

Avaliação de Modelos de Aprendizagem de Máquina para a Classificação da Adequação do Solo para o Cultivo do Milho

Maiquel Roberto Seidel^{a,b}, Rafael Roani Feijo^{a,b} and Taciano Ares Rodolfo^{a,b}

^aUniversidade Federal do Rio Grande do Sul, Porto Alegre, RS, Brasil

^bDisciplina Aprendizagem de Máquina, Prof. Dra Mariana Recamonde Mendoza

ARTICLE INFO

Keywords:

Adequação do solo

Classificação

Milho

Aprendizado de Máquina

Fertilidade do solo

Resumo

Este trabalho desenvolve e avalia modelos de classificação para determinar a adequação do solo ao cultivo de milho, com base em análises laboratoriais de 16 atributos físico-químicos do solo. Os modelos atribuem a cada amostra de solo uma das três classes de adequação: baixa, média ou alta, de acordo com critérios agrônômicos estabelecidos. O processo metodológico envolveu a limpeza e pré-processamento dos dados, seleção de atributos, experimentação com múltiplos algoritmos (Regressão Logística, k-NN, Árvore de Decisão, Random Forest, SVM e XGBoost), validação cruzada estratificada e otimização de hiperparâmetros visando maximizar o F1-score macro. Os resultados destacam os atributos mais influentes na classificação da qualidade do solo e fornecem recomendações práticas para melhorar a produtividade agrícola do milho.

1. Introdução


A avaliação da adequação do solo desempenha um papel central no planejamento agrícola, influenciando diretamente a produtividade, a sustentabilidade e a qualidade das culturas produzidas. Entre essas culturas, o milho destaca-se por sua expressiva importância econômica e estratégica, sendo amplamente utilizado na alimentação humana e animal, bem como como insumo essencial em diversos segmentos industriais. A seleção de solos apropriados para seu cultivo é fundamental para o uso racional dos recursos agrícolas, evitando o desperdício de insumos como fertilizantes e água, maximizando o rendimento das colheitas e minimizando impactos ambientais negativos.

Diante dos desafios contemporâneos relacionados à segurança alimentar, à escassez de recursos naturais e à necessidade de práticas agrícolas mais sustentáveis, torna-se imperativo aprofundar o conhecimento sobre os atributos do solo e sua influência no desempenho das culturas. Essa compreensão é particularmente relevante para o milho, dada sua sensibilidade às condições do solo e do clima e seu papel estratégico no cenário agroindustrial.

Nesse contexto, este trabalho tem como objetivo avaliar modelos de classificação capazes de determinar a melhor adequação de solo para o cultivo do milho, categorizando amostras em diferentes níveis de aptidão. Para isso, serão utilizados dados reais e sintéticos de propriedades físico-químicas do solo e técnicas de aprendizagem de máquina, buscando oferecer uma ferramenta eficiente para subsidiar decisões agrônômicas mais precisas e sustentáveis. Espera-se, com isso, contribuir para o aprimoramento do manejo agrícola, a redução de riscos produtivos e a valorização de práticas que promovam a resiliência e a eficiência no uso do solo.

2. Importância Relativa dos Atributos do Solo

A seleção dos atributos mais relevantes para avaliar a adequação do solo ao cultivo do milho baseia-se em evidências agrônômicas consolidadas. De acordo com estudos de referência Büll and Cantarella (1993) e Novais, V., Barros, Fontes, Cantarutti and Neves (2007), diferentes propriedades físico-químicas do solo exercem influência direta sobre o crescimento, desenvolvimento e produtividade do milho. Nesta seção, os atributos foram organizados em ordem decrescente de importância, considerando sua contribuição para a nutrição vegetal, processos fisiológicos e resposta produtiva da cultura.

 maiquel.seidel@hotmail.com (M.R. Seidel); rafael_feijo95@hotmail.com (R.R. Feijo); tacrodolfo@gmail.com (T.A. Rodolfo)

ORCID(s):

2.1. pH

O pH do solo regula a disponibilidade de macro e micronutrientes e interfere diretamente na atividade microbiana e nos processos de mineralização. Para o milho, o intervalo ideal situa-se entre 5,5 e 6,5. Valores fora dessa faixa reduzem a absorção de fósforo, zinco e manganês, comprometendo o desenvolvimento da cultura.

2.2. N-NO₃ (ppm)

O nitrogênio é o nutriente mais exigido pelo milho, sendo fundamental para o crescimento vegetativo, a formação da espiga e o enchimento dos grãos. O nitrato (NO₃⁻) é a forma mais absorvida pelas plantas, e sua disponibilidade adequada é essencial para altas produtividades.

2.3. P (ppm)

O fósforo estimula o crescimento radicular e a formação de estruturas reprodutivas, sendo especialmente importante nas fases iniciais do desenvolvimento do milho. Também está relacionado à eficiência na absorção de outros nutrientes.

2.4. K (ppm)

O potássio regula processos osmóticos, abertura e fechamento dos estômatos, e o transporte de fotoassimilados para os grãos. Sua deficiência pode resultar em má formação das espigas e baixa resistência a estresses hídricos e doenças.

2.5. Matéria Orgânica (%)

A matéria orgânica contribui para a melhoria da estrutura do solo, aumenta a capacidade de retenção de água e nutrientes, estimula a microbiota benéfica e promove liberação gradual de nitrogênio, fósforo e enxofre.

2.6. Mg (ppm)

O magnésio é componente central da molécula de clorofila e participa de importantes processos enzimáticos. Sua deficiência compromete a fotossíntese e a produtividade da planta.

2.7. CaCO₃ (%)

O carbonato de cálcio influencia o pH do solo e, por consequência, a disponibilidade de diversos nutrientes. Também afeta a estrutura física do solo, promovendo melhor aeração e desenvolvimento radicular.

2.8. CEC Aparente

A capacidade de troca de cátions (CEC) reflete a habilidade do solo em reter e disponibilizar nutrientes como Ca²⁺, Mg²⁺ e K⁺, essenciais ao crescimento do milho e ao enchimento dos grãos.

2.9. Textura do solo (Areia, Argila e Silte – %)

A composição granulométrica afeta a retenção de água, drenagem e disponibilidade de nutrientes. Solos com maior proporção de argila tendem a ter maior fertilidade natural, embora demandem atenção à compactação.

2.10. CE (mS/cm)

A condutividade elétrica indica a concentração de sais solúveis no solo. Valores elevados sugerem salinidade, que pode afetar a absorção de água e gerar estresse osmótico nas plantas, impactando negativamente o rendimento.

2.11. Micronutrientes (Fe, Zn, Mn, Cu, B – ppm)

Os micronutrientes desempenham funções metabólicas específicas e essenciais para o pleno desenvolvimento do milho. O ferro (Fe) participa da síntese de clorofila e de processos respiratórios fundamentais para a geração de energia na planta. O zinco (Zn) é crucial para a síntese de auxinas, hormônios responsáveis pelo crescimento, além de desempenhar papel importante no desenvolvimento embrionário dos grãos. O manganês (Mn) atua na fotólise da água durante a fotossíntese e em diversas reações enzimáticas que regulam o metabolismo vegetal. O cobre (Cu), por sua vez, está envolvido em sistemas antioxidantes, contribuindo para a defesa contra patógenos e o equilíbrio redox celular. Já o boro (B) é essencial para o desenvolvimento dos tecidos meristemáticos, que dão origem a novas estruturas vegetativas e reprodutivas, além de influenciar o transporte de açúcares dentro da planta. A deficiência de qualquer um desses elementos pode comprometer significativamente o desempenho fisiológico e o potencial produtivo da cultura, Google (2025).

3. Faixas Ideais dos Atributos do Solo para o Cultivo do Milho

O desempenho agrônômico do milho depende diretamente das condições físico-químicas do solo, que influenciam desde o estabelecimento inicial da cultura até o enchimento dos grãos. A definição de faixas ideais para cada atributo do solo é essencial para a interpretação dos dados e o desenvolvimento de modelos preditivos de adequação. Tais faixas refletem o intervalo em que os nutrientes e propriedades do solo favorecem o crescimento e a produtividade da planta, evitando tanto deficiências quanto excessos que possam comprometer o rendimento.

Para solos cultivados com milho, o pH ideal situa-se entre 5,5 e 6,5, valor que favorece a disponibilidade de nutrientes e a atividade microbiana benéfica, especialmente em solos ácidos que necessitam de correção com calcário (Infoteca Embrapa). A textura ideal corresponde a solos de composição média a franco-argilosa, com 30 a 50% de areia, 20 a 35% de argila e 20 a 40% de silte, o que assegura boa drenagem e adequada retenção de água (Embrapa Ainfo). A condutividade elétrica (CE) deve estar preferencialmente abaixo de 1 mS/cm, indicando baixa salinidade, condição favorável para a absorção de água e nutrientes pelas raízes.

A matéria orgânica (MO) deve ser superior a 2,0%, pois melhora a estrutura do solo, aumenta a CTC e fornece nutrientes de forma gradual (Infoteca Embrapa). Em relação ao carbonato de cálcio (CaCO_3), a faixa recomendada está entre 0 e 5%, evitando-se solos com excesso de carbonatos, que podem afetar negativamente a disponibilidade de fósforo e micronutrientes. O nitrogênio disponível na forma de nitrato (N-NO_3) deve ser superior a 20 ppm, valor comumente associado à meta de produtividade de 5 t/ha de grãos (Infoteca Embrapa). Os teores de fósforo (P) e potássio (K) também devem ser mantidos em níveis adequados, com valores mínimos de 12 ppm e 120 ppm, respectivamente, conforme os extratores Mehlich-1 (Embrapa Ainfo).

O magnésio (Mg) deve situar-se na faixa de 50 a 150 ppm, sendo desejável uma relação Ca:Mg entre 3:1 e 5:1, de modo a manter o equilíbrio iônico do solo. Os micronutrientes também possuem faixas ideais bem estabelecidas: ferro (Fe) entre 4 e 8 ppm, zinco (Zn) entre 1 e 2 ppm, manganês (Mn) de 5 a 20 ppm, cobre (Cu) de 0,5 a 2 ppm e boro (B) entre 0,5 e 1,5 ppm (Embrapa Ainfo). Manter os níveis desses elementos dentro das faixas de suficiência é crucial para garantir processos fisiológicos como a fotossíntese, a síntese de hormônios, o crescimento celular e o enchimento de grãos.

O conhecimento dessas faixas é fundamental para interpretar corretamente os dados de análise de solo, orientar intervenções agrônômicas e calibrar algoritmos de classificação e recomendação. Além disso, essas informações servem como referência para o desenvolvimento de modelos de aprendizado de máquina que visam prever a adequação do solo com base em atributos quantificáveis, contribuindo para uma agricultura mais precisa e eficiente. A Tabela 1 contém as faixas ideais dos atributos do solo para o cultivo do milho.

Tabela 1: Faixas ideais dos atributos do solo para o cultivo do milho.

Atributo	Faixa Ideal	Fonte
pH	5,5 – 6,5	Correção visando pH \approx 6,0 para milho em solos ácidos
Areia (%)	30 – 50	Textura média a franco (adaptado para boa drenagem)
Argila (%)	20 – 35	Textura média a franco-argilosa
Silte (%)	20 – 40	Textura equilibrada para retenção de água
CE (mS/cm)	0 – 1	Baixa salinidade ideal
MO (%)	\geq 2,0	Matéria orgânica moderada a alta
CaCO_3 (%)	0 – 5	Solos sem excesso de carbonato
N-NO_3 (ppm)	\geq 20	Necessidade típica para 5 t ha ⁻¹ de grãos
P (ppm)	\geq 12	Disponibilidade “boa” (P-Mehlich-1)
K (ppm)	\geq 120	Disponibilidade “boa” (K-Mehlich-1)
Mg (ppm)	50 – 150	Relação Ca:Mg de 3:1 a 5:1
Fe (ppm)	4 – 8	Faixa de suficiência
Zn (ppm)	1 – 2	Micronutriente limitante
Mn (ppm)	5 – 20	Faixa de suficiência
Cu (ppm)	0,5 – 2	Faixa de suficiência
B (ppm)	0,5 – 1,5	Faixa de suficiência

4. Algoritmos de Classificação

Este estudo avaliou cinco algoritmos de aprendizado supervisionado para a tarefa de classificação da adequação do solo: k-Nearest Neighbors (k-NN) - Cover and Hart (1967), Árvore de Decisão - Quinlan (1986), Floresta Aleatória (Random Forest) - Breiman (2001), Máquina de Vetores de Suporte (SVM) - Cortes and Vapnik (1995) e Extreme Gradient Boosting (XGBoost) - Chen and Guestrin (2016). Esses modelos foram escolhidos por suas características complementares e comprovada eficácia no tratamento de dados estruturados com alvos categóricos.

k-Nearest Neighbors (k-NN): O algoritmo k-NN é um método não paramétrico e baseado em instâncias, que classifica uma amostra com base no rótulo majoritário entre seus k vizinhos mais próximos no espaço das características. Sua simplicidade e natureza intuitiva o tornam uma linha de base sólida, mas seu desempenho pode ser sensível à escolha de k e à normalização das variáveis.

Árvore de Decisão: Árvores de decisão particionam o espaço das características por meio de divisões hierárquicas e alinhadas aos eixos, baseadas nos atributos mais informativos. São altamente interpretáveis e capazes de capturar relações não lineares, mas tendem a sobreajustar os dados, especialmente quando não são aplicadas restrições ao seu crescimento.

Random Forest: Random Forest é um método de ensemble que constrói múltiplas árvores de decisão usando amostras com reposição (bootstrap) e seleção aleatória de características, agregando suas previsões para melhorar a generalização. Essa abordagem mitiga o sobreajuste e frequentemente proporciona melhor desempenho e robustez em comparação a árvores de decisão individuais.

Máquina de Vetores de Suporte (SVM): SVMs são classificadores poderosos que buscam encontrar o hiperplano ótimo de separação entre as classes em um espaço de características transformado, frequentemente utilizando funções de kernel para lidar com fronteiras não lineares. São eficazes em contextos de alta dimensionalidade, embora seu custo computacional possa ser elevado em conjuntos de dados grandes.

XGBoost: Extreme Gradient Boosting (XGBoost) é um algoritmo de aprendizado por ensemble de última geração baseado em boosting de árvores de decisão. Constrói sequencialmente árvores para corrigir os erros das anteriores, otimizando uma função de perda diferenciável. O XGBoost é conhecido por seu alto desempenho preditivo, capacidade de lidar com dados ausentes e eficiência computacional, sendo amplamente utilizado em competições de aprendizado de máquina e aplicações práticas.

5. Materiais e Métodos

Esta seção descreve os dados utilizados, o processo de preparação das amostras, os critérios de classificação da adequação do solo ao cultivo do milho e os métodos empregados na construção e avaliação dos modelos preditivos. A definição criteriosa dos atributos do solo, baseada em literatura agrônoma consolidada, foi fundamental para orientar a análise e garantir a relevância dos parâmetros escolhidos. Foram utilizados dados físico-químicos reais de amostras de solo, complementados por dados sintéticos gerados para melhorar a representatividade das classes e a robustez dos modelos. Em seguida, aplicaram-se técnicas de aprendizado de máquina para classificar as amostras em diferentes níveis de adequação, conforme critérios previamente estabelecidos. As etapas metodológicas incluem a análise exploratória dos dados, pré-processamento, seleção de atributos, treinamento dos modelos, validação cruzada e avaliação de desempenho por métricas estatísticas.

5.1. Conjunto de Variáveis Predictoras

O modelo de classificação desenvolvido neste estudo baseia-se em um conjunto de variáveis predictoras que representam propriedades físico-químicas e texturais do solo, com potencial influência direta sobre a produtividade da cultura do milho. Essas variáveis foram selecionadas com base em literatura agrônoma consolidada e refletem os principais fatores que determinam a fertilidade do solo e a sua aptidão agrícola. O conjunto é composto por 16 variáveis predictoras numéricas e contínuas, além de um identificador único (ID) associado a cada amostra. As variáveis consideradas são: Areia, Argila, Silte, pH, CE, MO, CaCO_3 , N-NO_3 , P, K, Mg, Fe, Zn, Mn, Cu e B.

5.2. Conjunto de Dados Original

O presente estudo teve início a partir de um conjunto de dados intitulado `Dataset_OriginalComClass.csv`, proveniente de Tziachris (2022), o qual reúne amostras de solo acompanhadas de análises laboratoriais físico-químicas e uma classificação de adequação ao cultivo do milho. A análise exploratória foi conduzida com o apoio do

script `ML_Trabalho_EDA_v3.txt` e seu respectivo output (`RESPOSTA_ML_Trabalho_EDA_v3.txt`), que permitiram avaliar e preparar os dados para uso em modelos de aprendizado de máquina.

Inicialmente, o conjunto de dados continha 781 linhas e 23 colunas. No entanto, quatro colunas (Unnamed: 0 e Unnamed: 19 a Unnamed: 22) apresentavam apenas valores nulos e foram eliminadas, resultando em 18 colunas válidas para análise. As variáveis consideradas os atributos físico-químicos: areia, argila, silte, pH, condutividade elétrica, matéria orgânica, carbonato de cálcio (CaCO_3), nitrato disponível (N-NO_3), fósforo (P), potássio (K), magnésio (Mg), ferro (Fe), zinco (Zn), manganês (Mn), cobre (Cu) e boro (B). A coluna ID foi mantida apenas para fins de rastreamento das amostras e não foi utilizada como variável preditora.

A variável alvo, denominada `Adequacao_Milho`, é do tipo **categórica** e representa o nível de adequação do solo ao cultivo do milho. Sua definição baseou-se em uma abordagem fundamentada em critérios agrônômicos: para cada ponto de coleta, foi atribuído um *score* correspondente à quantidade de atributos do solo que se encontravam dentro de suas respectivas *faixas ideais*, conforme estabelecido na literatura técnica. Cada atributo em conformidade somava **1 ponto** à amostra, podendo totalizar até **16 pontos**. Em seguida, a pontuação obtida foi convertida em três classes, de acordo com a seguinte escala: amostras com **0 a 5 pontos** foram classificadas como de **baixa adequação**, aquelas com **6 a 11 pontos** como de **média adequação** e, por fim, amostras com **12 a 16 pontos** como de **alta adequação**. Essa estratégia permitiu transformar um conjunto de dados contínuos em uma variável categórica, adequada à tarefa de **classificação supervisionada** proposta neste estudo.

5.3. Análise Exploratória de Dados (EDA)

A Análise Exploratória de Dados (EDA) foi conduzida utilizando o *script* `ML_EDA.py`, com o objetivo de examinar a estrutura e as características estatísticas do conjunto de dados. Buscou-se também identificar eventuais inconsistências, padrões e necessidades de pré-processamento.

No carregamento inicial do conjunto de dados (*dataset*), foram identificadas colunas e linhas integralmente preenchidas com valores nulos. Adicionalmente, constatou-se a necessidade de ajustar os nomes das variáveis para a remoção de espaços em branco.

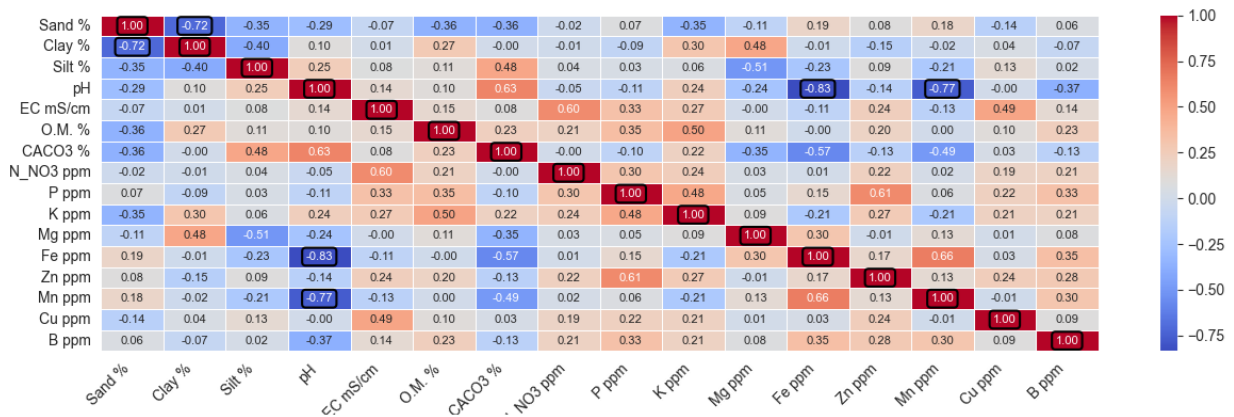
Em seguida, realizou-se uma análise descritiva básica, das dimensões do *dataset*, a visualização das primeiras entradas e a obtenção de informações gerais sobre os tipos de dados e a presença de valores nulos. Foram calculadas estatísticas descritivas — média, desvio padrão, valor mínimo, valor máximo e quartis — para as variáveis numéricas. Nesta etapa, verificou-se que os tipos de dados dos atributos foram interpretados erroneamente durante o carregamento inicial. Este problema foi corrigido mediante a alteração da rotina de leitura do arquivo CSV, configurando-a para reconhecer a vírgula como separador decimal.

A variável alvo, `Adequacao_Milho`, foi analisada isoladamente com o intuito de verificar o balanceamento das classes. A análise revelou um acentuado desbalanceamento: das 781 amostras, 158 (20,23%) foram classificadas como de *baixa adequação*, 622 (79,64%) como de *média adequação* e apenas 1 amostra (0,13%) como de *alta adequação*. Este resultado evidenciou a necessidade da aplicação de técnicas de reamostragem ou balanceamento para otimizar o desempenho dos modelos de aprendizado supervisionado.

Ademais, foi efetuada uma análise univariada dos atributos numéricos, com a geração de histogramas e diagramas de caixa (*boxplots*) para cada variável. Esta etapa permitiu a identificação possíveis *outliers*.

Por fim, conduziu-se uma análise de correlação, envolvendo o cálculo da matriz de correlação entre os atributos contínuos e a sua visualização por meio de um mapa de calor (*heatmap*). Este procedimento visou à identificação de relações lineares relevantes entre as variáveis. O *script* foi programado para destacar correlações com valor absoluto superior a 0,7. Conforme ilustrado na Figura 1, observa-se uma alta correlação linear entre as variáveis pH e Fe, assim como entre pH e Mn.

Figura 1: Mapa de calor da matriz de correlação entre os atributos numéricos.



Os resultados da EDA fundamentaram diretrizes importantes para a subsequente etapa de pré-processamento. Além de corroborar a necessidade de tratamento para dados ausentes, a EDA expôs um desbalanceamento extremo na classe *alta adequação*. Os *outliers* detectados, após criteriosa avaliação, foram interpretados como representativos da variabilidade intrínseca dos atributos do solo em condições reais. Assim, o pré-processamento contemplará verificações de consistência, para identificar os *outliers*, especificamente se a soma das frações (Areia, Argila e Silte) totalizarem 100% e se os valores de pH se encontram dentro do intervalo agronomicamente aceitável. Adicionalmente, foi identificada uma elevada correlação linear entre o pH e os teores de Ferro (Fe) e Manganês (Mn). Esta colinearidade sugere que Fe e Mn podem ser considerados para exclusão na fase de pré-processamento, com o intuito de reduzir a dimensionalidade e mitigar a redundância de informações no modelo.

5.4. Geração de Dados Sintéticos

Para lidar com o expressivo desbalanceamento na variável alvo, notadamente na classe *Alta adequação*, optou-se pela geração de dados sintéticos. O objetivo desta abordagem foi equilibrar a distribuição das classes, visando mitigar o viés de classificação e, consequentemente, aprimorar a robustez e a capacidade de generalização dos modelos supervisionados. O procedimento foi implementado em um script Python e a estratégia adotada é descrita na sequência.

Para corrigir o desbalanceamento entre as classes, foi empregada uma técnica de **Geração de Dados Sintéticos por Sobreamostragem Aleatória Restrita (GDSSAR)**. Esta abordagem é fundamentada em princípios da Análise Exploratória de Dados Tukey (1977), que visa gerar novas amostras mantendo a consistência estatística com o conjunto de dados original. O objetivo foi ampliar a representatividade das classes minoritárias: *Alta adequação* e *Baixa adequação*.

O cerne do método consiste em gerar valores sintéticos para cada atributo a partir de uma amostragem aleatória dentro de uma faixa estatisticamente robusta. Especificamente, os novos valores foram amostrados de uma distribuição uniforme definida pelos limites do Intervalo Interquartil (IQR), ou seja, entre o primeiro quartil (Q_1) e o terceiro quartil (Q_3) da distribuição do atributo na classe correspondente. A escolha por esta faixa garante que os dados sintéticos representem a região central e de maior densidade dos dados originais, evitando a criação de outliers artificiais e preservando a plausibilidade agrônômica.

Adicionalmente, um tratamento particular foi aplicado aos atributos de textura do solo (Sand, Clay, Silt). Para amostras da classe *Alta adequação*, o algoritmo busca não apenas a conformidade com os critérios ótimos, mas também assegura, por meio de um processo iterativo, que a soma de suas frações seja rigorosamente igual a 100%. Para as demais classes, a geração é seguida por uma etapa de normalização para assegurar a mesma restrição.

Finalmente, cada amostra sintética gerada é submetida a um processo de validação para confirmar sua aderência às regras de domínio que definem a classe alvo. Um mecanismo de nova tentativa (*retry mechanism*) é acionado caso a conformidade não seja atingida, garantindo que o conjunto de dados final seja composto por amostras sintéticas que respeitam tanto as restrições estatísticas quanto as do domínio.

O volume de amostras sintéticas geradas teve como meta equalizar o número de instâncias entre as três classes, equiparando as contagens das classes *Alta adequação* e *Baixa adequação* à da classe originalmente majoritária, *Média adequação*. Especificamente, para a classe *Alta adequação*, foram geradas 621 amostras sintéticas, as quais,

somadas à única amostra original, resultaram em 622 instâncias. Para a classe *Baixa adequação*, foram produzidas 464 amostras sintéticas, que, adicionadas às 158 originais, totalizaram igualmente 622 instâncias. A classe *Média adequação* permaneceu com suas 622 amostras originais, não sendo submetida ao processo de geração sintética.

O conjunto de dados resultante, passou a compreender um total de 1866 amostras, distribuídas equitativamente com 622 instâncias para cada uma das classes: *Alta adequação*, *Baixa adequação* e *Média adequação*. Este procedimento de enriquecimento de dados culminou em um *dataset* balanceado, condição fundamental para a mitigação de vies nos algoritmos de classificação e para o incremento da capacidade de generalização dos modelos resultantes.

5.5. Pré-processamento dos Dados

Inicialmente, o conjunto de dados combinado, resultante da união das amostras originais e sintéticas, foi submetido a uma nova verificação. Utilizando-se das abordagens e ferramentas previamente empregadas na Análise Exploratória de Dados (EDA), reavaliou-se a integridade e as características deste *dataset* ampliado. A análise dos gráficos de *boxplot* indicou que a distribuição de cada atributo preditor manteve-se consistente e muito próxima à observada no conjunto de dados original. Adicionalmente, confirmou-se que as classes da variável alvo alcançaram um balanceamento equilibrado, com cada classe representando aproximadamente 33,33% do total de amostras.

Com a validação da qualidade e do balanceamento do conjunto de dados aumentado, procedeu-se ao início da etapa de pré-processamento dos dados. O objetivo central desta fase foi refinar e preparar o *dataset* para otimizar o treinamento e a avaliação dos modelos de aprendizado de máquina, garantindo a qualidade, consistência e adequação dos dados.

A etapa inicial do pré-processamento consistiu na verificação de colunas e linhas integralmente preenchidas por valores nulos no *dataset* combinado. Colunas que satisfaziam esta condição foram identificadas e removidas. Adicionalmente, foram excluídas a coluna de identificação (ID) e as colunas Fe e Mn. A remoção destas duas últimas baseou-se na Análise Exploratória de Dados (EDA) realizada sobre o *dataset* original, que apontou uma alta correlação entre estes atributos e o pH.

Em seguida, foi realizado o tratamento e verificação de outliers, com foco em atributos agronomicamente sensíveis. A soma das frações texturais — areia, argila e silte — foi inspecionada para verificar se os valores totais por amostra estavam dentro de um intervalo aceitável (entre 99% e 101%). Nenhuma anomalia significativa foi identificada. Da mesma forma, os valores de pH foram avaliados quanto à coerência, sendo verificado se estavam dentro de uma faixa realista (de 0 a 14). Conforme o log de execução, também não foram encontrados valores fora do intervalo esperado, dispensando ações de remoção de outliers neste ponto.

A variável alvo categórica *Adequacao_Milho* foi então codificada em formato numérico ordinal, com o intuito de facilitar o processamento pelos algoritmos de aprendizado supervisionado. As classes foram mapeadas da seguinte forma: *Baixa adequação* foi codificada como 0,1, *Média adequação* como 0,5 e *Alta adequação* como 1,0.

A etapa final do pré-processamento consistiu na normalização de todos os atributos preditores numéricos, empregando duas técnicas distintas da biblioteca *Scikit-learn*: *MinMaxScaler* (MM) e *StandardScaler* (STD). O método *MinMaxScaler*, opera reescalando cada atributo para o intervalo específico de 0 a 1 (Equação (1)). Embora esta abordagem seja vantajosa para algoritmos sensíveis à magnitude dos dados, ela é suscetível a *outliers* extremos, os quais podem ocasionar a compressão da maioria dos valores em uma faixa reduzida do intervalo.

$$X_{\text{scaled}} = \frac{(X - X_{\min})}{(X_{\max} - X_{\min})} \quad (1)$$

Onde, X é o valor original do atributo, X_{\min} é o valor mínimo desse atributo no conjunto de dados, e X_{\max} é o valor máximo.

Por outro lado, o *StandardScaler*, padroniza os atributos ao remover a média e escalonar pela variância unitária, resultando em uma distribuição com média zero e desvio padrão unitário (Equação (2)). Este método tende a ser menos sensível a *outliers* e é frequentemente preferível para algoritmos que assumem uma distribuição próxima da normal para os dados de entrada.

$$X_{\text{scaled}} = \frac{(X - \mu)}{\sigma} \quad (2)$$

Onde, X é o valor original do atributo, μ representa a média dos valores desse atributo, e o σ é o desvio padrão do mesmo atributo.

Ambas as estratégias de escalonamento são aplicadas de forma independente aos dados de treinamento. Subsequentemente, os modelos de aprendizado de máquina são treinados e avaliados utilizando cada um dos conjuntos de dados normalizados, com o intuito de analisar comparativamente o impacto de cada técnica de normalização sobre o desempenho e a resposta dos algoritmos.

5.6. Treinamento e Avaliação dos Modelos

Esta seção detalha a metodologia de treinamento e avaliação dos modelos, desenhada para tratar o desbalanceamento de classes utilizando a técnica de GDSSAR e para obter uma estimativa de desempenho robusta. O protocolo adotado envolveu a separação de um conjunto de teste (com dados exclusivamente reais) e o treinamento de múltiplos algoritmos sobre um conjunto balanceado (dados reais + sintéticos) através de uma validação cruzada de 5 folds. A otimização de hiperparâmetros foi realizada com `RandomizedSearchCV` para maximizar o F1-Score Macro, e testes estatísticos foram empregados para validar decisões de pré-processamento. Todo o processo foi estruturado para garantir a reprodutibilidade e a validade das conclusões sobre a performance dos modelos.

O treinamento e a avaliação dos modelos foram desenhados para endereçar o desafio do desbalanceamento de classes e para garantir uma estimativa realista do desempenho preditivo. A estratégia central consistiu em treinar os modelos em um ambiente de dados balanceado e avaliá-los em um conjunto de dados que reflete a distribuição natural das classes. Para validar esta abordagem, a avaliação foi estruturada em duas configurações experimentais:

- **Configuração Baseline:** O treinamento e o teste foram realizados utilizando exclusivamente o dataset original desbalanceado, para estabelecer uma linha de base de desempenho.
- **Configuração Proposta:** Esta é a abordagem principal do estudo. Um conjunto de teste, contendo apenas amostras reais, foi inicialmente separado. Os modelos foram então treinados utilizando um conjunto de dados balanceado (dados de treino originais + amostras sintéticas via GDSSAR). A avaliação de desempenho foi conduzida sobre o conjunto de teste real e independente.

Conforme explicado na Subseção 5.5, os conjuntos de dados foram previamente normalizados por meio de duas técnicas, resultando em versões normalizadas por `MinMaxScaler` (MM) e `StandardScaler` (STD). Dentro da **Configuração Proposta**, foi conduzida uma análise para comparar o desempenho dos modelos em cada uma dessas versões. A fim de verificar se as diferenças de performance eram estatisticamente relevantes, o teste não paramétrico de Wilcoxon foi aplicado sobre os F1-Scores obtidos nos 5 folds da validação cruzada.

O protocolo de avaliação seguiu uma validação cruzada com 5 folds (*5-fold cross-validation*) sobre o conjunto de treinamento balanceado. Em cada iteração, um modelo foi treinado e seu desempenho aferido no conjunto de teste real e independente, gerando cinco estimativas de desempenho para cada algoritmo e permitindo a análise de sua estabilidade.

A busca pelos hiperparâmetros ótimos foi realizada via `RandomizedSearchCV` com validação cruzada interna, explorando 30 combinações para maximizar o F1-Score Macro. Os espaços de busca foram definidos da seguinte forma:

- **k-Nearest Neighbors (k-NN):** `n_neighbors` (3 a 50), `p` (1 para Manhattan, 2 para Euclidiana), `weights` ('uniform', 'distance').
- **Árvore de Decisão:** `max_depth` (3 a 50), `min_samples_split` (2 a 10), `criterion` ('gini', 'entropy').
- **Random Forest:** `n_estimators` (50 a 300), `max_depth` (5 a 50), e demais parâmetros seguindo a Árvore de Decisão.
- **SVM com kernel RBF:** `C` (0.1 a 10), `gamma` ('scale', 'auto', e log-escala de 0.001 a 100).
- **XGBoost:** `n_estimators` (50 a 300), `max_depth` (3 a 10), `learning_rate` (0.01 a 0.21).

Para garantir a reprodutibilidade, a semente aleatória (`random_state`) foi fixada em 42 em todos os processos. Os experimentos foram conduzidos em Python 3, com as bibliotecas `Pandas`, `NumPy`, `Scikit-learn`, `XGBoost`, `Seaborn` e `Matplotlib`. A análise estatística e o cálculo de intervalos de confiança de 95% para as métricas reportadas foram realizados com o auxílio da biblioteca `SciPy`. O código-fonte completo para este trabalho está disponível publicamente em <https://github.com/RafaelRoaniFeijo/AprendizadoMaquina>.

6. Resultados

Esta seção apresenta os resultados da avaliação de desempenho dos modelos de classificação, com foco em demonstrar o impacto positivo da técnica de GDSSAR. Inicialmente, são comparados os resultados da configuração *Baseline* (sem os dados sintéticos) com a *Proposta* (com os dados sintéticos), evidenciando a melhoria na capacidade de generalização dos modelos. Em seguida, é detalhada a performance comparativa dos cinco algoritmos testados. Por fim, são apresentadas as métricas detalhadas e a matriz de confusão deste modelo, ilustrando sua eficácia na classificação das três classes de adequação do solo.

A Tabela 2 apresenta um comparativo de desempenho dos modelos na configuração *Baseline*, avaliando o impacto de duas estratégias de normalização de atributos. São reportadas as métricas macro (F1, Precisão e Recall) para cada modelo sob as configurações *MinMaxScaler* (MM) e *StandardScaler* (STD).

Tabela 2: Resultados detalhados por classe na configuração *Baseline* (sem dados sintéticos), comparando o efeito dos normalizadores *MinMaxScaler* (MM) e *StandardScaler* (STD).

Modelo	Classe	MinMaxScaler (MM)			StandardScaler (STD)		
		F1-score (\pm DP)	Precisão (\pm DP)	Recall (\pm DP)	F1-score (\pm DP)	Precisão (\pm DP)	Recall (\pm DP)
k-NN	Baixa	0.570 \pm 0.034	0.664 \pm 0.092	0.506 \pm 0.035	0.604 \pm 0.042	0.644 \pm 0.101	0.576 \pm 0.018
	Média	0.897 \pm 0.018	0.869 \pm 0.006	0.928 \pm 0.036	0.895 \pm 0.021	0.882 \pm 0.006	0.909 \pm 0.038
	Alta	0.000 \pm 0.000	0.000 \pm 0.000	0.000 \pm 0.000	0.000 \pm 0.000	0.000 \pm 0.000	0.000 \pm 0.000
DecisionTree	Baixa	0.639 \pm 0.070	0.613 \pm 0.089	0.676 \pm 0.077	0.642 \pm 0.068	0.612 \pm 0.089	0.683 \pm 0.070
	Média	0.887 \pm 0.028	0.904 \pm 0.021	0.872 \pm 0.042	0.887 \pm 0.028	0.905 \pm 0.020	0.870 \pm 0.043
	Alta	0.144 \pm 0.198	0.129 \pm 0.194	0.200 \pm 0.245	0.144 \pm 0.198	0.129 \pm 0.194	0.200 \pm 0.245
RandomForest	Baixa	0.690 \pm 0.061	0.773 \pm 0.054	0.632 \pm 0.095	0.696 \pm 0.058	0.774 \pm 0.056	0.639 \pm 0.089
	Média	0.924 \pm 0.010	0.900 \pm 0.020	0.950 \pm 0.016	0.925 \pm 0.011	0.902 \pm 0.019	0.950 \pm 0.016
	Alta	0.133 \pm 0.267	0.200 \pm 0.400	0.100 \pm 0.200	0.133 \pm 0.267	0.200 \pm 0.400	0.100 \pm 0.200
SVM	Baixa	0.607 \pm 0.019	0.496 \pm 0.023	0.785 \pm 0.049	0.619 \pm 0.066	0.538 \pm 0.104	0.753 \pm 0.060
	Média	0.840 \pm 0.015	0.924 \pm 0.015	0.771 \pm 0.022	0.852 \pm 0.046	0.915 \pm 0.013	0.802 \pm 0.082
	Alta	0.200 \pm 0.267	0.150 \pm 0.200	0.300 \pm 0.400	0.057 \pm 0.114	0.040 \pm 0.080	0.100 \pm 0.200
XGBoost	Baixa	0.752 \pm 0.066	0.833 \pm 0.070	0.689 \pm 0.076	0.752 \pm 0.066	0.833 \pm 0.070	0.689 \pm 0.076
	Média	0.937 \pm 0.015	0.912 \pm 0.019	0.963 \pm 0.017	0.937 \pm 0.015	0.912 \pm 0.019	0.963 \pm 0.017
	Alta	0.000 \pm 0.000	0.000 \pm 0.000	0.000 \pm 0.000	0.000 \pm 0.000	0.000 \pm 0.000	0.000 \pm 0.000

Observa-se na Tabela 2 que, de forma geral, a escolha do normalizador teve baixo impacto no desempenho dos modelos na configuração *Baseline*. A principal conclusão, no entanto, é a dificuldade generalizada em classificar a classe 'Alta', onde a maioria dos modelos apresenta um F1-score próximo de zero. Este resultado evidencia o forte impacto do desbalanceamento de classes e justifica a necessidade da *Configuração Proposta* com a técnica SMOTE, cujos resultados são apresentados na sequência.

Após demonstrar as limitações da abordagem *Baseline*, a performance dos modelos foi avaliada na *Configuração Proposta*, que emprega a técnica de GDSSAR para o balanceamento de classes. A Tabela 3 apresenta os resultados detalhados desta abordagem, comparando novamente os dois pipelines de normalização.

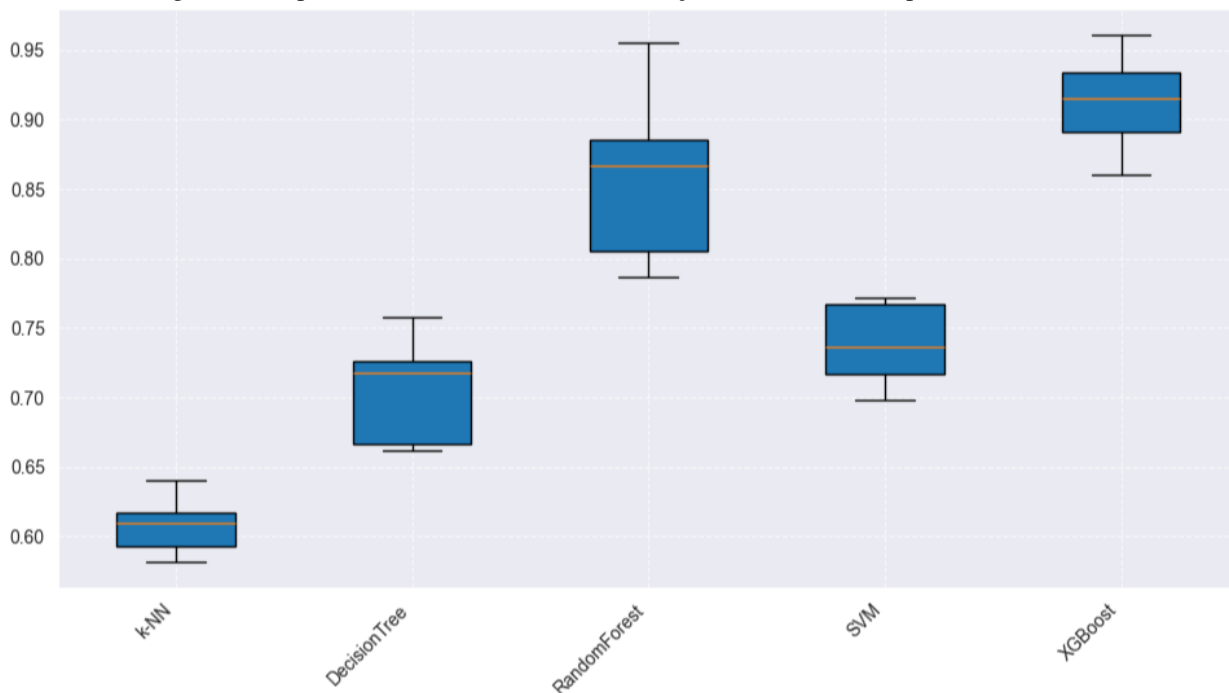
Tabela 3: Resultados detalhados por classe na configuração *Proposta* (com dados sintéticos), comparando o efeito dos normalizadores *MinMaxScaler* (MM) e *StandardScaler* (STD).

Modelo	Classe	MinMaxScaler (MM)			StandardScaler (STD)		
		F1-score (\pm DP)	Precisão (\pm DP)	Recall (\pm DP)	F1-score (\pm DP)	Precisão (\pm DP)	Recall (\pm DP)
k-NN	Baixa	0.689 ± 0.012	0.737 ± 0.022	0.647 ± 0.005	0.682 ± 0.017	0.728 ± 0.015	0.642 ± 0.025
	Média	0.912 ± 0.005	0.903 ± 0.002	0.920 ± 0.008	0.910 ± 0.004	0.901 ± 0.006	0.919 ± 0.003
	Alta	0.266 ± 0.074	0.209 ± 0.069	0.375 ± 0.079	0.233 ± 0.056	0.183 ± 0.050	0.325 ± 0.061
DecisionTree	Baixa	0.757 ± 0.036	0.690 ± 0.049	0.842 ± 0.040	0.757 ± 0.036	0.690 ± 0.049	0.842 ± 0.040
	Média	0.919 ± 0.012	0.951 ± 0.009	0.890 ± 0.021	0.919 ± 0.012	0.951 ± 0.009	0.890 ± 0.021
	Alta	0.442 ± 0.135	0.404 ± 0.146	0.550 ± 0.232	0.442 ± 0.135	0.404 ± 0.146	0.550 ± 0.232
RandomForest	Baixa	0.930 ± 0.011	0.929 ± 0.008	0.930 ± 0.019	0.930 ± 0.011	0.929 ± 0.008	0.930 ± 0.019
	Média	0.978 ± 0.004	0.978 ± 0.006	0.978 ± 0.004	0.978 ± 0.004	0.978 ± 0.006	0.978 ± 0.004
	Alta	0.672 ± 0.171	0.722 ± 0.178	0.650 ± 0.184	0.672 ± 0.171	0.722 ± 0.178	0.650 ± 0.184
SVM	Baixa	0.697 ± 0.023	0.748 ± 0.021	0.653 ± 0.031	0.750 ± 0.005	0.807 ± 0.021	0.701 ± 0.011
	Média	0.918 ± 0.005	0.906 ± 0.008	0.930 ± 0.006	0.935 ± 0.002	0.920 ± 0.003	0.950 ± 0.006
	Alta	0.348 ± 0.070	0.297 ± 0.055	0.425 ± 0.100	0.529 ± 0.088	0.516 ± 0.090	0.550 ± 0.100
XGBoost	Baixa	0.943 ± 0.009	0.935 ± 0.008	0.952 ± 0.018	0.943 ± 0.009	0.935 ± 0.008	0.952 ± 0.018
	Média	0.983 ± 0.003	0.985 ± 0.006	0.981 ± 0.002	0.983 ± 0.003	0.985 ± 0.006	0.981 ± 0.002
	Alta	0.810 ± 0.094	0.881 ± 0.106	0.775 ± 0.146	0.810 ± 0.094	0.881 ± 0.106	0.775 ± 0.146

A análise da Tabela 3 demonstra o impacto da geração de dados sintéticos (GDSSAR). O resultado mais significativo é na classe 'Alta', que na configuração *Baseline* não era aprendida, e agora alcança um F1-score de até **0.810** com o modelo XGBoost. Isso valida a eficácia do balanceamento em permitir que os modelos aprendam os padrões da classe minoritária. Novamente, a diferença entre os normalizadores MM e STD mostrou-se marginal para os modelos de melhor desempenho. Com base no alto desempenho geral e, especialmente, na sua capacidade de generalizar para a classe 'Alta', o modelo **XGBoost** foi selecionado como o de melhor performance para as análises subsequentes.

Para visualizar a performance e a estabilidade dos modelos na configuração *Proposta*, a Figura 2 apresenta os boxplots da distribuição dos scores de F1-Score Macro obtidos nos 5 folds da validação cruzada.

Figura 2: Boxplot dos Scores F1 Macro na Validação Cruzada Externa por Modelo (STD)



Tendo o XGBoost se destacado como o modelo mais promissor na etapa anterior, a Figura 3 apresenta sua curva de aprendizado para validar o comportamento do treinamento. A convergência entre a pontuação de validação e a de treinamento em um valor alto de F1-Score indica que o modelo foi capaz de generalizar bem, sem apresentar sinais de sobreajuste (*overfitting*) ou subajuste (*underfitting*) significativos.

Para avaliar a confiabilidade das probabilidades geradas pelo modelo, foi realizada uma análise de calibração. A Figura 4 apresenta a curva de calibração do XGBoost, que compara as probabilidades preditas com a frequência real dos resultados. Um modelo perfeitamente calibrado seguiria a linha diagonal.

A análise visual da Figura 4 mostra que a curva do modelo se aproxima da diagonal de referência, sugerindo uma boa calibração. Esta observação é confirmada quantitativamente pelo Brier Score. O Brier Score médio geral obtido foi de **0.0519**, um valor próximo do ideal (zero), o que indica que as probabilidades preditas pelo modelo são, de fato, confiáveis. A análise por classe revelou scores de 0.0226 para a classe 'Baixa', 0.0281 para 'Média' e um notável 0.0040 para a classe 'Alta', reforçando a robustez do modelo em todas as frentes.

Figura 3: Curva de aprendizado do modelo XGBoost na configuração *Proposta* (com dados sintéticos e STD), baseada no F1-Score Macro.

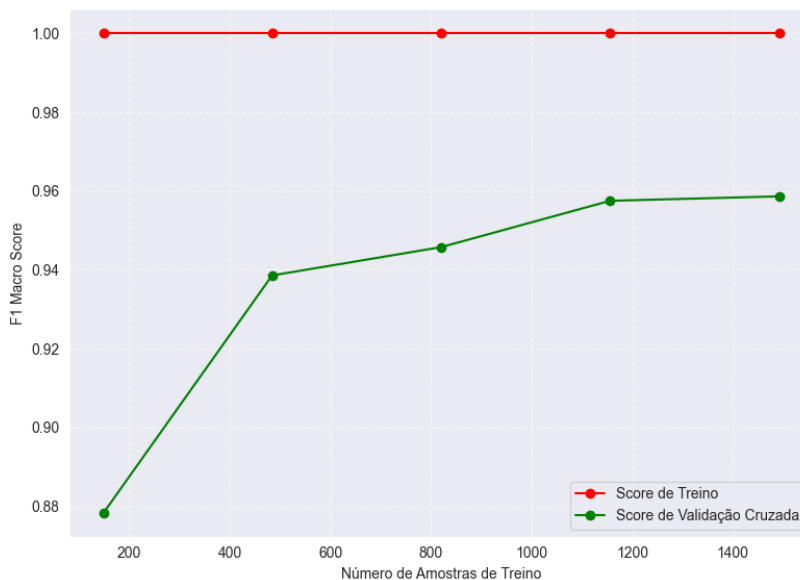
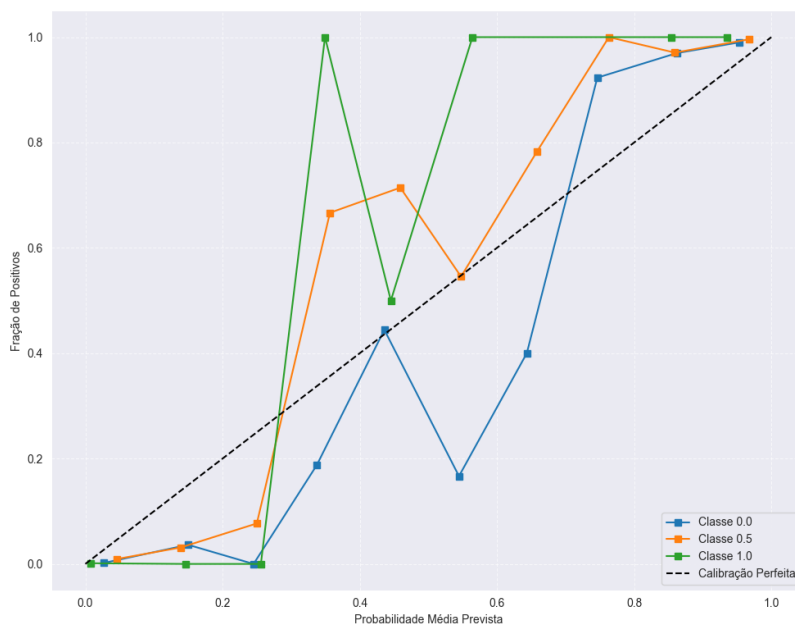


Figura 4: Curva de calibração para o modelo XGBoost na configuração *Proposta*.

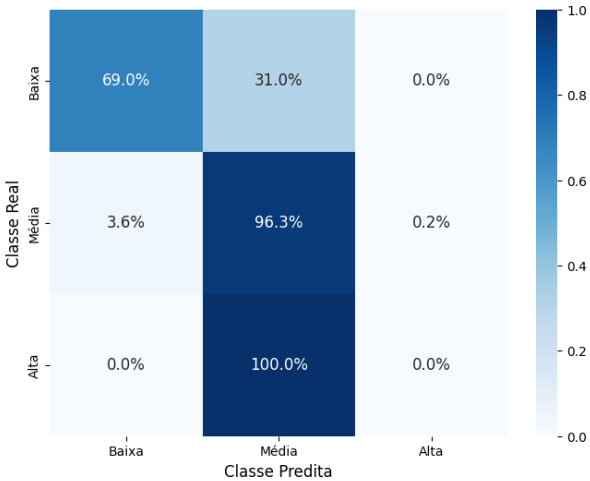


Para consolidar e visualizar o impacto da estratégia de balanceamento, esta análise final foca no desempenho do modelo campeão, o XGBoost. A Figura 5b apresenta uma comparação direta de sua performance nas configurações *Baseline* e *Proposta*. Para permitir uma avaliação visual justa e inequívoca, ambas as matrizes de confusão foram normalizadas para exibir a taxa de acerto (Recall) de cada classe em formato de porcentagem.

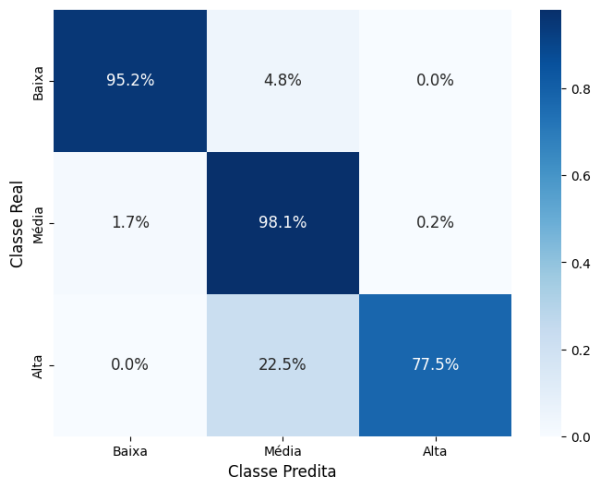
Em síntese, os resultados apresentados nesta seção demonstram uma trajetória clara desde a identificação de um desafio central — o severo desbalanceamento de classes que impedia a correta classificação da adequação 'Alta' do solo — até a validação de uma solução eficaz. A metodologia proposta, centrada na técnica de GDSSAR, permitiu que o modelo de melhor desempenho, o XGBoost, elevasse o F1-score da classe minoritária de um valor nulo para

Figura 5: Comparativo das matrizes de confusão do modelo XGBoost, normalizadas por Recall. A figura (a) mostra o desempenho no dataset original, enquanto a (b) ilustra a melhoria de performance após o treinamento com dados balanceados.

(a) Matriz de Confusão Normalizada (Recall).
Configuração *Baseline*, Sem dados sintéticos.



(b) Matriz de Confusão Normalizada (Recall).
Configuração *Proposta*, Sem dados sintéticos.



um patamar expressivo de 0.810. Análises subsequentes do modelo campeão confirmaram seu comportamento de aprendizado e a relevante confiabilidade de suas previsões.

7. Discussão

Os resultados obtidos fornecem uma base sólida para discutir o desempenho dos modelos avaliados e, mais importante, o impacto da estratégia de balanceamento de classes na classificação da adequação do solo para o cultivo de milho.

A principal conclusão deste estudo é a validação da metodologia proposta, centrada na Geração de Dados Sintéticos por Sobreamostragem Aleatória Restrita (GDSSAR), como uma solução eficaz para o severo desbalanceamento de classes presente no dataset. A comparação entre a configuração *Baseline* (Tabela 2) e a *Proposta* (Tabela 3) ilustra essa eficácia de forma inequívoca. Enquanto na abordagem *Baseline* os modelos foram incapazes de aprender os padrões da classe minoritária 'Alta' (com F1-score de 0.0 para os melhores modelos), a aplicação de GDSSAR no conjunto de treino permitiu que o modelo XGBoost alcançasse um F1-score de 0.810 para esta mesma classe. Este salto de desempenho demonstra que a geração de amostras sintéticas foi crucial para que o modelo pudesse generalizar seus aprendizados para a classe sub-representada.

Na comparação entre algoritmos, modelos baseados em ensemble de árvores, como Random Forest e XGBoost, se destacaram como as arquiteturas mais performáticas, o que é consistente com a literatura para dados tabulares complexos. Embora ambos tenham apresentado alto desempenho, o XGBoost foi selecionado como o modelo final devido à sua superioridade na classificação da classe 'Alta', o ponto mais crítico do problema. Os testes estatísticos (teste de Friedman, $p < 0.05$) confirmaram que as diferenças de desempenho entre os algoritmos eram estatisticamente significativas, validando a escolha do modelo.

Um achado secundário interessante foi o impacto limitado das estratégias de normalização. A comparação entre o uso de *MinMaxScaler* e *StandardScaler* (teste de Wilcoxon, $p > 0.05$) não revelou diferenças estatisticamente significativas para os modelos baseados em árvores. Isso é esperado, uma vez que estes algoritmos são inerentemente robustos à escala dos atributos, mas a validação empírica reforça a compreensão do comportamento dos modelos.

A análise aprofundada do modelo campeão, XGBoost, forneceu mais insights. A sua curva de aprendizado (Figura 3) demonstrou um comportamento de treinamento saudável, com as curvas de treino e validação convergindo para um patamar de alta performance, indicando boa generalização e ausência de sobreajuste (*overfitting*) severo. Adicionalmente, a análise de calibração (Figura 4) e o baixo Brier Score geral (0.0519) sugerem que as probabilidades preditas pelo modelo são confiáveis, o que é de grande valor para aplicações práticas onde a confiança na predição é importante, como em sistemas de recomendação de manejo do solo.

7.1. Limitações e Trabalhos Futuros

Apesar dos resultados promissores, este estudo possui limitações. A análise se baseou em um dataset de uma região geográfica específica, e a generalização dos modelos para solos com outras características precisa ser investigada. Além disso, o espaço de busca na otimização de hiperparâmetros, embora guiado por *RandomizedSearchCV*, não foi exaustivo.

Como trabalhos futuros, sugere-se a validação do modelo XGBoost com dados de diferentes safras e regiões para testar sua robustez. A exploração de outras técnicas de balanceamento, como ADASYN, ou de algoritmos de *deep learning* para dados tabulares, também representa um caminho promissor. Finalmente, a análise de importância dos atributos (*feature importance*) do modelo XGBoost pode ser aprofundada para gerar insights agrônômicos diretos, potencialmente levando ao desenvolvimento de uma ferramenta de auxílio à decisão para agricultores.

8. Conclusões

Este trabalho investigou a aplicação de múltiplos modelos de aprendizado de máquina para a tarefa de classificação da adequação do solo para o cultivo de milho, com um foco particular nos desafios impostos pelo desbalanceamento de classes.

A principal contribuição deste estudo foi demonstrar quantitativamente que o desbalanceamento de classes é um fator crítico que impede a correta identificação da classe minoritária ('Alta'). A conclusão central é que a aplicação da técnica de GDSSAR no conjunto de treinamento é uma estratégia fundamental e altamente eficaz para mitigar este problema. Isso foi evidenciado pela drástica melhoria de desempenho do modelo campeão, XGBoost, que elevou o F1-score da classe 'Alta' de um valor nulo para 0.810.

Dentre os algoritmos avaliados, os modelos de ensemble baseados em árvores, notadamente o XGBoost, provaram ser os mais adequados e robustos para este problema de classificação de dados tabulares. Como achado secundário,