função do uso do *one-hot-encoding*.

CD	BRT	RF	SVR	Inicial
1	45	44	40	46
2	49	37	44	50
3	44	58	58	58
4	28	52	52	53[1]

3.2 Importância de variáveis e análise de sensibilidade

Nos subconjuntos de dados 1 a 3, as variáveis com maior importância estavam ligadas à irrigação, número de dias para aplicação de vinhaça e variedade utilizada. Os valores intermediários estavam ligados a adubação, número de colheitas, forma de nitrogênio aplicada, queima anterior e condições meteorológicas. Os valores mais baixos foram atribuídos a propriedades do solo, aplicações de gesso e calcário, quantidade de vinhaça e ambiente de produção. No último subconjunto de dados, não houve uma distinção clara de grupos de variáveis ao longo da lista. Embora as variáveis mais importantes estivessem ligadas à variedade utilizada, as variáveis subsequentes são os grupos de classificação do solo, variáveis ligadas a propriedades químicas do solo, manejo do solo e fertilização, e algumas variáveis ligadas às condições meteorológicas. Tabelas completas com a importância das variáveis, análise de sensibilidade de primeira ordem e análise da sensibilidade na interação com nitrogênio são apresentadas no apêndice C.

Os valores de sensibilidade obtidos se mostraram menores do que os valores de importância das variáveis. Considerando que o modelo deveria aproximar os padrões do sistema modelado, e que são medidas na mesma escala, com métodos similares, era esperado que os valores fossem próximos. Comparando os valores de importância de variável e sensibilidade das entradas do modelo com a reta 1:1 na Figura 12, os modelos SVR apresentaram a dispersão de pontos mais próxima da reta, enquanto os modelos RF apresentaram os menores valores de sensibilidade, apresentando um padrão horizontal próximo a zero. É possível notar também que os pontos se aproximam mais da reta 1:1 para os modelos criados para bases com mais registros. Considerando o aspecto das curvas de real e predito, é possível que o sub-ajuste do modelo esteja conectado com os baixos valores de sensibilidade em relação aos valores de importância de variáveis encontrados. Recomenda-se que futuros trabalhos explorem essa relação entre análise de sensibilidade e importância de atributos em conjuntos de dados diversos.

Não só os valores de importância de variáveis foram maiores que os valores encontrados para análise de sensibilidade como a ordem das variáveis se mostrou diferente. Na Tabela 8, são apresentadas as variáveis com maior sensibilidade para os modelos gerados nos subconjuntos de dados. Para o primeiro subconjunto de dados, a temperatura diurna média no segundo período e variáveis ligadas à disponibilidade hídrica constavam entre as mais importantes. Nos modelos de *ensemble*, as variáveis mais importantes foram variáveis de soma de precipitação, e no geral, ocorreram mais variáveis relacionadas às condições meteorológicas, enquanto para SVR foram a variável de tempo para aplicação de vinhaça e variáveis de variedades que estiveram no topo. No segundo subconjunto, a temperatura no segundo período também constou na lista para as técnicas RF e SVR. O número de dias para aplicação de vinhaça aparece em primeiro para as técnicas BRT e SVR, subindo da terceira e segunda posições para as técnicas respectivamente. Para o modelo BRT, variáveis de manejo basicamente dominam as 5 primeiras posições, enquanto condições meteorológicas representam as variáveis mais importantes para RF. No terceiro subconjunto de dados, consta apenas uma variável meteorológica, havendo grande dominância de variáveis de manejo, em especial para RF e SVR. No quarto subconjunto de dados, 3 variáveis relacionadas a variedades ocupam o topo da lista, junto com características de solo. Analisando as variáveis mais importantes, não é possível identificar que uma técnica em especial favoreça variáveis de algum grupo (solo, manejo ou condições meteorológicas). Para os modelos SVR, para os cinco primeiros atributos, variáveis binárias passaram a ser mais importantes que variáveis contínuas, o que não aconteceu para os *ensembles*. De forma geral, as listas de sensibilidade a variáveis indicam que, conforme as técnicas de modelagem exploram o espaço de variáveis com estratégias diferentes, a saída é modelada com diferentes ênfases nas variáveis disponíveis. Ainda assim, algumas variáveis se destacam na modelagem, visto a frequência com que algumas variáveis se mostraram entre as cinco de maior sensibilidade. Entre essas variáveis, destacam-se o número de colheitas e a temperatura diurna média no segundo período, que ocorreram seis vezes no total de doze listas. A variável de dias de aplicação de vinhaça apareceu 5 vezes na lista, enquanto a precipitação no primeiro período apareceu três vezes. Embora as variáveis adicionadas nos subconjuntos 2 a 4 tenham aparecido nas listas, não houve destaque ou consenso entre as variáveis adicionadas.

Analisando a importância das variáveis de quantidade e tipo de adubação nitrogenada e sua posição relativa na lista (Tabela 9), a quantidade de adubo aplicada se mostrou menos importante que o tipo nos subconjuntos 1 e 3, com pequena diferença no subconjunto 2. No subconjunto de dados 1, a sensibilidade ao nitrogênio atingiu o maior valor para o modelo SVR, com 0,12 de importância, seguido de importância 0,06 e 0,04 para RF e BRT respectivamente (Tabela 14, apêndice C). A forma do adubo nitrogenado

Tabela 8 – Variáveis de maior sensibilidade (sens.) para os modelos gerados nos diferentes subconjuntos de dados (CD) e técnicas. BRT - *Boosted Regression Trees*, RF - *Random Forest*, SVR - *Support Vector Regression*. Nome das variáveis listadas sob o nome da técnica.

CD	BRT	sens.	RF	sens.	SVR	sens.
1	colheitas	0,174	colheitas	0,178	tsd_ii	0,296
1	tsd_ii	0,156	tsd_ii	0,077	d_vin	0,283
1	d_vin	0,149	ppt_ii	0,074	colheitas	0,212
1	ppt_i	0,135	qfert_k	0,072	RB72454	0,199
1	d_torta	0,128	ppt_iii	0,070	RB845210	0,184
2	d_vin	0,206	colheitas	0,183	d_vin	0,267
2	colheitas	0,134	RB855453	0,088	RB966928	0,250
2	ppt_i	0,123	ppt_ii	0,086	tsd_ii	0,240
2	q_calc	0,114	tsd_ii	0,063	SP803280	0,211
2	q_gesso	0,104	ppt_i	0,051	q_irrig	0,197
3	qfert_p	0,078	qfert_n	0,089	SP803280	0,293
3	SP801842	0,077	qfert_k	0,089	q_irrig	0,257
3	tsd_ii	0,075	ctce	0,088	RB966928	0,226
3	ambprod_D	0,069	colheitas	0,063	queima	0,213
3	SP803280	0,059	q_vin	0,062	q_torta	0,206
4	Outras	0,135	d_vin	0,062	SP813250	0,246
4	areia	0,102	qfert_p	0,061	SP801842	0,226
4	RB835054	0,102	RB835054	0,049	RB835054	0,182
4	ctcp	0,097	tsd_iii	0,049	solo_R	0,132
4	d_gesso	0,094	argila	0,047	solo_O	0,131

teve importância baixa, ocupando posições baixas na lista. Estes resultados são similares nos outros subconjuntos, alternando o modelo com a maior sensibilidade, sendo RF no subconjunto 3 e BRT no subconjunto 4. Seguindo a tendência observada na análise geral, os valores de sensibilidade foram menores que os valores de importância observados.

Tabela 9 – Importância e posição da variável quantidade e tipo de fertilizante nitrogenado aplicado. Valores para número de cortes e temperatura diurna da superfície no segundo período apresentados para comparação.

Subconjunto	qfert_n	n_form	estagio	tsd_ii
1	0,15/17	0,20/11	0,17/15	0,07/34
2	0,14/16	0,13/17	0,16/13	0,09/30
3	0,28/20	0,44/10	0,24/25	0,24/24
4	0,34/10	–	0,33/13	0,33/11

Para avaliação do efeito de interação, foram analisados os valores de sensibilidade em relação ao maior valor entre a sensibilidade à quantidade de fertilizante nitrogenado utilizado e a variável de interação. Foram analisadas para cada subconjunto de dados

e para cada técnica as cinco interações que apresentavam a maior diferença absoluta ou relativa (Tabela 10). Para muitas combinações de técnicas e subconjunto de dados, houve grande interseção entre as interações com maior diferença absoluta e relativa. Dos doze modelos gerados (três técnicas, quatro subconjuntos), as maiores diferenças absolutas e relativas entre a sensibilidade da interação e o maior valor individual coincidiram para cinco modelos, e houve acréscimo de um atributo para três modelos. Outros três modelos tiveram uma lista de sete atributos, e o modelo SVR no subconjunto 3 gerou uma lista de nove atributos, sendo os resultados para variável ppt_i e SP835037 suprimidos da Tabela 10. Embora para cada técnica/subconjunto as variáveis de interação com maior ganho absoluto ou relativo tenham coincidido, isso não se refletiu quando foram comparadas as listas geradas para modelos no mesmo subconjunto de dados ou listas geradas para modelos gerados pela mesma técnica. Conforme variáveis foram adicionadas nos subconjuntos de dados, houve uma tendência de que esses atributos constassem nas interações com maior magnitude para os modelos RF. Isso ocorreu para as variáveis de textura e fertilidade do solo no subconjunto 2, ctc efetiva e alumínio no subconjunto 3 e argila no subconjunto 4. Para as técnicas no subconjunto 4, a fração de argila foi um atributo de grande interação.

Considerando os resultados das análises de sensibilidade das variáveis, temos como destaque as variáveis número de colheitas, temperatura diurna média no segundo período, soma da precipitação no primeiro período, dias entre início do ciclo e aplicação de vinhaça e efeito das variedades de cana-de-açúcar na produtividade. Em função dos resultados de importância da interação das variáveis, temos como destaque a taxa de aplicação de nitrogênio com as variáveis de número de colheitas, quantidade de potássio aplicada, somas de precipitação no primeiro e segundo período, e teor de argila. Essas variáveis serão inspecionadas em gráficos de resposta parcial e condicional. Será inspecionada também a variável da forma de adubação.

Figura 11 – Valores reais e preditos da produtividade de cana-de-açúcar pelas diferentes técnicas (colunas) e subconjuntos de dados (linhas) para os diferentes subconjuntos de dados. Em cada gráfico, a linha contínua indica a reta 1:1 enquanto as linhas tracejadas indicam a faixa delimitada pelo erro de teste de cada modelo conforme apresentado na Tabela 5. BRT - *Boosted Regression Trees*, RF - *Random Forest*, SVR - *Support Vector Regression*. Os três primeiros gráficos correspondem aos modelos do subconjunto 1, os três seguintes aos modelos do subconjunto 2 e assim sucessivamente.

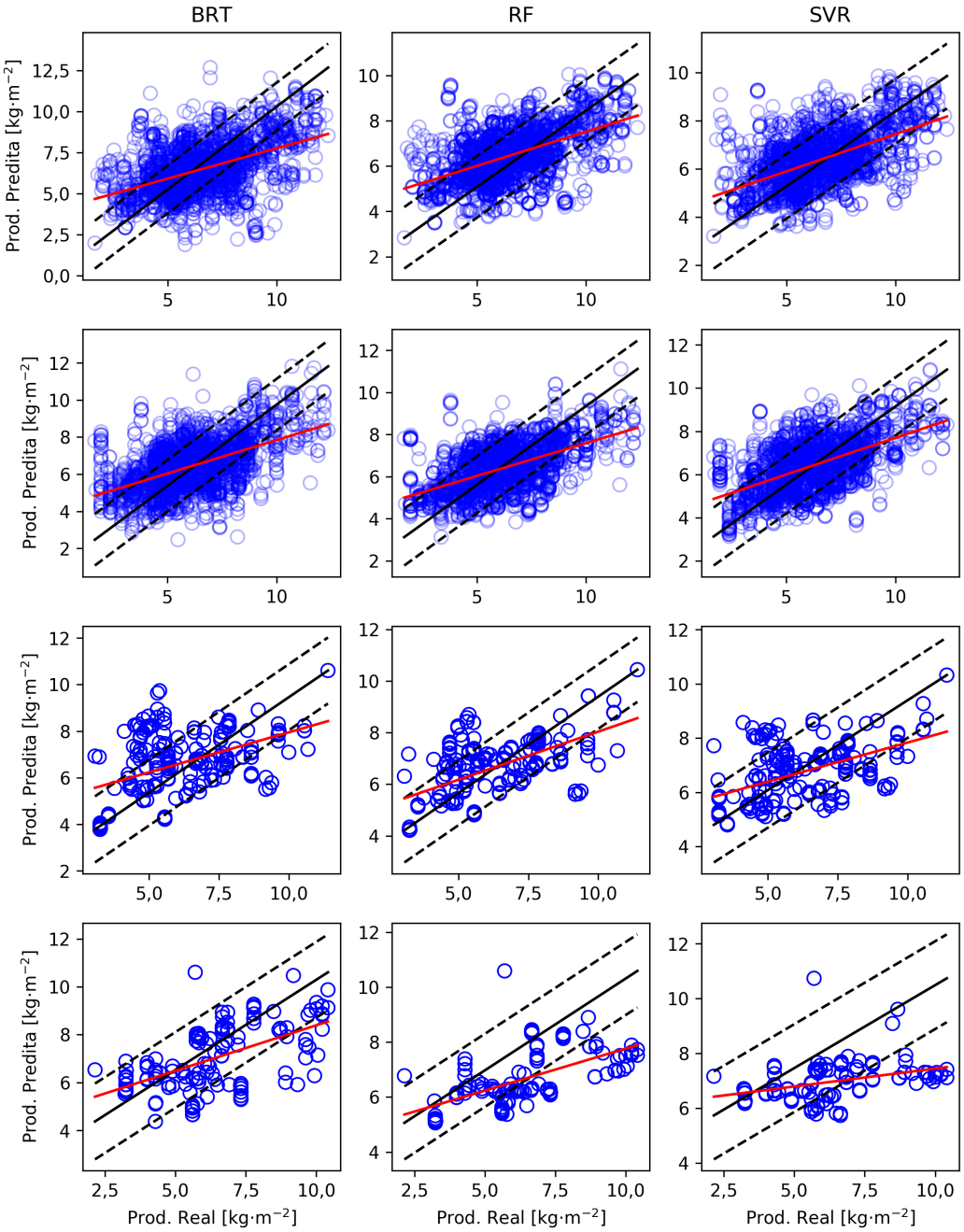


Figura 12 – Gráfico de dispersão dos valores de importância do atributo (eixo x) e sensibilidade aos atributos (eixo y) para as diferentes técnicas (colunas) e subconjuntos de dados (linhas). BRT - *Boosted Regression Trees*, RF - *Random Forest*, SVR - *Support Vector Regression*. Os três primeiros gráficos correspondem aos modelos do subconjunto 1, os três seguintes aos modelos do subconjunto 2 e assim sucessivamente.

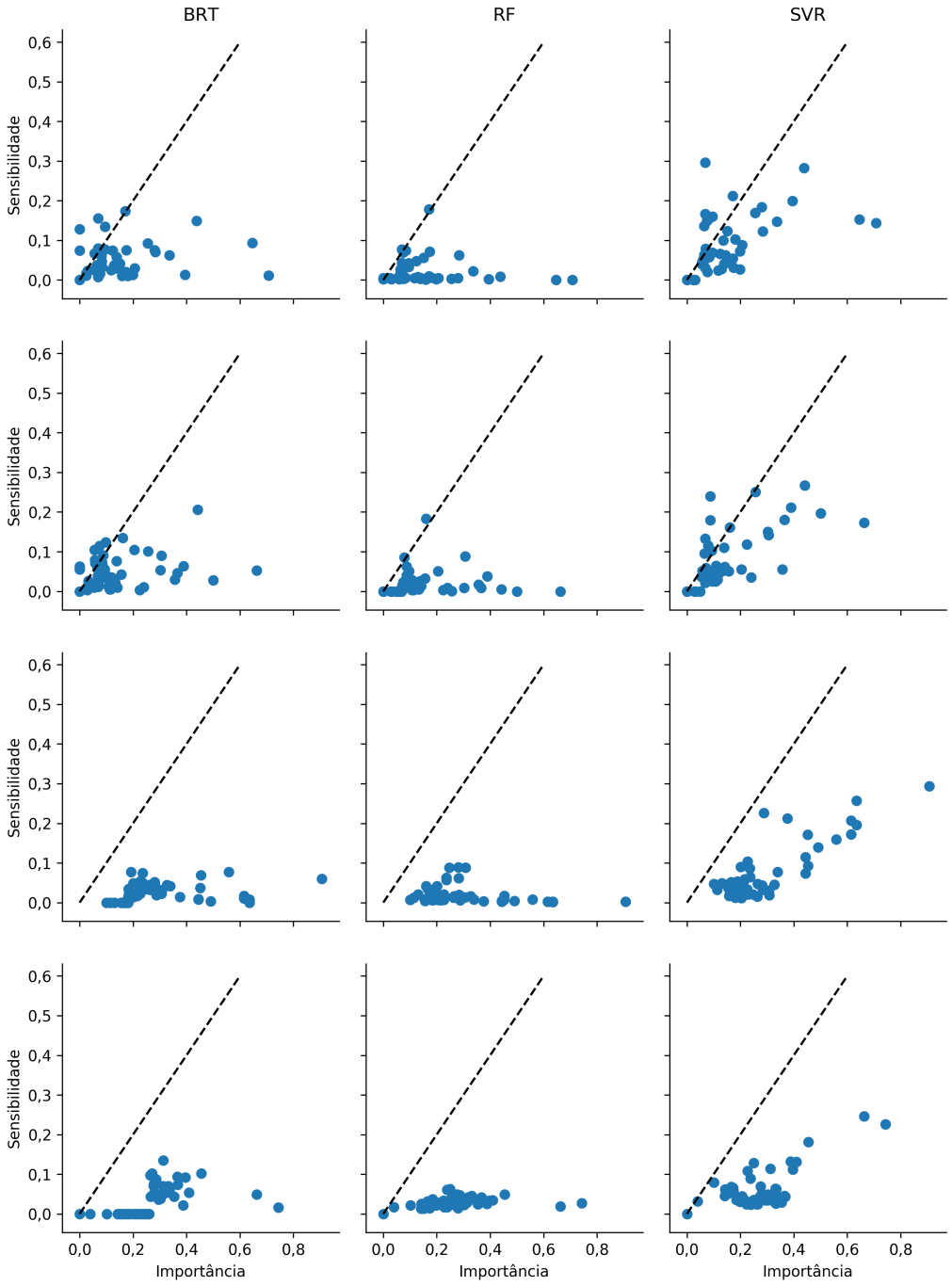


Tabela 10 – Sensibilidade (sens.) à variável e à interação (int.) para cinco maiores diferenças absolutas (abs.) ou relativas (rel.) de cada técnica e subconjunto (SC). BRT - *Boosted Regression Trees*, RF - *Random Forest*, SVR - *Support Vector Regression*. Foram suprimidas as linhas das variáveis ppt_i e SP835037 para o modelo SVR no subconjunto 3.

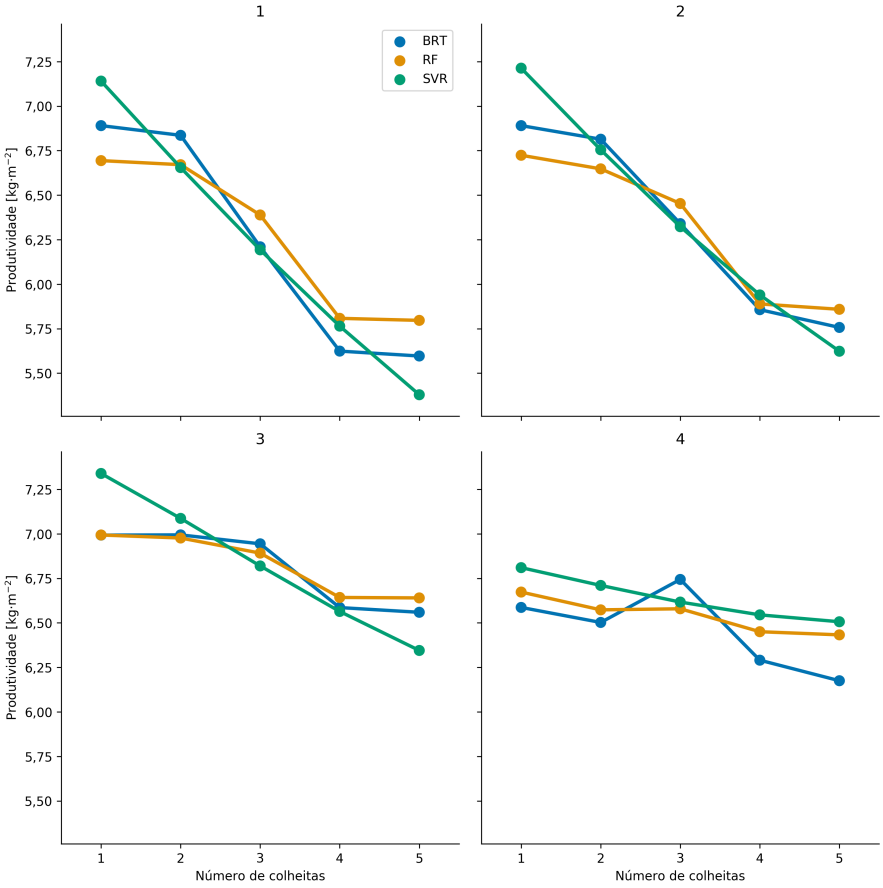
BRT					RF					SVR				SC
variavel	abs.	rel.	sens.	int.	variavel	abs.	rel.	sens.	int.	variavel	abs.	rel.	sens.	int.	
tsn_i	0,041	0,529	0,077	0,117	tsd_ii	0,027	0,358	0,077	0,104	d_irrig	0,089	0,620	0,143	0,232	1
d_calc	0,039	0,831	0,047	0,085	colheitas	0,022	0,123	0,178	0,200	q_irrig	0,087	0,567	0,153	0,240	1
d_torta	0,038	0,298	0,128	0,166	qfert_k	0,020	0,276	0,072	0,091	colheitas	0,050	0,235	0,212	0,262	1
tsd_i	0,032	0,771	0,038	0,074	ppt_ii	0,012	0,155	0,074	0,086	SP803280	0,040	0,234	0,169	0,209	1
q_gesso	0,031	0,467	0,067	0,098	ppt_i	0,008	0,151	0,043	0,064	SP801842	0,032	0,259	0,055	0,156	1
ppt_ii	0,025	0,601	0,033	0,067	-	-	-	-	-	-	-	-	-	-	1
tsn_iii	0,025	0,589	0,034	0,066	-	-	-	-	-	-	-	-	-	-	1
tsd_i	0,051	0,670	0,071	0,127	colheitas	0,054	0,295	0,183	0,237	q_irrig	0,117	0,594	0,197	0,314	2
q_torta	0,049	0,641	0,063	0,125	qfert_k	0,031	0,621	0,051	0,082	d_irrig	0,072	0,417	0,173	0,245	2
tsd_ii	0,047	0,514	0,091	0,137	txt_dmt	0,025	0,493	0,050	0,075	colheitas	0,041	0,252	0,161	0,202	2
RB845210	0,042	0,558	0,054	0,119	tsd_iii	0,021	0,458	0,046	0,068	queima	0,036	0,327	0,064	0,146	2
ppt_i	0,037	0,299	0,123	0,160	SP803280	0,019	0,495	0,038	0,056	d_vin	0,035	0,129	0,267	0,302	2
d_irrig	0,034	0,452	0,052	0,111	frt_dmt	0,013	0,457	0,028	0,040	d_calc	0,034	0,313	0,059	0,145	2
hal	0,035	0,654	0,054	0,089	ctce	0,049	0,547	0,088	0,138	q_torta	0,044	0,215	0,206	0,251	3
qfert_p	0,028	0,367	0,078	0,106	tsn_i	0,033	0,369	0,041	0,122	d_irrig	0,043	0,219	0,196	0,239	3
RB855453	0,027	0,525	0,044	0,079	q_vin	0,021	0,240	0,062	0,111	d_torta	0,038	0,220	0,172	0,210	3
SP803280	0,025	0,413	0,059	0,084	ppt_ii	0,021	0,234	0,035	0,110	qfert_p	0,037	0,763	0,048	0,084	3
colheitas	0,023	0,447	0,050	0,075	qfert_p	0,019	0,211	0,032	0,108	colheitas	0,030	0,348	0,086	0,116	3
frt_dmt	0,022	0,430	0,046	0,074	-	-	-	-	-	RB855453	0,026	0,575	0,045	0,072	3
-	-	-	-	-	-	-	-	-	-	ambprod_F	0,025	0,567	0,039	0,069	3
tsd_iii	0,054	0,765	0,071	0,125	qfert_k	0,023	0,617	0,037	0,061	mg	0,031	0,695	0,043	0,075	4
areia	0,053	0,518	0,102	0,154	argila	0,021	0,444	0,047	0,068	argila	0,025	0,573	0,044	0,070	4
qfert_k	0,043	0,617	0,070	0,114	solo_O	0,018	0,475	0,035	0,055	k	0,025	0,560	0,035	0,069	4
argila	0,037	0,416	0,088	0,124	h	0,018	0,487	0,031	0,055	v	0,025	0,570	0,043	0,070	4
ppt_i	0,033	0,465	0,070	0,103	RB835054	0,018	0,364	0,049	0,067	areia	0,024	0,533	0,043	0,068	4
-	-	-	-	-	ppt_ii	0,017	0,459	0,033	0,054	-	-	-	-	-	4
-	-	-	-	-	hal	0,017	0,457	0,033	0,054	-	-	-	-	-	4

3.3 Curvas de resposta parcial

Na Figura 13, pode ser visto o efeito do número de cortes na produtividade de cana-de-açúcar. O efeito é de decréscimo da produtividade em função do número de colheitas, o que corresponde ao observado em produção. A variação da produtividade é similar nos subconjuntos 1 e 2, sendo maior que no subconjunto 3, que por sua vez é maior que no subconjunto 4. Isso está de acordo com a sensibilidade ao atributo (ver Tabela 8) observada nos subconjuntos, onde a sensibilidade do número de colheitas é maior nos primeiros subconjuntos do que nos demais. No subconjunto 1, os valores de sensibilidade são próximos para BRT(0,174) e RF (0,178), e mais elevado para SVR (0,212), e ligeiramente menores para o subconjunto de dados 2 (0,134, 0,183 e 0,161 para BRT, RF e SVR respectivamente, conforme a Tabela 15 no apêndice C), exceto para RF, que teve uma sensibilidade pouco maior. No subconjunto 3, é possível notar uma menor amplitude do efeito do número de colheitas para o modelo gerado pela RF, seguido por BRT e SVR. A ordem entre a sensibilidade é a mesma, porém a diferença entre as sensibilidades é pequena (0,050, 0,063 e 0,086 para BRT, RF e SVR, respectivamente, conforme a Tabela 16). Enquanto existe uma diferença qualitativa entre as respostas dos modelos BRT e RF, os valores de sensibilidade são próximos. No subconjunto 4 os valores de sensibilidade também são próximos (0,067, 0,045 e 0,064 para BRT, RF e SVR respectivamente, conforme apresentado na Tabela 16). Chama atenção a produtividade maior para 3 cortes do que para 2 cortes no subconjunto 4. Embora isso se deva a uma idiossincrasia dos dados, os modelos gerados pela BRT e RF ajustaram à isso, enquanto a SVR não. Ainda que o comportamento do modelo SVR tenha capturado a relação esperada, esse modelo apresentou o maior erro de validação observado, superando o modelo nulo.

A produtividade é inversamente proporcional à temperatura diurna média da superfície no segundo período e para quase todas as técnicas e subconjuntos, com exceção dos modelos BRT e RF no subconjunto 4 (Figura 14). No subconjunto 3, não é observada resposta de produtividade no início e final das curvas. O efeito negativo da temperatura na produtividade não corresponde ao esperado. Para a cana-de-açúcar, entende-se que o crescimento ocorre para temperaturas do ar a partir de 20 ^{o}C, com temperatura do ar ótima próxima de 30 ^{o}C e crescimento cessando para temperaturas do ar de 40 ^{o}C (Liu et al., 1998). No geral, o efeito da transpiração das plantas faz com que suas folhas tenham temperatura inferior à do ar. Em condições de stress hídrico, a temperatura das folhas se eleva, dado o fechamento dos estômatos. Guardadas as diferenças entre a temperatura do ar e temperatura de superfície, as respostas dos modelos BRT e RF no subconjunto 4 apresentam o aspecto esperado.

Figura 13 – Curvas de resposta parcial do número de cortes na produtividade de cana-de-açúcar para os subconjuntos numerados de 1 a 4 e diferentes técnicas (cores).



Para a resposta da soma da precipitação no primeiro período, a resposta é basicamente inversamente proporcional para quase todos os modelos (Figura 15). Uma possível explicação para esse comportamento é que a precipitação no primeiro período é inversamente proporcional à precipitação nos outros períodos. Em uma janela de tempo delimitado, quanto maior a precipitação em um período, menor a quantidade de chuva observada nos demais. De certa forma, poderia ser discutido que dado o baixo potencial de crescimento da cana-de-açúcar no período inicial, essa precipitação ocorre em um período em que a cultura não consegue aproveitar a disponibilidade de água. Dessa forma, a influência negativa poderia estar mais relacionada à falta de precipitação ao longo do crescimento do que a um efeito negativo. Neste caso, a curva deveria indicar um efeito positivo no início que deveria ser seguido de um platô, potencialmente seguido de um efeito negativo. Entre as técnicas, os modelos gerados por BRT apresentam a maior sensibilidade nos subconjuntos 1,2 e 4, o que é refletido no aspecto das curvas. No subconjunto 3, a sensibilidade do modelo SVR apresenta o maior valor, enquanto para o modelo BRT essa variável não foi incluída pela seleção de atributos, apresentando valor constante para a média de produtividade do conjunto de treino.

Figura 14 – Curvas de resposta parcial da temperatura diurna média no segundo período na produtividade de cana-de-açúcar para os subconjuntos numerados de 1 a 4 e diferentes técnicas (cores).

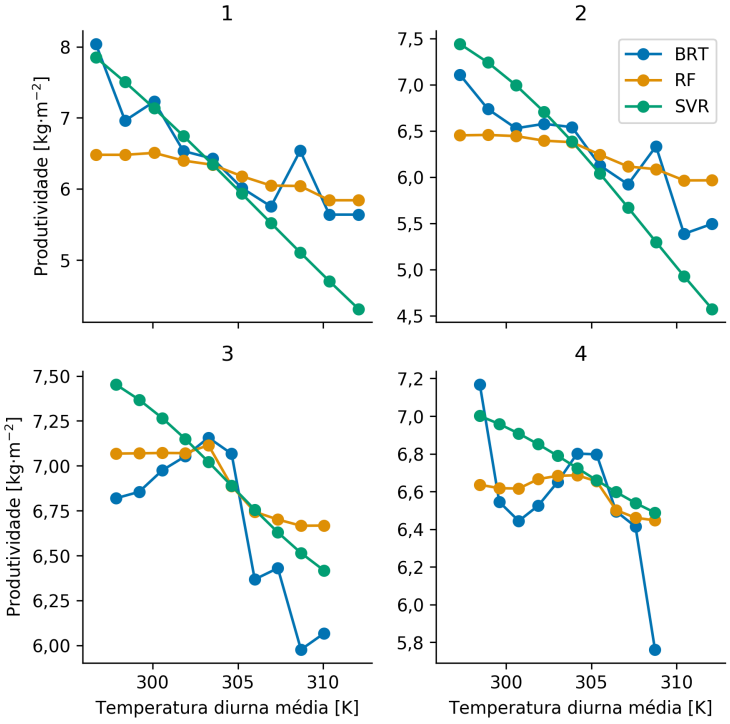
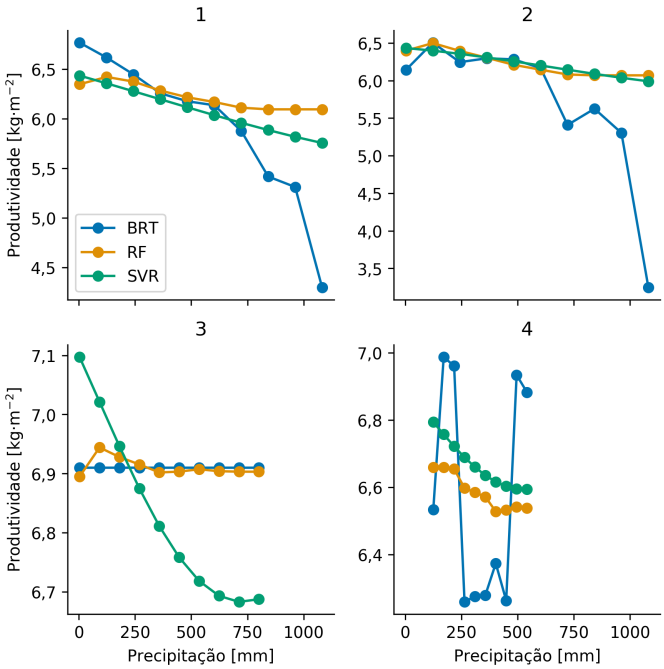


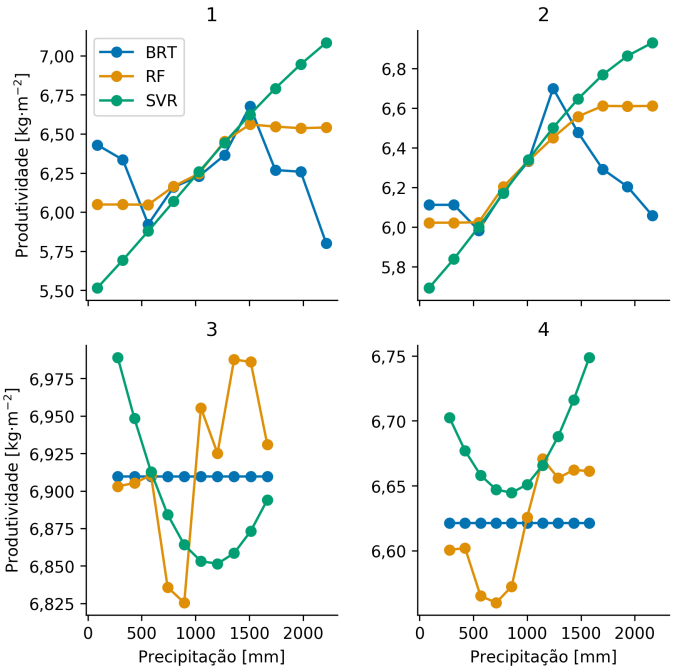
Figura 15 – Curvas de resposta parcial da precipitação acumulada no primeiro período na produtividade de cana-de-açúcar para os subconjuntos numerados de 1 a 4 e diferentes técnicas (cores).



Enquanto a resposta de produtividade esperada para a precipitação no segundo período deveria ser positiva no geral, este comportamento não é observado para a maioria

dos modelos (Figura 16). Os modelos da técnica RF nos subconjuntos 1 e 2 apresentam uma resposta geral mais próxima do esperado, sendo tal que pequenas somas tendem a não afetar a produtividade, valores intermediários tem efeito positivo e valores elevados deixam de ter efeito. Embora seja possível interpretar o platô final como uma região em que a disponibilidade hídrica deixou de ser o fator limitante, este comportamento pode ser mais facilmente explicado pela resposta de um modelo de árvore ser constante nos extremos de dados observados. Também nos subconjuntos 1 e 2, os modelos gerados pela SVR apresentaram resposta positiva, que parece diminuir nos valores mais elevados. A variável não foi escolhida para os modelos BRT nos subconjuntos 3 e 4, o que chama atenção, pois esta fase é crítica para o desenvolvimento da cana-de-açúcar e a disponibilidade hídrica deveria pautar a produtividade final. Os demais modelos apresentam respostas inverossímeis, sendo o aspecto dos modelos BRT no subconjunto 1 similar à RF nos subconjuntos 3 e 4. A região convexa desses modelos, junto da resposta dos modelos SVR não apresentam paralelo com o que poderia ser observado para o crescimento de plantas.

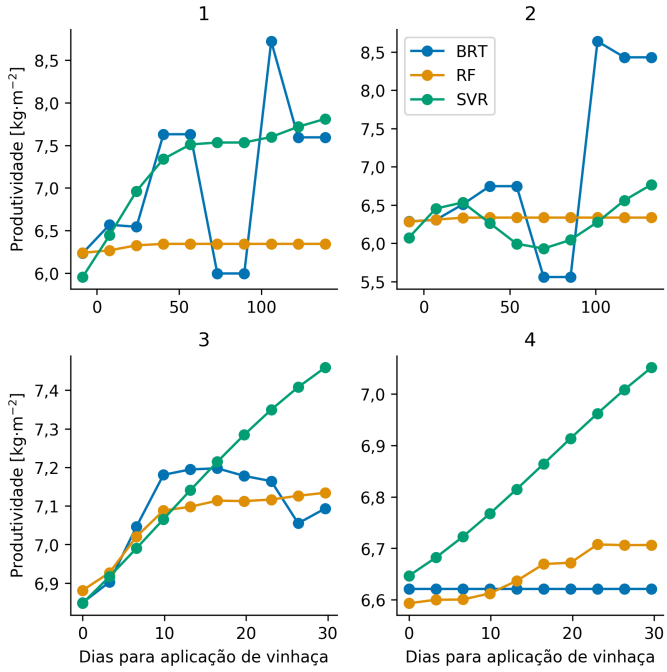
Figura 16 – Curvas de resposta parcial da precipitação acumulada no segundo período na produtividade de cana-de-açúcar para os subconjuntos numerados de 1 a 4 e diferentes técnicas (cores).



Conforme mostrado na Figura 17, quanto maior o tempo entre o início dos ciclos e a aplicação de vinhaça, maior a produtividade final. De forma geral, isso não é esperado, dado o uso nas usinas da irrigação com vinhaça como estratégia para viabilizar o início de ciclos de produção em épocas secas. Assim, o efeito esperado é que as intervenções mais próximas ao inicio do ciclo levassem a maiores produtividades, sendo o efeito das aplicações mais tardias menos benéfico. Curvas com maiores variações de produtividade

estão relacionadas a maiores coeficientes de sensibilidade, assim como menores variações estão relacionadas a menores valores de sensibilidade. Por outro lado, é possível que a vinhaça seja aplicada mais no início em regiões com baixo potencial, e mais tarde para regiões com maior potencial. A grande importância dessa variável representada nos primeiros lugares para cada subconjunto, assim como a grande sensibilidade não é esperada, e potencialmente representa alguma forma de idiossincrasia do conjunto de dados.

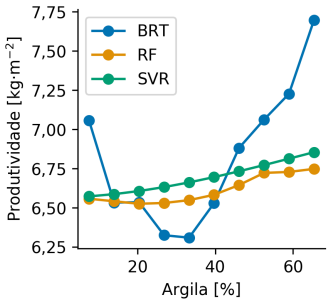
Figura 17 – Curvas de resposta parcial do número de dias entre o início do ciclo de crescimento e a aplicação de vinhaça na produtividade de cana-de-açúcar para os subconjuntos numerados de 1 a 4 e diferentes técnicas (cores).



A importância da fração de argila na produtividade é positiva para os três modelos, com exceção dos valores iniciais para o modelo BRT, que apresentou um aspecto convexo (Figura 18). Esse efeito geral é esperado, dado que maiores teores de argila estão relacionados a uma maior capacidade de retenção hídrica e potencialmente maior fertilidade. O efeito é pequeno nos modelos RF e SVR, com valores baixos de sensibilidade, e grande no modelo BRT em que existe grande sensibilidade.

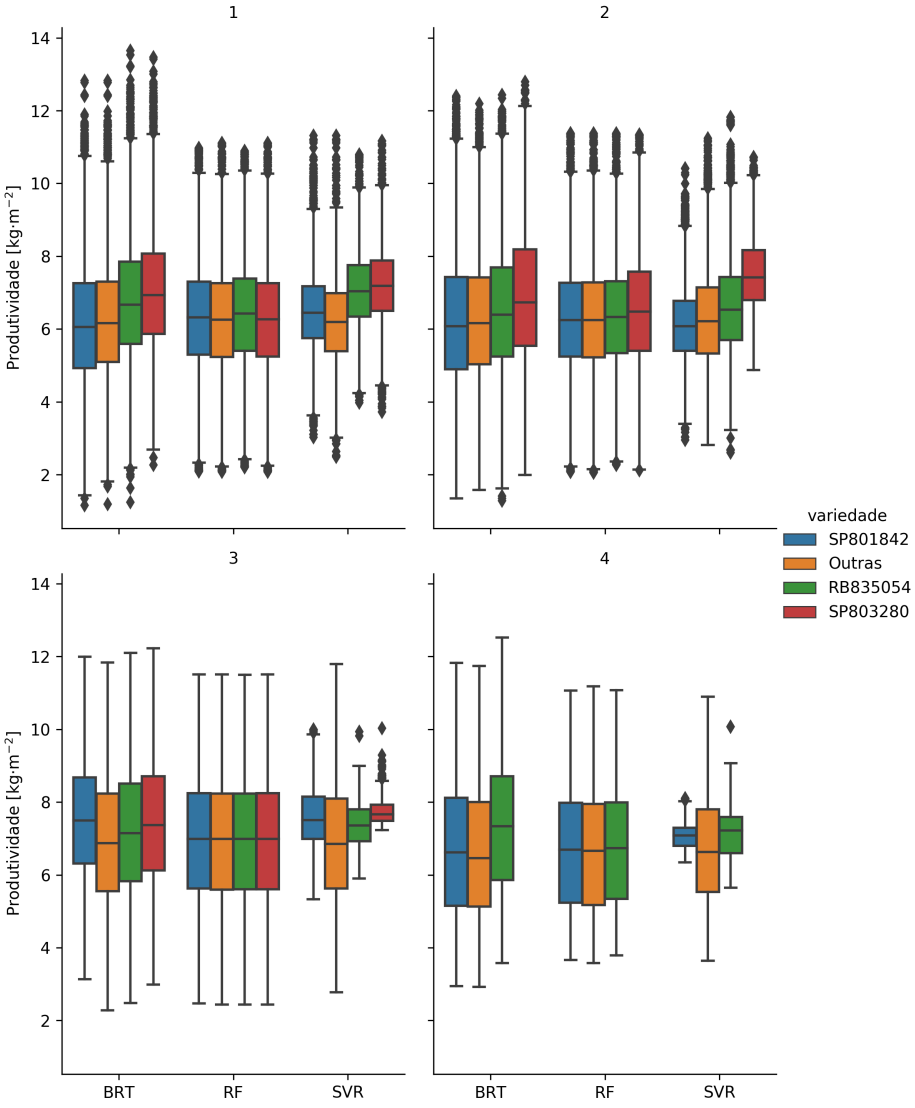
Na Figura 19, são mostrados os *boxplots* da produtividade de cana-de-açúcar em função de variedades que apresentaram valores altos de importância no geral. Dois efeitos podem ser notados, sendo a redução da variabilidade quando uma variedade é especificada, o que é observado para modelos SVR e deslocamentos da distribuição, observados nos modelos BRT e SVR. Nos subconjuntos 1 e 2, a ordenação das medianas das variedades especificadas é similar entre os modelos BRT e SVR, com pequena diferença para a distribuição das "outras variedades". É possível observar também que os modelos SVR possuem uma variabilidade menor que os demais, em especial nos subconjuntos 3 e

Figura 18 – Curvas de resposta parcial da fração de argila do solo na produtividade de cana-de-açúcar para o subconjunto 4 para as diferentes técnicas (cores). A fração de argila está disponível apenas no subconjunto 4.



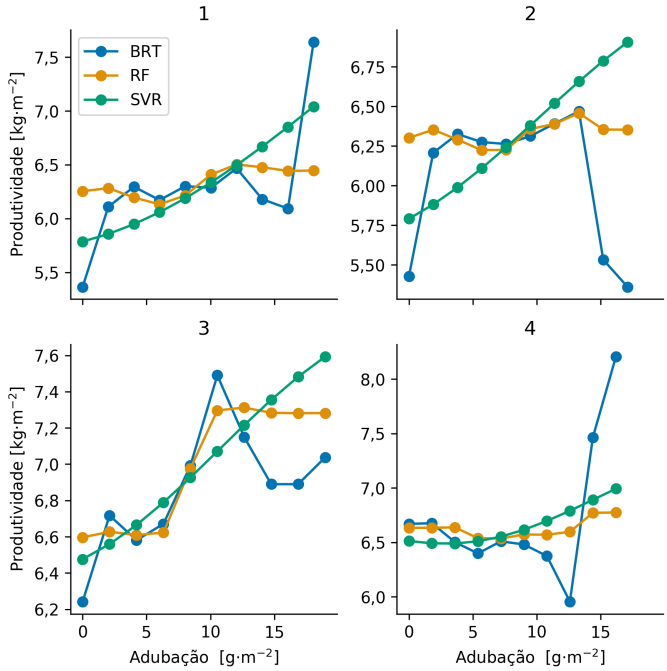
4. Conforme indicado pela sensibilidade dos modelos, existe pouca ou nenhuma diferença na saída dos modelos RF em função da especificação da variedade.

Figura 19 – Boxplots da resposta parcial das diferentes variedades na produtividade de cana-de-açúcar para o subconjunto 4 para as diferentes técnicas (cores).



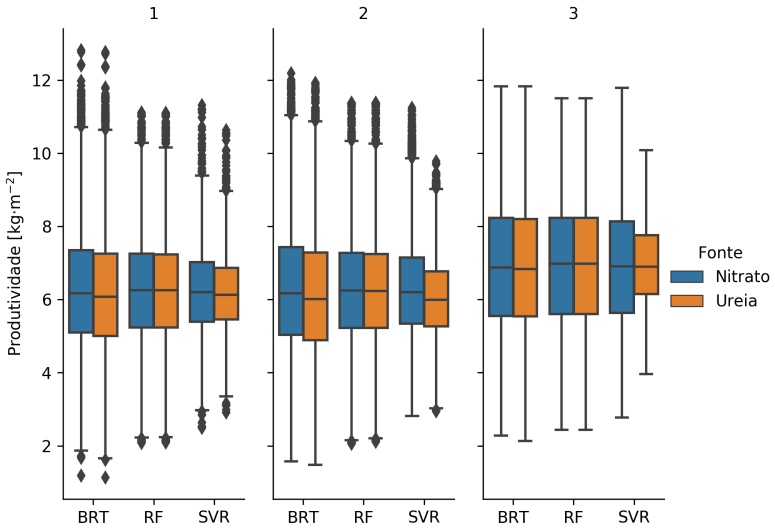
Assim como ocorrido para outras variáveis, a resposta da produtividade em função da quantidade de adubação nitrogenada (Figura 20) apresenta comportamentos mistos no que diz respeito ao esperado. As respostas dos modelos SVR e do modelo RF no subconjunto 3 têm o aspecto esperado para a resposta da cultura ao nitrogênio, sendo um efeito não-negativo, variando a taxa de resposta ao longo das diferentes quantidades aplicadas. Embora o platô encontrado para quantidades aplicadas superiores a 10 g $\cdot$m^{-2} (100 kg$\cdot$ha^{-1}) para o modelo RF no subconjunto 3 possa ser relacionado com a adubação deixar de ser limitante, isso não pode ser distinguido do fato de que modelos baseados em árvores terem uma resposta constante nos extremos, análogo ao ocorrido para a precipitação no segundo período para os modelos RF. As respostas dos modelos BRT e RF (exceto no subconjunto 3) não correspondem ao plausível para adubação, em especial o contraste nas regiões extremas (BRT nos subconjuntos 2 e 4 para doses próximas a 150 kg$\cdot$ha^{-1}) ou as regiões onde a resposta é convexa.

Figura 20 – Curvas de resposta parcial do efeito da adubação nitrogenada na produtividade de cana-de-açúcar para os subconjuntos numerados de 1 a 4 para as diferentes técnicas (cores).



Na Figura 21, pode ser visto o efeito da forma do adubo nitrogenado na produtividade de acordo com os modelos. Seguindo os resultados da análise de sensibilidade, existiu pouca diferença entre a saída dos modelos em função da forma da adubação. Considerando que a adubação com Nitrato corresponde à maior parte da base de dados, os resultados para adubação com Ureia para SVR são análogos aos apresentados para o efeito das variedades, em que especificando um subconjunto, a saída do modelo apresenta menor variabilidade.

Figura 21 – Boxplots da resposta parcial da forma de adubação na produtividade de cana-de-açúcar para os subconjuntos numerados de 1 a 3 para as diferentes técnicas.

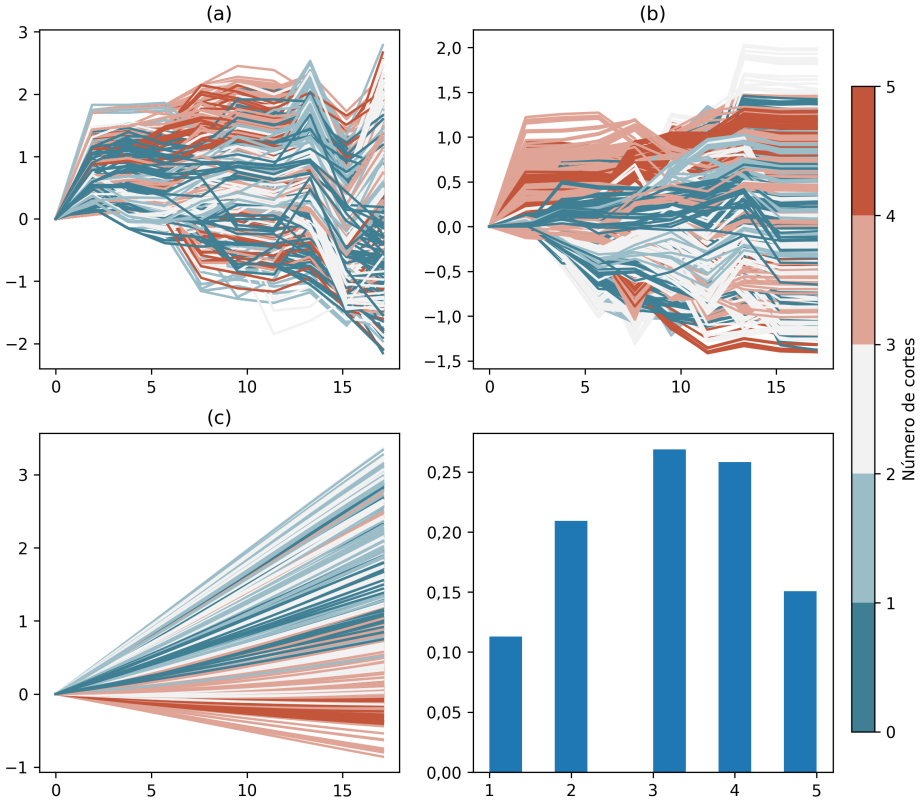


3.4 Curvas de resposta individual

Para apresentação das curvas de resposta condicional esperada, foi subtraído o valor de produtividade para zero de adubação. Assim, as curvas podem ser interpretadas como o ganho de produtividade em função da aplicação de nitrogênio.

Na Figura 22 são apresentadas as curvas de resposta esperada condicional da adubação com N e o número de colheitas. Embora a interação tenha tido uma maior diferença absoluta para a RF nesse subconjunto, a interação é mais clara na SVR. Chama atenção em todas as técnicas que existem curvas de resposta positivas e negativas, sendo que a resposta negativa não deveria ser esperada nessas faixas de quantidade de adubação. Enquanto a resposta identificada para a SVR é basicamente linear, a resposta obtida nos modelos BRT e RF apresenta diversas oscilações, o que não é esperado na resposta à adubação. Na resposta do modelo SVR, chama a atenção que as respostas mais positivas sejam para os cortes iniciais, enquanto as respostas negativas estejam relacionadas a cortes mais avançados, contradizendo o esperado para cana-de-açúcar, em que a cana planta não costuma apresentar resposta e sim as soqueiras. Cabe destacar que as regiões de alta adubação para soqueiras de número de cortes mais alto são regiões de extrapolação do modelo, dado que essas taxas são aplicadas para soqueiras mais novas. Dessa forma, a região de resposta positiva que é observada no início para os modelos BRT e RF está em uma região onde é esperado um melhor desempenho do modelo. Nas curvas da RF, é possível ver regiões de resposta constante para valores elevados, que quando agregadas, geram as regiões constantes observadas para a curva parcial observada na Figura 13. Para

Figura 22 – Efeito da interação entre número de colheitas(cores) e a taxa de aplicação de nitrogênio na produtividade de cana-de-açúcar para o subconjunto 2 para as técnicas BRT (a), RF (b) e SVR (c) junto do histograma do número de colheitas.

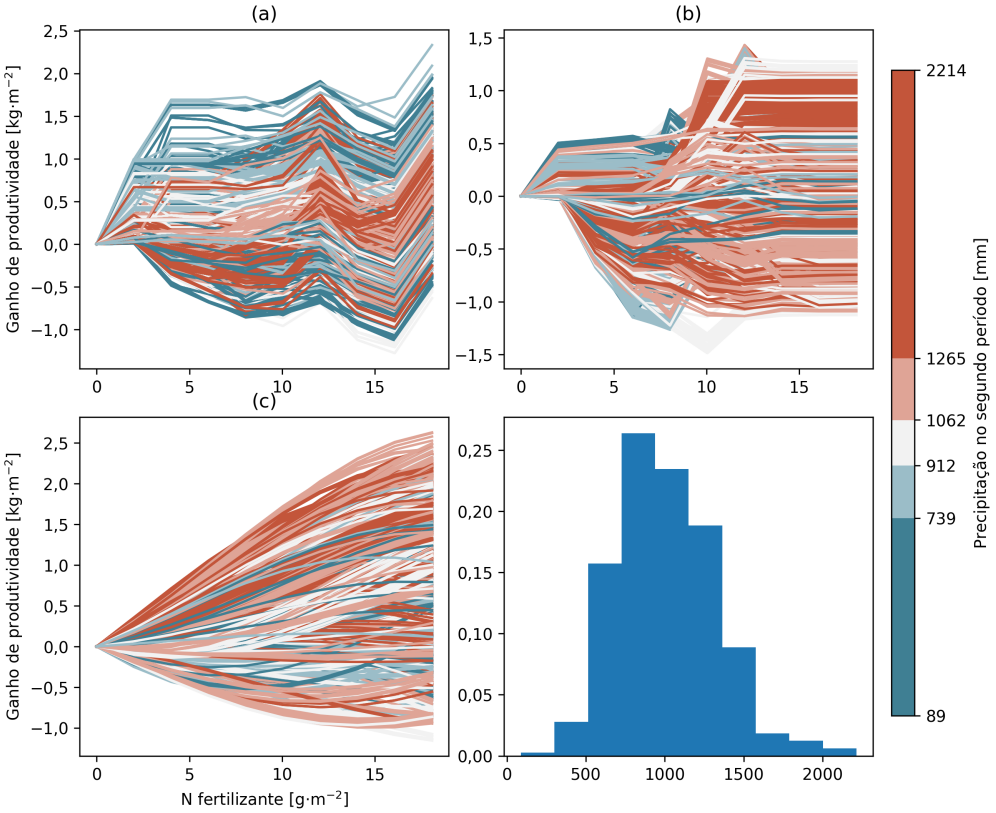


os valores iniciais, a resposta parcial constante observada na Figura 13 é explicada pela agregação de várias curvas, com efeitos positivos e negativos.

A Figura 23 mostra o efeito da interação entre a adubação com N e a precipitação no segundo período para o subconjunto 2. Essa interação não é esperada a princípio, dado o efeito do nitrogênio estar mais relacionado com o desenvolvimento inicial do cultivo. Embora essa variável seja responsável pela sexta e quarta maiores diferenças absolutas para as técnicas BRT e RF nesse subconjunto, respectivamente, não é possível notar uma tendência clara da interação. Nesse subconjunto, a resposta dos modelos SVR não tem aspecto linear, como ocorreu no subconjunto 1. Para o modelo BRT, chama atenção que as respostas mais positivas e mais negativas ocorrem para valores baixos de precipitação, enquanto o oposto ocorre para o modelo RF.

A interação entre a precipitação no primeiro período e a adubação de N é mostrada na Figura 24. Essa interação era esperada, porém as respostas dos modelos não são coerentes com o que devia ser encontrado. No caso do modelo SVR no subconjunto 3 (a), temos uma resposta positiva sempre para a adubação, porém as respostas mais extremas são para valores baixos de precipitação. A interação entre precipitação e nitrogênio deveria ser positiva, pois não adianta ampla disponibilidade de um fator se o outro for limitante.

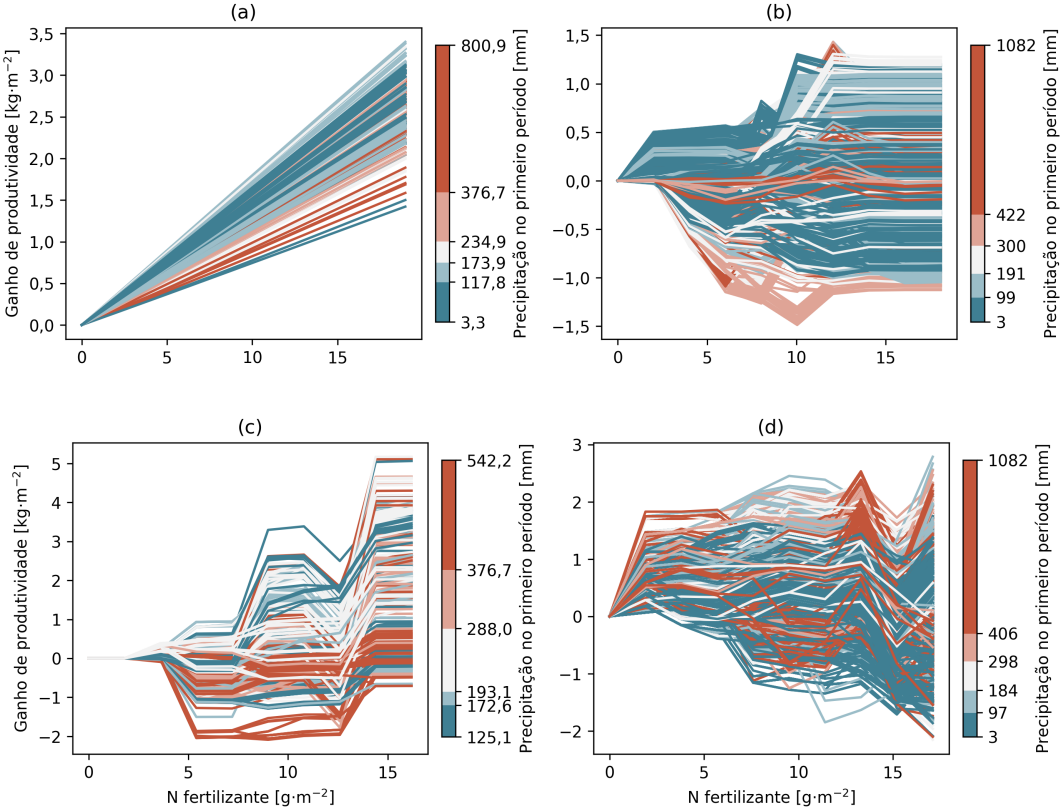
Figura 23 – Efeito da interação entre a precipitação no segundo período (cores) e a taxa de aplicação de nitrogênio na produtividade de cana-de-açúcar para o subconjunto 1 para as técnicas BRT (a), RF (b) e SVR (c).



Assim, respostas mais positivas deveriam ser observadas para quantidades maiores de precipitação. O modelo RF no subconjunto 1 (b) mostra uma grande dispersão de respostas para os valores mais baixos de precipitação, com curvas de resposta mais centrais para valores mais elevados. Comparando os modelos BRT gerados nos subconjuntos 4 (c) e 2 (d), podemos notar que o aspecto das curvas diverge entre os modelos. Enquanto no subconjunto 4 os modelos apresentam baixa responsividade para baixos valores de aplicação de N, a resposta é basicamente positiva para os valores baixos no modelo 2. Nos valores mais elevados de aplicação de N, o modelo do subconjunto 4 apresenta curvas estáveis, enquanto no modelo do subconjunto 2 existem respostas positivas e negativas.

Na Figura 25, podemos ver a interação entre a taxa de aplicação de nitrogênio e a taxa de aplicação de fósforo. Essa interação não é esperada do ponto de vista do manejo da cana-de-açúcar e não foi possível distinguir nos gráficos um padrão claro para resposta. Mesmo que a quantidade de P no solo possa ser um fator limitante para produção, o fósforo é um nutriente de baixa mobilidade, sendo mais razoável esperar uma interação entre a quantidade de fósforo no solo (também presente no subconjunto 3) do que uma relação com a adubação. De certa forma, a importância deste atributo pode estar mais relacionada com sua relação com o número de cortes do que o efeito da adubação de

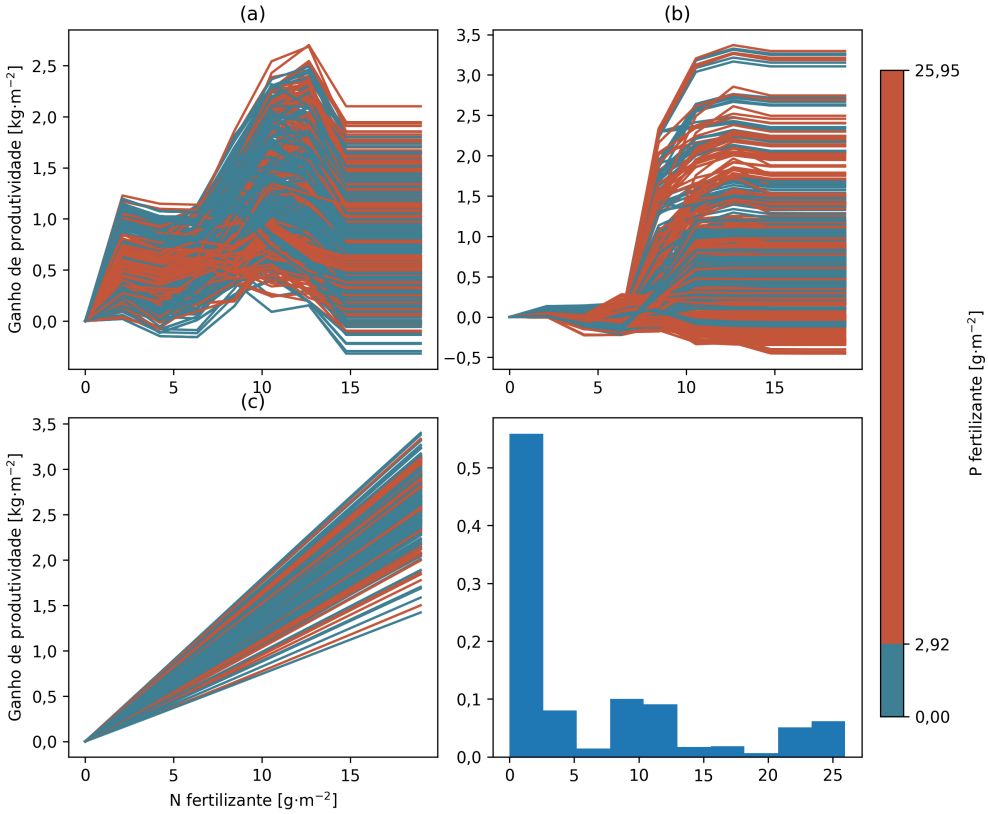
Figura 24 – Efeito da interação entre a precipitação no primeiro período (cores) e a taxa de aplicação de nitrogênio na produtividade de cana-de-açúcar. SVR no subconjunto 3 (a), RF no subconjunto 1 (b), e BRT nos subconjuntos 4 (c) e 2 (d).



fósforo na colheita seguinte, justificando sua inclusão na etapa de seleção de atributos. Assim, o efeito de interação poderia ser análogo ao capturado pelo número de colheitas, porém diferente do ocorrido no subconjunto 2 (Figura 22), não foi possível distinguir algum padrão neste caso.

Podemos ver a interação entre a taxa de aplicação de nitrogênio e taxa de aplicação de potássio na Figura 26. Essa interação é esperada do ponto de vista do manejo da cana-de-açúcar, dado que a partir do balanço de massa dos colmos, é esperada uma proporção de 1 a 1,2 partes de potássio para cada parte de nitrogênio, razão pela qual diversas formulações utilizadas no cultivo aproximam essas proporções. Assim, respostas mais elevadas a adubação com N são esperadas se houver uma contra-parte de adubação com K, que é o oposto do observado na Figura 26. No subconjunto 4, para os modelos modelo BRT (a) e RF (b), temos uma resposta maior para cultivos que recebem pouco potássio (abaixo de 5,4 g·m^{-2}). Nos subconjuntos 2 (c) e 1 (d), o padrão capturado pela RF é similar, mas dado o maior volume de dados, é possível visualizar uma maior diversidade de intensidades de aplicação de K. No subconjunto 2 (c), existe uma grande amplitude de respostas para os patamares elevados de aplicação (parte das maiores respostas positivas

Figura 25 – Efeito da interação entre fertilização de fósforo (cores) e a taxa de aplicação de nitrogênio na produtividade de cana-de-açúcar para o subconjunto 3 para as técnicas BRT (a), RF (b), SVR (c) e histograma da distribuição da adubação de fósforo no subconjunto de dados.

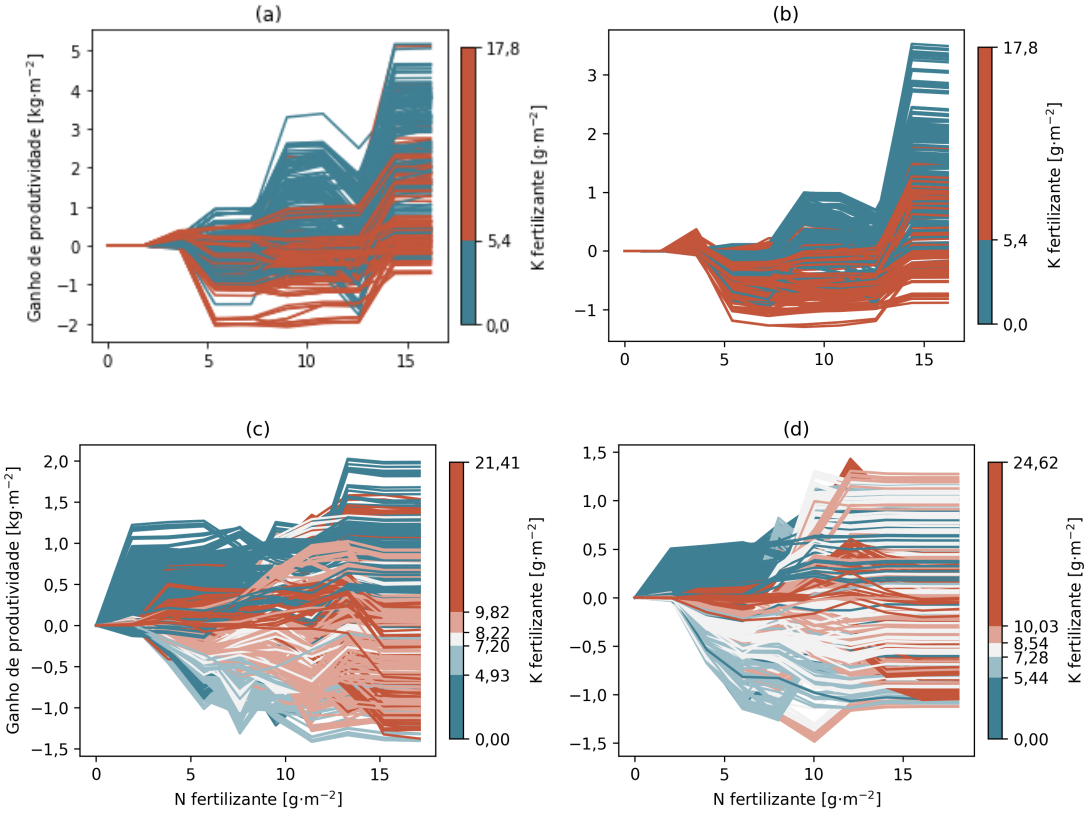


e negativas), porém as respostas positivas são dominadas por baixas aplicações de K. No subconjunto 1 (d), as maiores aplicações de K estão em conjunto com as respostas mais negativas, enquanto não há padrão claro para as demais taxas.

A interação entre o teor de argila no solo e a taxa de aplicação de nitrogênio pode ser visto na Figura 27. Esta interação era esperada, dado que um maior teor de argila no solo permite uma melhor retenção dos nutrientes aplicados, e também uma maior disponibilidade de água que, consequentemente, permitiria uma maior produtividade. As interações mostradas para todos os modelos na Figura 27 indicam as respostas positivas mais intensas para os maiores teores de argila, enquanto as respostas intermediárias e negativas ocorrem para os menores valores de argila. Enquanto o aspecto das curvas é mais consistente para SVR (c), as oscilações dos modelos BRT (a) e RF (b) não são condizentes com o esperado pelo sistema físico, sendo as discrepâncias maiores para a BRT que RF.

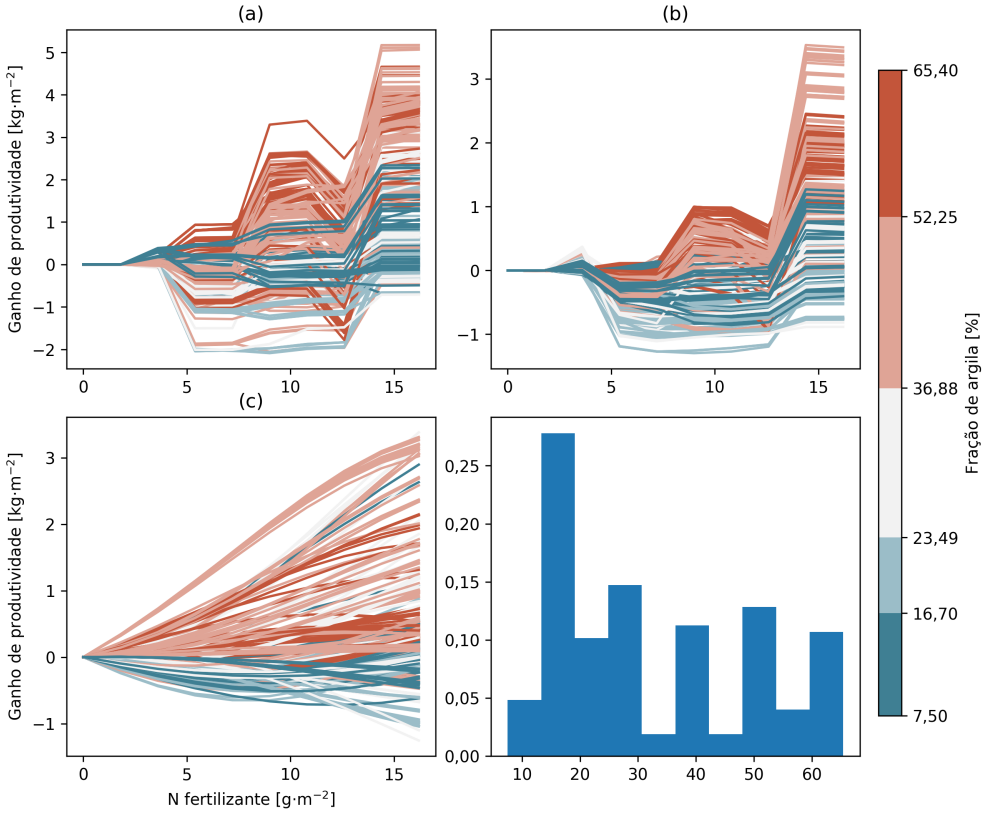
De forma geral, a análise de sensibilidade permitiu priorizar quais fatores inspecionar utilizando os gráficos de resposta parcial e os gráficos de resposta condicional independente. A inspeção dos gráficos permitiu explorar o comportamento dos modelos,

Figura 26 – Efeito da interação entre fertilização de potássio (cores) e a taxa de aplicação de nitrogênio na produtividade de cana-de-açúcar. BRT no subconjunto 4 (a), e RF nos subconjuntos 4 (b), 2 (c) e 1 (d).



e a representação do efeito de outras variáveis nos gráficos de resposta condicional independente permitiu explorar os efeitos de interação. Os padrões de resposta primários encontrados para o número de cortes nos diversos modelos e subconjuntos, precipitação no segundo período para o modelo RF nos subconjuntos 1 e 2, e e teor de argila no solo para o subconjunto 4 são coerentes com o esperado para produção de cana-de-açúcar, enquanto os demais padrões encontrados não são coerentes. Chama a atenção o efeito encontrado para o impacto do número de dias entre o início do desenvolvimento e a aplicação de vinhaça. A interação entre número de colheitas e aplicação de N encontrada corresponde ao oposto do esperado, assim como o na interação com potássio e para precipitação no primeiro período para três dos quatro modelos inspecionados. A interação com o teor de argila no solo foi a única interação que qualitativamente está alinhada com o esperado para o desenvolvimento da cana-de-açúcar, embora o comportamento individual das curvas de resposta seja prejudicado pelas oscilações e regiões convexas.

Figura 27 – Efeito da interação entre o teor de argila (cores) e a taxa de aplicação de nitrogênio na produtividade de cana-de-açúcar para o subconjunto 4 para as técnicas BRT (a), RF (b) e SVR (c) e histograma da fração de argila no subconjunto 4.



3.5 Considerações adicionais

Em função dos padrões encontrados, não é razoável utilizar os modelos obtidos neste trabalho com aprendizado de máquina para otimizar a recomendação da adubação nitrogenada para talhões de cana-de-açúcar. Considerando apenas o aspecto das curvas de resposta parcial, para as respostas dos modelos BRT e RF, haveria uma concentração de recomendações de baixos valores. Já para o declínio no começo da curva, um modelo de otimização não recomendaria o acréscimo de nitrogênio, dado que não há benefício para o custo implícito. Para o final das curvas, não há benefício, pois o mesmo valor de produtividade poderia ser atingido por um valor baixo de adubação. Para os modelos gerados por SVR, a resposta quase linear gera duas possibilidades. Na primeira, caso a taxa de resposta seja superior ao custo relativo do adubo, o modelo recomendaria adubação máxima, supondo alguma restrição ao modelo de otimização. Na segunda, caso a taxa de resposta seja inferior ao custo relativo, o modelo recomendaria adubação zero. Raciocínios similares podem ser empregados para as respostas individuais dos modelos.

O fato de que modelos empíricos podem estar captando apenas correlações nos dados é uma crítica comumente apresentada, e recentemente tem sido também discutido

o quanto as respostas desses modelos podem ser fisicamente inconsistentes. No caso deste estudo, as inconsistências encontradas para a resposta individual à adubação indicam que os modelos não permitem o uso prescritivo. A despeito desta limitação, o uso preditivo apresentou resultados preliminares positivos para fins de planejamento da produção de usinas de cana-de-açúcar (Mantelato, 2017).

Resultados de modelos empíricos já se mostraram inconsistentes em outras áreas de conhecimento. Modelos empíricos qualitativamente errados são apresentados por Wagner e Rondinelli (2016) para árvores de decisão que modelavam a formação de compostos em função das propriedades físicas dos materiais utilizados. Os autores sugerem um processo pautado por conhecimento do domínio para aumentar a robustez dos modelos criados por aprendizado de máquina. Nos resultados de Karpatne et al. (2017b), modelando a temperatura no perfil vertical de um lago usando dados de simulação, o modelo de rede neural encontrado apresentava violações de princípios físicos estabelecidos, em especial o efeito da temperatura na densidade da água. Para os autores, isso é um indício de que o modelo está aprendendo padrões nos dados que não são generalizáveis. Foram propostas mudanças na função de erro de treinamento da rede neural, incorporando o conhecimento das relações entre a densidade da água e sua profundidade (i.e. a água mais densa deve estar abaixo). Os modelos gerados com a modificação da função de erro não só apresentavam menos ocorrências de violação dos princípios físicos como também obtiveram menor erro de validação.

Recentemente, grandes avanços em aprendizado de máquina se deram com modelos criados por *deep learning*, mas esse tipo de avanço não está sendo tão efetivo em todos os domínios de ciências naturais. Esses modelos de redes neurais são caracterizados pelo grande número de camadas intermediárias, cujo treinamento era inviável até o início da década de 2010. Além dos avanços na capacidade computacional e nos algoritmos de treinamento, a existência de grandes conjuntos de dados é um dos fatores-chave para os novos patamares alcançados. Os domínios em que os avanços são mais marcantes são a visão computacional e a tradução e interpretação de linguagem natural (LeCun et al., 2015), em que conjuntos de dados extremamente grandes estão disponíveis em domínios digitais.

Karpatne et al. (2017a) apresentam dois aspectos que geram grandes diferenças na aplicação de técnicas de aprendizado de máquina nesses domínios e domínios científicos. O primeiro é a falta de grandes conjuntos de dados compatíveis com a complexidade dos problemas. Isso faz com que seja difícil a obtenção de conjuntos de dados representativos dos fenômenos modelados, muitas vezes complexos e não estacionários. Nestes casos, mesmo com a aplicação de metodologias como validação cruzada não é possível garantir que os padrões aprendidos não sejam apenas correlações nos dados usados. Outro aspecto

que os autores destacam é que em várias aplicações, o objetivo final é a criação de modelos acionáveis, o que não é necessariamente o objetivo final em domínios científicos. Nesses domínios, traduzir os padrões aprendidos para teorias e hipóteses que possam avançar o conhecimento científico é uma condição premente. A conclusão de Karpatne et al. (2017a) é de que um novo paradigma de ciência de dados baseada em teoria é necessário, visando gerar resultados que usufruam dos correntes avanços no uso de grandes bases de dados mas que seja compatível com o conhecimento científico estabelecido. Os autores apresentam uma diversidade de abordagens metodológicas para o novo paradigma apresentado, listando diversos domínios em que esse tipo de aplicação tem tido resultados positivos.

Considerando a diversidade de domínios nos quais a disponibilidade de dados está aumentando, aplicações relacionadas a ciências naturais irão demandar cuidados adicionais no uso de modelos de aprendizado de máquina. Recentemente, técnicas de aprendizado de máquina permitiram grandes avanços em diversos domínios nos quais é difícil expressar as relações entre as grandezas presentes nos dados, ou mesmo nos quais não se sabe o que esperar. Exemplos de domínios são comportamento de navegação *web*, classificação de imagens, entendimento de linguagem natural, fraudes ou sistemas de recomendação de conteúdo. Para domínios de ciências naturais, diversas relações entre as grandezas presentes nos conjuntos de dados são conhecidas, e embutir esse conhecimento para a modelagem se mostra necessário, visando garantir a consistência física da resposta dos modelos. Essa inclusão pode beneficiar os modelos gerados, seja no aumento da credibilidade do modelo como na melhora direta da performance dos modelos obtidos.

4 Conclusão

Investigando a resposta de modelos criados por diferentes técnicas de aprendizado de máquina, foi possível identificar que os padrões gerais aprendidos se mostraram coerentes com o esperado para grande parte dos fatores estudados. Isso não se mostrou verdade para os padrões locais analisados. Na análise de padrões locais, diversos padrões considerados inconsistentes foram encontrados. Para alguns fatores cujo efeito na produtividade é sabidamente não-negativo, foram encontradas respostas negativas, que em vários casos podem ser traçados diretamente a idiossincrasias do conjunto de dados. Exemplos disso são o efeito da irrigação no início do cultivo ou a adubação de potássio. Para outros fatores, chama a atenção a inversão do efeito sobre a produtividade. Esse foi o caso em especial para a oscilação da produtividade em função da adubação nitrogenada, mas ocorrendo também para o número de cortes ou idade do canavial. Em outros casos, a resposta não se mostra razoável, como a resposta extremamente positiva para modelos gerados por SVR.

Em função destas respostas locais inconsistentes, não foi considerado viável o uso destes modelos para fins prescritivos na escala de talhões. Enquanto o uso destes modelos para fins preditivos se mostra uma alternativa viável para fins como a estimativa de produção, isso parece se dar em casos em que o sistema produtivo não apresenta mudanças em relação aos dados de treinamento dos modelos. Dessa forma, considera-se que para os dados estudados, os modelos não foram capazes de gerar respostas coerentes para intervenções no sistema produtivo, o que é exatamente o objetivo do uso prescritivo dos modelos. É possível que isso seja superável com o aumento da diversidade de dados, ou mesmo simplesmente com o aumento do volume de dados disponíveis.

A inspeção dos modelos utilizando gráficos, seja de resposta global ou resposta geral, se mostrou uma forma efetiva para "abrir a caixa-preta"de técnicas como *ensembles* ou máquinas de vetor de suporte. Embora o presente resultado possa ser considerado negativo, a investigação dos modelos utilizando utilizando gráficos de resposta e análise de sensibilidade se mostram como uma alternativa viável para entender a resposta de técnicas avançadas de aprendizado de máquina. Essas práticas podem avançar o uso destas técnicas nas quais há demandas por entender o comportamento dos modelos que serão utilizados. Em casos positivos, entender a resposta dos modelos irá trazer um aumento da confiança em sua aplicação, enquanto para casos negativos, irá indicar a necessidade de rever os resultados de modelagem ou delimitar a aplicação dos modelos.

Referências

Abu-Mostafa, Y. S., Magdon-Ismail, M., e Lin, H.-T. (2012). *Learning from data*, volume 4. AMLBook New York, NY, USA:.

Bengio, Y. e Grandvalet, Y. (2004). No unbiased estimator of the variance of k-fold cross-validation. *Journal of machine learning research*, 5(Sep):1089–1105.

Bergstra, J. e Bengio, Y. (2012). Random search for hyper-parameter optimization. *Journal of Machine Learning Research*, 13(Feb):281–305.

Bergstra, J. S., Bardenet, R., Bengio, Y., e Kégl, B. (2011). Algorithms for hyper-parameter optimization. In Shawe-Taylor, J., Zemel, R. S., Bartlett, P. L., Pereira, F., e Weinberger, K. Q., editores, *Advances in Neural Information Processing Systems 24*, pages 2546–2554. Curran Associates, Inc.

Bocca, F. F. e Rodrigues, L. H. A. (2016). The effect of tuning, feature engineering, and feature selection in data mining applied to rainfed sugarcane yield modelling. *Computers and Electronics in Agriculture*, doi:10.1016/j.compag.2016.08.015.

Borgonovo, E. e Plischke, E. (2016). Sensitivity analysis: a review of recent advances. *European Journal of Operational Research*, 248(3):869–887.

Breiman, L. (1996). Bagging predictors. *Machine learning*, 24(2):123–140.

Breiman, L. (2001). Random forests. *Machine learning*, 45(1):5–32.

Breiman, L., Friedman, J., Olshen, R., e Stone, C. (1984). Classification and regression trees, 1984: Belmont. *CA: Wadsworth International Group*.

Cortes, C. e Vapnik, V. (1995). Support-vector networks. *Machine learning*, doi:10.1007/BF00994018.

Crosson, W. L., Al-Hamdan, M. Z., Hemmings, S. N., e Wade, G. M. (2012). A daily merged MODIS Aqua–Terra land surface temperature data set for the conterminous united states. *Remote Sensing of Environment*, doi:10.1016/j.rse.2011.12.019.

Demattê, J. L. I. e Demattê, J. A. M. (2009). Ambientes de produção como estratégia de manejo na cultura da cana-de-açúcar. *Informaçoes Agronômicas*, (127):10–18.

Dietterich, T. G. (2000). Ensemble methods in machine learning. In *International workshop on multiple classifier systems*, pages 1–15. Springer.

FAO (2018). Countries by commodity: Sugarcane. 'http://www.fao.org/faostat/en/#data/QC/visualize'. Data de acesso: 20 de Janeiro de 2018".

Ferraciolli, M. A., Bocca, F. F., e Rodrigues, L. H. A. (no prelo). Neglecting spatial autocorrelation causes underestimation of the error of sugarcane yield models. doi:10.1016/j.compag.2018.09.003.

Ferreira, D. A., Franco, H. C. J., Otto, R., Vitti, A. C., Fortes, C., Faroni, C. E., Garside, A. L., e Trivelin, P. C. O. (2015). Contribution of N from green harvest residues for sugarcane nutrition in brazil. doi:10.1111/gcbb.12292.

Franco, H. C. J., Otto, R., Faroni, C. E., Vitti, A. C., de Oliveira, E. C. A., e Trivelin, P. C. O. (2011). Nitrogen in sugarcane derived from fertilizer under Brazilian field conditions. *Field Crops Research*, doi:10.1016/j.fcr.2010.11.011.

Franco, H. C. J., Otto, R., Vitti, A. C., Faroni, C. E., Oliveira, E. C. d. A., Fortes, C., Ferreira, D. A., Kölln, O. T., Garside, A. L., e Trivelin, P. C. O. (2015). Residual recovery and yield performance of nitrogen fertilizer applied at sugarcane planting. *Scientia Agricola*, 72(6):528–534.

Franco, H. C. J. e Trivelin, P. C. O. (2010). Adubação nitrogenada em cana-de-açúcar: Reflexos do plantio à colheita. In Crusciol, C. A. C., Silva, M. d. A., e Rossetto, R., editores, *Tópicos em ecofisiologia da cana-de-açúcar*, pages 239–270. Fundação de Estudos e Pesquisas Agrícolas e Florestais, Botucatu.

Friedman, J. H. (2001). Greedy function approximation: A gradient boosting machine. *The Annals of Statistics*, 29(5):1189–1232.

Goldemberg, J., Coelho, S. T., e Guardabassi, P. (2008). The sustainability of ethanol production from sugarcane. *Energy Policy*, doi:10.1016/j.enpol.2008.02.028.

Goldemberg, J., Mello, F. F., Cerri, C. E., Davies, C. A., e Cerri, C. C. (2014). Meeting the global demand for biofuels in 2021 through sustainable land use change policy. *Energy Policy*, doi:10.1016/j.enpol.2014.02.008.

Goldstein, A., Kapelner, A., Bleich, J., e Pitkin, E. (2015). Peeking inside the black box: Visualizing statistical learning with plots of individual conditional expectation. *Journal of Computational and Graphical Statistics*, 24(1):44–65.

Guyon, I. e Elisseeff, A. (2003). An introduction to variable and feature selection. *Journal of machine learning research*, 3(Mar):1157–1182.

Hammer, R. G. (2016). *Modelagem da produtividade da cultura da cana de açúcar por meio do uso de técnicas de mineração de dados*. PhD thesis, Universidade de São Paulo, Piracicaba : Escola Superior de Agricultura Luiz de Queiroz.

Hastie, T., Tibshirani, R., e Friedman, J. (2009). *The Elements of Statistical Learning: Data Mining, Inference, and Prediction, Second Edition*. Springer Series in Statistics. Springer New York.

Head, T., MechCoder, Louppe, G., Shcherbatyi, I., fcharras, Vinícius, Z., cmmalone, chschroeder, nel215, Campos, N., Young, T., Cereda, S., Fan, T., Schwabedal, J., Hvass-Labs, Pak, M., SoManyUsernamesTaken, Callaway, F., Estève, L., Besson, L., Landwehr, P. M., Komarov, P., Cherti, M., Shi, K. K., Pfannschmidt, K., Linzberger, F., Cauet, C., Gut, A., Mueller, A., e Fabisch, A. (2018). scikit-optimize/scikit-optimize: v0.5rc1.

Huang, R., Zhang, C., Huang, J., Zhu, D., Wang, L., e Liu, J. (2015). Mapping of daily mean air temperature in agricultural regions using daytime and nighttime land surface temperatures derived from TERRA and AQUA MODIS data. *Remote Sensing*, doi:10.3390/rs70708728.

Huffman, G., Adler, R., Bolvin, D., e Nelkin, E. (2010). The TRMM multi-satellite precipitation analysis (tmpa). In Gebremichael, M. e Hossain, F., editores, *Satellite Rainfall Applications for Surface Hydrology*, pages 3–22. Springer Netherlands.

James, G., Witten, D., Hastie, T., e Tibshirani, R. (2013). *An introduction to statistical learning*, volume 112. Springer.

Karpatne, A., Atluri, G., Faghmous, J. H., Steinbach, M., Banerjee, A., Ganguly, A., Shekhar, S., Samatova, N., e Kumar, V. (2017a). Theory-guided data science: A new paradigm for scientific discovery from data. *IEEE Transactions on Knowledge and Data Engineering*, 29(10):2318–2331.

Karpatne, A., Watkins, W., Read, J., e Kumar, V. (2017b). Physics-guided neural networks (PGNN): an application in lake temperature modeling. *CoRR*, abs/1710.11431.

Kerns, G. J. (2010). *Introduction to probability and statistics using r*. Lulu. com.

Kingston, G. (2013). *Mineral Nutrition of Sugarcane*, pages 85–120. John Wiley & Sons Ltd.

Koizumi, T. (2014). Biofuels and food security in brazil. In *Biofuels and Food Security*, SpringerBriefs in Applied Sciences and Technology, pages 13–30. Springer International Publishing.

Lawes, R. e Lawn, R. (2005). Applications of industry information in sugarcane production systems. *Field Crops Research*, doi:10.1016/j.fcr.2005.01.033.

Leal, M. R. L., Galdos, M. V., Scarpare, F. V., Seabra, J. E., Walter, A., e Oliveira, C. O. (2013). Sugarcane straw availability, quality, recovery and energy use: A literature review. *Biomass and Bioenergy*, doi:10.1016/j.biombioe.2013.03.007. 20th European Biomass Conference.

LeCun, Y., Bengio, Y., e Hinton, G. (2015). Deep learning. *Nature*, 521(7553):436–444.

Liu, D., Kingston, G., e Bull, T. (1998). A new technique for determining the thermal parameters of phenological development in sugarcane, including suboptimum and supra-optimum temperature regimes. *Agricultural and Forest Meteorology*, 90(1-2):119–139.

Lopes, A. S. e Cox, F. R. (1977). A survey of the fertility status of surface soils under "Cerrado" vegetation in Brazil. *Soil Science Society of America Journal*, 41(4):742–747.

Mantelato, T. (2017). Efeito do uso de projeções climáticas na predição da produtividade de cana-de-açúcar. Trabalho de Conclusão Curso (Engenharia Agrícola), Faculdade de Engenharia Agrícola - Unicamp, Campinas - SP.

MAPA, M. d. A. P. e. A. (2018). *Valor bruto da produção*. Data de acesso: 20 de Janeiro de 2018.

Mariano, E., Otto, R., Montezano, Z. F., Cantarella, H., e Trivelin, P. C. (2017). Soil nitrogen availability indices as predictors of sugarcane nitrogen requirements. *European Journal of Agronomy*, 89:25–37.

McCray, J. M., Ji, S., Powell, G., Montes, G., e Perdomo, R. (2010). Sugarcane response to dris-based fertilizer supplements in Florida. *Journal of Agronomy and Crop Science*, doi:10.1111/j.1439-037X.2009.00395.x.

Meier, E. A., Thorburn, P. J., Wegener, M., e Basford, K. (2006). The availability of nitrogen from sugarcane trash on contrasting soils in the wet tropics of north queensland. *Nutrient Cycling in Agroecosystems*, doi:10.1007/s10705-006-9015-0.

Melo, D. d. C., Xavier, A. C., Bianchi, T., Oliveira, P. T., Scanlon, B. R., Lucas, M. C., e Wendland, E. (2015). Performance evaluation of rainfall estimates by trmm multi-satellite precipitation analysis 3b42v6 and v7 over brazil. *Journal of Geophysical Research: Atmospheres*, 120(18):9426–9436.

Oliveira, M. P. G. d., Bocca, F. F., e Rodrigues, L. H. A. (2017). From spreadsheets to sugar content modeling: A data mining approach. *Computers and Electronics in Agriculture*, doi:10.1016/j.compag.2016.11.012.

Otto, R., Castro, S., Mariano, E., Castro, S., Franco, H., e Trivelin, P. (2016). Nitrogen use efficiency for sugarcane-biofuel production: what is next? *BioEnergy Research*, 9(4):1272–1289.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., e Duchesnay, E. (2011). Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830.

Peloia, P. R., Bocca, F. F., e Rodrigues, L. H. A. (no prelo). Identification of patterns for increasing production with decision trees in sugarcane mill data . *Scientia Agricola*.

Peloia, P. R. e Rodrigues, L. H. (2016). Identification of commercial blocks of outstanding performance of sugarcane using data mining. *Engenharia Agrícola*, 36(5):895–901.

Penatti, C. P. (2013). Nitrogênio. In *Adubação de Cana-de-Açúcar: 30 anos de experiência*, pages 14–82. Ottoni Editora.

Pianosi, F. e Wagener, T. (2015). A simple and efficient method for global sensitivity analysis based on cumulative distribution functions. *Environmental Modelling & Software*, 67:1–11.

Raij, B. v., Cantarella, H., Quaggio, J., e Furlani, A. (1996). Cana-de-açúcar. In *Recomendações de adubação e calagem para o Estado de São Paulo.*, volume 100, pages 237–240. Campinas: Instituto Agronômico & Fundação IAC, 1996.

Reay, D. S., Davidson, E. A., Smith, K. A., Smith, P., Melillo, J. M., Dentener, F., e Crutzen, P. J. (2012). Global agriculture and nitrous oxide emissions. *Nature Climate Change*, 2(6):410–416.

Roberts, D. R., Bahn, V., Ciuti, S., Boyce, M. S., Elith, J., Guillera-Arroita, G., Hauenstein, S., Lahoz-Monfort, J. J., Schröder, B., Thuiller, W., et al. (2017). Cross-validation strategies for data with temporal, spatial, hierarchical, or phylogenetic structure. *Ecography*, 40:913–929.

Robertson, F. A. e Thorburn, P. J. (2007). Decomposition of sugarcane harvest residue in different climatic zones. *Soil Research*, 45(1):1–11.

Robertson, G. P. e Vitousek, P. M. (2009). Nitrogen in agriculture: Balancing the cost of an essential resource. *Annual Review of Environment and Resources*, doi:10.1146/annurev.environ.032108.105046.

Robinson, N., Brackin, R., Vinall, K., Soper, F., Holst, J., Gamage, H., Paungfoo-Lonhienne, C., Rennenberg, H., Lakshmanan, P., e Schmidt, S. (2011). Nitrate paradigm does not hold up for sugarcane. *PLoS ONE*, doi:10.1371/journal.pone.0019045.

Rudorff, B. F. T., Aguiar, D. A., Silva, W. F., Sugawara, L. M., Adami, M., e Moreira, M. A. (2010). Studies on the rapid expansion of sugarcane for ethanol production in São Paulo state (Brazil) using landsat data. *Remote Sensing*, doi:10.3390/rs2041057.

Rumelhart, D. E., Hinton, G. E., e Williams, R. J. (1985). Learning internal representations by error propagation. Technical report, California Univ San Diego La Jolla Inst for Cognitive Science.

Saltelli, A. e Annoni, P. (2010). How to avoid a perfunctory sensitivity analysis. *Environmental Modelling & Software*, doi:10.1016/j.envsoft.2010.04.012.

Saltelli, A., Ratto, M., Andres, T., Campolongo, F., Cariboni, J., Gatelli, D., Saisana, M., e Tarantola, S. (2008). *Global sensitivity analysis: the primer*. John Wiley & Sons.

Saltelli, A., Tarantola, S., e Chan, K.-S. (1999). A quantitative model-independent method for global sensitivity analysis of model output. *Technometrics*, 41(1):39–56.

Santos, H. d., Jacomine, P. K. T., dos Anjos, L., de Oliveira, V., de Oliveira, J. d., Coelho, M. R., Lumbreras, J. F., e Cunha, T. d. (2006). Sistema brasileiro de classificação de solos. *Embrapa Solos-Livros técnicos*.

Smola, A. J. e Schölkopf, B. (2004). A tutorial on support vector regression. *Statistics and computing*, doi:10.1023/B:STCO.0000035301.49549.88.

Sobol, I. M. (2001). Global sensitivity indices for nonlinear mathematical models and their monte carlo estimates. *Mathematics and computers in simulation*, 55(1):271–280.

Tartowski, S. L. e Howarth, R. W. (2001). Nitrogen, nitrogen cycle. In Levin, S. A., editor, *Encyclopedia of Biodiversity*, pages 377 – 388. Elsevier, New York.

Tedeschi, L. O. (2006). Assessment of the adequacy of mathematical models. *Agricultural systems*, 89(2-3):225–247.

Trivelin, P. C. O., Franco, H. C. J., Otto, R., Ferreira, D. A., Vitti, A. C., Fortes, C., Faroni, C. E., Oliveira, E. C. A., e Cantarella, H. (2013). Impact of sugarcane trash on fertilizer requirements for São Paulo, Brazil. *Scientia Agricola*, 70:345 – 352.

Vitti, A. C., Cantarella, H., Trivelin, P. C. O., e Rossetto, R. (2010). Nitrogênio. In Dinardo-Miranda, L. L., de Vasconcelos, A. C. M., e Landell, M. G. d. A., editores, *Cana-de-Açúcar*, pages 239–270. Instituto Agronômico, Campinas.

Wagner, N. e Rondinelli, J. M. (2016). Theory-guided machine learning in materials science. *Frontiers in Materials*, doi:10.3389/fmats.2016.00028.

Wan, Z., Hook, S., e Hulley., G. (2015a). *MODIS/Aqua Land Surface Temperature/Emissivity Daily L3 Global 1 km SIN Grid V006*.

Wan, Z., Hook, S., e Hulley., G. (2015b). *MODIS/Terra Land Surface Temperature/Emissivity Daily L3 Global 1 km SIN Grid V006*.

Williams, C. K. e Rasmussen, C. E. (2006). Gaussian processes for machine learning. *the MIT Press*, 2(3):4.

Zhao, D., Glaz, B., e Comstock, J. C. (2014). Physiological and growth responses of sugarcane genotypes to nitrogen rate on a sand soil. *Journal of Agronomy and Crop Science*, doi:10.1111/jac.12084.

APÊNDICE A – Códigos de textura e fertilidade do solo

Codificação do atributo *text_dmt* (tabela 11) e $fert\_dmt$ (tabela 12 conforme Demattê e Demattê (2009).

# Text.	Descrição
1	argila maior que 65 %
2	argila entre 36 e 65 %
3	argila entre 26 e 35 %
4	argila entre 16 e 25 %
5	argila entre 10 e 15 %
6	argila menor que 9 %
7	argila menor que 35 % e areia menor que 15 %

Tabela 11 – Categorias de textura (# Text.) do solo em função dos percentuais de argila e areia e sua descrição. Adaptado de Demattê e Demattê (2009).

# Fert.	Descrição
1	Eutrófico: saturação por bases maior que 50 % no perfil
2	Epieutrófico: saturação por bases maior que 50 % na superfície
3	Distrófico: saturação de bases inferior a 50 % em todo perfil mas não àlico
4	Álico: saturação por Al (m %) superior a 50 % em todo perfil ou nas camadas inferiores
5	Ácrico: solos contendo quantidade iguais ou menores que 1,5 $cmol_c \cdot kg^{-1}$ extraível por KCl 1N e pH KCl igual ou superior a 5,0 ou delta pH positivo
6	ou nulo Alumínico: m % maior que 50 e Al maior que 40 $mmol_c \cdot kg^{-1}$
7	Alítico: idem 6 e atividade de argila maior que 200 $mmol_c \cdot kg^{-1}$

Tabela 12 – Categorias de fertilidade (# Fert.) do solo e sua descrição. Adaptado de Demattê e Demattê (2009).

APÊNDICE B – Hiper-parâmetros escolhidos

A tebela 13 apresenta os hiper-parâmetros determinados para cada técnica em cada subconjunto de dados. Os valores correspondem a configuração que apresentou o menor erro medido por validação cruzada *K-fold* (5 *folds*)no conjunto de treino.

Técnica	Hiper-parâmetro	1	2	3	4
BRT	Número de árvores	533	1000	881	799
	Taxa de aprendizado	0,160656	0,027267	0,057686	0,045940
	Profundidade máxima	2	5	6	5
	Mínimo de amostras para split	0,015543	0,090671	0,100000	0,063717
	Fração de amostragem	0,762792	0,575759	0,836735	0,501249
	Fração de sorteio	0,121226	0,1	0,259919	0,189977
RF	Número de árvores	958	1000	1000	784
	Fração de sorteio	0,389497	0,295443	0,704271	0,050000
	Mínimo de amostras para split	0,001	0,001	0,001	0,001
SVR	Custo	16,72866	19,13268	45,04331	30,75387
	Gamma	0,005413	0,012115	0,022918	0,028767
	Epsilon[1]	9,723356	9,615603	4,254846	6,224210

Tabela 13 – Hiper-parâmetros para cada técnica (BRT - *Boosted Regression Trees*, RF - *Random Forest*, SVR - *Support Vector Regression*) nos subconjunto de dados 1 a 4. [1] Valores reportados na escala original dos dados, sendo o ajuste considerando o range de 0,05 a 0,3 conforme tabela 4.

APÊNDICE C – Importância de atributos e análise de sensibilidade

Atributo	Imp.	BRT_I	RF_I	SVR_I	BRT_I2	RF_I2	SVR_I2
d_irrig	0,708	0,012	–	0,143	0,035	–	0,232
q_irrig	0,646	0,094	–	0,153	0,112	–	0,240
d_vin	0,438	0,149	0,009	0,283	0,170	0,049	0,285
RB72454	0,394	0,014	0,002	0,199	0,043	0,056	0,201
RB835054	0,336	0,062	0,022	0,147	0,089	0,049	0,155
RB855453	0,284	0,071	0,062	0,122	0,095	0,067	0,136
RB845210	0,280	0,076	0,005	0,184	0,097	0,055	0,180
SP803280	0,255	0,092	0,003	0,169	0,113	0,057	0,209
SP835073	0,207	0,030	0,005	0,088	0,053	0,056	0,104
RB835486	0,199	0,014	0,002	0,072	0,043	0,055	0,102
n_tipo	0,198	0,013	0,003	0,027	0,042	0,056	0,116
SP832847	0,181	0,011	0,006	0,103	0,046	0,057	0,116
qfert_k	0,174	0,076	0,072	0,031	0,107	0,091	0,133
SP801842	0,172	0,020	0,009	0,055	0,042	0,057	0,156
colheitas	0,171	0,174	0,178	0,212	0,185	0,200	0,262
RB966928	0,159	0,010	0,001	0,048	0,043	0,056	0,119
qfert_n	0,151	0,042	0,056	0,124	–	–	–
ambprod_D	0,144	0,031	0,004	0,062	0,040	0,057	0,129
RB855156	0,139	0,058	0,003	0,041	0,073	0,057	0,147
queima	0,136	0,030	0,006	0,100	0,059	0,056	0,148
d_gesso	0,129	0,037	0,008	0,028	0,063	0,052	0,125
qfert_p	0,123	0,074	0,048	0,066	0,100	0,053	0,118
Outras	0,118	0,025	0,005	0,025	0,045	0,056	0,101
ppt_i	0,095	0,135	0,043	0,069	0,145	0,064	0,127
tsn_i	0,095	0,077	0,033	0,160	0,117	0,061	0,171
ppt_ii	0,084	0,033	0,074	0,156	0,067	0,086	0,174
d_calc	0,083	0,047	0,006	0,057	0,085	0,054	0,149
dfert	0,082	0,060	0,005	0,058	0,081	0,053	0,116
ambprod_F	0,076	0,016	0,003	0,021	0,044	0,056	0,127

Continua na próxima página.

Continuação da página anterior.

Atributo	Imp.	BRT_I	RF_I	SVR_I	BRT_I2	RF_I2	SVR_I2
ppt_iii	0,075	0,077	0,070	0,149	0,098	0,077	0,170
tsd_ii	0,069	0,156	0,077	0,296	0,168	0,104	0,298
q_calc	0,069	0,077	0,020	0,079	0,105	0,048	0,131
tsd_iii	0,069	0,080	0,043	0,167	0,107	0,059	0,179
espac	0,069	0,008	0,014	0,030	0,042	0,055	0,128
tsn_ii	0,067	0,077	0,017	0,052	0,107	0,055	0,127
tsd_i	0,064	0,038	0,017	0,064	0,074	0,056	0,123
tsn_iii	0,064	0,034	0,027	0,136	0,066	0,062	0,158
RB855536	0,060	0,023	0,002	0,052	0,042	0,056	0,129
q_gesso	0,056	0,067	0,007	0,040	0,098	0,052	0,137
ambprod_E	0,056	0,027	0,004	0,045	0,047	0,049	0,121
ambprod_C	0,030	0,023	0,002	–	0,042	0,056	–
RB867515	0,024	0,017	0,005	–	0,041	0,057	–
ambprod_G	0,024	0,011	0,004	–	0,040	0,057	–
d_torta	0,000	0,128	0,002	–	0,166	0,056	–
q_vin	0,000	–	0,005	–	0,042	–	–
q_torta	0,000	0,074	0,003	–	0,105	0,057	–

Tabela 14 – Importância (Imp.) dos atributos do subconjunto de dados 1. A sensibilidade aos atributos das diferentes técnicas (BRT - *Boosted Regression Trees*, RF - *Random Forest* e SVR - *Suppor Vector Regression*) é indicada por I enquanto a sensibilidade à interação do atributo com a adubação nitrogenada é indicada por I2. Atributos que não foram selecionados são marcados como '–' na tabela.

Atributo	Imp.	BRT_I	RF_I	SVR_I	BRT_I2	RF_I2	SVR_I2
d_irrig	0,663	0,052	–	0,173	0,111	–	0,245
q_irrig	0,500	0,028	–	0,197	0,094	–	0,314
d_vin	0,442	0,206	0,005	0,267	0,222	0,028	0,302
SP803280	0,389	0,063	0,038	0,211	0,096	0,056	0,232
RB72454	0,366	0,046	0,008	0,180	0,094	0,029	0,194
RB835054	0,357	0,030	0,017	0,055	0,079	0,034	0,113
RB855453	0,306	0,090	0,088	0,142	0,124	0,104	0,144
RB845210	0,302	0,054	0,009	0,150	0,119	0,028	0,147
RB966928	0,257	0,101	0,001	0,250	0,116	0,026	0,248
SP832847	0,240	0,010	0,009	0,035	0,081	0,030	0,089
RB835486	0,224	0,004	0,003	0,118	0,076	0,026	0,124
qfert_k	0,204	0,104	0,051	0,055	0,138	0,082	0,121
colheitas	0,161	0,134	0,183	0,161	0,122	0,237	0,202
qfert_p	0,156	0,043	0,032	0,051	0,100	0,045	0,100
solo_O	0,142	0,010	0,013	0,061	0,078	0,031	0,109
qfert_n	0,138	0,076	0,026	0,110	–	–	–
n_tipo	0,133	0,023	0,005	0,052	0,084	0,027	0,104
Outras	0,119	0,035	0,016	0,045	0,097	0,034	0,075
ambprod_F	0,115	0,005	0,004	0,029	0,079	0,027	0,119
SP801842	0,113	0,018	0,004	0,055	0,082	0,026	0,137
ambprod_D	0,112	0,005	0,005	0,044	0,077	0,027	0,130
queima	0,109	0,011	0,008	0,064	0,083	0,029	0,146
frt_dmt	0,109	0,029	0,028	0,052	0,086	0,040	0,101
SP835073	0,107	0,030	0,005	0,034	0,093	0,027	0,094
RB855156	0,107	0,037	0,003	0,025	0,085	0,025	0,128
ppt_i	0,097	0,123	0,051	0,049	0,160	0,062	0,118
txt_dmt	0,095	0,037	0,050	0,026	0,072	0,075	0,116
tsd_iii	0,093	0,055	0,046	0,102	0,090	0,068	0,130
d_gesso	0,088	0,060	0,014	0,039	0,090	0,030	0,110
tsd_ii	0,087	0,091	0,063	0,240	0,137	0,080	0,235
tsn_i	0,087	0,038	0,040	0,179	0,100	0,052	0,188
dfert	0,084	0,072	0,009	0,041	0,105	0,035	0,079
ppt_ii	0,079	0,034	0,086	0,115	0,094	0,097	0,139
q_calc	0,074	0,114	0,023	0,037	0,118	0,033	0,118
ambprod_E	0,072	0,015	0,005	0,034	0,081	0,027	0,113

Continua na próxima página.

Continuação da página anterior.

Atributo	Imp.	BRT_I	RF_I	SVR_I	BRT_I2	RF_I2	SVR_I2
d_calc	0,071	0,027	0,007	0,059	0,098	0,028	0,145
espac	0,070	0,018	0,015	0,025	0,084	0,033	0,114
ambprod_C	0,069	0,011	–	0,022	0,077	–	0,113
tsn_iii	0,068	0,039	–	0,133	0,098	–	0,150
ppt_iii	0,066	0,058	–	0,096	0,101	–	0,115
tsd_i	0,059	0,071	–	0,047	0,127	–	0,111
tsn_ii	0,057	0,079	–	0,052	0,082	–	0,102
q_gesso	0,056	0,104	–	0,036	0,080	–	0,115
RB855536	0,054	0,009	–	0,037	0,080	–	0,130
ambprod_G	0,050	0,012	–	–	0,078	–	–
RB867515	0,033	0,026	–	–	0,089	–	–
solo_R	0,028	0,003	–	–	0,078	–	–
q_vin	0,000	–	–	–	–	–	–
q_torta	0,000	0,063	–	–	0,125	–	–
d_torta	0,000	0,055	–	–	0,108	–	–

Tabela 15 – Importância (Imp.) dos atributos do subconjunto de dados 2. A sensibilidade aos atributos das diferentes técnicas (BRT - *Boosted Regression Trees*, RF - *Random Forest* e SVR - *Suppor Vector Regression*) é indicada por I enquanto a sensibilidade à interação do atributo com a adubação nitrogenada é indicada por I2. Atributos que não foram selecionados são marcados como '–' na tabela.

Atributo	Imp.	BRT_I	RF_I	SVR_I	BRT_I2	RF_I2	SVR_I2
SP803280	0,906	0,059	0,003	0,293	0,084	0,090	0,286
q_irrig	0,635	0,008	0,002	0,257	0,052	0,089	0,253
d_irrig	0,635	–	0,002	0,196	–	0,089	0,239
q_torta	0,615	0,017	0,003	0,206	0,048	0,089	0,251
d_torta	0,615	0,011	0,002	0,172	0,050	0,089	0,210
SP801842	0,558	0,077	0,008	0,160	0,097	0,089	0,166
RB835486	0,491	0,003	0,005	0,139	0,052	0,089	0,169
ambprod_D	0,453	0,069	0,017	0,093	0,090	0,091	0,111
RB835054	0,452	0,037	0,007	0,171	0,067	0,089	0,181
RB845210	0,444	0,009	0,003	0,115	0,053	0,089	0,136
n_tipo	0,444	0,008	0,004	0,074	0,061	0,089	0,099
queima	0,376	0,014	0,003	0,213	0,049	0,089	0,215
RB855536	0,339	0,041	0,008	0,078	0,066	0,087	0,097
RB855453	0,327	0,044	0,016	0,045	0,079	0,089	0,072
ctce	0,308	0,032	0,088	0,019	0,071	0,138	0,056
d_gesso	0,306	0,023	0,013	0,031	0,060	0,091	0,061
RB966928	0,287	0,019	0,006	0,226	0,055	0,087	0,219
p	0,284	0,032	0,020	0,033	0,069	0,102	0,063
q_vin	0,282	0,043	0,062	0,039	0,066	0,111	0,065
qfert_n	0,281	0,052	0,089	0,044	–	–	–
frt_dmt	0,265	0,046	0,016	0,047	0,074	0,093	0,071
sb	0,263	0,033	0,012	0,015	0,058	0,091	0,058
qfert_k	0,247	0,034	0,089	0,021	0,064	0,103	0,049
colheitas	0,236	0,050	0,063	0,086	0,075	0,077	0,116
tsd_ii	0,236	0,075	0,058	0,064	0,084	0,103	0,090
mg	0,228	0,034	0,011	0,020	0,059	0,087	0,050
hal	0,227	0,054	0,020	0,027	0,089	0,094	0,057
h	0,227	0,023	0,018	0,027	0,062	0,093	0,057
Outras	0,226	0,027	0,007	0,104	0,058	0,089	0,128
d_calc	0,225	0,034	0,018	0,033	0,068	0,094	0,057
v	0,216	0,045	0,018	0,059	0,069	0,083	0,085
ambprod_AB	0,215	0,017	0,006	0,056	0,054	0,088	0,074
k	0,204	0,022	0,018	0,012	0,056	0,090	0,055
ph	0,203	0,034	0,014	0,018	0,055	0,091	0,052
m	0,202	0,032	0,019	0,023	0,061	0,087	0,053

Continua na próxima página.

Continuação da página anterior.

Atributo	Imp.	BRT_I	RF_I	SVR_I	BRT_I2	RF_I2	SVR_I2
SP835073	0,202	0,014	0,007	0,045	0,051	0,088	0,069
d_vin	0,201	0,049	0,041	0,090	0,059	0,088	0,106
ctcp	0,199	0,036	0,011	0,015	0,063	0,089	0,057
qfert_p	0,193	0,078	0,032	0,048	0,106	0,108	0,084
ppt_iii	0,190	0,026	0,020	0,053	0,067	0,089	0,076
tsd_i	0,187	0,035	0,014	0,033	0,060	0,091	0,058
al	0,185	0,025	0,032	0,016	0,059	0,098	0,050
ambprod_F	0,183	0,016	0,006	0,039	0,052	0,089	0,069
tsn_iii	0,183	0,034	0,015	0,041	0,061	0,090	0,067
ambprod_C	0,182	0,015	0,009	0,043	0,053	0,090	0,059
tsd_iii	0,181	–	0,029	0,031	–	0,102	0,060
dfert	0,181	–	0,031	0,013	–	0,096	0,057
ppt_i	0,179	–	0,017	0,042	–	0,090	0,069
q_calc	0,171	–	0,016	0,025	–	0,086	0,051
ppt_ii	0,166	–	0,035	0,019	–	0,110	0,050
SP813250	0,166	–	0,010	0,051	–	0,088	0,062
tsn_i	0,160	–	0,041	0,045	–	0,122	0,065
q_gesso	0,157	–	0,008	0,017	–	0,089	0,052
SP832847	0,156	–	0,004	0,047	–	0,088	0,071
espac	0,155	–	0,007	0,036	–	0,089	0,057
tsn_ii	0,130	–	0,021	0,049	–	0,089	0,071
RB855156	0,114	–	0,012	0,033	–	0,089	0,056
ambprod_E	0,101	–	0,007	0,047	–	0,088	0,069

Tabela 16 – Importância (Imp.) dos atributos do subconjunto de dados 3. A sensibilidade aos atributos das diferentes técnicas (BRT - *Boosted Regression Trees*, RF - *Random Forest* e SVR - *Suppor Vector Regression*) é indicada por I enquanto a sensibilidade à interação do atributo com a adubação nitrogenada é indicada por I2. Atributos que não foram selecionados são marcados como '–' na tabela.

Atributo	Imp.	BRT_I	RF_I	SVR_I	BRT_I2	RF_I2	SVR_I2
SP801842	0,743	0,016	0,027	0,226	0,064	0,048	0,234
SP813250	0,663	0,049	0,019	0,246	0,061	0,037	0,261
RB835054	0,455	0,102	0,049	0,182	0,114	0,067	0,194
solo_O	0,409	0,053	0,035	0,131	0,066	0,055	0,134
SP835073	0,396	0,092	0,033	0,112	0,099	0,053	0,108
solo_R	0,388	0,021	0,025	0,132	0,060	0,044	0,130
mg	0,367	0,074	0,042	0,043	0,096	0,052	0,075
d_gesso	0,366	0,094	0,030	0,045	0,105	0,053	0,057
m	0,354	0,044	0,025	0,029	0,078	0,047	0,056
qfert_n	0,337	0,064	0,037	0,044	–	–	–
tsd_ii	0,334	0,055	0,038	0,063	0,090	0,050	0,084
qfert_k	0,332	0,070	0,037	0,026	0,114	0,061	0,057
colheitas	0,331	0,067	0,045	0,064	0,082	0,057	0,086
ctce	0,317	0,051	0,042	0,034	0,070	0,054	0,064
Outras	0,313	0,135	0,029	0,114	0,141	0,044	0,115
ppt_i	0,312	0,070	0,034	0,046	0,103	0,047	0,063
p	0,303	0,038	0,023	0,049	0,071	0,044	0,065
sb	0,301	0,043	0,027	0,034	0,074	0,047	0,065
v	0,298	0,048	0,029	0,043	0,089	0,048	0,070
hal	0,295	0,036	0,033	0,034	0,074	0,054	0,061
h	0,295	0,043	0,031	0,034	0,077	0,055	0,061
argila	0,286	0,088	0,047	0,044	0,124	0,068	0,070
ambprod_F	0,280	0,053	0,015	0,051	0,065	0,039	0,063
k	0,277	0,076	0,032	0,035	0,086	0,045	0,069
tsd_iii	0,276	0,071	0,049	0,069	0,125	0,058	0,087
areia	0,271	0,102	0,045	0,043	0,154	0,060	0,068
ph	0,265	0,044	0,023	0,024	0,072	0,042	0,056
ctcp	0,265	0,097	0,030	0,034	0,100	0,050	0,065
tsn_iii	0,260	–	0,024	0,025	–	0,045	0,052
dfert	0,257	–	0,036	0,026	–	0,051	0,050
silte	0,251	–	0,036	0,033	–	0,049	0,058
d_vin	0,249	–	0,062	0,128	–	0,051	0,125
tsd_i	0,249	–	0,022	0,035	–	0,043	0,057
ppt_iii	0,245	–	0,025	0,030	–	0,041	0,053
al	0,239	–	0,018	0,024	–	0,042	0,053

Continua na próxima página.

Continuação da página anterior.

Atributo	Imp.	BRT_I	RF_I	SVR_I	BRT_I2	RF_I2	SVR_I2
qfert_p	0,238	–	0,061	0,089	–	0,053	0,099
ambprod_G	0,227	–	0,017	0,108	–	0,043	0,106
ppt_ii	0,226	–	0,033	0,030	–	0,054	0,057
q_gesso	0,222	–	0,023	0,024	–	0,047	0,053
d_calc	0,213	–	0,024	0,036	–	0,047	0,066
RB855536	0,206	–	0,028	0,044	–	0,044	0,057
tsn_i	0,198	–	0,021	0,031	–	0,046	0,055
q_calc	0,185	–	0,037	0,035	–	0,035	0,062
tsn_ii	0,175	–	0,023	0,057	–	0,047	0,074
SP832847	0,173	–	0,015	0,066	–	0,044	0,072
ambprod_C	0,166	–	0,032	0,068	–	0,052	0,074
RB855453	0,155	–	0,021	0,052	–	0,038	0,062
ambprod_E	0,151	–	0,013	0,061	–	0,041	0,071
RB855156	0,143	–	0,025	0,061	–	0,043	0,077
ambprod_AB	0,142	–	0,013	0,045	–	0,039	0,066
q_vin	0,101	–	0,022	0,079	–	0,043	0,100
espac_1	0,040	–	0,017	0,032	–	0,044	0,056
SP803280	0,000	–	–	–	–	–	–

Tabela 17 – Importância (Imp.) dos atributos do subconjunto de dados 4. A sensibilidade aos atributos das diferentes técnicas (BRT - *Boosted Regression Trees*, RF - *Random Forest* e SVR - *Suppor Vector Regression*) é indicada por I enquanto a sensibilidade à interação do atributo com a adubação nitrogenada é indicada por I2. Atributos que não foram selecionados são marcados como '–' na tabela.