UNIVERSIDADE ESTADUAL DE CAMPINAS

INSTITUTO DE ECONOMIA

ELAINE PRISCILA DE ANDRADE GARCIA

AVALIAÇÃO DA SUSTENTABILIDADE DO SOLO ATRAVÉS DE APRENDIZADO DE MÁQUINA

CAMPINAS

2024

UNIVERSIDADE ESTADUAL DE CAMPINAS

INSTITUTO DE ECONOMIA

ELAINE PRISCILA DE ANDRADE GARCIA

AVALIAÇÃO DA SUSTENTABILIDADE DO SOLO ATRAVÉS DE APRENDIZADO DE MÁQUINA

Tese de Doutorado apresentada ao Programa de Pós-Graduação em Desenvolvimento Econômico do Instituto de Economia da Universidade Estadual de Campinas como parte dos requisitos para obtenção do título de Doutora em Desenvolvimento Econômico na área de concentração de Economia Agrícola e do Meio Ambiente.

Prof. Dr. Sérgio Gomes Tôsto - Orientador

Profa. Dra. Ivette R. Luna Huamaní - Coorientadora

ESTE TRABALHO CORRESPONDE À VERSÃO FINAL DA TESE DEFENDIDA PELA ALUNA ELAINE PRISCILA DE ANDRADE GARCIA, ORIENTADA PELO PROF. DR. SÉRGIO GOMES TÔSTO.

CAMPINAS

2024

Ficha catalográfica
Universidade Estadual de Campinas (UNICAMP)
Biblioteca do Instituto de Economia
Luana Araujo de Lima - CRB 8/9706

G165a

Garcia, Elaine Priscila de Andrade, 1976-
Avaliação da sustentabilidade do solo através de aprendizado de máquina / Elaine Priscila de Andrade Garcia. – Campinas, SP : [s.n.], 2024.

Orientador(es): Sérgio Gomes Tôsto.
Coorientador(es): Ivette Raymunda Luna Huamaní.
Tese (doutorado) – Universidade Estadual de Campinas (UNICAMP), Instituto de Economia.

1. Solos - Qualidade. 2. Solos - Conservação. 3. Desenvolvimento rural. 4. Aprendizado de máquina. 5. Inovações agrícolas. I. Tôsto, Sérgio Gomes, 1957-. II. Luna Huamaní, Ivette Raymunda, 1978-. III. Universidade Estadual de Campinas (UNICAMP). Instituto de Economia. IV. Título.

Informações complementares

Título em outro idioma: Soil sustainability assessment through machine learning
Palavras-chave em inglês:
Soils - Quality
Soil conservation
Rural development
Machine learning
Agricultural innovations
Área de concentração: Economia Aplicada, Agrícola e do Meio Ambiente
Titulação: Doutora em Desenvolvimento Econômico
Banca examinadora:
Sérgio Gomes Tôsto [Orientador]
Lauro Charlet Pereira
Romis Ribeiro de Faissol Attux
Lucas Ferreira Lima
Gisele Freitas Vilela
Data de defesa: 26-11-2024
Programa de Pós-Graduação: Desenvolvimento Econômico

Identificação e informações acadêmicas do(a) aluno(a)
- ORCID do autor: https://orcid.org/0000-0001-6225-2674
- Currículo Lattes do autor: https://lattes.cnpq.br/5887509054739295

UNIVERSIDADE ESTADUAL DE CAMPINAS

INSTITUTO DE ECONOMIA

ELAINE PRISCILA DE ANDRADE GARCIA

AVALIAÇÃO DA SUSTENTABILIDADE DO SOLO ATRAVÉS DE APRENDIZADO DE MÁQUINA

Prof. Dr. Sérgio Gomes Tôsto - Orientador

Profa. Dra. Ivette R. Luna Huamaní - Coorientadora

Defendida em 26/11/2024

COMISSÃO JULGADORA

Prof. Dr. Sérgio Gomes Tôsto - PRESIDENTE
Universidade Estadual de Campinas (UNICAMP)

Prof. Dr. Romis Ribeiro de Faissol Attux
Universidade Estadual de Campinas (UNICAMP)

Prof. Dr. Lucas Ferreira Lima
Universidade Estadual de Campinas (UNICAMP)

Prof. Dr. Lauro Charlet Pereira
Embrapa Meio Ambiente

Profa. Dra. Gisele Freitas Vilela
Embrapa Territorial

A Ata de defesa com as respectivas assinaturas dos membros encontra-se no SIGA/Sistema de Fluxo de Dissertação/Tese e na Secretaria do Programa da Unidade.

AGRADECIMENTOS

Ao longo da minha jornada, muitas pessoas contribuíram para a construção de mais um capítulo da minha vida. Divido essa felicidade com todos que, direta ou indiretamente, participaram da elaboração desta tese, da concepção à conclusão.

Neste momento, é com grande gratidão que expresso meus mais sinceros agradecimentos a todos que me apoiaram com incentivo, energia positiva, compreensão e conhecimentos técnicos. A contribuição de cada um foi essencial para a realização deste projeto. Muito obrigada!

Agradeço, imensamente,

À minha família pelo apoio incondicional ao longo desses anos de desafios e conquistas. Sou profundamente grata por toda a generosidade, bondade, desprendimento e, acima de tudo, pelo amor. Agradeço, em especial, ao meu marido, aos meus pais e à minha sogra, que nunca mediram esforços para estar ao meu lado nessa jornada, e ao meu filho, que com sua sabedoria infantil e amor puro, sempre me trouxe forças e alegria.

À Universidade Estadual de Campinas, pela oportunidade de aprimorar meus conhecimentos e realizar este curso, essencial para minha trajetória.

Ao Prof. Dr. Sérgio Gomes Tôsto pela orientação, confiança e ensinamentos em mais uma jornada.

À Profa. Dra. Ivette R. Luna Huamaní pela coorientação e apoio no desenvolvimento da tese.

Aos membros da banca examinadora de qualificação e de defesa pelos feedbacks recebidos na execução desta pesquisa.

"Nem tudo o que se vê pode ser mudado, mas nada pode ser mudado até que seja visto. **"**

James Baldwin

RESUMO

Até o presente momento, cerca de 20 a 40% da área terrestre total está degradada ou possui algum grau variado de degradação devido às atividades humanas. Esta condição preocupa, pois a mesma modifica as estruturas ecossistêmicas do solo, impondo limites à sua sustentabilidade. Neste sentido, as medidas para preservar o solo tornaram-se mais do que urgentes diante das perspectivas complexas da segurança alimentar global e do agravamento da crise climática. Frente aos desafios globais, esta pesquisa propõe a avaliação da sustentabilidade do solo agrícola através de Aprendizado de Máquina. Seu objetivo dentro da área de Desenvolvimento Econômico foi combinar diferentes critérios de qualidade do solo para fornecer novas percepções sobre a sua sustentabilidade no campo agrícola e, com isso, ajudar o produtor rural a minimizar as perdas (gestão de risco), antecipar as etapas da degradação do solo (uso sustentável do solo) e melhorar a autonomia financeira com redução de custos de produção, principalmente, daqueles com maior dificuldade ao acesso técnico especializado. A pesquisa foi realizada comparando quatro técnicas de Aprendizado de Máquina com 3 técnicas para seleção de variáveis da base de dados, sendo uma usando o processo sociotécnico M-MACBETH e duas delas usando Wrappers. Como resultado, a técnica *Random Forest* em conjunto com a RFE (*Recursive Feature Elimination)* mostrou maior robustez nas métricas de desempenho, apresentando 95% de sensibilidade, 97% de especificidade e 96% de precisão, indicando o potencial da técnica para ser utilizada como classificador para a avaliação da sustentabilidade do solo. Um modelo com boa sensibilidade, especificidade e precisão pode identificar corretamente solos que precisam de intervenções como adição de nutrientes ou manejo conservacionista. Assim, o diagnóstico da sustentabilidade do solo ajuda a melhorar a gestão, uma vez que apoia o produtor rural na tomada de decisão para que ele tenha ganhos socioeconômicos de forma sustentável e contribui com o meio ambiente e com o fornecimento de alimentos mais seguros ao estimular o manejo adequado.

Palavras-chave: Qualidade do solo, práticas agrícolas, autonomia financeira, *Random Forest*, M-MACBETH.

ABSTRACT

Currently, approximately 20 to 40% of the total terrestrial area is degraded or shows varying degrees of degradation due to human activities. This condition is concerning as it alters the soil's ecosystem structures, imposing limits on its sustainability. In this context, measures to preserve soil have become more than urgent in light of the complex prospects for global food security and the worsening climate crisis. In response to global challenges, this research proposes evaluating the sustainability of agricultural soil using Machine Learning. Its objective within the area of Economic Development was to combine different soil quality criteria to provide new insights into its sustainability in agricultural fields. This approach aims to help farmers minimize losses (risk management), anticipate soil degradation stages (sustainable soil use), and improve financial autonomy by reducing production costs, especially for those with limited access to specialized technical support. The study compared four Machine Learning techniques with three variable selection methods for the dataset: one using the socio-technical M-MACBETH process and two using Wrappers. As a result, the Random Forest technique, combined with RFE (Recursive Feature Elimination), demonstrated greater robustness in performance metrics, achieving 95% sensitivity, 97% specificity, and 96% accuracy. This indicates the potential of the technique to serve as a classifier for assessing soil sustainability. A model with good sensitivity, specificity and accuracy can correctly identify soils that need interventions such as nutrient addition or conservation management. Thus, soil sustainability diagnosis helps in management, as it supports rural producers in decision-making so that they can have sustainable socioeconomic gains and contribute to the environment and to the supply of safe food.

Keywords: Soil quality, agricultural practices, financial autonomy, Random Forest, M-MACBETH.

LISTA DE FIGURAS

LISTA DE TABELAS

SUMÁRIO

CAPÍTULO I - INTRODUÇÃO

A sustentabilidade do solo agrícola é um fator-chave para o desenvolvimento econômico. Isto porque a produtividade do solo é essencial para gerar renda a longo prazo e aumentar a quantidade e a qualidade dos produtos agrícolas. Tal sustentabilidade dentro do contexto desta pesquisa refere-se à capacidade do solo de manter sua estrutura física, química e biológica, evitando a erosão e a compactação, que podem afetar negativamente a produtividade das culturas e a qualidade do solo a longo prazo.

A valorização do solo pode transformar comunidades rurais em áreas de desenvolvimento econômico dinâmico. Quando o solo é visto como um recurso valioso e protegido, a terra pode ser usada para atividades que geram renda e crescimento econômico. O desenvolvimento de cadeias produtivas locais ajudam a gerar empregos e aumentar a renda de pequenos agricultores, criando oportunidades de negócios e estimulando o crescimento da economia local.

Outro aspecto econômico importante relacionado à sustentabilidade do solo é a prática do turismo rural. A preservação do solo e a adoção de práticas sustentáveis podem atrair turistas interessados em vivenciar experiências agrícolas autênticas, aumentando o fluxo de visitantes e impulsionando o desenvolvimento econômico das regiões agrícolas (Stone; Nyaupane, 2018).

Mas um dos principais fatores da sustentabilidade do solo é a sua grande influência na segurança alimentar e nutricional, garantindo a disponibilidade de alimentos saudáveis e nutritivos para uma população cada vez maior. Isso pode contribuir para a redução da fome e da pobreza, melhorando a qualidade de vida das pessoas e gerando benefícios econômicos a longo prazo (IPEA, 2023). Outro aspecto é sobre a mudança climática. No Brasil, as emissões de gases de efeito estufa (GEE) estão diretamente relacionadas ao desmatamento e à atividade agropecuária. Essas duas atividades juntas, representaram 73% das emissões totais no país em 2023, segundo o relatório da SEEG (2024).

É fato que, ao longo do tempo, as tecnologias têm ajudado a melhorar a sustentabilidade do solo agrícola uma vez que elas têm o poder de transformar o modo como o manejo do solo é estrategicamente pensado. As tecnologias

modernas permitem monitorar as condições do solo em tempo real, incluindo a umidade, pH, níveis de nutrientes e outros fatores que afetam a sustentabilidade do solo. Com essas informações, os agricultores podem tomar decisões mais assertivas sobre o manejo do solo, incluindo a aplicação de fertilizantes e agroquímicos que levam em conta as necessidades específicas das plantas em diferentes áreas do campo. Segundo Doran *et al.* (2018), o uso de fertilizantes de forma inteligente pode minimizar o seu desperdício e reduzir a poluição do solo, contribuindo para uma produção mais eficiente e sustentável.

Neste sentido, a agricultura de precisão usa tecnologias como o GPS e sensores para orientar a semeadura, a irrigação, a aplicação de fertilizantes e outros processos agrícolas de forma mais precisa e eficiente. Isso pode reduzir a compactação do solo, minimizar o uso de recursos e maximizar a produtividade das culturas. De acordo com o ISPA (2021), a implementação de práticas de Agricultura de Precisão pode levar a uma diminuição de até 20% nos gastos com insumos.

Tecnologias de precisão podem ajudar a implementar técnicas de plantio direto, que envolvem a semeadura de culturas sem o revolvimento do solo, preservando a matéria orgânica e os microrganismos benéficos no solo. A adoção de práticas de agricultura de conservação, como a rotação de culturas, a cobertura do solo com plantas de cobertura e a redução do uso de arados, aliada ao uso de tecnologias adequadas, portanto, contribui para a melhoria da qualidade do solo, promovendo a sua sustentabilidade e fertilidade a longo prazo (Stott; Moebius-Clune, 2017).

Diante dessas possibilidades, é possível reduzir o desperdício de insumos, minimizar a contaminação do solo e da água, além de diminuir a emissão de gases de efeito estufa (OECD/FAO, 2023). Isto ajuda na preservação da biodiversidade ao manter o ciclo natural de nutrientes e carbono no solo, minimizando os impactos ambientais e os riscos para a saúde humana. Portanto, as práticas sustentáveis não apenas beneficiam o meio ambiente, mas também agregam valor econômico ao atender os consumidores que estão cada vez mais exigentes em relação à procedência dos alimentos.

Dado a importância do solo e tudo o que ele representa para a vida terrestre, a motivação para esta pesquisa parte da preocupação em como alimentar uma

demanda crescente de pessoas no planeta ao mesmo tempo que é necessário ter ações que possam reduzir os efeitos da mudança climática, fatores relacionados aos ODS (Objetivos de Desenvolvimento Sustentável) "Fome Zero e Agricultura Sustentável" e "Ação contra mudança global do clima", estabelecidos pela ONU (2023) e aos compromissos de metas globais, como o Net Zero (reduzir as emissões de gases com efeito de estufa o mais próximo possível de zero) preconizados pelo Acordo de Paris (UNFCCC, 2015). Entende-se que através do aumento de consciência e da implementação de melhores práticas agrícolas, é viável sequestrar significativas quantidades de carbono atmosférico no solo, promovendo a regeneração de sua qualidade, melhorando a saúde das plantas e ampliando a restauração de ecossistemas como um todo.

Analisando o contexto brasileiro, há muito tempo, diversas regiões cultivam agricultura e pecuária altamente desenvolvidas devido às características do terreno, solo e clima, gerando vários pontos de atenção. Um deles é a degradação do solo, atualmente considerado um problema complexo ambiental, econômico e social diante ao futuro da segurança alimentar global. Essa degradação altera as estruturas e funções dos ecossistemas, acarretando sérios riscos para a preservação do solo, limitando sua capacidade de sustentabilidade e, por consequência, alterando a sua produtividade. Além disso, a crise climática tem impactado na quantidade, qualidade, rentabilidade e sustentabilidade da produção agrícola e na escassez de água, gerando uma emergência humanitária.

Os esforços para conciliar o manejo do solo e os aspectos da sua fertilidade iniciaram-se há muito tempo. Registros datados de 2500 a.C. trazem pela primeira vez, a fertilidade do solo e sua influência na produtividade da cevada em algumas regiões (Tisdale *et al.*, 1990). Os ingleses, desde 1600 d.C., e pesquisadores da França, Alemanha, Rússia e outros países, ajudaram a produção agrícola nos países desenvolvidos a alcançar altos patamares de produção.

No Brasil, Dafert (1895) publicou um dos primeiros trabalhos sobre a riqueza do solo, fornecendo detalhes sobre a análise química de fertilizantes orgânicos. Mas foi o estudo pioneiro de Primavesi (1979), publicado em seu livro "Manejo Ecológico do Solo" que indicou a necessidade de restabelecer o equilíbrio do solo e de suas relações com o ambiente.

Durante os últimos 30 anos, várias pesquisas compreenderam os problemas de fertilidade do solo. Contudo, a UNCCD (2022) afirma na segunda edição do Global Land Outlook (GLO2), que o avanço da atividade humana, entre elas, a agricultura moderna, gerou degradação em graus variados de 20-40% da área terrestre global.

De acordo com Lal e Stewart (1992), um solo degradado é aquele que sofreu alterações em sua composição física, química ou biológica, seja por fatores naturais ou como resultado da intervenção humana. Tal degradação reduz a capacidade produtiva dos solos, afetando o teor de matéria orgânica que é essencial para fornecer energia e nutrientes ao sistema (Adams; Attiwill, 1986; Swift; Woomer, 1993); bem como influencia nas condições físicas, no efeito tampão e no enriquecimento de nutrientes disponíveis à biota em geral (Aniétot, 1983), além da capacidade de troca catiônica (Raij, 1969; Beer, 1988), entre outros fatores.

Então, dado o cenário acima, emerge uma questão: quais medidas deveriam ser adotadas para avaliar os impactos do manejo do solo em relação à sua qualidade presente e futura visando ganhos socioeconômicos e ambientais?

Entende-se que a melhora na sustentabilidade do solo não seria possível sem o avanço da tecnologia e inovação. Inovar no contexto do desenvolvimento econômico é um processo fundamental que impulsiona o crescimento e a competitividade, desempenhando um papel crucial na geração de riqueza e no desenvolvimento sustentável.

A inovação gera novos produtos e serviços, cria mercados e oportunidades de negócios, estimula a produtividade e a competitividade das empresas e promove a geração de empregos de alta qualidade (Porter, 1990). Também contribui para o aumento da eficiência dos processos produtivos, a redução de custos, o aprimoramento da qualidade dos produtos e serviços, e a melhoria da experiência do cliente (Schumpeter, 2013). A inovação desempenha um papel fundamental na transformação de setores tradicionais, impulsionando a transição para economias baseadas no conhecimento e na tecnologia (Freeman, 1987). Nesse sentido, e com a crescente tecnologia, o uso de técnicas de *Machine Learning*, ou Aprendizado de Máquina, tem se destacado como uma ferramenta para a otimização do manejo agrícola, permitindo o uso eficiente de recursos e a redução de impactos ambientais.

Contudo, a sustentabilidade é um conceito muito mais amplo e de várias dimensões. Sachs (2002) propôs oito dimensões essenciais que deveriam ser consideradas de forma integrada: social, cultural, ecológica, ambiental, territorial, econômica, política nacional e política internacional. Essas dimensões refletem sua visão sobre o desenvolvimento como uma alternativa à ordem econômica global predominante, enfatizando a importância de modelos regionais que valorizem tecnologias apropriadas, especialmente em áreas rurais. Segundo Sachs, essa abordagem contribuiria para reduzir a dependência técnica e cultural, promovendo maior autonomia e resiliência regional. Essa abrangência reflete a natureza holística da sustentabilidade, o que torna sua conceituação e a definição de seu escopo um desafio significativo.

O meio ambiente, por sua vez, consiste em um complexo conjunto de interações, no qual a atuação humana desempenha um papel central. Representar essas interações é uma tarefa complexa, especialmente quando se trata de analisar a sustentabilidade do solo. Essas interações se alinham diretamente com as premissas da economia ecológica que busca uma integração equilibrada entre economia e ecossistemas naturais. A economia ecológica defende a sustentabilidade do solo como um recurso finito e crucial para a produção sustentável de alimentos e serviços ecossistêmicos (Costanza *et al.*, 1997). Ao aplicar as técnicas de Aprendizado de Máquina para monitorar e otimizar práticas agrícolas, é possível minimizar o uso excessivo de insumos, como água e fertilizantes, e evitar a degradação do solo, atendendo tanto às necessidades econômicas quanto à conservação ambiental.

Portanto, ao promover um uso mais eficiente e sustentável dos recursos naturais, as técnicas de Aprendizado de Máquina corroboram com as premissas da economia ecológica, que preconiza práticas sustentáveis de manejo do solo que não apenas garantam a viabilidade econômica das atividades agrícolas, mas também assegurem a conservação e recuperação dos recursos naturais (Cleveland *et al.*, 1997). Essa abordagem integrada reflete a perspectiva da economia ecológica de equilibrar o desenvolvimento econômico com a proteção do meio ambiente.

Por isso, a aplicação de técnicas de *Machine Learning* no monitoramento da sustentabilidade do solo tem o potencial de detectar precocemente problemas não

facilmente visíveis. Tal potencial possibilita que os produtores rurais evitem danos significativos ao solo e minimizem o impacto negativo na produtividade das culturas, além da redução dos custos de produção, uma vez que práticas sustentáveis podem diminuir a dependência de insumos externos, como fertilizantes sintéticos, por exemplo (Hanley *et al.*, 2009). Dessa forma, é possível maximizar a produtividade e reduzir custos econômicos.

Como a agricultura atua na cadeia alimentar de diversos seres vivos e se desenvolve sobre complexos nichos ecológicos ocasionando quebra do equilíbrio natural para o seu estabelecimento, foi possível perceber ao longo dos anos que as práticas conservacionistas sozinhas não são capazes de entender se há ou não otimização no uso do solo. Portanto, a sustentabilidade agrícola deve ser entendida como um processo complexo, utilizando-se de parâmetros que sejam capazes de avaliar, quantificar e indicar o nível de conservação de um dado sistema (De-Polli; Pimentel, 2005).

A hipótese aqui levantada, portanto, é que a utilização de algoritmos de Aprendizado de Máquina na análise de critérios de qualidade do solo pode potencialmente melhorar a tomada de decisão no manejo de áreas agrícolas, resultando em benefícios socioeconômicos e ambientais.

O principal objetivo da pesquisa dentro da área de Desenvolvimento Econômico foi propor uma nova abordagem para avaliar a sustentabilidade do solo agrícola, utilizando algoritmos de Aprendizado de Máquina, a fim de melhorar a tomada de decisão do produtor rural e promover desenvolvimento econômico local mais sustentável. Ao tornar visível a sustentabilidade do solo, acredita-se que seja possível minimizar as perdas (gestão de risco), antecipar as etapas da degradação do solo (uso sustentável do solo) e melhorar a autonomia financeira do produtor.

Como objetivos específicos, pretende-se:

a) Realizar uma análise exploratória dos dados coletados em uma área específica, identificando os indicadores chave de qualidade do solo e fatores associados à sua degradação.
b) Desenvolver e identificar um modelo de Aprendizado de Máquina para diagnosticar a sustentabilidade do solo, utilizando os indicadores identificados na análise exploratória.

c) Avaliar a precisão e a eficácia do modelo desenvolvido, comparando os resultados de diferentes técnicas.
d) Identificar e compreender os parâmetros que determinam os riscos econômicos associados à perda de sustentabilidade do solo, analisando como esses fatores influenciam a tomada de decisão do produtor rural e sua autonomia financeira.

Esta pesquisa é composta por cinco capítulos. O Capítulo I faz uma introdução ao tema. O Capítulo II trata de uma Revisão de Literatura, incluindo os aspectos gerais sobre a sustentabilidade do solo agrícola e atributos químicos, físicos e biológicos na ciclagem dos nutrientes. Também apresenta os conceitos de Aprendizado de Máquina e seu uso na economia e agricultura. O Capítulo III descreve os procedimentos metodológicos para as análises e construção do modelo, o Capítulo IV traz os resultados e a discussão sobre os aspectos econômicos e a adoção tecnológica e o Capítulo V faz as considerações finais sobre a pesquisa.

CAPÍTULO II – REVISÃO DE LITERATURA

1. A DINÂMICA DO SOLO

A sustentabilidade do solo agrícola depende do fornecimento contínuo e reciclagem de material orgânico. A matéria orgânica desempenha um papel fundamental ao servir como fonte de nutrientes e energia para diversos organismos, além de contribuir para a melhoria da estrutura e capacidade de retenção de água no solo, fatores essenciais para o desenvolvimento das culturas agrícolas e da vida microbiana do solo, sendo o parâmetro essencial para compor indicador da conservação do solo (Ronquim, 2020).

1.1. Aspectos gerais sobre a sustentabilidade do solo agrícola

Ao manejar o solo de forma a aumentar o teor de matéria orgânica (MOS), é possível aprimorar a produtividade e a qualidade ambiental, ao mesmo tempo em que se reduzem os impactos financeiros causados por eventos naturais adversos, como secas, inundações e doenças. Um manejo apropriado do solo também pode contribuir para a redução das concentrações de CO^2 na atmosfera, desempenhando um papel importante na mitigação dos efeitos das mudanças climáticas (USDA-NRCS, 2017).

Além da matéria orgânica, a erosão do solo também é considerada um indicador crucial de sua qualidade, pois pode resultar em mudanças na paisagem, perdas econômicas, desastres e, em situações extremas, perda de vidas. A Equação Universal de Perda de Solo (Universal Soil Loss Equation - USLE), desenvolvida em 1954 na Universidade de Purdue, EUA e adaptada para o Brasil por Bertoni e Lombardi Neto (1990) como Equação Universal de Perda de Solo (EUPS), é o modelo empírico mais utilizado para estimar as taxas de erosão.

$$A = R \cdot k \cdot L \cdot S \cdot C \cdot P$$

Onde:

A = representa a perda de solo por unidade de área ($t.ha^{-1}.ano^{-1}$);

R é o fator de erosividade da chuva ($MJ.mm.ha^{-1}.h^{-1}$);

K é a erodibilidade do solo em ($t.h.MJ^{-1}.mm^{-1}$);

L é o comprimento do declive (metros),

S é o grau do declive (porcentagem),

C representa o uso e o manejo do solo e o fator P representa as práticas conservacionistas, ambos adimensionais (Haan *et al.*, 1994).

A qualidade do solo também pode ser estimada por suas condições químicas. Estas, por sua vez, influenciam as interações entre o solo e as plantas, a qualidade da água, a capacidade de tamponamento, a disponibilidade de nutrientes para as plantas e outros organismos, a mobilidade de contaminantes, entre outros. O pH, a capacidade de troca de cátions, saturação por bases e disponibilidade de P, K, Mg, Ca, Al são considerados indicadores químicos (Vernetti Junior *et al.*, 2009).

Uma vez que a conexão entre a qualidade do solo e a sustentabilidade agrícola reside na produção em um solo que desempenhe suas funções de forma ambientalmente segura, economicamente viável e socialmente aceita, a qualidade do solo assume um papel fundamental no desenvolvimento econômico. Ela se torna a base para o manejo responsável do solo e das culturas. Assim, a sustentabilidade agrícola está diretamente vinculada à preservação da qualidade do solo, pois a complexidade do ecossistema faz a diferença para o desempenho eficiente das funções do sistema solo (Vezzani; Mielniczuk, 2009).

1.2. A importância da fertilidade do solo

Conforme mencionado, a população mundial continua crescendo e, por isso, é muito relevante o aumento na produção de alimentos de forma contínua. Diversas pesquisas, muitas delas já implementadas, têm apontado caminhos para aumentar a produtividade das culturas e a eficiência da produção agrícola. Por exemplo, o sistema de plantio direto auxilia no uso racional da água e na redução da erosão do solo, resultando em melhorias significativas na produtividade das culturas. Problemas relacionados ao solo, irrigação ou controle de pragas podem ser identificados e corrigidos em tempo hábil por meio do sensoriamento remoto. No entanto, todos esses avanços colocam uma pressão adicional sobre o solo,

ressaltando ainda mais a importância que a fertilidade do solo desempenha na produção das culturas.

No que se refere à agricultura, registros muito antigos citam a preocupação com a fertilidade do solo. Ao longo do tempo, pesquisadores de países desenvolvidos ajudaram a produção agrícola a alcançar altos níveis de produtividade. Em 1892, a *Lei do Mínimo* criada por Liebig, ajudou a melhorar as respostas das plantas com relação à adubação, sendo referência importante aos pesquisadores no manejo da fertilidade do solo (Novais *et al.*, 2007).

No Brasil, muitas décadas se passaram até que surgissem estudos abordando a fertilidade do solo e o uso de fertilizantes orgânicos e minerais, por meio de observações práticas, com o objetivo de estabelecer as bases para a adubação que permitisse a exploração contínua das propriedades rurais.

Foi o estudo pioneiro de Primavesi (1979), apresentado em seu livro "Manejo Ecológico do Solo", que destacou a necessidade de restabelecer o equilíbrio do solo e suas relações com o ambiente. Esta autora critica os princípios da revolução verde, focado no uso de NPK (Nitrogênio, Fósforo e Potássio) sem considerar os demais elementos, bem como o uso intensivo de máquinas. Isso se deve ao fato de que os grandes ciclos da cana-de-açúcar e do café eram baseados na fertilidade natural dos solos das matas, e quando essa fertilidade se esgotava, a exploração de novas áreas era realizada sem qualquer preocupação com práticas de manejo.

Apesar das várias pesquisas mundiais que ajudaram a compreender os problemas de fertilidade do solo, a ONU (2021) publicou que o avanço da agricultura moderna gerou 33% de degradação do solo mundial.

Atualmente, entende-se que essa degradação afeta negativamente o bem-estar de 3,2 bilhões de pessoas, prejudica a biodiversidade, contribui para as mudanças climáticas e aumenta o risco de surgimento de doenças infecciosas, como a COVID-19 (UNCCD, 2022). As projeções da ONU para os próximos 25 anos indicam que a degradação do solo pode reduzir a produtividade mundial de alimentos em até 12%, resultando em um aumento de 30% nos preços mundiais dos mesmos. Portanto, os problemas com a fertilidade do solo têm escala global.

Para o diagnóstico da sustentabilidade do solo, estudos relacionam os seus atributos com as mudanças entre diferentes manejos, sistemas e fragmentos

florestais (Martins, *et al.*, 2002; Bertol, *et al.*, 2004). Isto possibilita um melhor entendimento do uso e ocupação do solo, reduzindo as causas que levam à degradação ambiental.

Como parte dos resultados destes estudos, a Década das Nações Unidas sobre Restauração de Ecossistemas (2019) estimou que restaurar 350 milhões de hectares de terras degradadas até 2030 poderá reverter a presença de 13 a 26 gigatoneladas de gases de efeito estufa na atmosfera.

Apesar dos esforços, a avaliação da qualidade do solo é bem complexa devido a diversidade de uso, ao grande volume de inter-relações entre os atributos físicos, químicos e biológicos que controlam os processos e aos fatores inerentes a sua variação no tempo e no espaço (Mendes *et al.*, 2005).

Por isso, a análise dos riscos econômicos associados à qualidade do solo envolve vários parâmetros que afetam diretamente a produtividade agrícola, a sustentabilidade ambiental e o bem-estar socioeconômico das comunidades.

Isto porque, a produtividade agrícola é fortemente impactada quando há redução da fertilidade do solo, pois o solo diminui sua capacidade de fornecer nutrientes essenciais para as plantas e reduz a capacidade de retenção de água. Com isto, os custos de produção aumentam significativamente, tendo a necessidade de maior uso de insumos agrícolas, como fertilizantes e corretivos, para compensar a perda de nutrientes. Os riscos financeiros incluem o aumento do endividamento dos produtores, que precisam financiar insumos adicionais e tecnologias de recuperação, e a variabilidade de receita agrícola, uma vez que menor produtividade e aumento de custos levam a receitas menos previsíveis.

A sustentabilidade econômica de longo prazo é ameaçada pela redução da produtividade agrícola sustentável e pela desvalorização das terras agrícolas, que perdem valor econômico quando degradadas. Além disto, a erosão e a lixiviação de nutrientes e agroquímicos podem contaminar os recursos hídricos, enquanto a perda de biodiversidade afeta a flora e fauna locais, prejudicando os ecossistemas.

A segurança alimentar também é afetada, pois a menor produtividade do solo resulta em menor produção agrícola e, consequentemente, aumento dos preços dos alimentos, impactando a segurança alimentar das populações. Os impactos sociais são igualmente significativos, com a possibilidade de desemprego e migração

decorrente da menor produtividade agrícola, além de uma potencial ampliação da desigualdade econômica, visto que agricultores em regiões com solos degradados podem enfrentar maior pobreza em comparação com regiões de solo saudável. Por fim, a resiliência climática é comprometida, pois solos saudáveis são mais resilientes a secas e inundações. A degradação do solo, portanto, aumenta a vulnerabilidade a eventos climáticos extremos.

Existem técnicas alternativas para avaliar a qualidade e a fertilidade do solo. Uma delas é a pedometria que integra a Agricultura de Precisão e, segundo McBratney *et al.* (2019), é o ramo da Ciência do Solo que utiliza métodos matemáticos e estatísticos para estudar os atributos do solo. Atualmente, a pedometria está se desenvolvendo com o suporte da tecnologia da informação, impulsionada pelo avanço de computadores mais robustos e pela disponibilidade de grandes volumes de dados, o que demanda a utilização de métodos mais sofisticados (McBratney *et al.*, 2003).

Considerando que o desafio de suprir a demanda por alimentos sem a necessidade de invadir áreas naturais protegidas aumentará (SNA, 2018), isso implica que ferramentas da Agricultura de Precisão e técnicas para análises de dados na agricultura, podem promover um melhor gerenciamento da propriedade, redução de riscos, otimização do uso de recursos naturais e insumos, e consequentemente, o aumento da produtividade e renda para os agricultores.

Contudo, entre os principais desafios da Agricultura de Precisão, segundo o Grupo Kleffmann (2023) está o de fornecer ao produtor informações em tempo real para variar a taxa de insumos em pontos diferentes, conforme as necessidades do campo, sendo este um dos fatores da baixa adoção da Agricultura de Precisão no Brasil, devido ao alto custo e tempo das análises convencionais de solo (Ramaroson *et al.*, 2018).

A seguir, é apresentada uma detalhada descrição dos atributos químicos, físicos e biológicos à luz da Ciência do Solo. É importante destacar que, segundo Lepsch (2021), o solo é composto por três fases distintas: a) a fase sólida, formada por material rochoso e orgânico resultante da decomposição de vegetais e/ou animais; b) a fase líquida, que corresponde à água ou à solução do solo, contendo elementos orgânicos e inorgânicos em dissolução; e c) a fase gasosa, cuja

composição varia, dependendo dos gases (CO_2 e O_2) produzidos e consumidos pelas raízes das plantas e pelos animais.

1.2.1. Atributos químicos do solo

O solo possui um grande número de propriedades químicas que, juntamente com a atividade biológica, são responsáveis pelos principais mecanismos de atenuação de contaminantes nesse meio. Uma dessas propriedades é a superfície específica, que determina a extensão das reações entre as fases do solo e varia conforme o tamanho e o tipo dos minerais presentes.

Outra propriedade diz respeito às cargas elétricas, onde as negativas podem ser classificadas em cargas permanentes, encontradas em argilas, e cargas dependentes do pH, que aumentam à medida que o pH se eleva e diminuem quando ele se reduz, sendo essa característica mais abundante em solos tropicais. As cargas positivas do solo, por sua vez, são sempre dependentes do pH, aumentando à medida que ele diminui, e são responsáveis pela adsorção de ânions, que em muitos casos são os próprios nutrientes. A adsorção e a troca de íons no solo resultam da interação entre a fase líquida e a fase sólida do solo, definindo as transformações dos solos desde a sua formação.

A acidez do solo pode ser dividida em acidez ativa e acidez potencial, variando em uma escala de 0 a 14, sendo que nos solos o pH geralmente varia de 3 a 9. A acidez do solo impacta todas as suas propriedades químicas, físicas e biológicas, influenciando diretamente a nutrição das plantas. Esse processo é natural em regiões tropicais, onde o elevado volume de chuva lixivia quantidades significativas de cátions básicos. Portanto, a qualidade química do solo expressa sua importância ao influenciar diretamente a nutrição das plantas e, consequentemente, afetar a produtividade dos sistemas agrícolas (Alvarenga; Davide, 1999).

1.2.2. Atributos físicos do solo

Uma qualidade física do solo é a capacidade do mesmo fornecer as condições adequadas para o desenvolvimento de uma cultura. Isto porque, fatores como textura, estrutura, densidade, porosidade, permeabilidade e fluxo de água, ar e

calor são responsáveis pelos meios de redução física de poluentes, possibilitando também que os processos químicos e biológicos possam ocorrer.

As práticas de manejo podem causar alterações na estrutura do solo, o que tem um impacto direto na produtividade das culturas. Essas mudanças podem afetar a disponibilidade de água e oxigênio no solo, bem como a resistência do solo à penetração das raízes das plantas (Tormena *et al.*, 1998).

Neste contexto, a textura tem grande influência no comportamento físico-hídrico e químico do solo e é mensurada pela proporção dos diferentes componentes granulométricos da fase mineral do solo, como areia, silte e argila.

A estrutura do solo reflete a ligação das partículas primárias do solo entre si por outras substâncias como, por exemplo, matéria orgânica, sílica, óxidos de ferro e alumínio e, por isso, ela influi no desenvolvimento do sistema radicular de plantas, no armazenamento e disponibilidade de água e nutrientes e na resistência à erosão.

Já a porosidade é visualizada no perfil de solo e forma espaçamentos comunicantes ou não, favorecendo a movimentação da água ou a sua permanência, assim como a circulação do ar e a densidade do solo relaciona a porosidade total com a composição orgânica e mineralógica média do solo. O aumento na umidade do solo reduz a sua aeração, causando resistência à penetração de raízes no solo.

1.2.3. A influência dos microorganismos na ciclagem de nutrientes

A microbiologia é um dos cinco fatores que auxiliam na formação do solo, ajudando na manutenção da fertilidade do solo e na degradação de poluentes orgânicos. Isso porque, esses organismos funcionam como reservatórios de nutrientes para as plantas, capturando diversos elementos e micronutrientes em seus corpos. Após sua morte e decomposição, esses nutrientes são liberados no solo e assimilados pelas plantas, contribuindo para a ciclagem de nutrientes.

A atividade microbiana, portanto, representa a parte ativa da matéria orgânica do solo e é muito sensível às alterações provenientes do manejo agrícola, o que a torna um excelente indicador da qualidade e saúde do solo. Tais microrganismos ajudam na estrutura física do solo ao excretar substâncias poliméricas extracelulares (EPS) que atuam como aglutinantes naturais, unindo partículas do solo e formando agregados estáveis, e pela criação de uma rede capilar formada por hifas fúngicas e

outros microrganismos filamentosos, que melhora a porosidade e a capacidade de retenção de água do solo. Essas contribuições ajudam na prevenção da erosão e na manutenção de um ambiente propício para o crescimento das plantas.

Estudos recentes, como o trabalho de Pérez-Jaramillo, Mendes e Raaijmakers (2016), têm destacado como a domesticação de plantas afetou significativamente a composição da microbiota da rizosfera em comparação com seus parentes selvagens, o que pode ter implicações importantes para a saúde das plantas e a produtividade agrícola. A compreensão dessas interações microbianas complexas é essencial para o desenvolvimento de práticas agrícolas mais sustentáveis e eficientes que possam melhorar a resistência das plantas a doenças e aumentar a eficiência no uso de nutrientes. Portanto, a influência dos microrganismos na ciclagem de nutrientes e na saúde geral do solo é inegável e continua sendo um campo de estudo vital para a agricultura sustentável e a ecologia do solo.

2. APRENDIZADO DE MÁQUINA (*MACHINE LEARNING*)

Machine Learning, também conhecido como Aprendizado de Máquina, trata-se de um método de análise de dados que automatiza a criação de modelos analíticos (Turing, 1950). Por meio de algoritmos que aprendem de forma interativa a partir de dados, o aprendizado de máquina capacita os modelos computacionais a descobrirem padrões latentes nos dados sem a necessidade de serem programados explicitamente para procurar algo específico. A característica interativa do Aprendizado de Máquina é crucial, pois permite que os modelos se adaptem de forma autônoma quando expostos a novos dados. Eles aprendem com cálculos anteriores, resultando em decisões e resultados confiáveis e reproduzíveis.

Embora não seja uma ciência nova, os modelos de Aprendizado de Máquina têm ganhado um novo impulso significativo nos últimos anos, especialmente devido à sua aplicabilidade em diversos campos e à crescente disponibilidade de dados e poder computacional (Blum *et al.*, 2017).

O uso dos algoritmos de Aprendizado de Máquina está numa crescente. Tais algoritmos podem ser usados para reconhecer padrões, fazer previsão de falhas em

equipamento e sugerir novos modelos de precificação, por exemplo (Theobalt, 2021). Atualmente, Aprendizado de Máquina é utilizado em diferentes tarefas práticas como diagnósticos médicos, bioinformática, reconhecimento de voz e escrita, classificação de produtos, detecção de fraude financeira, perda de clientes (taxa de *churn*), automação de acesso de funcionários, análise de processos, assistentes pessoais, indicador de filmes ou músicas em plataformas de *streaming*, entre outros (Subhashini *et al.*, 2024).

Especificamente em Economia, o Aprendizado de Máquina desempenha um papel cada vez mais crucial na análise econômica, permitindo a extração de *insights* valiosos de grandes conjuntos de dados econômicos e financeiros. Algumas pesquisas se destacam como a de Colla (2009), que teve por objetivo avaliar a aplicação de um modelo gráfico probabilístico (Redes Bayesianas) para desenvolver modelos computacionais que pudessem auxiliar na compreensão de problemas e/ou na previsão de variáveis econômicas. Esta abordagem é significativa porque permite lidar com a complexidade de problemas econômicos que excedem a capacidade dos métodos tradicionais, oferece uma forma de modelar incertezas e relações complexas entre variáveis econômicas e pode ser aplicada em diversos contextos, desde diagnósticos médicos até análises econômicas.

Já Páscoa (2018), propôs uma abordagem baseada na análise de sentimentos para previsão do mercado financeiro, utilizando divulgações de notícias e combinando os dados históricos do mercado de ações com análise de sentimentos. Alcançou resultados superiores aos modelos de aprendizado profundo em termos de tempo e precisão.

Nogueira (2019), buscou verificar a existência de previsibilidade no mercado acionário brasileiro por meio de técnicas de *Machine Learning* e Kim *et al*. (2022) que verificou a importância da aplicação de tecnologias digitais como o uso de *Machine Learning* para modelar sistema ecoeficientes dentro da Economia Circular, demonstrando a aplicabilidade de Aprendizado de Máquina em questões de sustentabilidade e eficiência econômica ao integrar conceitos de Economia Circular com tecnologias avançadas de análise de dados.

Já na área de Agricultura, há o trabalho de Gómez (2020), que propõe um modelo de Aprendizado de Máquina para prever o estado da colheita com base em

dados de consumo de pesticidas e outras variáveis do cultivo; há também a pesquisa de Trivellato *et al.* (2020) que consiste no desenvolvimento de um índice centrado em avaliar a multifuncionalidade da agricultura (MFA) no contexto brasileiro com o apoio de técnicas de *Machine Learning*; e Santos *et al.* (2022), que apresenta um modelo para predição na gestão da água no plantio de lúpulo.

Os algoritmos podem ser agrupados segundo a técnica de aprendizado que usam: aprendizado supervisionado ou não supervisionado. O aprendizado supervisionado utiliza uma variável explícita (atributo-meta) como guia para as respostas e o aprendizado não supervisionado busca reconhecer agrupamentos ou tendências nos dados, sem a necessidade de um objetivo prévio a ser atingido (Müller e Guido, 2016). Os algoritmos de aprendizado supervisionado são empregados em aplicações que utilizam dados históricos para identificar relações e prever eventos futuros prováveis. Segundo Tan *et al.* (2009), eles recebem um conjunto de entradas juntamente com as saídas corretas correspondentes, aprendendo por meio da comparação entre a saída real observada e as saídas estimadas para identificar erros e, posteriormente, ajustar o modelo visando a melhora do seu desempenho.

As técnicas supervisionadas são as que possuem um atributo-meta, sendo chamadas de "Classificação" se o atributo-meta for categórico ou “Regressão”, se o mesmo for contínuo (Tan *et al.*, 2009). Entre essas técnicas, algumas são mais fáceis de interpretar devido ao uso de representações simbólicas como os algoritmos de indução de regras de classificação e árvores de decisão. Por outro lado, existem algoritmos que não permitem a geração de conhecimento inteligível, como regressão múltipla, regressão logística, redes neurais e SVM (Support Vector Machine) (Fayyad *et al.*, 1996). Hoje, reconhece-se que a interpretabilidade é um espectro, e mesmo algoritmos complexos podem ser tornados mais compreensíveis através de técnicas apropriadas (Coma-Puig, 2022). A busca por modelos que equilibrem desempenho e interpretabilidade continua sendo um tema central na pesquisa e aplicação da Mineração de Dados, refletindo a importância contínua do conhecimento inteligível na tomada de decisões baseada em dados.

2.1. Origem e crescimento

O Aprendizado de Máquina surge como conceito na pesquisa de Alan Turing por volta de 1950. Turing ficou conhecido como o pai da Ciência da Computação moderna, pois criou o primeiro teste que verificava a capacidade das máquinas de pensar como seres humanos. Na mesma década, o professor de matemática John McCarthy auxiliou na criação da inteligência artificial como campo de estudo. Contudo, a definição de Arthur Samuel de 1959 ganhou notoriedade dizendo que a aprendizagem de máquina é a área que permite aos computadores adquirir conhecimento e melhorar seu desempenho em tarefas específicas sem programação explícita, baseando-se em dados (Samuel, 1959).

Não existe uma data exata sobre quando a Ciência da Computação foi reconhecida como disciplina acadêmica, porém foi durante a década de 70 que o estudo de algoritmos foi adicionado como um componente importante da teoria (Garfinkel; Grunspan, 2018).

Ao longo dos anos, várias tecnologias para criar máquinas inteligentes foram desenvolvidas na tentativa de reduzir o esforço humano na obtenção de respostas para problemas, principalmente, os mais complexos. Estas tecnologias são baseadas em algoritmos que fazem a análise de um conjunto de dados e que apresentam as informações de forma mais precisa e específica, tornando compreensíveis os padrões e criando novas possibilidades para a tomada de decisão. Com isso, o uso de Aprendizado de Máquina está cada vez mais funcional. A verdadeira função de Aprendizado de Máquina é, portanto, ser parte integrante e dar sentido aos conceitos acerca da inteligência artificial (Mueller; Massaron, 2021).

É importante notar que o campo da inteligência artificial é mais amplo do que apenas o Aprendizado de Máquina. Larson (2021), argumenta que a inteligência artificial atual, incluindo o Aprendizado de Máquina, ainda está longe de alcançar uma verdadeira inteligência geral. Ele sugere que novas abordagens fundamentais possam ser necessárias para avançar além das limitações atuais. Isto ocorre quando novos padrões de diferentes fontes de dados são detectados, quando existem adaptações à novas circunstâncias não previstas inicialmente, quando são criados novos comportamentos baseados em padrões reconhecidos ou ainda,

quando são tomadas decisões com base nestes comportamentos. Para tanto, o Aprendizado de Máquina, baseado em matemática, estatística e programação computacional, utiliza uma série de procedimentos e técnicas capazes de resolver um problema, fornecendo respostas ao objetivo específico que se pretende alcançar (Thomas, 2018). Vale mencionar que ainda não existe um algoritmo mestre que seja capaz de aprender tudo e com autonomia, embora existam esforços e pesquisas neste sentido, como os estudos realizados por Domingos (2015).

Mesmo com todo o seu potencial, a inteligência artificial passou por períodos de descrédito (os maiores, na década de 70 e no final de 80) por não atingir os objetivos anunciados na época. Os "invernos da IA" foram períodos em que o entusiasmo e o financiamento para pesquisas em inteligência artificial diminuíram significativamente devido à incapacidade de cumprir promessas ambiciosas e expectativas infladas. Esses períodos de descrédito ocorreram principalmente porque as capacidades computacionais e os algoritmos disponíveis na época não eram suficientes para realizar as tarefas complexas prometidas pelos pesquisadores. Houve, portanto, uma tendência a superestimar o potencial da inteligência artificial em curto prazo e subestimar a dificuldade de replicar a inteligência humana, levando à previsões otimistas demais sobre o seu progresso (Mueller; Massaron, 2021).

2.2. A descoberta do conhecimento

A expressão KDD (*Knowledge Discovery in Databases*) foi cunhada em 1989 com o objetivo de extrair conhecimento de bases de dados, ou seja, desenvolver procedimentos e técnicas para dar sentido aos dados (Fayyad *et al.*, 1996). Portanto, o KDD é aplicado na identificação de padrões compreensíveis e potencialmente úteis.

O processo KDD envolve várias etapas que vão desde a coleta de dados até a apresentação do resultado final da extração do conhecimento. Ele é considerado iterativo, podendo ser repetido várias vezes na busca por melhores resultados, e interativo, pois envolve a participação de diferentes profissionais. Para usá-lo, é necessário ter os objetivos bem definidos, conhecer quais são os resultados desejados e ter o domínio da aplicação.

O estudo de Han *et al.* (2022) destaca como o KDD é uma disciplina crucial para a identificação de informações significativas em grandes volumes de dados, contribuindo para a tomada de decisões eficazes.

Apesar de cada fase do processo de KDD ser independente, existe uma forte dependência entre elas: a correta transformação dos dados depende da base de dados devidamente modelada, que depende da definição do objetivo a ser alcançado, da coleta e da mineração de dados, fornecendo resultados passíveis de interpretação e avaliação que garantam o conhecimento derivado dos dados (Kohavi; Sommerfield, 1995; Fayyad *et al.*, 1996).

No contexto econômico, o KDD é essencial para analisar tendências de mercado, identificar padrões de consumo e prever mudanças econômicas. De acordo com Fayyad *et al.* (1996), o KDD é uma abordagem abrangente para descobrir conhecimentos úteis a partir de dados, o que é particularmente relevante para a economia.

Na agricultura, o KDD é aplicado para otimizar o uso de recursos, melhorar a produção e aumentar a eficiência dos processos agrícolas. O trabalho de Patel e Patel (2014), explora como técnicas de mineração de dados podem ser empregadas para aprimorar o gerenciamento de culturas, monitorar condições climáticas e prever colheitas. Isso resulta em maior produtividade e sustentabilidade agrícola, contribuindo para a segurança alimentar global. Já o estudo de Gandhi e Armstrong (2016) destaca a importância do KDD na análise de dados agrícolas para melhorar o planejamento da produção, o monitoramento de doenças em plantações e a tomada de decisões informadas pelos agricultores.

Portanto, o processo KDD tem um impacto significativo na Economia e na Agricultura, capacitando a análise inteligente de dados complexos para tomar decisões estratégicas e impulsionar o desenvolvimento econômico e agrícola. Neste sentido, esta pesquisa se apoia nas premissas do KDD.

2.2.1. Relação de Aprendizado de Máquina com a Economia

Há um aumento crescente do uso de técnicas de Aprendizado de Máquina em todas as áreas do conhecimento e na Economia não é diferente. Mesmo que a estatística tradicional e/ou as técnicas econométricas funcionem satisfatoriamente

para estimar relações, testar hipóteses e avaliar o impacto de políticas econômicas (Wooldridge, 2012), existem questões em que os conjuntos de dados requerem outras ferramentas de análises e é para estes casos que o Aprendizado de Máquina pode ser útil quando aplicado em problemas econômicos. Isto porque, os algoritmos de Aprendizado de Máquina são capazes de trabalhar com novos e diferentes tipos de dados, mantendo sua capacidade de reconhecer padrões mesmo diante das estruturas mais complexas.

Na tentativa de diferenciar Aprendizado de Máquina da Econometria, Breiman (2001 b) defende que existem dois objetivos na análise dos dados: a) Predição que é capaz de prever quais serão as respostas para futuras variáveis de entrada e b) Informação serve para extrair algumas informações sobre como a natureza está associando as variáveis de resposta para as variáveis de entrada (livre tradução). Por esta definição, entende-se que as técnicas de Aprendizado de Máquina buscam a acurácia do resultado (predição), enquanto que as da Econometria buscam a interpretabilidade do problema observado (informação).

Já Varian (2014) argumenta que um dos principais objetivos da Econometria que usa métodos estatísticos para a modelagem econômica é a inferência causal, enquanto que nos algoritmos de Aprendizagem de Máquina o qual prevê uma variável em função de todas as outras, a inferência é preditiva. Para Varian, é possível ter mais preditores potenciais do que os adequados para estimação, de forma que seja necessário fazer alguma categoria de seleção de variáveis, permitindo análises mais robustas do que os modelos lineares.

Mullainathan e Spiess (2017) aprofundam a discussão, destacando a relação entre Aprendizado de Máquina e a Econometria. Para estes autores, o Aprendizado de Máquina é complementar à Econometria já que os seus objetivos são diferentes: Aprendizado de Máquina visa a projeção de determinada variável enquanto que a Econometria visa estimar parâmetros. As técnicas de Aprendizado de Máquina não foram desenvolvidas com o mesmo propósito da Econometria. Portanto, há um *trade-off* entre redução da incerteza das estimativas do modelo e viés dos coeficientes estimados não viesados (Vasconcelos, 2017).

Athey e Imbens (2019) também destacam que a Econometria e Aprendizado de Máquina têm objetivos diferentes: a Econometria foca em identificar causalidade

e entender os mecanismos subjacentes, enquanto o Aprendizado de Máquina se concentra em fazer previsões precisas com base em grandes conjuntos de dados. A Econometria tradicional busca explicações teóricas detalhadas para relações causais e é comum em políticas públicas, onde o entendimento do impacto de intervenções é essencial. Já o Aprendizado de Máquina usa algoritmos flexíveis que se destacam na precisão preditiva, embora nem sempre revelem causas específicas. Os autores acreditam que a combinação de Econometria com técnicas de Aprendizado de Máquina, como aprendizado supervisionado e métodos de regularização, pode proporcionar novos *insights*, principalmente ao explorar efeitos causais em contextos complexos. No entanto, eles advertem que métodos de Aprendizado de Máquina devem ser aplicados com cautela, uma vez que nem sempre garantem a validade causal dos resultados, diferentemente dos métodos econométricos mais tradicionais que se apoiam em fundamentos estatísticos rigorosos para evitar vieses.

Apesar disso, Aprendizado de Máquina é usado para prever tendências de mercado, otimizar estratégias de investimento e identificar padrões complexos nos mercados financeiros. O livro "Machine Learning for Asset Managers", de Prado (2020), explora aplicações específicas de Aprendizado de Máquina na gestão de ativos. Além disso, o trabalho de Scheidegger e Bilionis (2019) destaca o uso de métodos de Aprendizado de Máquina para modelagem econômica complexa. Esses exemplos respaldam a relevância de Aprendizado de Máquina na Economia e suas contribuições para a tomada de decisões mais assertivas e eficazes.

2.2.2. Aplicação de Aprendizado de Máquina na Agricultura

O principal desafio da agricultura moderna atualmente é agregar informações de diferentes fontes de dados (cultura, insumos, sensores pluviométricos, doenças, maquinário, aspectos econômicos, entre outros) e a partir disso, transferir valor para o conhecimento.

Existem algumas iniciativas tecnológicas que visam minimizar os problemas gerados pela agricultura moderna. Os modelos de simulação desempenham um papel muito significativo no desenvolvimento das condições agroecológicas e

socioeconômicas. Inúmeras aplicações estão em uso, baseadas em algoritmos de mineração de dados no campo da agricultura.

Gholap *et al*. (2012) usaram o algoritmo de Árvore de Decisão para prever a fertilidade do solo. Os autores coletaram o conjunto de dados do laboratório privado de testes de solo em Pune (Índia) e usaram técnicas de seleção e aprimoramento para ajustar o desempenho do algoritmo J48. O Adaboost foi uma das técnicas de aprimoramento utilizadas nesta pesquisa. Eles previram o nível de fertilidade do solo e classificaram como muito baixo, muito alto, baixo, alto, moderado e moderadamente alto. Outro exemplo inclui o uso de Aprendizado de Máquina para identificar ambientes de manejo de cana-de-açúcar, passando por análise estatística, seleção de variáveis úteis ao modelo, teste de multicolinearidade e classificação com árvore de decisão (Almeida, 2019). A principal crítica com relação a estes dois trabalhos é a falta de explicação sobre como as variáveis foram normalizadas ou padronizadas antes da aplicação dos algoritmos.

Outra aplicação foi sobre o uso do modelo de Análise de Componentes Principais (PCA) na avaliação de qualidade do solo com pastagens, criando um modelo matemático para envolver a correlação de quocientes microbianos e metabólicos com variáveis ambientais para entender os fatores que influenciam a atividade microbiana, atributos químicos e físicos do solo (Oikawa, 2023). Este trabalho utilizou a análise de dados com diferentes algoritmos, sendo o *Random Forest* o algoritmo com melhor desempenho. Apesar de todo o detalhamento técnico, no decorrer do trabalho fica demonstrado que as análises foram realizadas considerando variáveis com informações redundantes, o que pode prejudicar a interpretação dos resultados.

Apesar dos grandes avanços no campo de Aprendizado de Máquina, existe um distanciamento muito grande no uso de tecnologias entre os pequenos e grandes produtores. Grandes produtores geralmente têm acesso a tecnologias avançadas, como sistemas de monitoramento de solo baseados em Aprendizado de Máquina, enquanto pequenos produtores frequentemente carecem de recursos financeiros e conhecimento técnico para adotar essas inovações. Custos elevados, falta de treinamento e infraestrutura inadequada são obstáculos significativos que

impedem pequenos produtores de utilizarem estas tecnologias, ampliando o distanciamento.

Este distanciamento gera um *gap* ainda maior em relação à sustentabilidade, pois muitos destes pequenos produtores não sabem efetivamente o que está acontecendo com os solos da sua propriedade, principalmente, pela falta de frequente monitoramento. A análise de dados do solo permite práticas agrícolas mais sustentáveis, como o uso eficiente de insumos e a prevenção da erosão. Sem essas tecnologias, pequenos produtores podem não identificar problemas até que seja tarde demais, resultando em degradação ambiental.

2.3. Escolha das técnicas para a modelagem

O critério para escolher a técnica de Aprendizado de Máquina inclui o tipo de problema (classificação, regressão, aprendizado supervisionado ou não supervisionado), o volume e a qualidade dos dados, a necessidade de interpretabilidade, os recursos computacionais e a capacidade de generalização.

Por trabalhar com dados laboratoriais de análise de solo nesta pesquisa, a abordagem de aprendizado escolhida foi o Supervisionado, ou seja, o treinamento é realizado sobre um conjunto de dados pré-definido, garantindo que o algoritmo tome decisões precisas quando recebe novos dados.

Foi utilizado o RStudio (Version 1.1.463 – © 2009-2018 RStudio, Inc.), uma plataforma gratuita, de código aberto e disponível para vários sistemas operacionais que se destaca como uma ferramenta versátil e poderosa para análise de dados e programação estatística em R.

A partir disso, foram utilizadas quatro técnicas de Aprendizado de Máquina para efeitos comparativos e de interpretabilidade dos resultados.

a) Regressão Logística (*rg*):

Trata-se de um algoritmo amplamente usado para solucionar problemas de classificação binária. Seu objetivo é prever uma variável de saída categórica que possui somente duas classes possíveis, geralmente representadas como 0 e 1 ou "negativo" e "positivo" (Goodfellow *et al.*, 2016).

O modelo paramétrico é um método de classificação supervisionada e

representa um caso específico dos modelos lineares generalizados (Mccullagh, 2019). Ele é definido pela distribuição binomial com a função de ligação canônica (*logit*) e desempenha um papel crucial na transformação de probabilidades para uma escala linear, permitindo uma análise mais eficaz de variáveis de resposta binárias.

A *logit* (função de ligação) proporciona a interpretação mais natural dos coeficientes estimados. Quanto maiores forem as chances logísticas, maior a probabilidade de ocorrência do evento de referência. Assim, coeficientes positivos indicam um aumento na probabilidade do evento, enquanto coeficientes negativos indicam uma redução dessa probabilidade (Menard, 2002).

A regressão logística utiliza a função logística ou sigmoide para transformar uma combinação linear das variáveis independentes em uma probabilidade que varia entre 0 e 1. O modelo *logit* é matematicamente representado por:

$$P(Y = 1|X) = \frac{1}{1+e^{-(\beta_0+\beta_1 X_1+\beta_2 X_2+\cdots+\beta_n X_n)}}$$

Onde:

$P(Y = 1|X)$ é a probabilidade do evento $Y = 1$ dado o vetor de preditores X. A expressão $\beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_n X_n$ é a combinação linear dos preditores, β_0 é o intercepto e $\beta_1, \beta_2, \ldots, \beta_n$ são os coeficientes dos preditores $X_1, X_2, \ldots, X_n$.

O vetor de preditores no modelo *logit* desempenha um papel importante na modelagem da relação entre as variáveis independentes e a probabilidade de ocorrência do evento de interesse.

Os parâmetros do modelo (β) são tipicamente estimados pelo método de máxima verossimilhança, que encontra os valores de β que maximizam a probabilidade de observar os dados amostrais. A função de verossimilhança para a regressão logística é:

$$L(\beta) = \prod_{i=1}^{N} P(y_i|x_i)^{y_i}(1 - P(y_i|x_i))^{1-y_i}$$

As premissas da função de verossimilhança inclui a independência das observações, ou seja, o resultado de uma instância não deve afetar o resultado de outra e variável dependente binária (possui apenas dois possíveis resultados).

Embora a regressão logística não exija uma relação linear entre as variáveis independentes e a variável dependente (Wasserman, 2004), ela assume uma relação linear entre as variáveis independentes e o log-odds (logaritmo das chances). Os coeficientes da regressão logística representam o log-odds de um evento ocorrer. Para um dado coeficiente *βj*:

$$\text{log-odds}(Y = 1|X) = \beta_0 + \beta_1 X_1 + \cdots + \beta_j X_j + \cdots + \beta_n X_n$$

Durante o treinamento, o modelo ajusta os coeficientes para minimizar a discrepância entre as probabilidades previstas e os rótulos reais dos dados de treinamento. Essa otimização é realizada por meio de uma função de perda adequada, como a entropia cruzada ou a função de erro logístico. Após o treinamento, o modelo é capaz de prever a classe de um novo exemplo com base em suas características de entrada, fornecendo a probabilidade de pertencer à classe positiva (classe 1) ou negativa (classe 0). Com o auxílio de um limiar de decisão adequado (geralmente 0,5), a regressão logística classifica o exemplo em uma das duas classes (Goodfellow *et al.*, 2016).

Nesse contexto, essa técnica possibilita uma análise que estima a probabilidade relacionada à ocorrência de um evento específico com base em um conjunto de variáveis explicativas. Em outras palavras, ela é utilizada para modelar a probabilidade de um evento acontecer em função de outros parâmetros. As suas principais vantagens incluem a facilidade em lidar com variáveis independentes categóricas ou binárias, o fornecimento de resultados em termos probabilísticos, simplicidade na classificação de indivíduos em categorias e um alto nível de confiabilidade. Nesta pesquisa, foi configurado o hiperparâmetro *cost* que regula o *trade-off* entre regularização e classificação correta dos dados (Figueira, 2006).

b) Árvore de Decisão - *Decision Tree* (*dt*):

Este algoritmo se baseia na ideia de divisão dos dados em grupos homogêneos que se assemelha a um fluxograma, onde cada nó interno representa um teste em uma característica específica, cada ramo representa um resultado possível do teste e cada folha representa a decisão final (Braz, 2022). Essa estrutura

permite uma interpretação das decisões tomadas pelo modelo durante o processo de classificação ou regressão (Hastie *et al.*, 2009).

Uma das principais vantagens desta técnica é a capacidade de lidar com dados de forma não linear e de realizar seleção automática de características, ou seja, a árvore escolhe quais características são mais relevantes para fazer as divisões e tomar as decisões. Além disso, tais árvores são eficientes e de fácil interpretação, tornando-as úteis em cenários onde é importante entender os critérios utilizados pelo modelo para chegar a uma conclusão (James *et al.*, 2013).

Apesar da fácil interpretação da sua estrutura, o modelo tem a tendência de sofrer *overfitting* (sobreajuste), ajustando-se muito bem aos dados de treino sem ter uma boa performance com os dados de teste. Outro desafio desta técnica é a tendência a criar modelos muito complexos para conjuntos de dados grandes e com muitas características. Isso pode levar a árvores profundas e difíceis de interpretar, bem como aumentar o risco de *overfitting*. Matematicamente, as árvores de decisão são construída da seguinte forma:

- Seleção do Atributo de Divisão: Para cada nó da árvore, o algoritmo seleciona o atributo que melhor divide o conjunto de dados *D*. A seleção do atributo é feita por uma métrica de divisão como Ganho de Informação ou índice de Gini.

  Ganho de Informação (ΔE):

$$\Delta E(D, A) = E(D) - \sum_{v \in \text{Valores}(A)} \frac{|D_v|}{|D|} E(D_v)$$

  Onde:
  E(D) é a entropia do conjunto *D*,
  Dv é o subconjunto de *D* para qual o atributo *A* tem valor *v* e
  Valores(A) são os possíveis valores de *A*.

  Índice de Gini:

$$Gini(D) = 1 - \sum_{i=1}^{m} p_i^2$$

  Onde p_i é a proporção de exemplos da classe *i* no conjunto *D*.

- Divisão Recursiva: A árvore é construída de forma recursiva dividindo *D* em subconjuntos *Dv*, baseados no melhor atributo *A* selecionado. A divisão continua até que um critério de parada seja alcançado, como um número mínimo de exemplos em um nó ou a entropia mínima.
- Critérios de Parada: Todos os exemplos em um nó pertencem à mesma Classe. Não há mais atributos para dividir. A divisão não melhora significativamente a pureza dos subconjuntos.
- Podamento (Pruning): Para evitar o sobreajuste, a árvore pode ser podada. Isso pode ser feito removendo nós que têm pouca importância estatística.

Formalização do Processo

Considerando um conjunto de dados *D* com *n* atributos *{X1,X2,…,Xn}* e uma variável alvo *Y*.

1. Entropia:

$$E(D) = -\sum_{i=1}^{m} p_i \log_2(p_i)$$

Onde *m* é o número de classes e p_i é a proporção de exemplos da classe *i*.

2. Ganho de Informação:

$$\Delta E(D, A) = E(D) - \sum_{v \in \text{Valores}(A)} \frac{|D_v|}{|D|} E(D_v)$$

3. Índice de Gini:

$$Gini(D) = 1 - \sum_{i=1}^{m} p_i^2$$

4. Função de Divisão:

$$D = \bigcup_{v \in \text{Valores}(A)} D_v$$

Onde $D_v = \{(x, y) \in D | x_A = v\}$

5. Construção Recursiva: A árvore é construída recursivamente aplicando a função de divisão em cada subconjunto D_v.

Para resolver o *overfitting*, nesta pesquisa foram realizados ajustes de hiperparâmetros *minbucket*, *maxdepth, cp* e validação cruzada com o mesmo número de *folds* (dados de treinamento divididos em várias partes), iniciando sempre do mesmo *set.seed* (este assegura que, a cada execução do código, os dados sejam divididos da mesma forma, criando os mesmos subconjuntos de treinamento e teste). O set.seed é importante para garantir que os resultados da validação cruzada sejam consistentes e comparáveis mesmo que o processo de treinamento seja repetido várias vezes, sendo bastante utilizado ao ajustar os hiperparâmetros, comparar modelos ou quando há necessidade de um controle sobre a aleatoriedade. O *minbucket* fornece o menor número de observações permitidas num nó terminal. Se o valor de *minbucket* for muito pequeno, a árvore pode se tornar muito complexa (*overfitting*), capturando ruídos do conjunto de dados. Por outro lado, valores muito altos podem simplificar excessivamente o modelo, levando a *underfitting* (perda de capacidade preditiva). O *maxdepth* evita que a árvore cresça além de uma certa profundidade / altura. Uma árvore mais profunda (com valores mais altos de *maxdepth*) terá mais divisões, o que pode melhorar a capacidade de captura de padrões complexos nos dados, mas também aumentar o risco de *overfitting*. Já o *cp* é o parâmetro de complexidade que define a melhoria mínima necessária em cada nó. Um valor de *cp* maior significa uma árvore mais simples (com menos nós e folhas), enquanto um valor menor pode resultar em uma árvore mais complexa. A partir do conjunto de dados foi feito uma divisão (*split*), sendo que cada *split* representa um nó da árvore. O nó raiz é onde começa a árvore e ao nó folha é atribuída a classe para os casos de classificação.

c) *Support Vector Machine (svm)*:

Trata-se de um método poderoso que busca maximizar a margem de separação das diferentes classes dos dados em cenários onde os dados são linearmente separáveis ou quase linearmente separáveis. Neste sentido, o *SVM* busca encontrar o hiperplano que melhor divide as classes no espaço de características para maximizar a margem entre os pontos mais próximos de cada classe. Esses pontos são chamados de vetores de suporte (*support vectors*) e desempenham um papel fundamental na construção do modelo (Hastie *et al.*, 2009).

A principal vantagem é sua capacidade de lidar com problemas de alta dimensionalidade, nos quais o número de características é muito maior do que o número de exemplos de treinamento. Além disso, o *SVM* é menos suscetível a *overfitting*, especialmente em comparação com modelos mais complexos, como redes neurais profundas. Essa propriedade o torna uma escolha popular em aplicações com conjuntos de dados pequenos ou de tamanho moderado (James *et al.*, 2013).

Apesar das suas vantagens, o *SVM* pode ser computacionalmente exigente, especialmente em grandes conjuntos de dados ou com a utilização de kernels não lineares. O modelo é um classificador linear binário não probabilístico que analisa um conjunto de dados de entrada e determina para cada elemento, a qual de duas classes possíveis ele pertence, utilizando um hiperplano de separação no espaço de características (Brito, 2020), ou seja, ele não fornece probabilidades diretas para suas classificações, apenas a classe prevista.

O *SVM* trabalha com um problema bem comum da matemática: otimização quadrática com restrições lineares. Na separação linear, são necessários somente os pontos na “borda” de separação e os planos são definidos pelos vetores de suporte. Quando o SVM se depara com dados que não são linearmente separáveis, a utilização de *kernels* se torna uma solução poderosa e eficaz. Os *kernels* possibilitam que o *SVM* realize classificações não lineares, expandindo significativamente sua aplicabilidade e eficácia (Lorena; Carvalho, 2007).

Os *kernels* permitem mapear o espaço de características original em um espaço de maior dimensão, onde os dados podem se tornar linearmente separáveis.

O modelo matemático do *SVM* é representado por:

1. Hiperplano Separador

Dado um conjunto de treino $\{(X_i, Y_i)\}^n_{i=1}$, onde $X_i \in R^d$ são vetores de características e $Y_i \in \{-1,1\}$ são os rótulos de classe, o SVM busca encontrar um hiperplano $\mathbf{w}^\top\mathbf{x}+b=0$ que maximiza a margem entre as classes. O vetor **w** é o vetor normal ao hiperplano e *b* é o termo de viés.

2. Margem

A margem *γ* é definida como a distância entre o hiperplano e os pontos mais próximos de cada classe. Os vetores de suporte são os pontos mais próximos do hiperplano, para os quais:

$$y_i(\mathbf{w}^\top\mathbf{x}_i + b) = 1$$

A largura da margem é $\frac{2}{\|\mathbf{w}\|}$.

3. Função Objetivo

O problema de otimização do *SVM* consiste em maximizar a margem, que equivale a minimizar $\|\mathbf{w}\|^2$, sujeito às restrições de classificação correta dos pontos:

$$\min_{\mathbf{w},b} \frac{1}{2}\|\mathbf{w}\|^2$$

sujeito a

$$y_i(\mathbf{w}^\top\mathbf{x}_i + b) \geq 1, \quad \forall i = 1, \ldots, n$$

4. Problema Dual

Utilizando os multiplicadores de Lagrange α_i, o problema primal é transformado em um problema dual:

$$\max_{\alpha} \sum_{i=1}^{n} \alpha_i - \frac{1}{2} \sum_{i=1}^{n} \sum_{j=1}^{n} \alpha_i \alpha_j y_i y_j \mathbf{x}_i^\top \mathbf{x}_j$$

sujeito a

$$\sum_{i=1}^{n} \alpha_i y_i = 0, \quad \alpha_i \geq 0, \quad \forall i = 1, \ldots, n$$

Os vetores de suporte são aqueles para os quais $\alpha_i > 0$.

5. *Kernel Trick*

Em vez de calcular explicitamente as coordenadas dos dados no espaço de alta dimensão, computa diretamente o produto escalar entre os vetores nesse espaço. Ele realiza uma transformação não linear dos dados de entrada para um espaço de características de maior dimensão. No espaço de características transformado, o *SVM* pode encontrar um hiperplano linear que separe as classes de forma eficaz, onde:

$$K(\mathbf{x}_i, \mathbf{x}_j) = \phi(\mathbf{x}_i)^\top \phi(\mathbf{x}_j)$$

Os *kernels* comuns incluem o *kernel* Linear, Polinomial e *Radial Base Function* (RBF). O Linear, o mais simples, realiza um produto escalar entre os vetores de entrada, sendo eficaz para dados já linearmente separáveis. O Polinomial eleva o produto escalar dos vetores a uma potência d, permitindo a captura de relações não lineares de ordem superior entre as variáveis. Já o RBF, também conhecido como *kernel* Gaussiano, mapeia os dados para um espaço de dimensão infinita, utilizando uma função exponencial da distância euclidiana entre os pontos, o que o torna extremamente flexível e capaz de lidar com relações altamente não lineares (Burges, 1998). A escolha entre esses *kernels* depende da natureza dos dados e da complexidade do problema de classificação, com o RBF sendo frequentemente a opção mais versátil, enquanto o Linear é preferido para problemas mais simples e o Polinomial oferece um meio-termo entre simplicidade e capacidade de modelar relações não lineares (Schölkopf; Smola, 2002). Nesta pesquisa, foi configurado o *gamma* (parâmetro para o *kernel* RBF) e *cost* (custo de violação de restrições).

d) *Random Forest* - Floresta Aleatória (*rf*):

É um poderoso algoritmo que pertence à família dos métodos *Ensemble* ou comitê de máquinas. Esse método combina várias árvores de decisão em um único modelo, onde cada árvore é construída utilizando uma amostra aleatória dos dados de treinamento e características selecionadas de forma aleatória em cada divisão. A abordagem permite que o Random Forest obtenha um desempenho superior ao de uma única árvore de decisão, aumentando a precisão, reduzindo o *overfitting* e fornecendo uma maior robustez (Hastie *et al.*, 2009).

Uma das principais vantagens desta técnica é sua capacidade de lidar com dados de alta dimensionalidade e grande quantidade de características. É também menos suscetível a *overfitting*, pois a média das previsões das várias árvores reduz os erros individuais, garantindo uma generalização mais precisa para novos dados. A aleatoriedade introduzida no processo de construção das árvores também evita a formação de padrões viciosos que poderiam ocorrer em uma única árvore de decisão (James *et al.*, 2013).

Outra característica importante deste modelo é a capacidade de medir a importância das características para o processo de classificação. Isto é obtido através da avaliação de quanto cada característica contribuiu para a redução da impureza (como o índice Gini ou a Entropia) ao longo de todas as árvores. (Breiman, 2001 a).

O primeiro passo realizado pelo algoritmo é selecionar aleatoriamente algumas amostras dos dados de treinamento, em vez de usar o conjunto completo. Nessa etapa, utiliza-se o método de *bootstrap*, que permite a repetição das amostras selecionadas. Com as amostras iniciais, a primeira árvore de decisão é construída. O algoritmo, em seguida, escolhe aleatoriamente duas ou mais variáveis e realiza os cálculos com base nessas amostras para definir qual variável será usada no primeiro nó. Para o próximo nó, novas duas (ou mais) variáveis serão escolhidas, excluindo aquelas já selecionadas anteriormente. Este processo se repete até o último nó.

É possível selecionar a quantidade de variáveis na criação do modelo. Com o modelo devidamente criado, apresenta-se novos dados para obter o resultado da previsão. Cada árvore criada irá indicar o seu resultado, sendo que na Classificação,

será escolhido o resultado que mais vezes foi apresentado. Matematicamente, o Random Forest é representado por:

1. Árvores de Decisão Individuais

Cada árvore de decisão T_m em uma Random Forest é construída a partir de um subconjunto de dados, selecionado aleatoriamente com reposição (técnica conhecida como *bootstrap*).

Seja $X=\{x1,x2,\ldots,xn\}$ o conjunto de treinamento com n amostras e $Y=\{y1,y2,\ldots,yn\}$ os rótulos associados. Cada árvore T_m é treinada em um subconjunto $X_m \subseteq X$ com amostras (X_i, Y_i) selecionadas aleatoriamente com reposição.

2. Seleção de Atributos Aleatórios

Durante a construção de cada árvore, para cada nó, o algoritmo seleciona aleatoriamente um subconjunto de k atributos $\{X_1,X_2,\ldots,X_k\}$ do total de d atributos disponíveis $\{X_1,X_2,\ldots,X_d\}$. A melhor divisão para aquele nó é escolhida com base nesse subconjunto aleatório.

3. Comitê de Árvores (Ensemble)

Dado um conjunto de M árvores de decisão $T_1, T_2,\ldots,T_m$, o *Random Forest* faz a previsão agregando as predições individuais das árvores:

Na Classificação:

$$\hat{y} = \text{mode}\{T_1(\mathbf{x}), T_2(\mathbf{x}), \ldots, T_M(\mathbf{x})\}$$

Onde *mode* é a classe mais votada entre as árvores.

Na Regressão

$$\hat{y} = \frac{1}{M}\sum_{m=1}^{M} T_m(\mathbf{x})$$

Onde a previsão final é a média das predições das árvores.

Nesta pesquisa, para minimizar o efeito do *overfitting*, foram realizados ajustes de hiperparâmetros *ntree*, *mtry* e *cross-validation* com o mesmo número de *folds*, iniciando sempre do mesmo *set.seed*. O *ntree* define o número de árvores que serão geradas durante a modelagem e este número não deve ser muito pequeno para garantir que cada linha de entrada seja prevista pelo menos algumas vezes e o *mtry* define o número de variáveis aleatoriamente em cada *split* (divisão).

CAPÍTULO III – MATERIAIS E MÉTODOS

Nesta pesquisa, foi realizada a pesquisa bibliográfica para obter o suporte conceitual e utilizou-se o RStudio para a modelagem com os algoritmos supervisionados de Aprendizado de Máquina.

3.1. Descrição da base de dados

A base de dados (Tabela 1) utilizada como referência para o desenvolvimento da pesquisa corresponde a 2255 registros e 27 atributos das análises de solo de cana-de-açúcar dos talhões da unidade Alcídia na região de Presidente Prudente/ SP. De acordo com Brunini (2010), a região não tem restrições para o plantio da cana-de-açúcar, com temperatura anual média anual entre 20°C e 24° e temperatura do mês mais frio superior a 17°C, apresentando uma boa distribuição de chuvas durante o ano.

Tabela 1 - Amostra da base de dados

CodFaz	Talhao	Area	Areia1	Silte1	Argila1	Areia2	Silte2	Argila2	Areia3	Silte3	Argila3	AmbP	pH	P	K	Ca	Mg	HAl	Al	SB	CTC	V	m	SiBCS	GradText	MO
110001	1	25,59	86,3	1,2	12,5	80	3,8	16,2	78,8	3,7	17,5	8	6	8	1,2	6	2	16	1	9,2	25,6	36	9,8	PV	1,3	4
110001	2	19,74	86,3	1,2	12,5	80	3,8	16,2	78,8	3,7	17,5	8	6	8	1,2	6	2	16	1	9,2	25,6	36	9,8	LVA	1,3	4
110001	3	18,3	86,3	1,2	12,5	80	3,8	16,2	78,8	3,7	17,5	8	6	8	1,2	6	2	16	1	9,2	25,6	36	9,8	LVA	1,3	4
110001	4	1,91	86,3	1,2	12,5	80	3,8	16,2	78,8	3,7	17,5	8	6	9	0,8	9	3	14	0	12,8	26,3	49	0	LVA	1,3	7
110001	5	22,62	86,3	1,2	12,5	80	3,8	16,2	78,8	3,7	17,5	8	6	9	0,8	9	3	14	0	12,8	26,3	49	0	LVA	1,3	7
110001	6	25,97	86,3	1,2	12,5	80	3,8	16,2	78,8	3,7	17,5	8	6	8	1,2	6	2	16	1	9,2	25,6	36	9,8	LVA	1,3	4
110001	7	18,3	82,5	1,3	16,2	81,2	3,8	15	77,5	5	17,5	5	6,3	3	0,5	9	3	11	0	12,5	23,8	53	1	LVA	0,93	6
110001	8	5,8	82,5	1,3	16,2	81,2	3,8	15	77,5	5	17,5	5	6,3	5	0,5	11	3	14	0	14,5	28,5	51	0	LVA	0,93	5
110001	9	28,72	82,5	1,3	16,2	81,2	3,8	15	77,5	5	17,5	8	6,3	5	0,5	11	3	14	0	14,5	28,5	51	1	LVA	0,93	5

Fonte: Dados da pesquisa

Segue o detalhamento sobre as variáveis da base de dados relacionadas à área de estudo (Tabela 2):

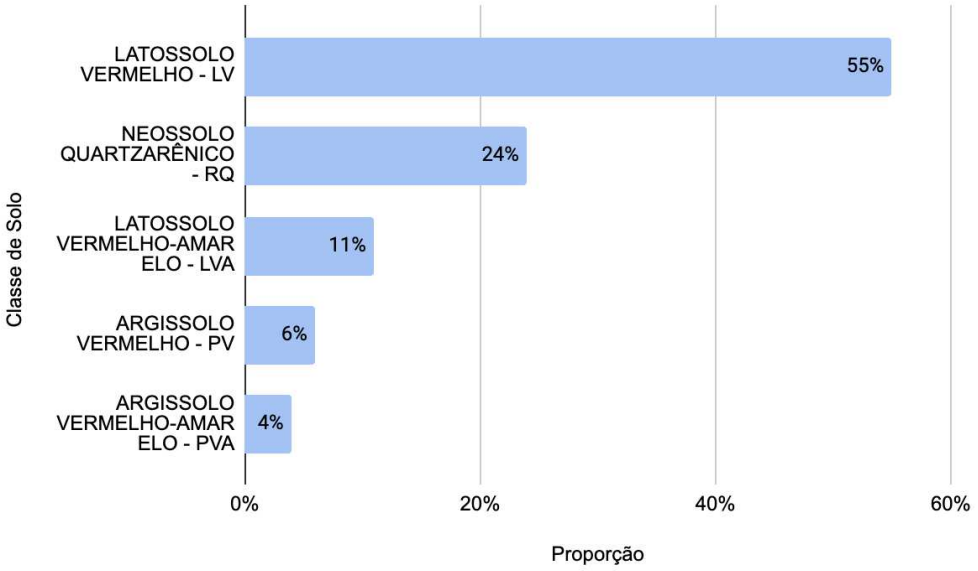
Tabela 2 - Variáveis da base de dados

Variáveis	Unidade	Descrição
CodFaz	—	Código identificador da fazenda.
Talhao	—	Número do talhão na fazenda.
Area	m^2	Área do talhão em estudo.
Areia1	g/dm^3	Percentual de areia na camada 1 do solo (0-20 cm).
Silte1	g/dm^3	Percentual de silte na camada 1 do solo (0-20 cm).
Argila1	g/dm^3	Percentual de argila na camada 1 do solo (0-20 cm).
Areia2	g/dm^3	Percentual de areia na camada 2 do solo (20-40 cm).
Silte2	g/dm^3	Percentual de silte na camada 2 do solo (20-40 cm).
Argila2	g/dm^3	Percentual de argila na camada 2 do solo (20-40 cm).
Areia3	g/dm^3	Percentual de areia na camada 3 do solo (40-60 cm).
Silte3	g/dm^3	Percentual de silte na camada 3 do solo (40-60 cm).
Argila3	g/dm^3	Percentual de argila na camada 3 do solo (40-60 cm).
AmbProd_num	t/ha	Produtividade do Talhão.
pH	—	Nível de pH na camada 0-20cm do solo.
P	mg/dm^3	Concentração de Fósforo na camada 0-20cm do solo.
K	$mmol_c/dm^3$	Concentração de Potássio na camada 0-20cm do solo.
Ca	$mmol_c/dm^3$	Concentração de Cálcio na camada 0-20cm do solo.
Mg	$mmol_c/dm^3$	Concentração de Magnésio na camada 0-20cm do solo.
HAl	$mmol_c/dm^3$	Acidez Potencial na camada 0-20cm do solo.
Al	$mmol_c/dm^3$	Concentração de Alumínio na camada 0-20cm do solo.
SB	$mmol_c/dm^3$	Soma de Bases no solo.
CTC	$mmol_c/dm^3$	Capacidade de Troca Catiônica no solo.
V	%	Saturação de Bases.
m	%	Saturação por Alumínio.
SiBCS	—	Proporção de registros encontrados na base de dados referentes ao Sistema Brasileiro de Classificação dos Solos em relação ao número total de registros.
GradText	%	Textura do solo (granulometria) = corresponde a % Argila da Camada 2 / % Argila da Camada 1
MO	g/dm^3	Matéria orgânica na camada 0-20 cm do solo.

Fonte: Dados da pesquisa. As Unidades foram baseadas no Boletim Técnico no.100 do Instituto Agronômico de Campinas (IAC, 2016)

Especificamente sobre a variável "SiBCS" da Tabela 2, o Gráfico 1 mostra a proporção de registros encontrados na base de dados em relação ao número total de registros. As informações sobre a classificação do solo foram baseadas em Santos *et al.*, (2018).

Gráfico 1 - Proporção de registro por Classe de Solo



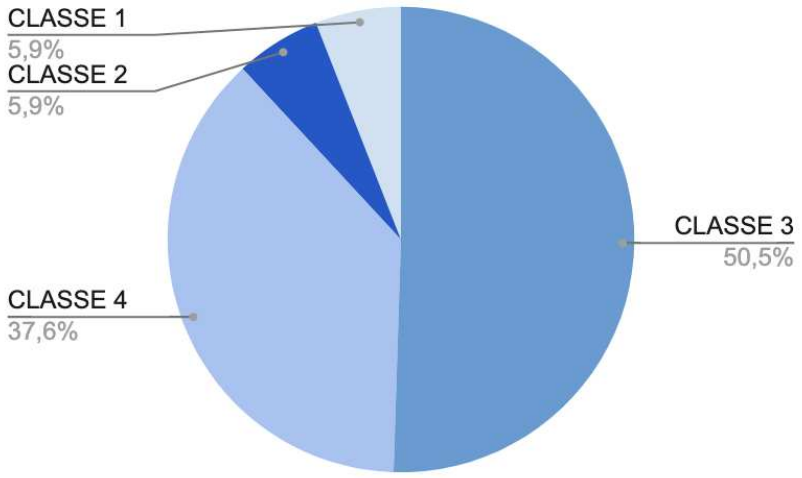
Fonte: Dados da pesquisa

No Gráfico 1 é observado a predominância do Latossolo Vermelho (LV) com 55% dos registros encontrados na base de dados. Este tipo de solo é caracterizado por uma textura argilosa, boa drenagem e alta profundidade. É típico de regiões tropicais e subtropicais e apresenta baixa fertilidade natural devido à intensa lixiviação. Contudo, ele é favorável à agricultura após correções, como a adição de fertilizantes e calcário. O Neossolo Quartzarênico ocupa o segundo lugar em abundância, com 24%. Esse solo, formado majoritariamente por partículas de areia, é altamente drenável e apresenta baixa capacidade de retenção de nutrientes e água, exigindo cuidados adicionais para atividades agrícolas, como irrigação e suplementação de nutrientes. Em seguida, temos o Latossolo Vermelho-Amarelo (LVA), com 11%. Este solo é similar ao Latossolo Vermelho, mas apresenta uma coloração amarela, que indica uma composição com menor teor de óxidos de ferro. Ele também é um solo profundo e bem drenado, porém, assim como o Latossolo Vermelho, apresenta baixa fertilidade natural e requer manejo adequado para usos agrícolas. Os Argissolos, tanto o Vermelho quanto o Vermelho-Amarelo, estão em menor quantidade, com 6 e 4%, respectivamente. Esses tipos de solo possuem uma camada argilosa mais profunda, que pode dificultar a penetração das raízes em

profundidade. Em contrapartida, sua capacidade de retenção de água e nutrientes é melhor que a dos solos arenosos, o que pode ser benéfico para algumas culturas. Contudo, a presença de alumínio e a acidez natural são fatores que exigem correções para uso agrícola. Portanto, o Gráfico 1 indica uma diversidade de solos, o que sugere a necessidade de estratégias variadas de manejo para maximizar a sustentabilidade do uso do solo na região em questão.

Seguindo o mesmo critério, o Gráfico 2 mostra a relação da Classe de Fertilidade, sendo a CLASSE 1 (mais fértil) e a CLASSE 4 (menos fértil).

Gráfico 2 - Proporção de registro por Classe de Fertilidade



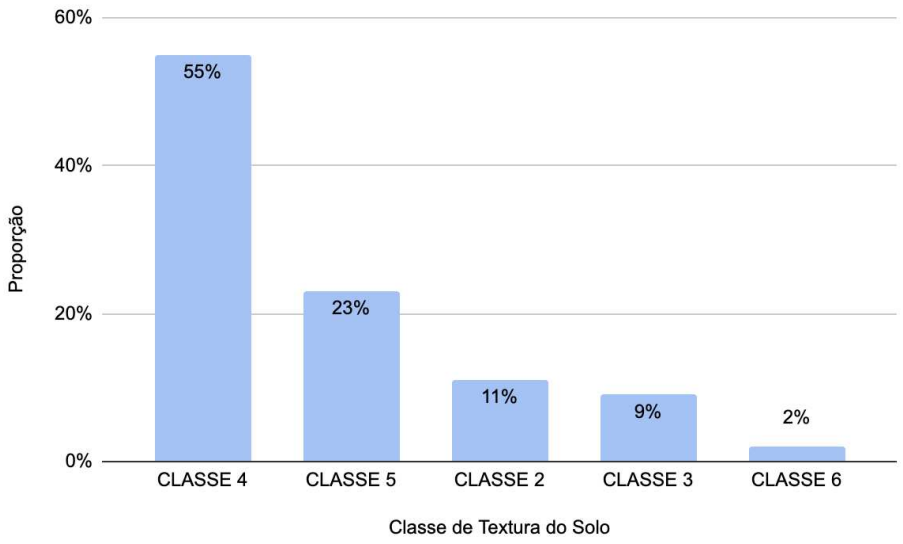
Fonte: Dados da pesquisa

O Gráfico 2 revela que a maioria do solo analisado pertence às classes intermediárias. A Classe 3 representa 50,5% do total, indicando que metade da área estudada possui um nível de fertilidade moderado, adequado para atividades agrícolas, mas provavelmente necessitando de algum tipo de manejo para maximizar a produtividade. A Classe 4, com 37,6%, está ligeiramente abaixo da Classe 3 em fertilidade, o que sugere que essa área também é relativamente produtiva, mas com uma capacidade de suporte de nutrientes um pouco inferior. Ambas as Classes 3 e 4 representam solos que, com o manejo adequado, podem ser usados de maneira eficiente para cultivos. As Classes 1 e 2, cada uma com 5,9%, são as mais férteis do grupo, destacando-se como solos de alta produtividade natural, ideais para culturas que demandam solos ricos em nutrientes. A ausência de

Classes mais baixas (como 5, 6 ou 7) sugere que, nesta área, não há solos com baixa fertilidade. Isso é um indicador positivo para o uso agrícola, pois a área não possui solos pobres que exigem intervenções intensivas.

Finalmente, para Classes de Textura do Solo (CLASSE 1 – Mais Argilosa; CLASSE 6 – Mais Arenosa; CLASSE 7 – Siltosa), o Gráfico 3 mostra a proporção encontrada na área em estudo.

Gráfico 3 - Proporção de registro por Classe de Textura do Solo



Fonte: Dados da pesquisa

O Gráfico 3 mostra a predominância de solos nas Classes 4 e 5, sugerindo uma área com textura majoritariamente intermediária, o que facilita o manejo agrícola, pois solos intermediários geralmente permitem boa retenção de nutrientes e drenagem, essenciais para uma ampla variedade de culturas. As Classes 2 e 3, mais argilosas, complementam essa área oferecendo locais com maior retenção de água, úteis em períodos de seca. A baixa presença de solos extremamente arenosos (Classe 6) minimiza a necessidade de manejo intensivo de irrigação e fertilização. Essa distribuição de texturas indica um solo bem adaptado para uso agrícola diversificado com manejo moderado.

Nesta pesquisa, optou-se por utilizar dados relativos à uma única cultura de uma única região para aumentar a probabilidade de se encontrar regras passíveis de ação ou inesperadas (Han *et al.*, 2022) para a otimização das análises.

3.1.1. Variáveis de interesse

Entende-se que o nível de Matéria Orgânica (MO) e pH devem estar dentro do padrão nutricional recomendado para o bom desenvolvimento da cultura, pois estes influenciam a disponibilidade dos demais nutrientes, aumentando a sustentabilidade do solo. Níveis adequados de MO e pH favorecem a disponibilidade de macronutrientes como N, P, K, Ca, Mg e S e de micronutrientes como B, Cu, Fe, Mn, Mo e Zn. Além disso, a MO aumenta a Capacidade de Troca Catiônica (CTC) do solo, melhorando a retenção de nutrientes. Esses fatores combinados contribuem para um ambiente de solo mais equilibrado e propício ao desenvolvimento saudável da cultura, ressaltando a importância da gestão adequada da matéria orgânica e do pH para a sustentabilidade e produtividade agrícola.

Após pesquisa bibliográfica, orientação encontrada em Centros de Referência em solos e utilizando os conceitos do KDD, a modelagem foi realizada a partir das variáveis MO e pH, criando os atributos-meta binário MO_P e pH_P separados, cada um contendo duas classes: 0 fora do padrão nutricional e 1, dentro do padrão nutricional. Para a base de dados analisada e referenciando o Sistema Brasileiro de Classificação do Solo (Santos *et al*, 2018), foi definido o teor do MO_P dentro do padrão 1, os registros contendo os valores de MO de 11 a 25 e foi definido o teor do pH_P dentro do padrão 1, os registros com os valores de pH de 5,5 a 6,5 (IAC - Boletim Técnico, 216, 2016). Para o atributo-meta MO_P, a base de dados encontra-se balanceada, sendo 45% fora do padrão nutricional (1006 registros) e 55% dentro do padrão nutricional (1249 registros). Para o atributo-meta pH_P, a base de dados encontra-se desbalanceada, sendo 38% fora do padrão nutricional (879 registros) e 62% dentro do padrão nutricional (1376 registros).

3.2. Estruturação do ambiente de Aprendizado de Máquina (Machine Learning)

Nesta pesquisa, foram desenvolvidos os seguintes passos metodológicos:

1. Pré-processamento dos dados: Foi verificado se existiam na base de dados de referência valores redundantes, ausentes ou nulos.

2. Entendimento dos dados: análise exploratória para compreender os dados que

foram usados para alcançar o propósito da pesquisa. Para isso, foi verificado as estatísticas básicas de cada variável (Tabela 3).

Tabela 3 - Resumo estatístico das variáveis

	Areia1	Silte1	Argila1	Areia2	Silte2	Argila2	Areia3	Silte3	Argila3
Min.	35.00000	0.100000	4.50000	35.00000	0.10000	4.50000	32.50000	0.400000	5.50000
1st Qu.	73.00000	1.800000	11.00000	70.00000	2.40000	13.50000	66.20000	2.500000	17.00000
Median	82.40000	3.000000	15.00000	78.80000	3.70000	17.50000	75.00000	3.300000	20.00000
Mean	78.97632	3.823636	17.20044	75.70989	3.96816	20.32195	72.55313	3.678625	23.76825
3rd Qu.	86.30000	5.000000	22.50000	82.50000	5.00000	25.25000	80.00000	5.000000	30.00000
Max.	94.70000	18.800000	55.00000	95.00000	14.60000	57.50000	93.10000	17.500000	60.00000

	pH	P	K	Ca	Mg	Al	SB	CTC
Min.	3.600000	1.000000	0.100000	1.70000	0.700000	0.0000000	2.70000	10.70000
1st Qu.	5.200000	4.200000	0.500000	9.00000	3.000000	0.0000000	12.90000	26.10000
Median	5.600000	6.500000	0.790000	12.00000	4.700000	0.0000000	17.60000	32.20000
Mean	5.518891	9.790776	0.926541	13.32785	5.059734	0.5996896	19.30213	35.17131
3rd Qu.	5.900000	10.900000	1.100000	17.00000	7.000000	0.5000000	24.30000	40.80000
Max.	7.200000	89.700000	8.350000	38.00000	25.000000	19.0000000	56.40000	92.70000

	V	m	GradText	MO
Min.	7.00000	0.000000	0.620000	2.00000
1st Qu.	46.00000	0.000000	1.070000	7.00000
Median	57.00000	0.000000	1.220000	10.00000
Mean	54.67149	4.366297	1.227211	10.54412
3rd Qu.	65.20000	3.000000	1.380000	13.00000
Max.	100.00000	61.400000	2.900000	44.00000

Fonte: Dados da pesquisa

Cada variável possui as seguintes medidas:

Min.: Valor mínimo observado.

1st Qu. (primeiro quartil): Representa o valor abaixo do qual 25% dos dados estão.

Median: Mediana (valor central), onde 50% dos dados estão abaixo e 50% acima.

Mean: Média aritmética, o valor médio dos dados.

3rd Qu. (terceiro quartil): Representa o valor abaixo do qual 75% dos dados estão.

Max.: Valor máximo observado.

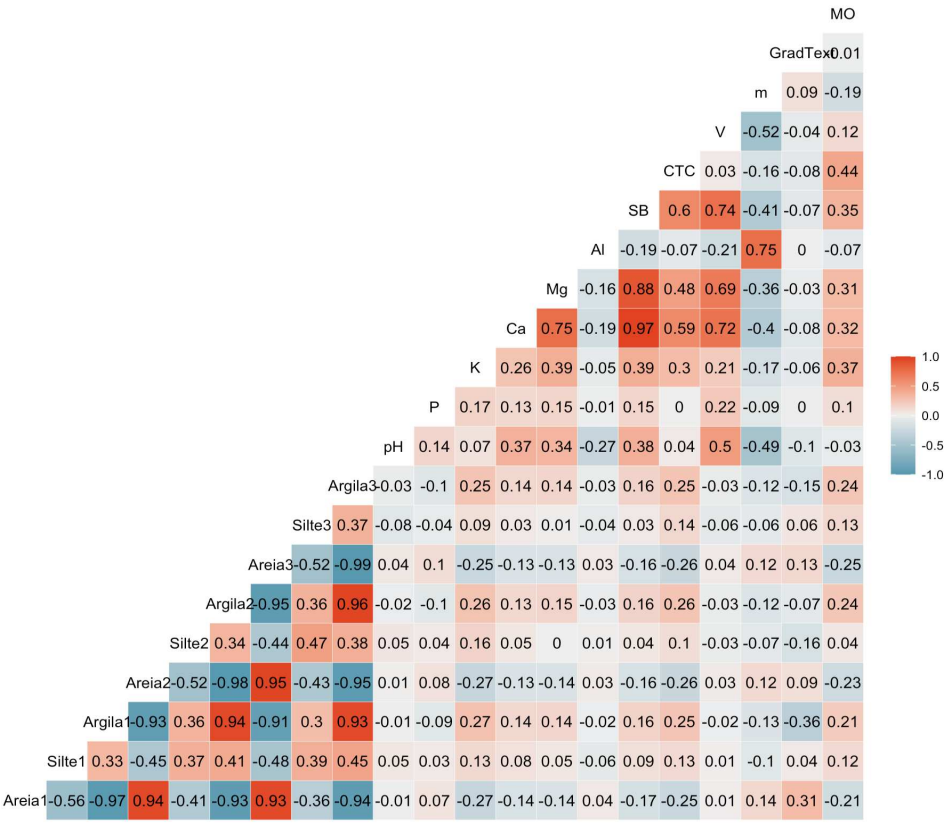
O resumo estatístico é usado para identificar padrões e variações nos dados, como solos mais arenosos ou argilosos, solos ácidos ou alcalinos e a disponibilidade de nutrientes. Fazendo uma comparação entre os valores médios de Argila e Areia, pode-se dizer que a proporção de Areia é significativamente mais alta em todas as camadas do solo, com valores médios variando de 72,55% a 78,98%, enquanto que a proporção de Argila é bem mais baixa, variando de 17,2% a 23,77%. Portanto, o solo é predominantemente arenoso. O solo arenoso tem alta drenagem, mas pode reter menos nutrientes em comparação com o solo argiloso. Sobre o solo ser ácido

ou alcalino, como o valor de pH está com a média abaixo de 7 (5,5) significa que o solo é moderadamente ácido. A acidez pode tornar menos disponíveis os nutrientes como Cálcio (Ca), Magnésio (Mg) e Potássio (K). A presença de Matéria Orgânica (MO) com valores que variam de 2,0 a 44,00 g/dm^3, é importante para a fertilidade do solo, pois melhora a capacidade de retenção de água e nutrientes e ajuda na formação de agregados de solo, melhorando a estrutura do solo.

Também foi visualizado como duas ou mais variáveis que se relacionam por meio da análise de correlação. Existem dois tipos principais de correlação: a positiva e a negativa (Figura 1).

a) Correlação Positiva: Quando duas variáveis apresentam uma correlação positiva, isso significa que, à medida que uma variável aumenta, a outra tende a aumentar também.
b) Correlação Negativa: Em contraste, a correlação negativa ocorre quando uma variável aumenta enquanto a outra diminui.

Figura 1 - Correlação entre as variáveis



Fonte: Dados da pesquisa

Na Figura 1, os diferentes tons de cores estão associados aos valores das correlações, sendo que os tons vermelhos indicam correlações positivas e os tons azuis indicam correlações negativas. As cores mais intensas indicam correlações mais fortes, enquanto tons mais claros representam correlações fracas ou próximas de zero.

A Argila é uma variável fortemente correlacionada com outras variáveis de textura do solo, como Siltes e Areias, sendo este um resultado esperado já que essas variáveis estão relacionadas entre si pela composição do solo. CTC (Capacidade de Troca de Cátions) tem uma correlação positiva moderada com variáveis como Ca e Mg e este resultado faz sentido já que esses cátions contribuem diretamente para a CTC do solo. Entre as propriedades químicas, variáveis como Ca, Mg, K, e P apresentam correlações interessantes entre si e com CTC, indicando como a fertilidade do solo pode ser inter-relacionada. Por exemplo, há uma correlação positiva entre Ca e Mg o que pode refletir num solo rico em nutrientes. O pH também está correlacionado com várias variáveis químicas, como P e Ca, refletindo como o pH pode influenciar a disponibilidade de nutrientes. A Saturação de Bases (SB) também mostra correlação com várias variáveis, especialmente aquelas relacionadas à fertilidade do solo como Ca e Mg. A Saturação de Bases é uma medida que indica a proporção de cátions básicos (Ca^{2+}, Mg^{2+}, K^{+} e Na^{+}) em relação à Capacidade de Troca de Cátions (CTC) do solo. Um valor mais alto de SB indica que o solo possui uma maior quantidade de cátions benéficos que podem ser utilizados pelas plantas.

É importante notar que correlação não implica causalidade, ou seja, mesmo que duas variáveis estejam correlacionadas, isso não significa necessariamente que uma causa a outra (Bussab; Morettin, 2010).

3. Gerenciamento do conjunto de dados: Realização da separação dos dados que serão utilizados para treinamento do modelo e os dados para realização dos testes com os algoritmos de Aprendizado de Máquina. Nesta pesquisa, a divisão é de 70% para treinamento e 30% para teste.

4. Técnicas de validação e otimização de modelos (*holdout*, validação cruzada, ajustes de hiperparâmetros): serve para ajustar o modelo, minimizar os erros e verificar se o algoritmo possui capacidade de prever com o maior nível de assertividade. A validação cruzada é uma técnica usada para detectar sobreajuste, dividindo os dados em várias partições para treinar e avaliar o modelo de forma mais robusta, o que melhora a capacidade de generalização, sendo útil em cenários com poucos dados. Já a técnica de holdout separa os dados em três partes (treinamento, validação e teste), mantendo o conjunto de teste isolado para uma avaliação final do modelo em dados não vistos. Os ajustes de hiperparâmetros envolvem otimizar parâmetros do modelo que não são ajustados durante o treinamento com a validação cruzada, ajudando a escolher a melhor configuração para evitar sobreajuste e maximizar o desempenho. Cada técnica de Aprendizado de Máquina permite a configuração de hiperparâmetros próprios.

5. Aprendizado: Identificação dos critérios adequados para minimizar o erro do algoritmo e entender o seu comportamento para as fases posteriores.

6. Aplicação: Nesta fase foi realizada a análise para avaliar se o modelo estava ou não performando. Estando em conformidade, o modelo foi validado. Do contrário, voltou-se ao passo 4.

A sequência dos itens 3 a 6 descrita no passo a passo acima foi realizada para as tarefas supervisionadas de Aprendizado de Máquina Regressão Logística, Árvore de Decisão, *Support Vector Machine* (SVM) e *Random Forest,* com o objetivo de descobrir as relações entre os atributos preditivos e o atributo-meta, usando dados históricos conhecidos.

3.3. Técnicas para seleção das variáveis

A complexidade para modelar o comportamento de algumas culturas torna difícil a seleção das variáveis (Corrêa *et al.*, 2020). Modelos com grandes conjuntos de dados nem sempre são sinônimos de alta performance. Esta pesquisa utilizou diferentes técnicas para selecionar as variáveis e avaliar o melhor desempenho.

3.3.1. Wrappers

O modelo Wrappers é uma abordagem de seleção de variáveis (*feature selection*) utilizada em Aprendizado de Máquina para melhorar a performance dos modelos preditivos. Ao contrário de métodos embutidos (*Embedded*) e de filtragem (*Filter*), que selecionam atributos com base em critérios estatísticos ou na contribuição direta dos atributos no modelo, os Wrappers envolvem o treinamento de um modelo e a avaliação do desempenho preditivo para diferentes subconjuntos de variáveis.

O método testa diferentes subconjuntos de atributos. Cada subconjunto de atributos é usado para treinar um modelo preditivo, e o desempenho do modelo é avaliado (geralmente, usando validação cruzada). O desempenho é medido utilizando métricas como acurácia, precisão, recall, F1-score, etc. O subconjunto de variáveis que resulta no melhor desempenho do modelo é selecionado.

Nesta pesquisa, foram utilizados dois tipos de Wrappers para seleção de variáveis:

a) *Forward Selection*, que começa com um conjunto vazio de atributos e adiciona iterativamente os atributos que mais aumentam a performance do modelo.

b) *Recursive Feature Elimination* (RFE), que iterativamente treina o modelo, remove o atributo menos importante, e reavalia o modelo até que o conjunto de atributos ideal seja encontrado.

3.3.2. Processo sociotécnico M-MACBETH

A Análise de Decisão Multicritério (MCDA) é uma ferramenta que permite aos atores sociais refletirem e compreenderem melhor seus problemas e, desse modo, a estruturar seus objetivos e tomar melhores decisões em cenários complexos. Seus diferenciais incluem a possibilidade de extração dos valores, aspirações e percepções de todos os agentes envolvidos na tomada de decisão; a incorporação dos aspectos subjetivos da tomada de decisão, envolvendo múltiplos critérios e mensurando-os quantitativa e qualitativamente (avaliação ordinal e cardinal) e o ranqueamento das alternativas analisadas em função de seus impactos nos objetivos estabelecidos.

A análise multicritério tem origem na teoria da decisão, que surgiu na década de 1960. Com o crescimento da Pesquisa Operacional (Operations Research - OR) nos anos 70, os pesquisadores matemáticos utilizaram técnicas avançadas para obter resultados analíticos, evoluindo de "análise elementar de problemas complexos e mal-estruturados" para "análise avançada de problemas bem-estruturados" (Raiffa, 2007), mas foi o estudo de Keeney e Raiffa (1976), que incorporou múltiplos critérios na análise de decisão, resultando na análise multicritério de decisão. Na década de 80, Roy (1986) e outros pesquisadores desenvolveram um procedimento analítico para incluir a participação dos decisores, ou seja, das pessoas com conhecimento sobre o problema em questão. Assim, surgiu a base dos métodos "Multicritérios de Apoio à Decisão Construtivista".

O método M-MACBETH (Measuring Attractiveness by a Categorical Based Evaluation Technique) de Bana e Costa e Vansnick (1997) é uma abordagem interativa de metodologia de apoio multicritério à decisão (MCDA) que visa auxiliar na identificação e estruturação de problemas complexos de gestão e na construção de opções para abordá-los. Essa abordagem ajuda a entender o problema e suas causas, identificar se o problema imediato é sintomático de um problema de longo prazo e também a identificar as principais incertezas junto aos interessados no problema, estabelecendo valores importantes considerando múltiplos critérios. A ponderação dos critérios e a avaliação das opções são baseadas em julgamentos qualitativos sobre diferenças de atratividade. Esse processo de estruturação do processo de decisão, além de superar as lacunas dos métodos multicritérios alternativos (AHP, ELECTRE e PROMETHEE), organiza o processo de decisão, tornando-o mais claro e assertivo.

Dentro desse contexto, a aplicação do processo sociotécnico da análise multicritério M-MACBETH auxilia na seleção das variáveis relevantes para a pesquisa, contribuindo para a avaliação estruturada do cenário, pois contempla os julgamentos de valores objetivos e subjetivos de um agente decisor (pessoa com informações relevantes sobre o problema) dentro de uma realidade específica, com robustez e consistência científica, aumentando a confiabilidade dos resultados. Trata-se de um método construtivista que incorpora tanto a componente social quanto a técnica.

A MCDA, ao contrário dos métodos "monocritérios" da pesquisa operacional tradicional, visa incorporar múltiplos aspectos no processo de decisão, auxiliando na escolha, ordenação ou classificação das possíveis ações (Ensslin; Montibeller, 1998). Em suma, a abordagem MCDA vai além de escolher entre alternativas pré-selecionadas, permitindo que os decisores criem alternativas mais adequadas à solução do problema com base em seus valores, por ser uma metodologia construtivista (Keeney, 1992; Ensslin *et al.*, 2001).

A implementação do processo MCDA envolve, basicamente, três fases principais: estruturação do problema, avaliação dos dados e recomendações.

1) Fase de Estruturação: Nesta etapa, define-se o contexto a ser avaliado, identificando o problema, os envolvidos e o ambiente de decisão. É nessa fase que se mapeiam os valores dos responsáveis pelo processo, resultando na construção da árvore de critérios[1]. Essa estrutura contempla escalas ordinais e não ambíguas, assegurando interpretações consistentes dos níveis de análise no contexto. Cada ponto de vista reflete um aspecto essencial da decisão real, sendo relevante para criar um modelo de avaliação que considere as ações existentes ou em desenvolvimento. Esse aspecto deriva tanto do sistema de valores quanto da estratégia de intervenção de um ator no processo decisório, agregando elementos fundamentais que moldam as preferências desse ator (Bana e Costa, 1993).
2) Fase de Avaliação: Após a construção dos descritores no mapa cognitivo, procede-se à avaliação das ações potenciais mensuráveis, definindo-se as funções de valor para cada critério. A criação dos descritores exige a identificação de níveis de impacto claros, que representem os possíveis desempenhos de uma ação, ordenados conforme a preferência dos decisores. Esses níveis devem ser consistentes com os sistemas de valores envolvidos (Bana e Costa; Silva, 1994). A função de valor, por sua vez, é uma representação matemática que expressa a intensidade das preferências (diferenças de atratividade) entre os níveis de impacto de um descritor (Ensslin *et al.*, 2001).

[1] Na literatura sobre MCDA, os termos "Critério", "Atributo" e "Ponto de Vista Fundamental" (PVF) são frequentemente utilizados de forma correlata.

3) Fase de Recomendações: Nesta etapa final, analisa-se o desempenho de cada ação potencial em relação aos descritores construídos. Com base nesse perfil de desempenho, identifica-se a necessidade de implementar ações de melhoria que possam elevar os resultados das opções avaliadas (Ensslin *et al.*, 2001).

Existem diversas metodologias multicritério que dão suporte à tomada de decisão (Belton; Stewart, 2002), e elas podem ser divididas em dois grandes grupos: modelos compensatórios e modelos não-compensatórios. No primeiro grupo, aceita-se que uma má performance de uma alternativa em um critério pode sempre ser "compensada", em termos de avaliação geral, por uma boa performance dessa alternativa em outro ou outros critérios distintos. Já no segundo grupo, entende-se que uma má performance de uma alternativa em um critério pode afetar sua avaliação geral, não sendo "compensada" por bons desempenhos em outros critérios. O modelo compensatório é o mais utilizado, e dentro dele, o modelo aditivo tem ampla aceitação por parte dos tomadores de decisão que utilizam as metodologias multicritérios.

O método M-MACBETH utiliza o modelo aditivo, mas de uma maneira diferente da maioria dos métodos disponíveis no mercado, pois não solicita ao decisor expressar juízos quantitativos. Em vez disso, pede apenas que o decisor se expresse em termos qualitativos, mantendo o rigor e a consistência científica e facilitando o julgamento. Inicialmente, aplica-se um protocolo interativo de perguntas, onde os elementos são comparados dois a dois, solicitando apenas um julgamento qualitativo. Com base nesses julgamentos, uma escala de pontuações é gerada. Por meio de um processo semelhante, determinam-se os pesos para cada critério selecionado.

A atribuição correta de significado aos pesos é importante para evitar interpretações equivocadas. Autores como Von Winterfeldt e Edwards (1986), Keeney (1992) e Lourenço (2002) discutem o assunto em detalhes. Toda essa abordagem intelectual floresceu a partir da teoria de decisão e da pesquisa comportamental de decisão, com o objetivo de compreender a natureza das preferências e valores humanos, desenvolvendo uma forma de avaliá-los, dada a

complexidade, quantidade e natureza das informações (Gregory *et al.*, 1993).

Em resumo, o processo sociotécnico M-MACBETH auxilia na estruturação do processo de decisão ao medir o grau de preferência de um decisor em relação a um conjunto de alternativas. Dessa forma, permite verificar inconsistências nos juízos de valores e possibilita revisões para a elaboração de recomendações. Sua maior vantagem é a interatividade (Bana e Costa *et al.*, 2012). Ele utiliza categorias semânticas para determinar a função de valor por meio de modelos de Programação Linear, onde o decisor é questionado sobre a diferença de atratividade entre dois elementos de forma par a par. Nesta pesquisa, tal processo foi utilizado como um seletor de variáveis da base de dados e o seu efeito foi comparado com o de outras técnicas para conhecer a decisão humana frente às decisões dos algoritmos.

De acordo com Bana e Costa (2013, p. 5), o método M-MACBETH combina informações ordinais e cardinais para criar uma escala numérica. Em outras palavras, ele gera valores qualitativos com base em critérios quantitativos.

> "Basicamente, considerando X um conjunto finito de opções, medir ordinalmente a atratividade das opções x do conjunto X consiste em associar a cada x um número real v(x) – em uma escala numérica que satisfaça as condições de preferência estrita (1), de indiferença (2) e de cardinalidade (3):
>
> (1) x, y ∈ X: [x é mais atrativa do que y (xPy) v(x) > v(y)]
>
> (2) x, y ∈ X: [x e y são igualmente atrativas (xIy) v(x) = v(y)]
>
> (3) x, y, w, z ∈ X, com x mais atrativo que y e w mais atrativo que z: o quociente [v(x) – v(y) / v(w) – v(z)] mede a diferença de atratividade entre x e y quando a diferença na atratividade entre w e z é tomada como unidade de medida.
>
> Essa nova escala numérica v: XR : xv(x) pode ser definida posicionando as opções de X sobre um eixo vertical de forma que:
>
> x, y ∈ X: x é posicionado acima de y se e somente se x é mais atrativa do que y (informação do valor ordinal).
>
> As distâncias relativas entre as opções no eixo vertical refletem as diferenças relativas de atratividade entre elas (informação de valor cardinal)" (Bana e Costa *et al.*, 2013 p. 5).

Após definir o conjunto de critérios, realiza-se a estruturação do modelo multicritério aditivo de valor. Para isso, é necessário criar uma estrutura que

represente a relação de preferência entre os diferentes elementos (x's). A combinação dos atributos em um modelo de valor é baseada no conceito de independência entre os critérios.

No processo M-MACBETH, ao comparar dois elementos de forma par a par, sendo um mais preferido que o outro, o decisor é questionado sobre a diferença de atratividade entre eles, tomando como referência o elemento mais preferido. Para expressar essa diferença de atratividade, o método propõe categorias semânticas:

C0: nenhuma diferença
C1: diferença muito fraca
C2: diferença fraca
C3: diferença moderada
C4: diferença forte
C5: diferença muito forte
C6: diferença extrema

À medida que o decisor expressa seus julgamentos, o método verifica a consistência entre eles, garantindo que haja uma representação numérica na escala de intervalos. Caso haja inconsistência, o método sugere correções.

Os algoritmos matemáticos usados para verificar a consistência e para construir uma escala de intervalos consistente, com base nas categorias semânticas, são detalhadamente apresentados por Bana e Costa *et al.* (2011). A escala de valores derivada da matriz de julgamentos semânticos só existirá se esses julgamentos forem consistentes.

Existem três conceitos principais de independência em métodos multicritérios: independência de utilidade, independência preferencial e independência aditiva. Bana e Costa (2011) apontam que a má compreensão da independência preferencial leva à confusão entre independência aditiva e estatística.

Keeney e Raiffa (1976) afirmam que a independência de utilidade facilita a definição da função de utilidade multicritério e a análise de sensibilidade. A independência preferencial verifica se os critérios são mutuamente independentes, exigindo que o desempenho de um descritor não afete a atratividade entre níveis de outro (Roy, 1986; Ensslin *et al.*, 2011).

A independência aditiva permite que a função utilidade seja aditiva,

estabelecendo como impactos menos atrativos podem ser compensados por outros mais atrativos (Bana e Costa; Beinat, 2011).

O modelo utilizado neste trabalho é o compensatório aditivo. As condições que devem ser satisfeitas para que o modelo aditivo possa ser aplicado estão descritas em Lourenço (2002), representado matematicamente por:

$$V(a) = \sum_{j=0}^{n} w_j * v_j\ (a), \text{em que } \sum_{j=0}^{n} w_j = 1, \text{e } 0 < wj < 1\ (j=1,...,n),$$

Onde:

V(a), o valor global da performance da alternativa "a" diante dos 'n' critérios.

$V_j(a)$, o valor da performance alternativa "a" diante do critério 'j'.

w_j, constantes de escalas ou coeficientes de ponderação dos 'n' critérios que permitem que as diferentes alternativas 'a' em cada critério sejam adicionáveis. Tais coeficientes são conhecidos por 'pesos'. (Lourenço, 2002).

Lourenço (2002) explica que essas condições podem ser interpretadas como os axiomas do método, sendo elas:

1) Capacidade de decisão: para cada critério, assume-se que o decisor consegue determinar qual das duas alternativas prefere, ou se as considera indiferentes.
2) Transitividade: se a alternativa A é preferida à alternativa B e a alternativa B é preferida à alternativa C, então, de acordo com este axioma, a alternativa A deve ser preferida à alternativa C.
3) Aditividade: se o decisor prefere a alternativa A à B e a B à C, a intensidade da preferência pela alternativa A em relação à C deve ser maior que a intensidade da preferência de A sobre B (ou de B sobre C).
4) Monotonia: as funções que operacionalizam os critérios são sempre crescentes ou decrescentes.
5) Limites superiores e inferiores com valores finitos: ao atribuir valores às alternativas com base nos critérios, assume-se que a melhor alternativa não é infinitamente boa e a pior não é infinitamente ruim, evitando atribuir-lhes valores de +∞ ou -∞.

O modelo aditivo hierárquico proposto por Lourenço (2002) consiste em uma combinação de modelos aditivos simples, ajustados a uma estrutura hierárquica de critérios. Nesse modelo, o valor global de uma alternativa é calculado de forma ascendente, começando pela agregação das pontuações ponderadas nos critérios mais básicos. Essas pontuações são, então, utilizadas para calcular os valores agregados nos critérios hierarquicamente superiores, repetindo-se o processo até se alcançar o valor global. Para que o modelo aditivo seja aplicado corretamente, é fundamental que haja independência entre os critérios analisados (Lourenço, 2002).

Para eliminar as informações redundantes, na fase estruturação do processo sociotécnico M-MACBETH, foi perguntado ao agente decisor se algum dos critérios da base de dados analisada continha as mesmas informações. Pelo seu conhecimento e experiência, o mesmo eliminou vários critérios, mantendo somente MO, pH, P, K, Ca, Mg, Al e GradText. Foi perguntado ao decisor porque cada critério era importante até a exaustão. Em seguida, para verificar a independência entre os critérios, o decisor respondeu aos testes ordinal (onde ele definiu qual é a sua preferência entre dois critérios, até esgotar todos os critérios) e cardinal (onde é avaliado o grau de atratividade entre esses mesmos critérios, segundo a escala semântica). De acordo com Keeney (1992, p. 133):

> "o par de atributos $\{X_1, X_2\}$ é preferencialmente independente de outros os atributos X_3, ... , X_n, se a ordem de preferência para as consequências envolvendo apenas as alterações nos níveis de X_1 e X_2 não depende dos níveis nos quais os atributos X_3, ... , X_n são fixados. Independência preferencial implica que as curvas de indiferença sobre X_1 e X_2 não dependem de outros atributos" (Keeney, 1992, p. 133).

Como não houve alteração no julgamento do decisor, os testes indicaram que os critérios possuem independência aditiva. Assim, apenas os critérios considerados importantes, independentes entre si, controláveis e essenciais foram selecionados para fazerem parte da modelagem. Detalhes metodológicos podem ser vistos em Ensslin *et al.* (2001).

O próximo passo consistiu em definir os Descritores de Impacto (Tabela 4) para os critérios selecionados, bem como estabelecer os seus níveis de

performance. Em resumo, este descritor desempenha um papel fundamental ao tornar o critério mais claro e compreensível. Ele facilita a identificação de ações de melhoria e permite a mensuração dos impactos das ações em relação a cada critério. Essas ações são avaliadas com base em uma ordem de preferência, utilizando dois níveis de referência nesta pesquisa: nível Bom (que indica um impacto satisfatório), nível Mínimo aceitável (nível de impacto que não é nem positivo e nem negativo) e o Indesejável (nível negativo). Os descritores são ordenados de maior para menor preferência. O descritor para cada critério foi criado com base no conhecimento do decisor, tendo como referência os dados da literatura existente (IAC, 2016), padronizando as unidades de medida. Vale destacar que, ao incluir posteriormente os níveis “Bom” e “Mínimo aceitável” no sistema M-MACBETH, observou-se que a escala gerada apresentou valores negativos. Para evitar escala negativa, decidiu-se considerar apenas os níveis “Bom” e “Indesejado”.

Tabela 4 - Exemplo de Descritor de Impacto

Matéria Orgânica (MO)	É a fonte primordial de Carbono. Constitui a principal fonte de energia e nutrientes para a atividade microbiana.
Bom	25 g/dm3
Mínimo aceitável	11 g/dm3
Indesejado	<10 e > 30g/dm3

Fonte: Dados da Pesquisa

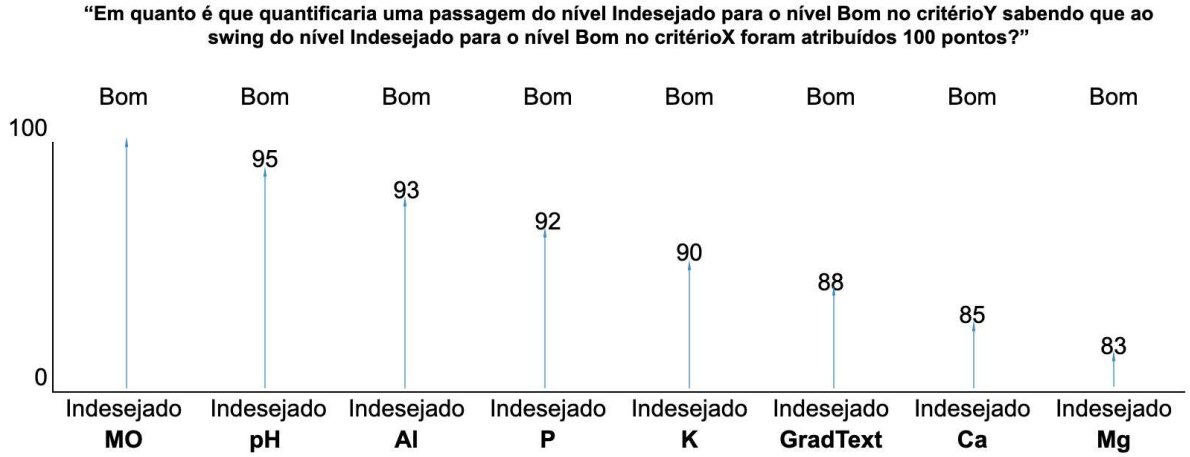
Seguindo o procedimento metodológico, a fase seguinte determinou os pesos dos critérios utilizando o método Swing que, segundo as orientações de Lourenço (2002), o decisor, com a sua experiência e conhecimento, precisa responder a duas perguntas: a primeira é “Se fosse possível passar do nível Indesejado para o nível Bom um único critério, qual critério selecionaria para esta mudança?” (Figura 2). O decisor optou pela Matéria Orgânica. Fazendo a mesma pergunta, o pH foi escolhido e assim sucessivamente até o término de todos os critérios selecionados pelo decisor.

Figura 2 - Determinação do peso pelo método Swing

Fonte: Dados da Pesquisa

Ordenando os critérios da Figura 2, a segunda pergunta é "Em quanto é que quantificaria uma passagem do nível Indesejado para o nível Bom no critérioY sabendo que ao swing do nível Indesejado para o nível Bom no critérioX foram atribuídos 100 pontos?" Segundo o decisor, "Passar do pior nível para o melhor em pH é equivalente a 95% do swing em MO". Novamente, esta pergunta foi feita para os demais critérios. A Figura 3 apresenta a representação gráfica do valor atribuído à passagem (Swing) do nível Indesejado para o nível Bom em todos os critérios.

Figura 3 - Representação gráfica do método Swing



Fonte: Dados da Pesquisa

Segundo o julgamento de valor do decisor, a Matéria Orgânica é um critério fundamental para entrar na análise, uma vez que ela indica a fertilidade do solo e o pH é limitante, pois se o mesmo não estiver adequado, ele impossibilita a absorção dos outros nutrientes pelas plantas.

Para Lourenço (2002), a determinação do peso pelo método Swing deve-se

proceder à normalização dos coeficientes de ponderação obtidos na etapa anterior para que a sua soma seja igual a 1, utilizando a seguinte expressão:

$$k_j = \frac{k'_j}{\sum_{j=1}^{n} k'_j} \text{ , com } (j = 1, \ldots, n)$$

Onde:

k'_j é o coeficiente de ponderação não normalizado do ponto de vista j;

k_j é o coeficiente de ponderação normalizado do ponto de vista j.

$$kMO = \frac{100}{100+95+93+92+90+88+85+83} = 0,1377$$

$$kpH = \frac{95}{100+95+93+92+90+88+85+83} = 0,1309$$

$$kAl = \frac{93}{100+95+93+92+90+88+85+83} = 0,1281$$

$$kP = \frac{92}{100+95+93+92+90+88+85+83} = 0,1267$$

$$kK = \frac{90}{100+95+93+92+90+88+85+83} = 0,1240$$

$$radText = \frac{88}{100+95+93+92+90+88+85+83} = 0,1212$$

$$kCa = \frac{85}{100+95+93+92+90+88+85+83} = 0,1171$$

$$kMg = \frac{83}{100+95+93+92+90+88+85+83} = 0,1143$$

Após a normalização do peso, o processo sociotécnico apresentou a seguinte ordem de importância dos critérios selecionados (Tabela 5).

Tabela 5 - Importância dos critérios

	MO	pH	Al	P	K	Grad TExt	Ca	Mg
Pesos	0,1377	0,1309	0,1281	0,1267	0,124	0,1212	0,1171	0,1143

Fonte: Dados da Pesquisa

E como última etapa, foi realizada a análise de sensibilidade e robustez dos critérios selecionados no processo sociotécnico M-MACBETH o qual oferece funcionalidades específicas para conduzir estas análises (Bana e Costa *et al.*, 2017). A análise de sensibilidade avalia como as mudanças nos pesos dos critérios ou nas pontuações das opções afetam os resultados finais do modelo. Isso ajuda a identificar quais critérios ou opções são mais sensíveis a mudanças, indicando a estabilidade do modelo.

A análise de robustez é um processo abrangente que examina como diferentes combinações de pesos e pontuações influenciam os rankings finais dos critérios. O objetivo principal é identificar conclusões robustas, ou seja, aquelas que permanecem estáveis e consistentes mesmo quando submetidas a diferentes cenários e condições. Esse método permite aos tomadores de decisão ter uma compreensão mais profunda da confiabilidade e estabilidade das recomendações geradas pelo M-MACBETH, fornecendo uma base mais sólida para decisões estratégicas em ambientes complexos e incertos.

Com base nas análises de sensibilidade e robustez realizadas através do sistema M-MACBETH, conclui-se que o processo sociotécnico apresentou uma estrutura sólida e confiável para usar os critérios selecionados pelo decisor na modelagem com as técnicas de Aprendizado de Máquina.

3.4. Tabela comparativa entre as técnicas utilizadas para seleção das variáveis

A Tabela 6 apresenta a seleção de critérios que foi usada para a análise comparativa e avaliação de desempenho da modelagem.

Tabela 6 - Seleção de Variáveis

	Wrappers		
Variável	**RFE**	**Forward Selection**	**M-Macbeth**
K	Sim	Sim	Sim
pH	Sim	Não	Sim
P	Sim	Sim	Sim
CTC	Sim	Sim	Não
V	Sim	Não	Não
m	Sim	Não	Não
Al	Sim	Não	Sim
Mg	Sim	Não	Sim
SB	Sim	Sim	Não
Areia1	Não	Não	Não
Silte1	Não	Não	Não
Argila1	Não	Não	Não
Areia2	Não	Não	Não
Silte2	Não	Não	Não
Argila2	Não	Não	Não
Areia3	Não	Não	Não
Silte 3	Não	Não	Não
Argila3	Não	Não	Não
Ca	Não	Não	Sim
GradText	Não	Não	Sim
MO	Sim	Sim	Sim

Fonte: Dados da Pesquisa

3.5. Avaliação de desempenho dos modelos de Classificação

A qualidade do modelo foi avaliada por meio da precisão preditiva, utilizando a matriz de confusão (Figura 4).

Figura 4 - Matriz de Confusão

		Valor Predito	
		Sim	**Não**
Real	**Sim**	Verdadeiro Positivo (TP)	Falso Negativo (FN)
	Não	Falso Positivo (FP)	Verdadeiro Negativo (TN)

Fonte: Freitas (2015)

Segue o detalhamento das métricas utilizadas para avaliação dos modelos:

1. *tpr* - *True Positive Rate* (Sensibilidade ou Recall): proporção de exemplos positivos corretamente identificados como positivos. Mede a habilidade do modelo de detectar verdadeiros positivos entre todos os casos positivos. A *tpr* alta indica que poucos positivos foram perdidos.

$$TPR = \frac{TP}{TP + FN}$$

2. *tnr* - *True Negative Rate* (Especificidade): proporção de exemplos negativos corretamente identificados como negativos. Avalia o quanto o modelo é eficaz em identificar corretamente os negativos. A *tnr* alta significa poucos falsos positivos.

$$TNR = \frac{TN}{TN + FP}$$

3. *fpr* - *False Positive Rate* (Fall-out): proporção de exemplos negativos incorretamente classificados como positivos. Mede a frequência de falsos positivos. A *frp* baixa indica que o modelo comete poucos erros ao identificar falsos positivos.

$$FPR = \frac{FP}{FP + TN}$$

4. *fnr* - *False Negative Rate* (Miss rate): proporção de exemplos positivos incorretamente classificados como negativos. Mede a frequência de falsos negativos. A *fnr* baixa indica que poucos positivos reais foram classificados erroneamente como negativos.

$$FNR = \frac{FN}{FN + TP}$$

5. *ppv* - *Positive Predictive Value* (Precisão): proporção de exemplos positivos preditos corretamente entre todas as previsões positivas. Indica a confiança que se

pode ter quando o modelo prevê um positivo. Uma *ppv* alta significa poucas previsões de falsos positivos.

$$PPV = \frac{TP}{TP + FP}$$

6. *for* - *False Omission Rate*: proporção de exemplos negativos preditos incorretamente entre todas as previsões negativas. Mede a frequência de falsos negativos entre as previsões negativas. A *for* baixa significa que poucos negativos foram previstos incorretamente como positivos.

$$FOR = \frac{FN}{FN + TN}$$

7. LR+ - *Positive Likelihood Ratio* (Razão de Verossimilhança Positiva): mede o quão mais provável é obter um resultado positivo em um caso verdadeiro positivo em comparação com um caso verdadeiro negativo. Uma LR+ maior que 1 indica que o teste é útil para distinguir positivos reais.

$$LR+ = \frac{TPR}{FPR} = \frac{TP/(TP + FN)}{FP/(FP + TN)}$$

8. *fdr* - *False Discovery Rate*: proporção de exemplos positivos preditos incorretamente entre todas as previsões positivas. A *fdr* baixa indica que poucas das previsões positivas feitas pelo modelo são falsos positivos.

$$FDR = \frac{FP}{FP + TP}$$

9. *npv* - *Negative Predictive Value*: proporção de exemplos negativos preditos corretamente entre todas as previsões negativas. Mede a confiança que se pode ter quando o modelo prevê um negativo. Uma *npv* alta significa poucas previsões de falsos negativos.

$$NPV = \frac{TN}{TN + FN}$$

10. acc - *Accuracy* (Acurácia): proporção de exemplos corretamente classificados entre todos os exemplos. Mede a proporção geral de classificações corretas feitas pelo modelo, tanto para positivos quanto para negativos.

$$ACC = \frac{TP + TN}{TP + TN + FP + FN}$$

11. LR- - *Negative Likelihood Ratio* (Razão de Verossimilhança Negativa): mede o quão menos provável é obter um resultado negativo em um caso positivo verdadeiro em comparação com um negativo verdadeiro. Uma LR- menor que 1 indica que o teste é útil para excluir a possibilidade de um caso positivo real. Quanto menor o LR-, melhor o teste é para essa função.

$$LR- = \frac{FNR}{TNR} = \frac{FN/(TP + FN)}{TN/(FP + TN)}$$

12. *DOR* - *Diagnostic Odds Ratio*: mede a razão entre a chance de um teste diagnóstico dar um resultado correto e a chance de dar um resultado incorreto. Ela combina sensibilidade e especificidade em uma única métrica. Um valor alto de *dor* indica que o teste tem boa capacidade de discriminar entre os indivíduos com e sem a condição.

$$DOR = \frac{TP/FP}{FN/TN} = \frac{LR+}{LR-}$$

13. Curva ROC: A Curva ROC - *Receiver Operating Characteristic Curve* é usada para avaliar a capacidade de predição do modelo, medindo as predições da sensibilidade (*tpr*) e da especificidade (*fpr*) para diferentes limiares de Classificação e apresentando os resultados de forma gráfica. Conforme explicado por Fawcett (2006), ela permite visualizar, organizar e classificar o modelo com base na sua performance preditiva. O desempenho do modelo é geralmente quantificado pela área sob a curva (AUC - *Area Under the Curve*). Quanto mais próxima a AUC estiver de 1, melhor é o desempenho do classificador. Um AUC de 0,5 indica que o modelo não está melhor do que um chute aleatório. Modelos com curvas mais elevadas também apresentam maior precisão ao lidar com dados desbalanceados.

CAPÍTULO IV – RESULTADOS E DISCUSSÃO

O uso de algoritmos por si só tem sido facilitado ao longo dos anos, contudo, saber interpretar e transformar a informação extraída dos seus resultados para apoiar a tomada de decisão ainda requer análise e pesquisa. Um dos desafios é estabelecer um conjunto consistente de variáveis que garantam resultados robustos e satisfatórios para todas as técnicas analisadas.

4.1. Apresentação dos resultados

Nesta pesquisa, a modelagem em Aprendizado de Máquina para o problema de Classificação tendo como atributo-meta a variável MO_P ou pH_P iniciou com a Regressão Logística por ser o método estatístico mais utilizado para modelar variáveis categóricas e/ou binárias. Em seguida, a modelagem foi realizada por Árvore de Decisão, *SVM* e *Random Forest*. A modelagem com o atributo-meta MO_P apresentou melhor performance que com o atributo-meta pH_P e, por isso, foi detalhada a seguir. As análises realizadas com o atributo-meta pH_P, para efeitos comparativos, encontram-se no item 4.1.1 deste documento.

Todas as técnicas utilizaram *holdout* e validação cruzada com 10 *folds*. Este número foi escolhido por representar um bom equilíbrio entre viés e variância e tem o objetivo de minimizar o efeito do *overfitting*. Para efeitos de comparação entre os algoritmos, foi mantida a mesma semente do gerador pseudo-aleatório em todas as simulações (*set.seed*). Foram realizadas 20 iterações para todas as técnicas.

A Tabela 7 apresenta as métricas geradas pelas técnicas Regressão Logística (rg), Árvore de Decisão (dt), Support Vector Machine (svm) e Random Forest (rf):

Tabela 7 - Métricas geradas pelas técnicas de Aprendizado de Máquina

Métrica	rg	dt	svm	rf
	true 0 1 err 0 TP=189 FP=96 96 1 FN=104 TN=288 104 err 104 96 200	true 0 1 err 0 TP=221 FP=64 64 1 FN=115 TN=227 115 err 115 64 179	true 0 1 err 0 TP=231 FP=54 54 1 FN=29 TN=363 29 err 29 54 83	true 0 1 err 0 TP=264 FP=21 21 1 FN=11 TN=381 11 err 11 21 32
tpr	0,66	0,78	0,81	0,93
tnr	0,73	0,71	0,93	0,97
fpr	0,27	0,29	0,07	0,03
fnr	0,34	0,22	0,19	0,07
ppv	0,65	0,66	0,89	0,96
for	0,25	0,19	0,13	0,05
lrp	2,5	2,64	10,96	33,01
fdr	0,35	0,34	0,11	0,04
npv	0,75	0,81	0,87	0,95
acc	0,7	0,74	0,88	0,95
lrm	0,46	0,32	0,2	0,08
dor	5,45	8,32	53,55	435,43

Legenda

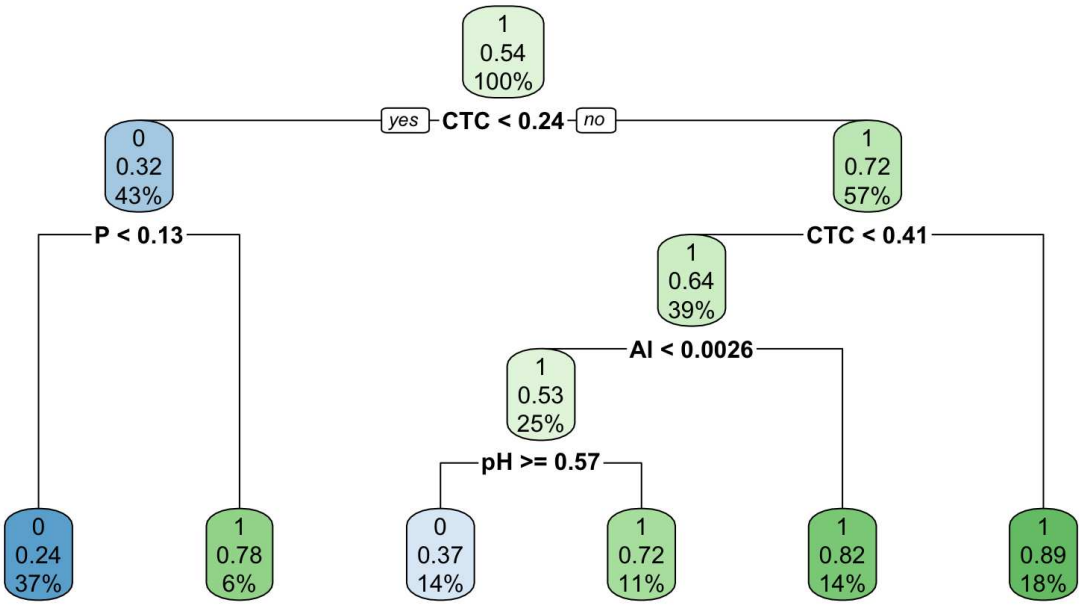
tpr - True positive rate (Sensitivity, Recall)
tnr - True negative rate (Specificity)
fpr - False positive rate (Fall-out)
fnr - False negative rate (Miss rate)
ppv - Positive predictive value (Precision)
for - False omission rate
lrp - Positive likelihood ratio (LR+)
fdr - False discovery rate
npv - Negative predictive value
acc - Accuracy
lrm - Negative likelihood ratio (LR-)
dor - Diagnostic odds ratio

Fonte: Dados da Pesquisa

Para a Regressão Logística (rg) foi utilizada a função de ligação *logit*. Nesta pesquisa, o *cost* variou de 0,01 a 0,5, apresentando como melhor resultado 0,423. O modelo produziu a acurácia de 0,70. Foram realizadas as análises de Sensibilidade (*tpr* - proporção de casos positivos que foram identificados corretamente) e Especificidade (*tnr* - proporção de casos negativos que foram identificados corretamente) para avaliar a performance do modelo, tendo como resultados, 066 e 0,73, respectivamente. Uma alta taxa de *tpr* significa que o modelo consegue identificar adequadamente solos de alta qualidade ou com boa sustentabilidade e a alta especificidade indica que o modelo não rotula erroneamente solos pobres como sustentáveis, o que é essencial para evitar diagnósticos equivocados de qualidade do solo.

Para a Árvore de Decisão (dt), os hiperparâmetros *minbucket* variou de 25 a 150, o *cp* de 0,01 a 0,05 e o *maxdepth* de 2 a 5, apresentando como melhor configuração, *minbucket*=65, *cp*=0,02309 e *maxdepth*=5. O modelo produziu a acurácia de 0,74, *tpr* de 0,78 e *tnr* de 0,71. Até aqui, a Árvore de Decisão apresentou melhor resultado que a Regressão Logística, pois previu menos falso negativo (*fnr* = 0,22), aumentando a margem de segurança. A *fnr* indica que o modelo é eficaz em identificar corretamente áreas que realmente têm problemas de sustentabilidade. Se um solo pobre for classificado como fértil, medidas de recuperação podem ser ignoradas, levando à degradação continuada. Portanto, um baixo valor de *fnr* é o ideal, pois minimiza o erro de tratar solos inadequados como se fossem produtivos. Contudo, esta técnica produziu uma maior *fpr* (0,29) quando comparado com a Regressão Logística, o que significa que Árvore de Decisão tem a tendência de sinalizar problemas onde eles não existem, podendo dificultar a priorização das áreas que realmente necessitam de ação. A Figura 5 representa o desenho da árvore gerada pela modelagem.

Figura 5 - Árvore gerada pela modelagem



Fonte: Dados da Pesquisa

A técnica Árvore de Decisão classificou como primeiro ponto de decisão um determinado registro do conjunto de dados com base na variável CTC (Capacidade de Troca Catiônica) do solo, que está associado à capacidade de retenção de nutrientes. Se CTC < 0,24, o modelo segue para a esquerda, sugerindo um solo com menor capacidade de retenção de nutrientes, o que pode indicar um estado menos sustentável.

No próximo passo, o modelo verifica o nível de P (Fósforo). O fósforo é um nutriente essencial para o crescimento das plantas. Se P < 0,13, o solo é classificado como 0 (fora do padrão nutricional), com uma proporção de 0,24 e representando 37% dos dados nessa ramificação. Esse resultado sugere que solos com baixa CTC e baixos níveis de Fósforo são considerados insustentáveis, talvez pela dificuldade de sustentação de culturas agrícolas sem suplementação. Se P ≥ 0,13, o solo é classificado como 1 (dentro do padrão nutricional), com uma probabilidade de 0,78, o que representa 6% dos dados. Isso indica que mesmo com baixa CTC, um nível de Fósforo adequado pode sustentar a produtividade.

Agora, se CTC ≥ 0,24, o modelo segue para a direita. Se a CTC está entre 0,24 e 0,41, a árvore continua avaliando outras variáveis. Solos com CTC nesta faixa ainda possuem uma capacidade moderada de troca catiônica, o que pode influenciar a sustentabilidade dependendo de outros nutrientes. Neste ponto, a árvore verifica o nível de Alumínio (Al) no solo. Altos níveis de alumínio podem ser tóxicos para muitas culturas, inclusive, para a cana-de-açúcar e são comuns em solos ácidos. Se Al < 0,0026, o solo segue para um próximo nível de verificação (mais propenso a ser sustentável). Neste caso, a árvore avalia o pH do solo. Um pH mais elevado pode reduzir a toxicidade do alumínio e melhorar a absorção de nutrientes. Se o pH é maior ou igual a 0,57, o solo é classificado como 0 (fora do padrão nutricional), com probabilidade de 0,37 e representando 14% dos dados. Isso sugere que mesmo com baixos níveis de Alumínio, um pH inadequado pode afetar a sustentabilidade do solo. Caso contrário (pH < 0,57), o solo é classificado como 1 (dentro do padrão nutricional), com uma probabilidade de 0,72 e representando 11% dos dados, indicando que um pH ideal contribui para a sustentabilidade. Este resultado reflete a maior taxa de *fpr* encontrada por esta técnica.

Se a CTC é > 0,41, o solo é classificado diretamente como 1 (dentro do padrão nutricional), com uma probabilidade de 0,89 e representando 18% dos dados. Isso sugere que solos com alta CTC tendem a ser mais sustentáveis, pois podem reter melhor os nutrientes necessários para o crescimento saudável das plantas. O percentual (%) encontrado em cada nó representa a proporção do total de observações do conjunto de dados que atingiu o nó específico. Portanto, esta árvore demonstra que a CTC é o fator mais influente para classificar um registro da base de dados, seguido do Fósforo (P), especialmente quando a CTC é baixa.

No SVM (*Support Vector Machine*), o hiperparâmetro *cost* variou entre 0,1 e 1 e o hiperparâmetro *gamma* variou entre 0,1 e 2, apresentando como melhor configuração 0,728 e 0,959, respectivamente. O modelo produziu a acurácia de 0,88. Após a análise de sensibilidade (0,81) e especificidade (0,93), o SVM apresentou melhor performance entre as técnicas avaliadas até este momento, pois possui menor *fpr* (0,07) e maior *ppv* (0,89) que as técnicas anteriores. A *ppv* refere-se à precisão do modelo em prever quais solos são férteis ou sustentáveis de fato. Uma alta precisão significa que, quando o modelo classifica um solo como sustentável, ele realmente o é, sendo importante para a alocação eficiente de recursos de cultivo. Semelhante a *fpr*, mas calculado a partir de uma perspectiva diferente, a *for* de 0,13 indica a classificação correta. Um alto valor de *for* indica que muitos solos que não são férteis estão sendo rotulados incorretamente, o que prejudica a tomada de decisões agrícolas.

Já no *Random Forest*, o hiperparâmetro *ntree* variou de 100 e 500, o *mtry* variou de 2 a 5, apresentando como melhor configuração, 237 e 5, respectivamente. O modelo apresentou a melhor acurácia (0,95) entre as técnicas avaliadas e tem a tendência de prever melhor os verdadeiros positivos e verdadeiros negativos que o SVM. A análise de sensibilidade (0,93) e especificidade (0,97) apresenta resultados elevados, comprovando a boa performance do modelo. O alto valor da *lrp* (33,01) indica confiança nas previsões de fertilidade e o baixo valor de *lmr* (0,08) e de *for* (0,05) indicam que, se o modelo disser que o solo não é fértil, essa previsão tem alta probabilidade de estar correta. Além disso, possui um nível elevado de *dor* (435,43) sugerindo que o modelo é eficaz em distinguir solos sustentáveis de não sustentáveis. Isso indica que o modelo seria uma ferramenta útil para guiar decisões

agrícolas sobre a recuperação ou a gestão de solos.

Para efeito comparativo entre a escolha das variáveis realizada pelos algoritmos e pelo homem (processo sócio-técnico M-MACBETH), foi realizada a modelagem usando as mesmas técnicas de Aprendizado de Máquina, com as mesmas configurações para o atributo-meta MO_P. Segue o resultado (Tabela 8):

Tabela 8 - Métricas geradas pelo M-MACBETH como seletor de variáveis

Métrica	rg	dt	svm	rf
	true 0 1 err 0 TP=184 FP=101 101 1 FN=103 TN=289 103 err 103 101 204	true 0 1 err 0 TP=188 FP=97 97 1 FN=68 TN=324 68 err 68 97 165	true 0 1 err 0 TP=238 FP=47 47 1 FN=37 TN=355 37 err 37 47 84	true 0 1 err 0 TP=262 FP=23 23 1 FN=14 TN=378 14 err 14 23 37
tpr	0,65	0,66	0,84	0,92
tnr	0,74	0,83	0,91	0,96
fpr	0,26	0,17	0,09	0,04
fnr	0,35	0,34	0,16	0,08
ppv	0,64	0,73	0,87	0,95
for	0,26	0,23	0,12	0,06
lrp	2,46	3,8	8,85	25,74
fdr	0,36	0,27	0,13	0,05
npv	0,74	0,77	0,88	0,94
acc	0,7	0,76	0,88	0,95
lrm	0,48	0,41	0,18	0,08
dor	5,11	9,23	48,59	307,57

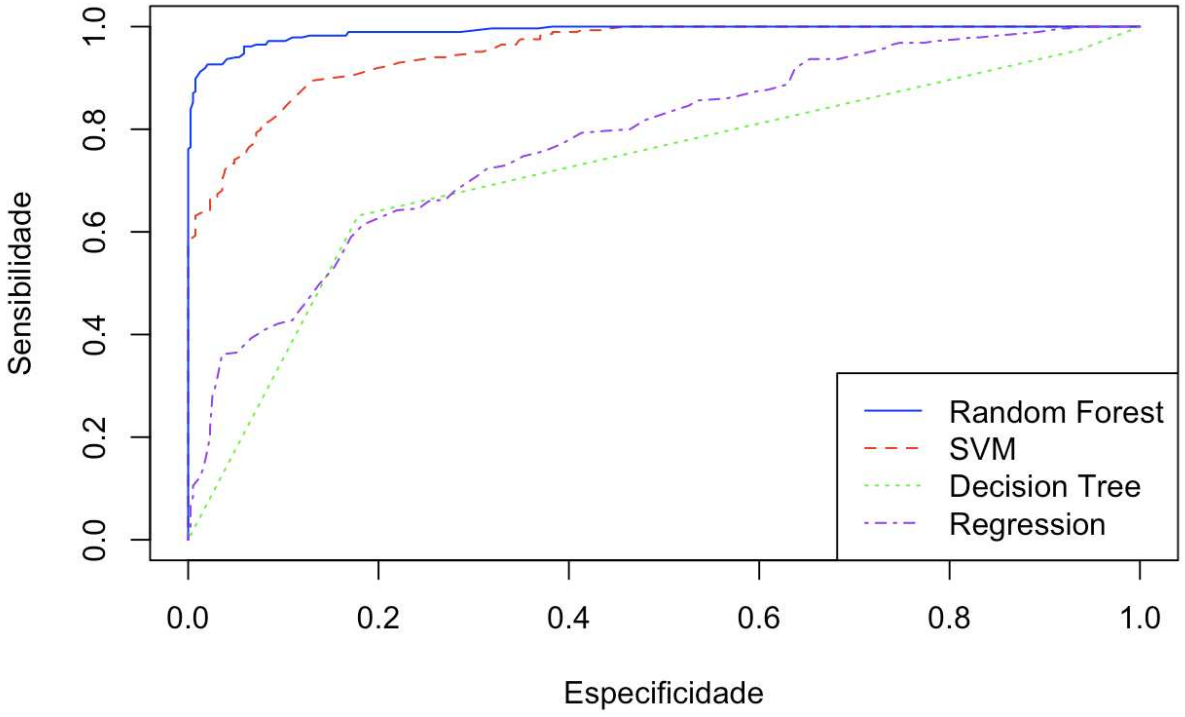
Legenda

tpr - True positive rate (Sensitivity, Recall)
tnr - True negative rate (Specificity)
fpr - False positive rate (Fall-out)
fnr - False negative rate (Miss rate)
ppv - Positive predictive value (Precision)
for - False omission rate
lrp - Positive likelihood ratio (LR+)
fdr - False discovery rate
npv - Negative predictive value
acc - Accuracy
lrm - Negative likelihood ratio (LR-)
dor - Diagnostic odds ratio

Fonte: Dados da Pesquisa

Novamente, o *Random Forest* apresentou o melhor desempenho. O método M-MACBETH teve uma boa performance, com *tpr* de 0,92 e *tnr* de 0,96. A *dor* de 307,57 sugere que o modelo tem uma boa capacidade de distinguir entre as classes positivas e negativas. A boa performance do *Random Forest* também pode ser observada pela comparação das Curvas ROC dos modelos (Figura 6).

Figura 6 - Comparação de Curvas ROC



Fonte: Dados da Pesquisa

Quanto mais elevada e distante da linha diagonal estiver a Curva ROC (*Receiver Operating Characteristic Curve)*, melhor é a performance do modelo, indicando uma maior capacidade de discriminar corretamente entre as Classes. Uma curva que se aproxima do canto superior esquerdo do gráfico demonstra uma alta taxa de verdadeiros positivos (*tpr*) e uma baixa taxa de falsos positivos (*fpr*), sugerindo que o modelo separa as classes com alta precisão. O eixo horizontal da Curva ROC representa o complemento da especificidade, ou seja, a taxa de falsos positivos (*fpr*); valores mais próximos de zero indicam que o modelo evita falsos positivos com eficiência.

Considerando o *Random Forest* o melhor modelo, fez-se a modelagem com as técnicas de seleção de variáveis Wrappers descritas no item 3.3.1 deste documento para compreender se havia mudança significativa na performance do modelo. Seguem os resultados (Tabela 9):

Tabela 9 - Métricas geradas pelas técnicas Wrappers

	Wrappers	
Métrica	**RFE**	**Forward Selection**
	true 0 1 err 0 TP=271 FP=14 14 1 FN=11 TN=381 11 err 11 14 25	true 0 1 err 0 TP=265 FP=20 20 1 FN=15 TN=377 15 err 15 20 35
tpr	0,95	0,93
tnr	0,97	0,96
fpr	0,03	0,04
fnr	0,05	0,07
ppv	0,96	0,95
for	0,04	0,05
lrp	33,89	24,3
fdr	0,04	0,05
npv	0,96	0,95
acc	0,96	0,95
lrm	0,05	0,07
dor	670,46	333,02

Legenda

tpr - True positive rate (Sensitivity, Recall)	lrp - Positive likelihood ratio (LR+)
tnr - True negative rate (Specificity)	fdr - False discovery rate
fpr - False positive rate (Fall-out)	npv - Negative predictive value
fnr - False negative rate (Miss rate)	acc - Accuracy
ppv - Positive predictive value (Precision)	lrm - Negative likelihood ratio (LR-)
for - False omission rate	dor - Diagnostic odds ratio

Fonte: Dados da Pesquisa

Sobre as técnicas de seleção de variáveis, o RFE (*Recursive Feature Elimination)* apresenta os melhores resultados na maioria das métricas, com uma taxa de verdadeiros positivos (*tpr*) de 0,95 e uma taxa de verdadeiros negativos (*tnr*) de 0,97, o que sugere um excelente equilíbrio entre a capacidade de detectar casos positivos e negativos. Também possui a menor taxa de falsos positivos (*fpr* = 0,03) e uma baixa taxa de falsos negativos (*fnr* = 0,05). Com um *ppv* de 0,96 e um *npv* de 0,96, as previsões positivas e negativas são muito confiáveis e a razão de odds (possibilidades) diagnóstica (*dor*) de 670,46 indica que este modelo tem a maior capacidade de distinguir entre casos da classe positiva e negativa. Portanto, o RFE

é o melhor método em termos de performance geral, especialmente quando a precisão e a sensibilidade são importantes.

Já o método *Forward Selection* tem um bom desempenho com *tpr* de 0,93 e *tnr* de 0,96, o que significa que também faz um bom trabalho em distinguir positivos e negativos, embora ligeiramente inferior ao RFE. A *fpr* de 0,04 e a *fnr* de 0,07 deste métodos são maiores que as do RFE, o que indica que ele comete mais erros. A *dor* de 333,02 ainda é bastante alta, indicando que o modelo é confiável, mas inferior ao RFE. Neste sentido, o *Forward Selection* é uma boa alternativa ao RFE, mas com leve redução na acurácia e na capacidade de discriminação.

Comparando o M-MACBETH com Wrappers, em resumo, o método M-MACBETH performou bem, mas seu desempenho foi inferior ao RFE. Ele se assemelha ao *Forward Selection*, sendo uma alternativa viável quando é necessário um bom equilíbrio de performance entre sensibilidade e especificidade.

4.1.1 Modelagem com o atributo-meta pH_P

Para uma análise mais completa, nesta pesquisa, foi realizada a modelagem para o problema de Classificação tendo como atributo-meta a variável pH com as mesmas técnicas. Na Regressão Logística (rg), o *cost* apresentou como melhor resultado 0,262. Para a Árvore de Decisão (dt), os hiperparâmetros *minbucket* = 54, o *cp* = 0,0593 e o *maxdepth* = 3, apresentaram a melhor configuração. Para o *Support Vector Machine* (svm), o *cost* = 0,6999 e *gamma* = 1,026 tiveram o melhor desempenho. Para o *Random Forest* (rf), os hiperparâmetros *ntree* = 427 e o *mtry* = 2 apresentaram a melhor configuração. A Tabela 10 mostra as principais métricas:

Tabela 10 - Métricas geradas para avaliação do modelo utilizando pH_P

Técnica	Sensibilidade (tpr)	Especificidade (tnr)	Precisão (ppv)	Acurácia (acc)
Regressão Logistica	0,24	0,95	0,73	0,68
Árvore de Decisão	0,61	0,87	0,74	0,77
Support Vector Machine	0,73	0,95	0,91	0,87
Random Forest	0,85	0,97	0,94	0,92

Legenda: tpr - True positive rate tnr - True negative rate ppv - Positive predictive value acc - Accuracy

Fonte: Dados da Pesquisa

A Tabela 9 mostra que o *Random Forest* se destaca com uma sensibilidade de 0,85, indicando que ele é o mais eficaz em detectar corretamente os solos com pH favorável ao uso do solo. Isso pode ser importante para garantir que solos identificados como sustentáveis tenham as condições adequadas para o cultivo. O *Random Forest* e o *Support Vector Machine* têm altas taxas de especificidade (0,97 e 0,95, respectivamente), o que significa que esses modelos são precisos ao evitar falsos positivos, identificando corretamente solos com pH inadequado, o que permite fazer a correção do solo para o plantio. O *Random Forest* também se destaca na análise da *ppv* com 0,94, indicando que quase todas as vezes que o modelo prevê que o solo está com pH adequado, essa previsão é correta, sendo importante para evitar desperdício de recursos. Também lidera em termos de acurácia (0,92), tornando o modelo mais confiável entre os quatro testados, tanto na identificação de solos sustentáveis quanto não sustentáveis. Portanto, o *Random Forest* com o atributo-meta pH_P e com os mesmos hiperparâmetros configurados também teve melhor performance, mostrando-se o melhor modelo entre os quatro avaliados, porém teve um desempenho ligeiramente inferior que as métricas encontradas com o atributo-meta MO_P.

4.2. Discussão

Para o conjunto de dados analisado e parâmetros das técnicas configuradas, o *Random Forest* com o atributo-meta binário MO_P (Matéria Orgânica) apresentou o melhor desempenho entre todas as técnicas. Um modelo com boa sensibilidade, especificidade e acurácia pode identificar corretamente solos que precisam de intervenções como adição de nutrientes ou manejo conservacionista.

A Regressão Logística, apesar de ser uma técnica conhecida e amplamente utilizada em problemas de Classificação, apresentou o pior desempenho. Isto porque ela possui limitações: a Regressão Logística encontrou uma linha de separação entre as Classes (dentro ou fora do padrão), porém essa linha pode não ser a mais adequada ou ótima para o conjunto de dados analisado.

As Árvores de Decisão são intuitivas e fáceis de visualizar e interpretar, porém são instáveis, o que significa que pequenas variações nos dados de treinamento

podem resultar em árvores completamente diferentes (alta variância). Isso torna o modelo sensível à ruído nos dados, podendo aprender padrões específicos em vez de capturar tendências gerais.

Para superar as limitações anteriores, foi recorrido a outro algoritmo mais avançado, como o SVM (*Support Vector Machine*) que é capaz de lidar com fronteira de separação mais complexa e encontrar soluções mais precisas e robustas para problemas de Classificação. Para o conjunto de dados em questão, esta técnica apresentou o segundo melhor desempenho com 88% de acurácia, 81% de sensibilidade, 93% de especificidade e 89% de precisão.

Com os resultados, a abordagem proposta nesta pesquisa tem potencial para analisar o solo com diferentes culturas desenvolvidas por produtores em várias regiões do país, o que possibilitaria ações preventivas relacionadas a sua degradação e, consequentemente, na disponibilidade de alimentos. O diagnóstico da sustentabilidade do solo ajuda a melhorar a gestão, uma vez que apoia o produtor rural na tomada de decisão para que ele tenha ganhos socioeconômicos de forma sustentável e ajuda na conservação do meio ambiente com o manejo adequado do solo e fornecimento de alimentos mais seguros.

A relação entre a análise de variáveis do solo e o desenvolvimento econômico por meio do uso de dados pode ser visualizada através da melhor utilização dos insumos agrícolas e aumento da eficiência produtiva, uma vez que estes aspectos contribuem diretamente para o crescimento econômico, fornecimento de ferramentas e conhecimento aos agricultores para promover o desenvolvimento sustentável.

A abordagem desta pesquisa apresenta não apenas uma dinâmica sustentável, como também pode fornecer estrutura para a formulação de políticas públicas e para a captação de recursos pelos produtores agrícolas. Isto porque ela fornece uma base consistente para ajudar os produtores a acessarem recursos financeiros de programas nacionais de crédito, pois tais produtores podem embasar suas solicitações de crédito com informações mais precisas sobre suas atividades agrícolas, tornando o processo mais eficiente e transparente.

Outro aspecto relevante é o impacto positivo que essa abordagem pode ter no compromisso Net Zero. Ao contribuir para a adoção de práticas sustentáveis no manejo do solo agrícola, essa abordagem pode desempenhar um papel significativo

na redução das emissões e no avanço em direção às metas de neutralidade de carbono.

Assim, a pesquisa não só impulsiona o desenvolvimento econômico do setor agrícola, mas também oferece ferramentas valiosas para a gestão eficiente do solo, o acesso a recursos financeiros e a promoção da sustentabilidade ambiental. Com a aplicação prática desses resultados, espera-se uma contribuição importante para um futuro mais equilibrado e resiliente para a agricultura nacional.

Com o objetivo de tornar visível a sustentabilidade do solo e ter aplicabilidade prática, esta pesquisa construiu um protótipo de um software. O desenvolvimento do protótipo foi realizado numa plataforma No-Code (plataforma de desenvolvimento sem código que permite a criação de aplicativos utilizando interface gráfica e configurações). De forma intuitiva, tal protótipo apresenta a sustentabilidade do solo (Solo em Dia, 2020) partindo de uma análise laboratorial isolada de amostra de solo e, ao mesmo tempo, permite a coleta de dados para posterior análise conforme a abordagem proposta nesta pesquisa. Num estudo preliminar, acredita-se que a adoção tecnológica desta ferramenta, principalmente, pelos pequenos produtores poderia ser realizada via associação ou cooperativa, pois assim eles não só veriam como está a sustentabilidade do solo da sua propriedade, mas teriam condições de fazer os ajustes necessários seguindo as recomendações técnicas dos profissionais especializados pertencentes a estas instituições. Além disso, a distribuição da ferramenta desta maneira aumentaria a capilaridade na captação dos dados, permitindo uma visualização mais ampla nas áreas de atuação dessas instituições, o que facilitaria o planejamento de compra de insumos, por exemplo.

Em resumo, na análise das técnicas de classificação da qualidade do solo, foi comparada a Regressão Logística, largamente utilizada com outras 3(três) técnicas de Aprendizado de Máquina. A melhora no resultado foi significativa, aumentando de 70% de acurácia da Regressão Logística para 95% para *Random Forest*. Esta última, utilizando RFE como seletor de variável, mostrou robustez nos resultados, apresentando 95% de sensibilidade, 97% de especificidade e 96% de precisão, resultados estes que mostram a potencialidade da técnica para ser utilizada como classificador na abordagem proposta. Estes resultados também mostram que, por ora, não é necessária a busca por novas técnicas, sendo sugerido, como próximo

passo, a validação para novas culturas e/ou novas classes de solo.

Portanto, os resultados obtidos nesta pesquisa sugerem que a adoção da abordagem proposta pode representar um impulso significativo para o desenvolvimento econômico sustentável. Ao promover práticas agrícolas ecologicamente conscientes, o estudo não apenas destaca a importância da sustentabilidade do solo, mas também revela o potencial de integração entre esse conceito e técnicas avançadas de Aprendizado de Máquina para promover o desenvolvimento econômico local. Essa integração pode resultar em maior eficiência sem comprometer a qualidade ambiental, beneficiando, assim, não apenas a produtividade agrícola, mas também a qualidade de vida da sociedade em geral.

As técnicas de Aprendizado de Máquina oferecem a capacidade de analisar grandes volumes de dados ambientais e agronômicos, permitindo uma tomada de decisão mais precisa e eficiente. Por meio da análise de dados climáticos, padrões de uso do solo e a resposta das culturas à diferentes práticas agrícolas, algoritmos de Aprendizado de Máquina podem prever cenários futuros, identificar práticas sustentáveis e propor melhorias que equilibram crescimento econômico e conservação ambiental. Isso cria um ambiente favorável para maximizar a produção e minimizar os impactos ambientais, gerando resultados positivos para produtores, consumidores e para o meio ambiente.

Com isso, o uso de tecnologias baseadas em Aprendizado de Máquina pode facilitar a colaboração entre diferentes atores (de agricultores e especialistas ambientais até governos e empresas), permitindo a criação de soluções integradas e adaptadas às realidades locais. Isso não só promove o compartilhamento de conhecimento, como também aumenta a capacidade de inovação e a implementação de práticas agrícolas sustentáveis, impulsionando a eficiência e a responsabilidade ecológica.

Em última análise, essa abordagem metodológica contribui para uma agricultura mais resiliente que equilibra o desenvolvimento econômico e a proteção dos recursos naturais, beneficiando a sociedade atual e as futuras gerações.

CAPÍTULO V - CONSIDERAÇÕES FINAIS

Foi realizado um estudo bibliográfico para contextualizar a importância do solo e seu manejo adequado, com ênfase na sustentabilidade. O solo é um recurso vital, não renovável a curto prazo, e seu manejo responsável é fundamental para a manutenção da produtividade agrícola e a preservação ambiental. Além disso, o estudo explorou o uso do Aprendizado de Máquina (*Machine Learning*) na Agricultura e na Economia, evidenciando seu potencial como uma ferramenta de análise de sustentabilidade. A aplicação dessas técnicas permite identificar padrões em conjuntos de dados, oferecendo percepções para a gestão eficiente dos recursos naturais, monitoramento das condições do solo e otimização da produção agrícola.

A maior contribuição dessa pesquisa para a sustentabilidade do solo é capacidade de análise preditiva e otimização na tomada de decisão, possibilitando identificar padrões e tendências que seriam difíceis de detectar por métodos tradicionais e permitindo que produtores e especialistas tomem decisões mais assertivas sobre o manejo do solo.

Para o avanço deste estudo, optou-se pelas técnicas Supervisionadas de Aprendizado de Máquina, pois estas visam à elaboração de modelos para encontrar uma função que seja capaz de predizer rótulos desconhecidos. Com as análises geradas pela modelagem, foi possível identificar se o solo é sustentável perante aos critérios de qualidade do solo. Isto porque, modelos preditivos podem antecipar a degradação do solo com base em dados históricos e em tempo real. Isso possibilita a adoção de medidas preventivas antes que os problemas se agravem, como a rotação de culturas, cobertura do solo ou adubação adequada.

A perda de sustentabilidade do solo pode se traduzir em menor rendimento por área, maior custo de insumos (fertilizantes, corretivos) e maior necessidade de práticas de manejo mais complexas e caras, o que diminui a margem de lucro do produtor e eleva os riscos econômicos. Esses fatores também podem reduzir a competitividade do produtor no mercado, já que a perda de produtividade pode torná-lo menos eficiente em relação a outros agricultores que conseguem manter solos mais saudáveis.

A abordagem proposta nesta pesquisa demonstra um grande potencial para auxiliar na elaboração de políticas públicas eficazes e facilitar a captação de recursos pelos produtores. Com a integração de tecnologias avançadas e a análise dos dados, torna-se possível mapear as culturas que estão sendo desenvolvidas pelos produtores em diferentes regiões do país. Essa capacidade de monitoramento contínuo permite a tomada de ações proativas no que diz respeito à prevenção da degradação do solo, à mitigação de impactos ambientais e à garantia de uma oferta estável de alimentos. Além disso, os dados podem embasar estratégias de longo prazo, como a recuperação de áreas degradadas e a promoção de práticas agrícolas regenerativas.

A utilização dessas informações também serve como um suporte robusto para os produtores que buscam acessar recursos de programas nacionais de crédito, como o Pronaf (Programa Nacional de Fortalecimento da Agricultura Familiar). Ao dispor de dados concretos e análises detalhadas sobre o estado do solo e a sustentabilidade das suas práticas agrícolas, os produtores podem se apresentar com mais segurança e precisão nas demandas de crédito rural, aumentando suas chances de sucesso na obtenção de financiamentos. Ademais, essa abordagem contribui diretamente para o compromisso global de redução de emissões, como o objetivo Net Zero até 2050, ao promover práticas agrícolas que minimizam a emissão de carbono e maximizam a eficiência na utilização dos recursos naturais.

Como produto desta pesquisa, houve uma preocupação em operacionalizar a visualização da sustentabilidade do solo de forma acessível e intuitiva. Nesse sentido, foi desenvolvido um protótipo de aplicativo digital baseado em tecnologia No-Code, ou seja, sua criação foi feita utilizando uma plataforma de desenvolvimento sem código, permitindo a construção de interfaces gráficas amigáveis e a configuração simples. Esse protótipo oferece aos usuários a capacidade de visualizar a sustentabilidade do solo a partir de uma análise laboratorial isolada de solo e, ao mesmo tempo, consegue coletar dados para utilizar a abordagem proposta nesta pesquisa, facilitando a adoção de decisões mais fundamentadas e assertivas.

As contribuições metodológicas realizadas durante o desenvolvimento deste estudo proporcionaram um entendimento do cenário em estudo. Análises detalhadas

de atributos do solo não se limitam a aspectos técnicos ou científicos isolados; elas têm implicações diretas e amplas no desenvolvimento econômico, social e ambiental. Ao melhorar a produtividade agrícola por meio de práticas sustentáveis e o uso eficiente de tecnologias de análise de dados, criam-se condições para transformar economias rurais, aumentar a resiliência das cadeias produtivas e promover um crescimento econômico robusto e sustentável. Essa abordagem holística, que conecta sustentabilidade ambiental e desenvolvimento econômico, é essencial para enfrentar os desafios globais da atualidade, garantindo a segurança alimentar, a preservação dos recursos naturais e o fortalecimento das economias locais e regionais.

Para continuar este estudo, sugere-se a integração de variáveis climáticas, agrícolas e socioeconômicas para projetar cenários futuros da agricultura, ajudando os produtores a fazerem escolhas sustentáveis sobre o uso e manejo do solo. Isso pode incluir a seleção de culturas mais adequadas para o tipo de solo, práticas de conservação específicas ou até, mudanças na distribuição de áreas de cultivo para maximizar a produtividade sem comprometer a sustentabilidade a longo prazo.

REFERÊNCIAS

ADAMS, M. A.; ATTWILL, P. M. Nutrient cycling and nitrogen mineralizacion in eucalyptus forests in southeastern Australia: I. Nutrient cycling and nitrogen turnover. **Plant and soil,** 92:319-339. 1986.

ALMEIDA, G. M. **Aprendizagem de máquina na determinação de ambientes de produção de cana-de-açúcar**. 2019. Disponível em: https://repositorio.unesp.br/server/api/core/bitstreams/6798b6cd-021d-4022-ba0b-1a050761fbcb/content. Acesso em 28 de Abr 2024.

ALVARENGA, D, 1999 Características físicas e químicas de um latossolo vermelho-escuro e a sustentabilidade de agroecossistemas Manejo e Conservação do Solo e da Água. **Revista Brasileira de Ciência do Solo,** 23 (4), Dez 1999.

ANIÉTOT, N. **Les Fertilizants Organiques A.D.A.S.** - Sciences et Techniques de L'An 2000. Paris-France: Le courier du Livre, 1983. 124 p.

ATHEY, S.; IMBENS, G. **Machine Learning Methods Economists Should Know About**. 2019 Disponível em https://www.gsb.stanford.edu/gsb-box/route-download/476281 Acesso em 24 de Set 2024.

BANA E COSTA, C. A. As três convicções fundamentais na prática do apoio à decisão. **Revista Pesquisa Operacional**, v. 13, n. 1, 1993.

BANA E COSTA, C.A.; ANGULO-MEZA, L.; OLIVEIRA, M. D. O Método MACBETH e aplicação no Brasil. **ENGEVISTA**, V. 15, n. 1. p. 3-27, 2013

BANA e COSTA, C.A.; BEINAT, E. Estruturação de Modelos de Análise Multicritério de Problemas de Decisão Pública. In: S. Costa, P. Nijkamp, T.P. Dentinho (eds.), **Compêndio de Economia Regional**. Volume II: Métodos e Técnicas de Análise Regional, Capítulo 20 (611- 645), 2011.

BANA E COSTA C.A.; DE CORTE J. M.; VANSNICK, J.C. **International Journal of Information Technology & Decision Making** Vol. 11, No. 2, 2012.

_____. **Guia do Utilizador 2017.** Disponível em: https://m-macbeth.com/wp-content/uploads/2017/10/M-MACBETH-Guia-do-utilizador_BETA.pdf Acesso em: 27 de Nov 2024.

_____. On the mathematical foundations of MACBETH. In: Figueira J, Greco S, Ehrgott M, editors. **Multiple criteria decision analysis: state of the art surveys**. New York: Springer. p. 409– 442, 2005.

BANA e COSTA, C.A.; SILVA, F. N. Concepção de uma 'boa' alternativa de ligação ferroviária ao porto de Lisboa: uma aplicação da metodologia multicritério de apoio à decisão e à negociação. **Investigação Operacional** v. 14 p. 115-131, 1994.

BANA E COSTA, C.A.; VANSNICK, J.C. A theoretical framework for Measuring Attractiveness by a Categorical Based Evaluation Technique (MACBETH), In J. Clímaco (ed.) **Multicriteria Analysis**, Springer-Verlag, 1997, p. 15-24.

BEER, J. Litter production and nutrient cycling in coffee (*Coffea arabica*) or cacao (*Theobroma cacao*) plantations with shade trees. **Agrofor Systems,** 7:103-114, 1988.

BELTON, V.; STEWART, T. J. **Multiple Criteria Decision Analysis: An Integrated Approach.** Kluwer Academic Publishers, Boston/Dordre, 2002.

BERTOL, I. *et al.* Propriedades físicas do solo sob preparo convencional e semeadura direta em rotação e sucessão de culturas, comparadas às do campo nativo. **Revista Brasileira de Ciência do Solo**, v. 28, n. 1, p. 155-163, 2004.

BERTONI, J.; LOMBARDI NETO, F. **Conservação do solo**. Piracicaba: Livroceres, 1990. 392 p.

BLUM, A.; HOPCROFT, J.; KANNAN, R. **Foundations of Data Science.** 2017 Disponível em: https://www.microsoft.com/en-us/research/wp-content/uploads/2017/11/book-June-14-2017pdf.pdf Acesso em 27 de Ago 2023.

BRAZ, D. C. **Aprendizado de máquina aplicado em dados de biossensores para diagnóstico de câncer e COVID-19**. 2022. Tese (Doutorado em Física Computacional) - Instituto de Física de São Carlos, Universidade de São Paulo, São Carlos, 2022. doi:10.11606/T.76.2022.tde-16112022-161304. Acesso em: 02 Ago 2023.

BREIMAN, L. (a) Random forests. **Machine Learning**, Boston, v.45, n.1, p.5-32, 2001.

BREIMAN, L. (b) Statistical Modeling: The Two Cultures (with comments and a rejoinder by the author). **Statist. Sci.** 16 (3) 199 - 231, August 2001. https://doi.org/10.1214/ss/1009213726

BRITO, P. C. O. **Plano de desenvolvimento de uma ferramenta de aprendizado de máquina para previsão e análises de dados em estudos de casos de obesidade.** 2020. Dissertação em Ciência e Tecnologia em Saúde da Universidade Estadual da Paraíba. Disponível em: http://tede.bc.uepb.edu.br/jspui/bitstream/tede/4151/2/PDF%20-%20Paulo%20C%c3%a9sar%20Oliveira%20Brito.pdf Acesso em 02 de Ago 2023.

BRUNINI, R. G.; SILVA, M. C. da; PISSARRA, T. C. T. Efeito do sistema de produção de cana-de-açúcar na qualidade da água em bacias hidrográficas. **Agrarian**, *[S. l.]*, v. 10, n. 36, p. 170–180, 2017. DOI: 10.30612/agrarian.v10i36.4309. Disponível em: https://ojs.ufgd.edu.br/index.php/agrarian/article/view/4309. Acesso em: 23 Jul. 2023.

BURGES, C. J. C. A tutorial on support vector machine for pattern recognition. **Data mining and knowledge discovery**, v. 2, n. 2, p. 955-974, 1998. Disponível em: https://www.cs.princeton.edu/courses/archive/fall12/cos402/readings/burgesSVM.pdf Acesso em 20 de Fev 2023.

BUSSAB, W. DE O.; MORETTIN, P. A. **Estatística Básica** 6ª ed., Saraiva, 2010.

CLEVELAND, C. J. *et al.* Energy and the US economy: a biophysical perspective. **International library of critical writings in economics**, v. 75, p. 295-302, 1997.

COLLA, E. C. **Aplicação de modelos gráficos probabilísticos computacionais em economia.** 2009. Tese de Doutorado. Disponível em: https://bibliotecadigital.fgv.br/dspace/handle/10438/4261 Acesso em: 24 de Ago 2023.

COMA-PUIG, B. **Human-aware application of data science techniques.** 2022 Disponível em: https://upcommons.upc.edu/handle/2117/365522. Acesso em: 12 de Nov 2022.

CORRÊA, S. T. R. *et al.* Aplicações e limitações da modelagem em agricultura: revisão. **Revista de Agricultura**, v. 86, n. 1, p. 1-13, 2011.

COSTANZA, R. *et al.* The value of the world's ecosystem services and natural capital. **Nature**, v. 387, n. 6630, p. 253-260, 1997.

DAFERT, F.W. **Sobre estrumes nacionaes - Relatório de 1893.** São Paulo, Instituto Agronômico do Estado de São Paulo, 1895. p.154-166. (Collecção de Trabalhos Agrícolas, Relatórios Annuaes de 1888-1893).

DÉCADA DAS NAÇÕES UNIDAS SOBRE RESTAURAÇÃO DE ECOSSISTEMAS (2019) Disponível em https://www.decadeonrestoration.org/pt-br/sobre-decada-da-onu Acesso em 18 de Mar 2023.

DE-POLLI, H.; PIMENTEL, M. S. Indicadores de qualidade do solo. In: AQUINO, A. M. de; ASSIS, R. L. de (Ed.). **Processos biológicos no sistema solo-planta: ferramentas para uma agricultura sustentável.** Brasília, DF: Embrapa Informação Tecnológica; Seropédica: Embrapa Agrobiologia, 2005. cap. 1. p. 17-28.

DOMINGOS, P. **The Master Algorithm: How the Quest for the Ultimate Learning Machine Will Remake Our World.** Basic Books, New York, 2015.

DORAN, J. W. *et al.* Determinants of soil quality and health. In: **Soil quality and soil erosion**. CRC Press, 2018. p. 17-36.

ENSSLIN, L.; MONTIBELLER, G. N. **Quais Critérios Deve-se Considerar em uma Avaliação?** Anais do XVIII ENEGEP, Niterói/RJ, 21-25 de Setembro, 1998.

ENSSLIN, L. *et al.* **Apoio à Decisão**: **Metodologias para Estruturação de Problemas e Avaliação Multicritério de Alternativas.** Insular, Florianópolis, 2001.

FAWCETT, T. An introduction to ROC analysis. **Pattern Recognition Letters** Volume 27, Issue 8, June 2006, Pages 861-874. Disponível em https://www.sciencedirect.com/science/article/abs/pii/S016786550500303X. Acesso em 23 de Jul 2023.

FAYYAD, U. M. *et al.* Knowledge Discovery and Data Mining: Towards a Unifying Framework. In: **KDD**. 1996. p. 82-88.

FIGUEIRA, C. V. **Modelos de regressão logística.** 2006. Disponível em: https://www.lume.ufrgs.br/handle/10183/8192. Acesso em 13 de Ago 2022.

FREEMAN, C. Technical innovation, diffusion, and long cycles of economic development. In: **The Long-Wave Debate: Selected Papers from an IIASA (International Institute for Applied Systems Analysis) International Meeting on Long-Term Fluctuations in Economic Growth: Their Causes and Consequences, Held in Weimar, GDR, June 10–14, 1985**. Berlin, Heidelberg: Springer Berlin Heidelberg, 1987. p. 295-309.

FREITAS, N. **Machine Learning: 2014-2015.** University of Oxford - Department of Computer Science, 2015 Disponível em https://www.cs.ox.ac.uk/people/nando.defreitas/machinelearning/. Acesso em 27 de Abr 2023.

GANDHI, N.; ARMSTRONG, L. J. A review of the application of data mining techniques for decision making in agriculture. In: **2016 2nd International Conference on Contemporary Computing and Informatics (IC3I).** IEEE, 2016. p. 1-6.

GARFINKEL, S.; GRUNSPAN, R. H. **The Computer Book : from the Abacus to Artificial Intelligence, 250 Milestones in the History of Computer Science**. New York, NY: Sterling, 2018.

GHOLAP, J. *et al.* **Soil data analysis using classification techniques and soil attribute prediction**. arXiv preprint arXiv:1206.1557, 2012. Disponível em: https://arxiv.org/abs/1206.1557 Acesso em 13 de Ago 2023.

GÓMEZ, C. A. R.. Aplicación del machine learning en agricultura de precisión. **Revista Cintex**, v. 25, n. 2, p. 14-27, 2020. Disponível em: https://revistas.pascualbravo.edu.co/index.php/cintex/article/view/356 Acesso em: 22 de Ago 2023.

GOODFELLOW, I.; BENGIO, Y.; COURVILLE, A.. **Deep learning**. MIT press, 2016.

GREGORY, R.; LICHTENSTEIN, S.; SLOVIC, P. Valuing environmental resources: A constructive approach. **Journal of Risk and Uncertainty**, 7, 177–197, 1993.

GRUPO KLEFFMANN. Disponível em https://www.kynetec.com/news Acesso em 22 de Jul 2023.

HAN, J.; PEI, J.; TONG, H. **Data mining: concepts and techniques.** Morgan kaufmann, 2022.

HAAN, C.T.; BARFIELD, B.J.; HAYES, J.C. 1994. **Design Hydrology and Sedimentology for Small Catchments**. Academic Press, San Diego,588p.

HANLEY, N.; BARBIER, E. B.; BARBIER, E. **Pricing nature: cost-benefit analysis and environmental policy**. Edward Elgar Publishing, 2009.

HASTIE, T., TIBSHIRANI, R., FRIEDMAN, J. **The elements of statistical learning: data mining, inference, and prediction**. New York: springer, 2009.

IAC - Boletim Técnico IAC, 216, 2016. Disponível em: https://www.iac.sp.gov.br/publicacoes/arquivos/iacbt126.pdf. Acesso em 20 de Jan 2023.

IPEA, 2023. Disponível em: https://www.ipea.gov.br/ods Acesso em: 15 de Jul 2023.

ISPA. International Society of Precision Agriculture. Available, 2021 Disponível em: https://www.ispag.org/. Acesso em: 11 de Jul 2023.

JAMES, G. *et al.* **An introduction to statistical learning**. New York: springer, 2013.

KEENEY, R. L. **Value-Focused thinking: A path to creative decision making.** Cambridge, MA: Harvard University Press, 1992.

KEENEY, R. L.; RAIFFA, H. **Decisions with multiple objectives: preferences and value trade-offs.** New York: John Wiley, 1976.

KIM, V. N. P.; CONTI, D. M.; GONZALEZ, E. S. A inter-relação entre Indústria 4.0 e Economia Circular: Revisão Sistemática da Literatura. **Revista Gestão & Tecnologia**, v. 22, n. 4, 2022.

KOHAVI, R.; SOMMERFIELD, D. **Feature subset selection using the wrapper model**: Overfitting and dynamic search space topology. Proceedings of the First International Conference on Knowledge Discovery and Data Mining (KDD-95), 1995.

LAL, R.; STEWART, B.A. **Need for land restoration**. Adv. Soil Sci., 17:1-11, 1992.

LARSON, E.J. **The Myth of Artificial Intelligence**, 2021 Disponível em: https://www.degruyter.com/document/doi/10.4159/9780674259935/html Acesso em 04 de Set 2024.

LEPSCH, I. F. **19 lições de pedologia**. Oficina de textos, 2021.

LORENA, A. C.; CARVALHO, A. C. P. L. F. de. Uma introdução às support vector machines. **Revista de Informática Teórica e Aplicada**, v. 14, n. 2, p. 43-67, 2007.

LOURENÇO, J.C. Modelo Aditivo hierárquico: exemplo de métodos de ponderação e problemas associados. **Artigo de investigação no 13/2002**, CEG-IST, Lisboa, 2002.

MARTINS, I.C.M. *et al.* Diagnóstico ambiental no contexto da paisagem de fragmentos florestais naturais" ipucas" no município de Lagoa da Confusão, Tocantins. **Revista Árvore**, v. 26, n. 3, p. 299-309, 2002.

MCBRATNEY, A., GRUIJTER, J., BRYCE, A., 2019. Pedometrics timeline. **Geoderma** 338, 568–575. Disponível em: https://doi.org/10.1016/j.geoderma.2018.11.048 Acesso em 22 de Mar 2024.

MCBRATNEY, A.B., MENDONÇA SANTOS, M.L., MINASNY, B., 2003. On digital soil mapping, **Geoderma**. Disponível em: https://doi.org/10.1016/S0016-7061(03)00223-4. Acesso em 22 de Mar 2024.

MCCULLAGH, P. **Generalized linear models**. Routledge, 2019.

MENARD, S. **Applied logistic regression analysis**. Sage, 2002.

MENDES, S. M. *et al*. Custo da produção de *Orius insidiosus* como agente de controle biológico. **Pesquisa Agropecuária Brasileira**, v. 40, p. 441-446, 2005.

MUELLER, J. P., MASSARON, L. **Machine Learning for Dummies**; 2nd ed. edição. John Wiley & Sons, 2021.

MULLAINATHAN, S.; SPIESS, J.. Machine Learning: An Applied Econometric Approach. **Journal of Econometric Perspectives**, v. 31, n. 2, p. 87-106, 2017.

MÜLLER, A. C.; GUIDO, S. **Introduction to machine learning with Python: a guide for data scientists.** O'Reilly Media, Inc, 2016.

NOGUEIRA, C. C. **Previsibilidade no mercado acionário utilizando machine learning**. 2019. Tese de Doutorado. Disponível em: https://bibliotecadigital.fgv.br/dspace/handle/10438/27999 Acesso em: 24 de Ago de 2023.

NOVAIS, R.F.; ALVAREZ V., V.H.; BARROS, N.F.; FONTES, R.L.F.; CANTARUTTI, R.B. & NEVES, J.C.L. **Fertilidade do Solo**. Viçosa, MG, Sociedade Brasileira de Ciência do Solo, 2007.

OECD/FAO, **OECD-FAO Agricultural Outlook 2023-2032**, OECD Publishing, Paris, 2023. Disponível em: https://doi.org/10.1787/08801ab7-en. Acesso em 23 de Jul 2023.

OIKAWA, R. T. **Aprendizagem de máquina e análise de componentes principais na avaliação de qualidade do solo com pastagens**. 2023. Disponível em http://bdtd.unoeste.br:8080/jspui/handle/jspui/1582. Acesso em 30 de Abr 2024

ONU (2021). Disponível em https://news.un.org/pt/story/2021/06/1753932 Acesso em 17 de Jun 2023.

ONU (2023). Disponível em: https://brasil.un.org/pt-br/sdgs Acesso em 27 de Ago 2023.

PATEL, H.; PATEL, D. A brief survey of data mining techniques applied to agricultural data. **International Journal of Computer Applications**, v. 95, n. 9, 2014.

PÁSCOA, M. I. F. **Os desafios da Machine Learning: Aplicação ao Mercado Financeiro**. 2018. Dissertação de Mestrado. Disponível em: https://estudogeral.uc.pt/handle/10316/84617 Acesso em: 24 de Ago 2023.

PÉREZ-JARAMILLO, J. E.; MENDES, R.; RAAIJMAKERS, J. M. Impact of plant domestication on rhizosphere microbiome assembly and functions. **Plant molecular biology,** v. 90, p. 635-644, 2016.

PORTER, M. E. New global strategies for competitive advantage. **Planning Review**, v. 18, n. 3, p. 4-14, 1990.

PRADO, M. M. L. **Machine learning for asset managers**. Cambridge University Press, 2020.

PRIMAVESI, A. **Manejo ecológico do solo**: a agricultura em regiões tropicais. São Paulo: Nobel, 1979. 579 p.

RAIFFA, H. Decision Analysis: A Personal Account of How It Got Started and Evolved. In: EDWARDS, W.; MILES, J.R.F., WINTERFELDT, D.V. (eds.). **Advances in Decision Analysis**. New York, NY: Cambridge University Press; p. 57–70, 2007.

RAIJ, B. V. A capacidade de troca de cátions das frações orgânica e mineral em solos. **Bragantia,** 28(8):85-112. 1969.

RAMAROSON, V.H., BECQUER, T., SÁ, S.O., RAZAFIMAHATRATRA, H., DELARIVIÈRE, J.L., BLAVET, D., VENDRAME, P.R.S., RABEHARISOA, L., RAKOTONDRAZAFY, A.F.M., 2018. Mineralogical analysis of ferralitic soils in Madagascar using NIR spectroscopy. **Catena** (Amst) 168, 102–109. Disponível em: https://doi.org/10.1016/j.catena.2017.07.016. Acesso em 22 de Mar 2024.

RONQUIM, C. C. **Conceitos de fertilidade do solo e manejo adequado para as regiões tropicais** - 2.ed. - Campinas: Embrapa Territorial, 2020. Disponível em: https://www.infoteca.cnptia.embrapa.br/infoteca/bitstream/doc/1128267/1/5840.pdf Acesso em 27 de Nov 2024.

ROY, B. **Multicriteria Methodology for Decision Aiding**. Dordrecth: Kluwer Academic Publishers. 1986,167p.

SACHS, I. **Caminhos para o Desenvolvimento Sustentável**. Rio de Janeiro: Garamond, 2002.

SAMUEL, A. L. (1959). Some studies in machine learning using the game of checkers. **IBM Journal of research and development**, 3(3), 210-229.

SANTOS, H. G. *et al.* **Sistema Brasileiro de Classificação de Solos**, 5. ed., rev. e ampl. − Brasília, DF : Embrapa, 2018. Disponível em: https://www.agroapi.cnptia.embrapa.br/portal/assets/docs/SiBCS-2018-ISBN-9788570358004.pdf. Acesso em 20 de Jan 2023.

SANTOS, R. P.; BEKO, M.; LEITHARDT, V. R. Q. Modelo de machine learning em tempo real para agricultura de precisão. In: **Anais da XXII Escola Regional de Alto Desempenho da Região Sul. SBC,** 2022. p. 69-70. Disponível em: https://sol.sbc.org.br/index.php/eradrs/article/view/19166 Acesso em 24 de Ago de 2023.

SCHEIDEGGER, S.; BILIONIS, I. Machine learning for high-dimensional dynamic stochastic economies. **Journal of Computational Science**, v. 33, p. 68-82, 2019. Disponível em: https://www.sciencedirect.com/science/article/abs/pii/S1877750318306161 Acesso em: 27 de Ago 2023.

SCHÖLKOPF, B.; SMOLA, A. J. **Learning with kernels: support vector machines, regularization, optimization, and beyond,** 2002.

SCHUMPETER, J. A. **Capitalism, socialism and democracy**. routledge, 2013.

SEEG. **Análise das emissões de gases de efeito estufa e suas implicações para as metas climáticas do Brasil,** 2024 Disponível em: https://seeg.eco.br/wp-content/uploads/2024/11/SEEG-RELATORIO-ANALITICO-12.pdf Acesso em 27 de Nov 2024.

SNA. 2018 Disponível em https://www.sna.agr.br/o-papel-da-tecnologia-na-evolucao-da-agricultura/ Acesso em 01 de Abr 2023

SOLO EM DIA (2020). Disponível em https://estatera.glideapp.io/. Acesso em 02 de Fev 2024.

STONE, M. T.; NYAUPANE, G. P. Protected areas, wildlife-based community tourism and community livelihoods dynamics: Spiraling up and down of community capitals. **Journal of Sustainable Tourism**, v. 26, n. 2, p. 307-324, 2018.

STOTT, D. E.; MOEBIUS-CLUNE, B. N. Soil health: Challenges and opportunities. **Global soil security**, p. 109-121, 2017.

SUBHASHINI, S.; BEGAM, Y. S.; UMAMAHESWARI, P. A Study on Machine Learning Algorithms and its Applications. **International Journal Of Scientific Research In Engineering And Management**. Disponível em https://ijsrem.com/download/a-study-on-machine-learning-algorithms-and-its-applications/. Acesso em 17 de Out 2024.

SWIFT, M. J.; WOOMER, P. Organic matter and the sustainability of agricultural systems: Definition and measurement. In: Mulongoy, K.; Merckx, R. (Ed.). **Soil Organic Matter Dynamics and Sustainability of Tropical Agriculture.** IITA/K.U.Leuven, 1993. p 3-18.

TAN, P.; STEINBACH, M.; KUMAR, V. **Introdução ao Data Mining.** Rio de Janeiro: Ciência Moderna, 2009.

THEOBALT, O. **Machine Learning for absolute beginners: a plain english introduction**. Scatterplot Press, 2021

THOMAS, G. **Mathematics for Machine Learning**. University of California, Berkeley, 2018. Disponível em: https://gwthomas.github.io/docs/math4ml.pdf Acesso em 27 de Ago 2023.

TISDALE, S.L., NELSON, W.L. AND BEATON, J.D. **Soil Fertility and Fertilizer**. 4th Edition, Macmillan Publishing Company, New York, 1990.

TORMENA, C. A.; ROLOFF, G.; SÁ, J. C. M. Propriedades físicas do solo sob plantio direto influenciadas por calagem, preparo inicial e tráfego. **Revista Brasileira de Ciência do Solo**, v. 22, p. 301-309, 1998.

TRIVELLATO, G. M. L.; SARRIÉS, G. A.; FURLAN, G. N. Sistema de Avaliação Ponderada da Multifuncionalidade da Agricultura: Machine Learning, Índices de Sustentabilidade Ambiental e Serviços Ecossistêmicos. **University Rankings for Brazilian Universities.** Webinar 2020. Disponível em: https://www.depi.unicamp.br/wp-content/uploads/2020/10/ANAIS_NWGM.pdf#page=16 Acesso em 27 de Ago 2023.

TURING, A. M. I. COMPUTING MACHINERY AND INTELLIGENCE, **Mind**, Volume LIX, Issue 236, October 1950, Pages 433–460, Disponível em https://doi.org/10.1093/mind/LIX.236.433 Acesso em 17 de Dez 2023.

UNCCD, 2022. Summary for Decision Makers. Global Land Outlook, second edition. United Nations Convention to Combat Desertification, Bonn. Disponível em: https://www.unccd.int/sites/default/files/2022-04/GLO2_SDM_low-res_0.pdf Acesso em 23 de Set 2023.

UNFCCC. United Nations Framework Convention on Climate Change Convenção. 21º Conference of the Parties. Acordo de Paris, 2015. Disponível em: https://nacoesunidas.org/cop21/. Acesso em: 27 de Nov 2024.

USDA-NRCS **A Matéria Orgânica no Solo (MOS)**, 2017. Disponível em: https://www.nrcs.usda.gov/sites/default/files/2022-09/NotaTécnicaDoSolo12A.pdf Acesso em 17 de Jul 2023.

VARIAN, H. R. Big Data: New Tricks for Econometrics. **Journal of Economic Perspectives***,* v. 28, n. 2, p. 3–28, Spring 2014.

VASCONCELOS, B. F. B. de. **Poder preditivo de métodos de Machine Learning com processos de seleção de variáveis: uma aplicação às projeções de produto de países**. 2017.

VERNETTI JUNIOR, F. J.; GOMES, A. S.; SCHUCH, L. O. B. Sustentabilidade de sistemas de rotação e sucessão de culturas em solos de várzea no Sul do Brasil. **Ciência Rural**, v. 39, n. 6, p. 1708-1714, 2009.

VEZZANI, F. M.; MIELNICZUK, J.. Uma visão sobre qualidade do solo. **Revista Brasileira de Ciência do Solo**, v. 33, n. 4, p. 743-755, 2009.

VON WINTERFELDT, D., e EDWARDS, W. **Decision analysis and behavioral research**. Cambridge, UK: Cambridge University Press, 1986.

WASSERMAN, L. **All of Statistics. A Concise Course in Statistical Inference New York**: Springer, 2004.

WOOLDRIDGE, J. M. **Introductory Econometrics: A Modern Approach**. [S.l.]: Cengage South-Western, 2012.