CYNTHIA APARECIDA VALIATI BARRETO

Predição genômica em ensaios de múltiplos ambientes em milho utilizando abordagens estatística e de aprendizado de máquina

Tese apresentada à Universidade Federal de Viçosa, como parte das exigências do Programa de Pós-Graduação em Genética e Melhoramento para o título de *Doctor Scientiae.*

Orientador: Moysés Nascimento

Coorientadores: Camila Ferreira Azevedo
Eveline Teixeira Caixeta
Kaio Olimpio das Graças Dias

VIÇOSA - MINAS GERAIS
2024

Ficha catalográfica elaborada pela Biblioteca Central da Universidade Federal de Viçosa - Campus Viçosa

T
B273p
2024

Barreto, Cynthia Aparecida Valiati, 1994-
Predição genômica em ensaios de múltiplos ambientes em milho utilizando abordagensestatística e de aprendizado de máquina: Predição genômica em ensaios de múltiplos ambientes em milho / Cynthia Aparecida Valiati Barreto. – Viçosa, MG, 2024.
1 tese eletrônica (53 f.): il. (algumas color.).

Texto em português e inglês.
Orientador: Moysés Nascimento.
Tese (doutorado) - Universidade Federal de Viçosa, Departamento de Estatística, 2024.
Inclui bibliografia.
DOI: https://doi.org/10.47328/ufvbbt.2024.207
Modo de acesso: World Wide Web.

1. Genômica - Métodos estatísticos. 2. BLUP. 3. Amostragem (Estatística). 4. Aprendizado do computador. 5. *Zea mays*. I. Nascimento, Moysés, 1979-. II. Universidade Federal de Viçosa. Departamento de Estatística. Programa de Pós-Graduação em Genética e Melhoramento. III. Título.

CDD 22. ed. 576.5

Bibliotecário(a) responsável: Alice Regina Pinto Pires CRB-6/2523

CYNTHIA APARECIDA VALIATI BARRETO

Predição genômica em ensaios de múltiplos ambientes em milho utilizando abordagens estatística e de aprendizado de máquina

Tese apresentada à Universidade Federal de Viçosa, como parte das exigências do Programa de Pós-Graduação em Genética e Melhoramento para o título de *Doctor Scientiae.*

APROVADA: 01 de março de 2024

Assentimento:

gov.br Documento assinado digitalmente
CYNTHIA APARECIDA VALIATI BARRETO
Data: 21/07/2024 16:25:10-0300
Verifique em https://validar.iti.gov.br

Cynthia Aparecida Valiati Barreto

gov.br Documento assinado digitalmente
MOYSES NASCIMENTO
Data: 23/07/2024 09:26:23-0300
Verifique em https://validar.iti.gov.br

Moysés Nascimento
Orientador

DEDICO

À minha família.

AGRADECIMENTOS

Agradeço à Deus, cuja orientação divina e bênçãos estiveram presentes em cada etapa desta jornada acadêmica, fortalecendo minha fé e me dando forças para superar os desafios.

Agradeço do fundo do coração à minha família, por todo o amor, apoio incondicional e compreensão ao longo desta jornada. Vocês foram meu alicerce e minha fonte de força em cada passo do caminho.

Aos meus amigos, que estiveram ao meu lado nos momentos de alegria, desafios e celebrações, meu profundo agradecimento por serem minha rede de apoio e companhia ao longo dessa jornada.

Ao Laboratório de Inteligência Computacional e Aprendizado Estatístico (LICAE) pelo apoio e colaboração.

Ao meu orientador, Moysés Nascimento, pela orientação, sabedoria, apoio incansável ao longo deste percurso acadêmico. Sua orientação foi fundamental para o sucesso deste trabalho.

Aos meus coorientadores, Camila Ferreira Azevedo e Kaio Olimpio das Graças Dias, por sua orientação adicional, apoio e contribuições valiosas para o desenvolvimento deste estudo.

Aos membros da banca examinadora, por dedicarem seu tempo e expertise na avaliação deste trabalho, meu sincero agradecimento pelas valiosas contribuições para o aprimoramento desta tese.

À Embrapa Milho e Sorgo, pela parceria e disponibilização dos dados para a realização deste trabalho.

À empresa LongPing Hight-Tech, pela oportunidade de Residência durante meu Doutorado, na qual sou imensamente grata por todo conhecimento e vivência adquirida durante o processo.

À Universidade Federal de Viçosa e ao Programa de Pós-Graduação em Genética e Melhoramento, por proporcionarem um ambiente acadêmico estimulante, recursos e oportunidades para meu desenvolvimento como pesquisadora.

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

Por fim, a todos aqueles que contribuíram de alguma forma para este trabalho e para minha jornada acadêmica, meu mais profundo agradecimento. Este trabalho não teria sido possível sem o apoio e colaboração de cada um de vocês. MUITO OBRIGADO!

“Nothing in life is to be feared, it is only to be understood.”
(Marie Curie)

RESUMO

BARRETO, Cynthia Aparecida Valiati, Universidade Federal de Viçosa, março de 2024. **Predição genômica em ensaios de múltiplos ambientes em milho utilizando abordagens estatística e de aprendizado de máquina.** Orientador: Moysés Nascimento. Coorientadores: Camila Ferreira Azevedo, Eveline Teixeira Caixeta e Kaio Olimpio das Graças Dias

No contexto de ensaios de múltiplos ambientes (MET - *multiple environment trials*), a predição genômica é proposta como uma ferramenta que permite a predição de híbridos que não foram avaliados em campo, devido à relação captada por marcadores moleculares, com os híbridos avaliados fenotipicamente, economizando tempo, área experimental ou custos em comparação com a fenotipagem. Diversos métodos já foram propostos para as análises de predição genômica em ensaios MET, e o método mais comumente utilizado é o *Genomic Best Linear Unbiased Predictor* (GBLUP). No entanto, as metodologias de aprendizado de máquina têm ganhado atenção devido a sua capacidade de reconhecer estruturas de interação complexas nos dados. Entre as metodologias de aprendizado de máquina não paramétricas utilizadas na predição genômica podem-se citar as árvores de decisão e seus refinamentos. No entanto, apesar de suas potencialidades, ainda são escassos os estudos que avaliaram como as análises MET podem se beneficiar dessas metodologias. Assim, este estudo teve como objetivo avaliar a predição genômica de híbridos simples de milho não testados em MET para produção de grãos e tempo de florescimento feminino. Além disso, uma aplicação de metodologias de aprendizado de máquina em MET na predição de híbridos e comparamos seu desempenho com o GBLUP modelado para efeitos não aditivos é apresentada. Os resultados destacam que ambas as metodologias são eficientes e podem ser utilizadas em programas de melhoramento de milho para predizer com precisão o desempenho de híbridos em ambientes específicos. A melhor metodologia é caso dependente, especificamente, para explorar o potencial do GBLUP, é importante realizar uma modelagem precisa dos componentes de variância para otimizar a predição de novos híbridos. Por outro lado, as metodologias de aprendizagem de máquina podem capturar efeitos não aditivos sem fazer quaisquer pressuposições quanto ao modelo, já que seus resultados dependem do processo de aprendizado e não da distribuição das variáveis em si. No geral, predizer o desempenho de novos híbridos que não foram avaliados em nenhum ensaio de campo foi mais desafiador do que predizer híbridos em ensaios de campo desbalanceados.

Palavras-chave: GBLUP; Boosting; Bagging; Random Forest; Predição Genômica; *Zea mays*.

ABSTRACT

BARRETO, Cynthia Aparecida Valiati, Universidade Federal de Viçosa, March 2024. **Genomic prediction in multi-environment trials in maize using statistical and machine learning methods.** Advisor: Moysés Nascimento. Co-advisors: Camila Ferreira Azevedo, Eveline Teixeira Caixeta e Kaio Olimpio das Graças Dias

In the context of multiple environment trials (MET), genomic prediction is proposed as a tool that allows the prediction of single-cross hybrids that have not been evaluated in the field, due to the molecular markers relationship, with the phenotypically evaluated hybrids, bypassing the problem of sparse test designs, in addition to saving time, experimental area or costs compared to phenotyping. Several methods have been proposed for genomic prediction analyzes in MET assays, and the most commonly used is Genomic Best Linear Unbiased Predictor (GBLUP). However, in recent years, machine learning methodologies have gained attention due to their ability to recognize complex interaction structures in data. Among the non-parametric machine learning methodologies used in genomic selection are decision trees and their refinements. However, despite its potential, there are still few studies that have evaluated how MET analyzes can benefit from these methodologies. Thus, this study aimed to evaluate the genomic prediction of single cross maize hybrids not tested in MET for grain yield and female flowering time. We also aimed to propose an application of machine learning methodologies in MET in the prediction of hybrids and compare their performance with Genomic best linear unbiased prediction (GBLUP) with non-additive effects. Our results highlight that both methodologies are efficient and can be used in maize breeding programs to accurately predict the performance of hybrids in specific environments. The best methodology is case-dependent, specifically, to explore the potential of GBLUP, it is important to perform accurate modeling of the variance components to optimize the prediction of new hybrids. On the other hand, machine learning methodologies can capture non-additive effects without making any assumptions at the outset of the model, as their results depend on the learning process and not on the distribution of the variables themselves. Overall, predicting the performance of new hybrids that were not evaluated in any field trials was more challenging than predicting hybrids in sparse test designs.

Keywords: GBLUP; Boosting; Bagging; Random Forest; Genomic Prediction; *Zea mays*.

SUMÁRIO

1. INTRODUÇÃO GERAL

A relevância econômica do milho (*Zea mays*) é delineada pela variedade de suas aplicações, abrangendo desde a alimentação animal, onde representa a principal fonte de energia para diversos setores da pecuária, até a indústria de alta tecnologia, onde seus derivados são utilizados em produtos como biocombustíveis, biomateriais e até mesmo na produção de medicamentos.

Diante da tendência para temperaturas globais mais elevadas e possíveis mudanças climáticas mais extremas, o desenvolvimento de variedades de milho com maior tolerância ao estresse hídrico e maior eficiência no uso da água tornou-se uma meta prioritária para programas de melhoramento, tanto no setor privado quanto no setor público (RIBAUT *et al.*, 2009).

Entre as diversas cultivares de milho encontradas no mercado, desde a década de 1960, os agricultores de milho têm predominantemente cultivado híbridos simples (SHULL, 1908) que são obtidos pelo cruzamento entre duas linhagens puras. Os híbridos simples exibem uma heterose superior em comparação com seus pais homozigotos, além de apresentar maior uniformidade em relação às variedades de polinização aberta (LI *et al.*, 2022). Entretanto, a recombinação entre um conjunto relativamente pequeno de linhagens parentais tem o potencial de gerar um grande número de híbridos, e com o advento da tecnologia duplo-haplóide, linhagens totalmente homozigotas podem ser geradas rapidamente, com baixo custo e em grande número (WĘDZONY *et al.*, 2009).

A avaliação de híbridos em campo sofre com dificuldades operacionais, pois, o grande número de cruzamentos possíveis demanda extensas áreas de testes, mão de obra, equipamento e cuidados que inflacionam os custos experimentais, inviabilizando financeiramente a avaliação desses híbridos (DOS SANTOS *et al.*, 2016). Assim, o desafio para o melhorista é: encontrar uma combinação promissora entre os pares de linhagens a partir da grande quantidade possível de híbridos simples que podem ser gerados dispondo de recursos limitados (BERNARDO, 1994; SCHRAG; SCHIPPRACK; MELCHINGER, 2019).

Além disso, outro complicador, é a necessidade de se avaliar tais híbridos em diferentes ambientes, pois estes podem induzir respostas fenotípicas distintas ao longo de ambientes, o que significa que um genótipo pode ser produtivo em um ambiente,

mas não em outro (CROSSA, JOSE, 1990; CRUZ; REGAZZI; CARNEIRO, 2012; MALOSETTI; RIBAUT; VAN EEUWIJK, 2013).

Para superar esse desafio, os melhoristas devem avaliar os híbridos desenvolvidos em ensaios de múltiplos ambientes (MET - *multiple environment trials*), e seus principais objetivos incluem a análise da interação entre genótipos e ambientes, bem como a avaliação da adaptabilidade e estabilidade genotípica (BURGUEÑO *et al.*, 2011). Contudo, a fenotipagem MET enfrenta obstáculos como a escassez de sementes, o alto número de genótipos a serem avaliados nos ensaios preliminares e os custos associados, resultando em desenhos experimentais desequilibrados em diferentes ambientes (JARQUIN *et al.*, 2020; KRAUSE *et al.*, 2020). Em ensaios desbalanceados, nos quais os híbridos não são avaliados em todos os ambientes, pode ser desafiador selecionar com precisão híbridos superiores para o próximo ciclo, pois alguns podem não ser estáveis em muitos ambientes e, genótipos descartados, podem ter um desempenho superior em ambientes não avaliados. Ademais, se considerarmos a necessidade de seleção de híbridos sob condições de estresse hídrico, a dificuldade aumenta, pois há uma baixa correlação entre a produção de grãos para híbridos cultivados em condições bem irrigadas e sob estresse hídrico (RIBAUT *et al.*, 2009).

Devido as dificuldades operacionais e ao elevado custo de fenotipagem em extensas áreas, a predição genômica vem ganhando cada vez mais espaço nos programas de melhoramento de plantas, pois ela permite a predição do fenótipo de indivíduos não testados em campo, podendo ser utilizada na identificação de híbridos promissores baseado apenas na informação genômica do indivíduo (MEUWISSEN; HAYES; GODDARD, 2001), levando ao aumento da intensidade de seleção e a redução de custos em condições experimentais. Com isso, é possível reduzir o tamanho da área de teste ao se reduzir o número de materiais avaliados em campo por meio da pré-seleção de híbridos via predição genômica, realizar a predição de híbridos em ambientes distintos e ainda manter uma alta concordância com os materiais que seriam selecionados apenas com base no fenótipo.

Até o momento, a predição do valor genético de indivíduos em MET resultou no surgimento de vários modelos, abrangendo desde modelos de interação entre genótipos e ambientes, efeitos aditivos e de dominância e covariáveis ambientais (BURGUEÑO *et al.*, 2012; CROSSA, JOSÉ; PÉREZ-RODRÍGUEZ; CUEVAS; MONTESINOS-LÓPEZ; JARQUIN; *et al.*, 2017; FERNANDES *et al.*, 2018; JARQUIN *et al.*, 2014,

2020). Porém, nestes modelos citados, para se obter uma estimativa mais precisa do real valor genético do indivíduo, todos os efeitos que podem afetar o fenótipo devem ser conhecidos a priori e modelados.

Uma abordagem que têm sido bem sucedida em reconhecer padrões complexos e tomar decisões corretas com base em dados é o aprendizado de máquina (CROSSA, JOSÉ; PÉREZ-RODRÍGUEZ; CUEVAS; MONTESINOS-LÓPEZ; JARQUÍN; *et al.*, 2017), que possibilita captar fatores complicadores tais como dominância e epistasia no modelo de predição. Além disso, tais algoritmos não possuem pressuposições quanto as variáveis de entrada no modelo, já que seus resultados dependem do processo de aprendizado e não da distribuição das variáveis em si.

Neste sentido, este estudo teve como objetivo avaliar e comparar a eficiência de metodologias estatísticas (GBLUP) e aprendizado de máquina para predição genômica para híbridos simples de milho avaliados para características de tolerância ao estresse hídrico em MET. Foram considerados dois cenários distintos de predição, que imitam duas situações que os melhoristas de plantas podem encontrar: (i) predizer o desempenho de híbridos simples de milho recentemente desenvolvidos e, para os quais não existem registros fenotípicos; (ii) predizer o desempenho de híbridos simples de milho em ensaios de design esparso, onde alguns híbridos foram avaliados em alguns ambientes, mas não em outros.

REFERÊNCIAS

BERNARDO, Rex. Prediction of maize single-cross performance using RFLPs and information from related hybrids. *Crop Science*, v. 34, n. 1, p. 20–25, 1994.

BURGUEÑO, Juan *et al.* Genomic Prediction of Breeding Values when Modeling Genotype × Environment Interaction using Pedigree and Dense Molecular Markers. *Crop Science*, v. 52, n. 2, p. 707–719, mar. 2012. Disponível em: <http://doi.wiley.com/10.2135/cropsci2011.06.0299>.

BURGUEÑO, Juan *et al.* Prediction Assessment of Linear Mixed Models for Multienvironment Trials. *Crop Science*, v. 51, n. 3, p. 944–954, maio 2011. Disponível em: <http://doi.wiley.com/10.2135/cropsci2010.07.0403>.

CROSSA, Jose. Statistical Analyses of Multilocation Trials. [S.l: s.n.], 1990. p. 55–85. Disponível em: <https://linkinghub.elsevier.com/retrieve/pii/S0065211308608184>.

CROSSA, José; PÉREZ-RODRÍGUEZ, Paulino; CUEVAS, Jaime; MONTESINOS-

LÓPEZ, Osval; JARQUIN, Diego; *et al.* Genomic Selection in Plant Breeding: Methods, Models, and Perspectives. *Trends in Plant Science*, v. 22, n. 11, p. 961–975, nov. 2017. Disponível em: <https://linkinghub.elsevier.com/retrieve/pii/S136013851730184X>.

CROSSA, José; PÉREZ-RODRÍGUEZ, Paulino; CUEVAS, Jaime; MONTESINOS-LÓPEZ, Osval; JARQUÍN, Diego; *et al.* Genomic Selection in Plant Breeding: Methods, Models, and Perspectives. *Trends in Plant Science*, v. 22, n. 11, p. 961–975, nov. 2017. Disponível em: <https://linkinghub.elsevier.com/retrieve/pii/S136013851730184X>.

CRUZ, C D; REGAZZI, A J; CARNEIRO, P C S. Modelos biométricos aplicados ao melhoramento. *UFV, Viçosa*, 2012.

DOS SANTOS, Jhonathan Pedroso Rigal *et al.* Inclusion of dominance effects in the multivariate GBLUP model. *PLoS One*, v. 11, n. 4, p. e0152045, 2016.

FERNANDES, Samuel B. *et al.* Efficiency of multi-trait, indirect, and trait-assisted genomic selection for improvement of biomass sorghum. *Theoretical and Applied Genetics*, v. 131, n. 3, p. 747–755, 7 mar. 2018. Disponível em: <http://link.springer.com/10.1007/s00122-017-3033-y>.

JARQUIN, Diego *et al.* A reaction norm model for genomic selection using high-dimensional genomic and environmental data. *Theoretical and Applied Genetics*, v. 127, n. 3, p. 595–607, 12 mar. 2014. Disponível em: <http://link.springer.com/10.1007/s00122-013-2243-1>.

JARQUIN, Diego *et al.* Genomic Prediction Enhanced Sparse Testing for Multi-environment Trials. *G3 Genes|Genomes|Genetics*, v. 10, n. 8, p. 2725–2739, 1 ago. 2020. Disponível em: <https://academic.oup.com/g3journal/article/10/8/2725/6048674>.

KRAUSE, Matheus Dalsente *et al.* Boosting predictive ability of tropical maize hybrids via genotype-by-environment interaction under multivariate GBLUP models. *Crop Science*, v. 60, n. 6, p. 3049–3065, 19 nov. 2020. Disponível em: <https://onlinelibrary.wiley.com/doi/10.1002/csc2.20253>.

LI, Chunhui *et al.* Genomic insights into historical improvement of heterotic groups during modern hybrid maize breeding. *Nature Plants*, v. 8, n. 7, p. 750–763, 18 jul. 2022. Disponível em: <https://www.nature.com/articles/s41477-022-01190-2>.

MALOSETTI, Marcos; RIBAUT, Jean-Marcel; VAN EEUWIJK, Fred A. The statistical analysis of multi-environment data: modeling genotype-by-environment interaction and its genetic basis. *Frontiers in Physiology*, v. 4, 2013. Disponível em: <http://journal.frontiersin.org/article/10.3389/fphys.2013.00044/abstract>.

MEUWISSEN, Theo H E; HAYES, Ben J; GODDARD, ME1461589. Prediction of total genetic value using genome-wide dense marker maps. *Genetics*, v. 157, n. 4, p. 1819–1829, 2001.

RIBAUT, Jean-Marcel *et al.* Drought Tolerance in Maize. *Handbook of Maize: Its Biology*. New York, NY: Springer New York, 2009. p. 311–344. Disponível em: <http://link.springer.com/10.1007/978-0-387-79418-1_16>.

SCHRAG, Tobias A; SCHIPPRACK, Wolfgang; MELCHINGER, Albrecht E. Across-years prediction of hybrid performance in maize using genomics. *Theoretical and Applied Genetics*, v. 132, n. 4, p. 933–946, 2019.

SHULL, George H. The composition of a field of maize. *Journal of Heredity*, n. 1, p. 296–301, 1908.

WĘDZONY, M *et al.* Progress in doubled haploid technology in higher plants. *Advances in haploid production in higher plants*, p. 1–33, 2009.

2. REFERENCIAL TEÓRICO

2.1. Híbrido simples

Híbridos simples (SHULL, 1908) são obtidos pelo cruzamento entre duas linhagens puras de grupos heteróticos diferentes. Esses grupos heteróticos são conjuntos de linhagens ou variedades que exibem forte heterose quando cruzadas entre si. Essas linhagens ou variedades são selecionadas com base em suas características genéticas complementares, de modo que a prole resultante do cruzamento entre elas demonstre heterose significativa. A formação e a utilização de grupos heteróticos são estratégias essenciais no melhoramento genético de milho e em programas de produção de sementes (BORÉM; MIRANDA; FRITSCHE-NETO, 2021).

Desde a sua proposta por Shull (1908), o principal obstáculo encontrado para a produção de híbridos simples, é avaliar todas as combinações híbridas possíveis à medida que o número de linhagens aumenta. Devido a inviabilidade de avaliação de todas as combinações híbridas em campo, os melhoristas podem optar por diferentes tipos de delineamentos genéticos, como as análises dialélicas propostas por Gardner e Eberhart (1966) e Hayman (1954) e os *top cross* (BORÉM; MIRANDA; FRITSCHE-NETO, 2021; CRUZ; REGAZZI; CARNEIRO, 2004). Essas abordagens têm o objetivo de avaliar os possíveis bons híbridos. Neste caso, o ideal seria que todos as linhagens e combinações híbridas fossem avaliadas, com o intuito de se realizar uma seleção mais acurada dos genótipos nos programas de melhoramento, entretanto, apenas uma parte dos cruzamentos são testados.

Com o advento da tecnologia duplo-haplóide, o problema de encontrar uma combinação híbrida promissora tornou-se ainda mais desafiador, pois linhagens totalmente homozigotas podem ser geradas rapidamente, a baixo custo e em grande número (WĘDZONY *et al.*, 2009). Com isso, muitas linhagens são produzidas em um curto espaço de tempo, e a grande maioria das linhagens de cada grupo heterótico passaram a ser linhagens “novas” sem nenhum registro fenotípico na progênie híbrida de ciclos de melhoramento anteriores.

2.2. Predição genômica

A Predição Genômica foi proposta por Meuwissen *et al.* (2001) e consiste na utilização de milhares de marcadores distribuídos ao longo de todo o genoma para predizer do mérito genético dos indivíduos. A predição genômica permite reduzir a

duração do ciclo de seleção, enquanto aumenta o ganho genético esperado e a resposta de seleção por unidade de tempo (HEFFNER; SORRELLS; JANNINK, 2009). Além de reduzir significativamente o custo de desenvolvimento de linhagens e híbridos (CROSSA; PÉREZ-RODRÍGUEZ; CUEVAS; MONTESINOS-LÓPEZ; JARQUÍN; *et al.*, 2017).

A predição genômica utiliza uma "população de treinamento" de indivíduos que foram genotipados e fenotipados para desenvolver um modelo que obtém informações genotípicas de uma população de indivíduos não testados ("população de teste") e produz valores genéticos estimados genômicos (GEBV – *genomic breeding values*). Para maximizar a acurácia do GEBV, a população de treinamento deve ser representativa da população de teste no programa de melhoramento que a predição genômica será aplicada.

Crossa *et al.* (2017) afirmam que a predição genômica, quando comparada com a seleção fenotípica, em termos de redução de custos no melhoramento de milho, permite que o melhorista cruze 50% de todas as linhagens disponíveis, avaliando-as em ensaios de múltiplos ambientes, e então use os dados fenotípicos para predizer os 50% restantes pela predição genômica, reduzindo assim o custo por ciclo e o tempo necessário para o desenvolvimento da variedade. Porém, vários fatores genéticos e estatísticos podem complicar a aplicação prática da predição genômica. Dificuldades genéticas surgem do tamanho e diversidade da população de treinamento e da herdabilidade das características que serão preditas. Os desafios estatísticos estão relacionados, principalmente, à alta dimensionalidade dos dados dos marcadores, na qual o número de marcadores (p) é muito maior que o número de observações (n) ($p >> n$) e a presença de multicolinearidade entre marcadores (CROSSA; PÉREZ-RODRÍGUEZ; CUEVAS; MONTESINOS-LÓPEZ; JARQUÍN *et al.*, 2017; JANNINK; LORENZ; IWATA, 2010).

Para superar esses problemas, muitos métodos estatísticos foram desenvolvidos para predizer indivíduos não observados. Entre eles, destaca-se o *Genomic Best Linear Unbiased Predictor* – GBLUP, que reduz as estimativas de coeficientes na direção de zero em relação às estimativas de mínimos quadrados ordinários, gera economia nos graus de liberdade e conduz a estimativas estáveis, permitindo estimar os parâmetros no caso $n >> N$, em que n é o número de covariáveis e N é o número de observações e quando há multicolinearidade entre as variáveis.

Outro algoritmo que têm sido bem-sucedido em reconhecer padrões complexos e tomar decisões corretas com base em dados é o aprendizado de máquina (SILVA *et al.*, 2017; SOUSA *et al.*, 2021), que possuem como atrativos a o fato do número de marcadores poder exceder o número das observações; podem acomodar interações complexas entre marcadores e não fazem suposições sobre as variáveis de entrada no modelo, uma vez que seus resultados são dependentes do processo de aprendizado e não da distribuição das variáveis em si.

No contexto de ensaios em múltiplos ambientes (MET), a predição genômica desempenha um papel fundamental no avanço dos programas de melhoramento genético de plantas. Essa abordagem utiliza informações genômicas para predizer o desempenho fenotípico de genótipos em diferentes ambientes, proporcionando diversos benefícios, como eficiência no melhoramento, reduz a dependência de ensaios extensivos e demorados em diferentes ambientes, além de antecipar resultados, pois os melhoristas conseguem tomar decisões mais precisas em estágios iniciais do processo de melhoramento, acelerando o desenvolvimento de novas variedades.

Neste sentido, a maioria das pesquisas que utilizam modelos de validação de predição genômica seguem alguns cenários propostos na literatura como por exemplo, os cenários CV1 e CV2 propostos por Burgueño *et al.* (2012), que simulam problemas que o melhorista pode enfrentar. No CV1, é avaliada a capacidade dos algoritmos em predizer o desempenho de genótipos que ainda não foram avaliados em nenhum ensaio de campo. Assim, as predições derivadas do cenário CV1 são inteiramente baseadas em registros fenotípicos e genotípicos de outros genótipos aparentados. Já em CV2, é avaliada a capacidade dos algoritmos de predizer o desempenho dos genótipos usando dados coletados em outros ambientes. Ele simula o problema de predição encontrado em ensaios MET incompletos. Aqui as informações de indivíduos aparentados são usadas, e a avaliação de predição pode se beneficiar do relacionamento genético entre genótipos e das correlações entre ambientes. Outros cenários conhecidos como CV0 e CV00 são mais desafiadores e simulam, respectivamente, a predição de genótipos avaliados em ambientes não avaliados e a predição de genótipos não avaliados para um ambiente também não avaliado (CROSSA; PÉREZ-RODRÍGUEZ; CUEVAS; MONTESINOS-LÓPEZ; JARQUIN *et al.*, 2017; FILHO *et al.*, 2023).

2.3. Aprendizado estatístico

2.3.1. *Genomic Best Linear Unbiased Predictor* - GBLUP

As equações de modelos mistos desenvolvidas por Henderson (1975) é uma abordagem muito utilizada para modelar a média em função das variáveis dependentes e que permitem a estimação dos efeitos fixos e a predição dos efeitos aleatórios do modelo conjuntamente.

O método GBLUP é um modelo linear misto que realiza a predição direta dos efeitos genéticos genômicos dos indivíduos. Para valores genéticos aditivos individuais ($\boldsymbol{u}_a$), tem-se:

$$\boldsymbol{y} = \boldsymbol{Xb} + \boldsymbol{Zu_a} + \boldsymbol{e},$$

em que, $\boldsymbol{y}$ $(i \times 1)$ é o vetor de observações fenotípicas de i indivíduos; $\boldsymbol{b}$ $(k \times 1)$ é o vetor de efeitos fixos em que k é o número de efeitos fixos considerados; $\boldsymbol{X}$ $(i \times k)$ é a matriz de incidência para os efeitos fixos. Para o caso de fenótipos corrigidos a matriz $\boldsymbol{X}$ se resume a 1 (vetor de mesma dimensão de $\boldsymbol{y}$ com todos os elementos iguais a 1) e $\boldsymbol{b}$ se resume a μ que é a média da característica de interesse; $\boldsymbol{u_a}$ é o vetor de valores genéticos genômicos aditivos individuais com matriz de incidência $\boldsymbol{Z}$ $(i \times i)$, com estrutura de variância dada por $\boldsymbol{u_a} \sim N(0, \boldsymbol{G_a}\sigma_a^2)$; $\boldsymbol{G_a}$ é a matriz de parentesco genômica entre os indivíduos para efeitos aditivos, σ_a^2 é a variância aditiva do caráter; $\boldsymbol{e}$ é o vetor de erros aleatório do modelo com $\boldsymbol{e} \sim N(0, \boldsymbol{I}\sigma_e^2)$, $\boldsymbol{I}$ é uma matriz identidade; σ_e^2 é a variância residual. Os componentes de variância são obtidos via o método de máxima verossimilhança restrita (REML)(CORBEIL; SEARLE, 1976).

No contexto da GWS, a inclusão de efeitos de dominância pode ser realizada no modelo GBLUP com a matriz de parentesco de dominância usando informações de marcadores (VITEZICA; VARONA; LEGARRA, 2013). Para o modelo GBLUP aditivo-dominante, os valores genéticos genômicos devido aos desvios de dominância ($\boldsymbol{u_d}$) são ajustados no modelo de predição e sua inclusão é dada por:

$$\boldsymbol{y} = \boldsymbol{Xb} + \boldsymbol{Zu_a} + \boldsymbol{Zu_d} + \boldsymbol{e}$$

em que, $\boldsymbol{y}$ $(i \times 1)$ é o vetor de observações fenotípicas de i indivíduos; $\boldsymbol{b}$ $(k \times 1)$ é o vetor de efeitos fixos em que k é o número de efeitos fixos considerados; $\boldsymbol{X}$ $(i \times k)$ é a matriz de incidência para os efeitos fixos. Para o caso de fenótipos corrigidos a matriz $\boldsymbol{X}$ se resume a 1 (vetor de mesma dimensão de $\boldsymbol{y}$ com todos os elementos iguais a 1) e $\boldsymbol{b}$ se resume a μ que é a média da característica de interesse; $\boldsymbol{u_a}$ é o vetor de valores genéticos genômicos aditivos individuais com matriz de incidência $\boldsymbol{Z}$ $(i \times i)$, com

estrutura de variância dada por $\boldsymbol{u_a} \sim N(0, \boldsymbol{G_a}\sigma_a^2)$; $\boldsymbol{G_a}$ é a matriz de parentesco genômica entre os indivíduos para efeitos aditivos, σ_a^2 é a variância aditiva do caráter; $\boldsymbol{u_d}$ é o vetor de valores genéticos devido à dominância dos indivíduos com matriz de incidência $\boldsymbol{Z}$ $(i \times i)$, com estrutura de variância dada por $\boldsymbol{u_d} \sim N(0, \boldsymbol{G_d}\sigma_d^2)$; $\boldsymbol{G_d}$ é a matriz de parentesco genômica para os efeitos devido à dominância e σ_d^2 é a variância devido à dominância.

As matrizes de parentesco genômica aditiva e devido aos desvios de dominância são dadas, respectivamente, por:

$$G_a = \frac{WW'}{\sum_{j=1}^{n} 2p_j(1-p_j)}$$

$$G_d = \frac{SS'}{\sum_{j=1}^{n} [2p_j(1-p_j)]^2}$$

Em que, n corresponde ao número de marcadores.

A parametrização para as matrizes de incidência W e S são apresentadas a seguir e está de acordo com a teoria clássica de genética quantitativa (FALCONER; MACKAY, 1996):

$$W = \begin{cases} Se\ MM, então\ 2 - 2p_j \rightarrow 2q_j \\ Se\ Mm, então\ 1 - 2p_j \rightarrow q_j - p_j \\ Se\ mm, então\ 0 - 2p_j \rightarrow -2p_j \end{cases}$$

$$S = \begin{cases} Se\ MM, então\ 0 \rightarrow -2q_j^2 \\ Se\ Mm, então\ 1 \rightarrow 2p_j q_j \\ Se\ mm, então\ 0 \rightarrow -2p_j^2 \end{cases}$$

Em que p_j e q_j são frequências alélicas de M e m, respectivamente.

Além dos modelos apresentados acima, a literatura apresenta também os modelos que conseguem estimar o valor genético dos genótipos em ensaios de múltiplos ambientes (MET). Dentre estes, encontra-se o modelo de Simetria Composta (SC), que prediz cada genótipo individualmente dentro de um local, juntamente com uma matriz de variância-covariância desses efeitos genéticos entre pares de locais (GEZAN; DE CARVALHO; SHERRILL, 2017). Neste caso, a matriz de variância-covariância considera a não heterogeneidade de variâncias entre os diferentes locais e uma única covariância entre eles.

O modelo linear de simetria composta é descrito abaixo:

$$\boldsymbol{y} = \mu\boldsymbol{1} + \boldsymbol{Xb} + \boldsymbol{Z_1 u_a} + \boldsymbol{Z_2 u_d} + \boldsymbol{e},$$

em que $\boldsymbol{y}$ $(ik \times 1)$ é o vetor de observações fenotípicas de i indivíduos em k ambientes; μ é a média geral; $\boldsymbol{b}$ $(k \times 1)$ é o vetor de efeitos ambientais (fixo); $\boldsymbol{u_a}$ $(ik \times 1)$ é o vetor de valores genéticos genômicos aditivos individuais aninhado aos ambientes (aleatório), com $\boldsymbol{u_a} \sim MVN(\boldsymbol{0}, [\boldsymbol{I_k}\sigma^2_{u_a} + \rho_a(\boldsymbol{J_k} - \boldsymbol{I}_k)] \otimes \boldsymbol{A})$, em que $\boldsymbol{A}$ é a matriz de parentesco genômica entre indivíduos para efeitos aditivos, ρ_a é o coeficiente de correlação genética aditiva entre os ambientes, $\boldsymbol{I_q}$ $(k \times k)$ é uma matriz identidade, $\boldsymbol{J_q}$ $(k \times k)$ é uma matriz de uns, e $\otimes$ denota o produto de Kronecker; $\boldsymbol{u_d}$ $(ik \times 1)$ é o vetor de valores genéticos genômicos de dominância individuais aninhado aos ambientes (aleatório), com $\boldsymbol{u_d} \sim MVN(\boldsymbol{0}, [\boldsymbol{I_k}\sigma^2_{u_d} + \rho_d(\boldsymbol{J_k} - \boldsymbol{I_k})] \otimes \boldsymbol{D})$, onde $\boldsymbol{D}$ que é a matriz de parentesco genômica entre os indivíduos para efeitos devido aos desvios de dominância, $\rho_{\boldsymbol{d}}$ é o coeficiente de correlação genética para efeitos devido aos desvios de dominância entre os ambientes; $\boldsymbol{e}$ $(ik \times 1)$ é o vetor de erros aleatórios $\boldsymbol{e} \sim MVN(\boldsymbol{0}, \boldsymbol{I}\sigma^2_e)$. As letras $\boldsymbol{X}$ $(ik \times k)$, $\boldsymbol{Z_1}(ik \times ik)$ e $\boldsymbol{Z_2}$ $(ik \times ik)$ representam as matrizes de incidência para seus respectivos efeitos, $\boldsymbol{1}$ $(ik \times 1)$ é um vetor de uns.

2.4. Aprendizado de máquina

2.4.1. Árvore de decisão e seus refinamentos

A árvore de decisão (AD) é um método de aprendizado supervisionado não paramétrico utilizado para classificação e regressão (GARETH *et al.*, 2013). A estrutura da AD é composta pelos nós internos, ramos e nós externos/folhas. O nó é dito interno, quando os dados contidos neste nó são divididos de acordo com um critério de divisão, formando assim dois novos grupos de dados, sendo estes novos grupos ligados ao grupo antigo pelos ramos, já o nó é dito externo (folha) quando não ocorre mais divisões dos indivíduos pertencentes a este nó. A AD pode ser classificada como árvore de regressão quando a variável resposta é do tipo quantitativa e, quando a variável dependente assume valores qualitativos, a AD é chamada de árvore de classificação.

A árvore de regressão é uma metodologia que particiona o espaço preditor em sub-regiões através de alguns critérios, para cada sub-região formada é atribuído um valor que será utilizado como valor predito para os novos indivíduos que serão alocados a essas sub-regiões. Para a construção da árvore de regressão, o objetivo é

construir regiões R_1, R_2, ..., R_J que minimizam a Soma de Quadrados dos Resíduos dado por:

$$\sum_{j=1}^{J}\sum_{i\in R_j}\left(y_i - \hat{y}_{R_j}\right)^2,$$

em que $\hat{y}_{R_j}$ é a média da variável resposta das observações de treinamento pertencente a *j*-ésima região. Entretanto, o custo computacional para se obter a menor soma de quadrados dos resíduos (SQR), é muito alto quando se considera cada partição possível do espaço em J regiões. Por esta razão, a literatura propõe realizar um procedimento baseado em divisões binárias recursivas, que tem como objetivo, selecionar o preditor X_j e o ponto s, que divida o espaço preditor em duas regiões:

$$R_1(j,s) = \{X|X_j < s\}\ e\ R_2(j,s) = \{X|X_j \geq s\}$$

tal que o ponto s que divide a *j*-ésima variável em duas regiões é aquele que apresenta a menor soma de quadrados dos resíduos, por fim se utiliza a variável que obteve o menor SQR para a primeira divisão. Em seguida, em vez de se dividir todo o espaço preditor, divide-se então, uma das duas regiões previamente identificadas e o processo é repetido para cada região gerada. Para evitar o superajustamento do modelo, indica-se que nenhuma região obtenha mais que 5 indivíduos e em seguida deve-se podar a árvore usando o *custo complexidade* da poda.Este, envolve atribuir um custo a cada subárvore da árvore totalmente crescida e, em seguida, selecionar a subárvore com o menor custo como a árvore podada. O custo de uma subárvore é determinado por um parâmetro de complexidade e pelo número de nós folha na subárvore (HASTIE *et al.*, 2009).

O *bagging* é uma metodologia que tem como objetivo reduzir a variância da AD. Em outras palavras, ele consiste em obter B amostras com reposição da amostragem disponível, obtendo assim B modelos $\hat{f}^1(x), \hat{f}^2(x), \ldots, \hat{f}^B(x)$ e, por fim utilizar os modelos gerados para obter uma média das predições, dado por:

$$\hat{f}_{medio}(x) = \frac{1}{B}\sum_{b=1}^{B}\hat{f}^b(x)$$

desta forma, busca-se diminuir a variabilidade obtida nas árvores de decisão.

A quantidade de árvores utilizadas no *bagging* não é uma parâmetro que irá resultar num superajustamento do modelo, na pratica é utilizado uma quantidade onde o erro tenha estabilizado (GARETH *et al.*, 2013).

O *random forest* foi proposto por Ho (1995) e é um aperfeiçoamento do *bagging*, com o intuito de evitar a alta correlação das AD e de melhorar a acurácia na seleção dos indivíduos. O *random forest* altera somente o número de variáveis preditoras ($m < p$) utilizadas em cada partição, obtendo os valores preditos mais independentes, ocasionando assim na redução da variabilidade encontrada nas AD. Hastie *et al*. (2009) sugerem que o número de variáveis preditoras utilizadas em cada partição seja dada da seguinte forma para árvores de regressão $m = p/3$.

REFERÊNCIAS

BORÉM, Aluízio; MIRANDA, Glauco V; FRITSCHE-NETO, Roberto. *Melhoramento de plantas*. [S.l.]: Oficina de Textos, 2021.

BURGUEÑO, Juan *et al.* Genomic Prediction of Breeding Values when Modeling Genotype × Environment Interaction using Pedigree and Dense Molecular Markers. *Crop Science*, v. 52, n. 2, p. 707–719, mar. 2012. Disponível em: <http://doi.wiley.com/10.2135/cropsci2011.06.0299>.

CORBEIL, R. R.; SEARLE, S. R. Restricted Maximum Likelihood (REML) Estimation of Variance Components in the Mixed Model. *Technometrics*, v. 18, n. 1, p. 31, fev. 1976. Disponível em: <https://www.jstor.org/stable/1267913?origin=crossref>.

CROSSA, José; PÉREZ-RODRÍGUEZ, Paulino; CUEVAS, Jaime; MONTESINOS-LÓPEZ, Osval; JARQUÍN, Diego; *et al.* Genomic Selection in Plant Breeding: Methods, Models, and Perspectives. *Trends in Plant Science*, v. 22, n. 11, p. 961–975, nov. 2017. Disponível em: <https://linkinghub.elsevier.com/retrieve/pii/S136013851730184X>.

CROSSA, José; PÉREZ-RODRÍGUEZ, Paulino; CUEVAS, Jaime; MONTESINOS-LÓPEZ, Osval; JARQUIN, Diego; *et al.* Genomic Selection in Plant Breeding: Methods, Models, and Perspectives. *Trends in Plant Science*, v. 22, n. 11, p. 961–975, nov. 2017. Disponível em: <https://linkinghub.elsevier.com/retrieve/pii/S136013851730184X>.

CRUZ, C D; REGAZZI, A J; CARNEIRO, P C S. Modelos biométricos aplicados ao melhoramento genético (volume 1. *Viçosa, Editora UFV*, v. 1, p. 480p, 2004.

FALCONER, D S; MACKAY, T F C. Introduction to quantitative genetics. Essex. *UK: Longman Group*, 1996.

FILHO, Claudio Carlos Fernandes *et al.* Genomic prediction for complex traits across multiples harvests in alfalfa (Medicago sativa L.) is enhanced by enviromics. *The Plant Genome*, v. 16, n. 2, 22 jun. 2023. Disponível em: <https://acsess.onlinelibrary.wiley.com/doi/10.1002/tpg2.20306>.

GARETH, James *et al. An introduction to statistical learning: with applications in R*.

[S.l.]: Spinger, 2013.

GEZAN, Salvador Alejandro; DE CARVALHO, Melissa Pisaroglo; SHERRILL, Josh. Statistical methods to explore genotype-by-environment interaction for loblolly pine clonal trials. *Tree Genetics & Genomes*, v. 13, n. 1, p. 1, 6 fev. 2017. Disponível em: <http://link.springer.com/10.1007/s11295-016-1081-0>.

HASTIE, Trevor *et al. The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. 2. ed. New York: Springer-Verlag New York, 2009.

HEFFNER, Elliot L.; SORRELLS, Mark E.; JANNINK, Jean-Luc. Genomic Selection for Crop Improvement. *Crop Science*, v. 49, n. 1, p. 1–12, jan. 2009. Disponível em: <http://doi.wiley.com/10.2135/cropsci2008.08.0512>.

HENDERSON, Charles R. Best linear unbiased estimation and prediction under a selection model. *Biometrics*, p. 423–447, 1975.

HO, Tin Kam. Random decision forests. 1995, [S.l.]: IEEE, 1995. p. 278–282.

JANNINK, J.-L.; LORENZ, A. J.; IWATA, H. Genomic selection in plant breeding: from theory to practice. *Briefings in Functional Genomics*, v. 9, n. 2, p. 166–177, 1 mar. 2010. Disponível em: <https://academic.oup.com/bfg/article-lookup/doi/10.1093/bfgp/elq001>.

MEUWISSEN, Theo H E; HAYES, Ben J; GODDARD, ME1461589. Prediction of total genetic value using genome-wide dense marker maps. *Genetics*, v. 157, n. 4, p. 1819–1829, 2001.

SHULL, George H. The composition of a field of maize. *Journal of Heredity*, n. 1, p. 296–301, 1908.

SILVA, Gabi Nunes *et al.* Artificial neural networks compared with Bayesian generalized linear regression for leaf rust resistance prediction in Arabica coffee. *Pesquisa Agropecuária Brasileira*, v. 52, n. 3, p. 186–193, mar. 2017. Disponível em: <http://www.scielo.br/scielo.php?script=sci_arttext&pid=S0100-204X2017000300186&lng=en&tlng=en>.

SOUSA, Ithalo Coelho De *et al.* Genomic prediction of leaf rust resistance to Arabica coffee using machine learning algorithms. *Scientia Agricola*, v. 78, n. 4, 2021. Disponível em: <http://www.scielo.br/scielo.php?script=sci_arttext&pid=S0103-90162021000401102&tlng=en>.

VITEZICA, Zulma G; VARONA, Luis; LEGARRA, Andres. On the additive and dominant variance and covariance of individuals within the genomic selection scope. *Genetics*, v. 195, n. 4, p. 1223–1230, 2013.

WĘDZONY, M *et al.* Progress in doubled haploid technology in higher plants. *Advances in haploid production in higher plants*, p. 1–33, 2009.

3. ARTIGO DE PESQUISA: Genomic prediction in multi-environment trials in maize using statistical and machine learning methods

Scientific Reports: Doi: 10.1038/s41598-024-51792-3

Cynthia Aparecida Valiati Barreto[1], Kaio Olimpio das Graças Dias[2], Ithalo Coelho de Sousa[3], Camila Ferreira Azevedo[1], Ana Carolina Campana Nascimento[1], Lauro José Moreira Guimarães[4], Claudia Teixeira Guimarães[4], Maria Marta Pastina[4], Moysés Nascimento[1]

[1] Universidade Federal de Viçosa, Department of Statistics, Viçosa, Minas Gerais, Brazil.
[2] Universidade Federal de Viçosa, Department of General Biology, Viçosa, Minas Gerais, Brazil.
[3] Universidade Federal de Rondônia, Department of Mathematics and Statistics, Ji-Paraná, RO, Brazil.
[4] Embrapa Maize and Sorghum, Sete Lagoas, Minas Gerais, Brazil.

*Corresponding author:
e-mail: moysesnascim@ufv.br

3.1. Abstract

In the context of multi-environment trials (MET), genomic prediction is proposed as a tool that allows the prediction of the phenotype of single cross hybrids that were not tested in field trials. This approach saves time and costs compared to traditional breeding methods. Thus, this study aimed to evaluate the genomic prediction of single cross maize hybrids not tested in MET for grain yield and female flowering time. We also aimed to propose an application of machine learning methodologies in MET in the prediction of hybrids and compare their performance with Genomic best linear unbiased prediction (GBLUP) with non-additive effects. Our results highlight that both methodologies are efficient and can be used in maize breeding programs to accurately predict the performance of hybrids in specific environments. The best methodology is case-dependent, specifically, to explore the potential of GBLUP, it is important to perform accurate modeling of the variance components to optimize the prediction of new hybrids. On the other hand, machine learning methodologies can capture non-additive effects without making any assumptions at the outset of the model. Overall, predicting the performance of new hybrids that were not evaluated in any field trials was more challenging than predicting hybrids in sparse test designs.

3.2. Introduction

Maize (*Zea mays*) has emerged as an important crop for food, feed production, and various industrial applications, providing livelihoods for millions of people around the world[1,2]. However, its production is affected by several factors, with drought being one of the most common causes of agricultural shortages in rainfed systems[3]. This fact, combined with the high demand for this crop and the prospect of a worldwide growth of more than 2 billion people over the next 20 years[4], makes it necessary to cultivate increasingly productive crops, as well as more adapted to climate change and also to different planting regions, such tropical conditions.

Cultivars can exhibit differentiated phenotypic responses between environments, and it is possible that a genotype may perform well in one environment but not in another[5–7]. To address this, breeders must submit the developed hybrids to multiple environment trials (MET). In MET, the main objectives are to study the interaction between genotypes and environments and to evaluate genotypic overall performance

and stability [8]. However, MET phenotyping faces challenges such as the limited seeds availability, a high number of genotypes to be tested in the preliminary trials, and the associated costs, resulting in unbalanced experimental designs in different environments (Jarquin *et al.*, 2020; Krause *et al.*, 2020a). In sparse designs, where hybrids are not evaluated in all environments, accurately selecting superior hybrids for the next cycle can be difficult, as some hybrids may not be stable in many environments, and other genotypes that are discarded may outperform in untested environments.

To address these challenges, genomic prediction (GP) is proposed as a tool to predict the genetic value of individuals that were not evaluated in the field[11,12]. Several GP methods have been proposed, with Genomic Best Linear Unbiased Predictor (GBLUP) being one of the most commonly used methods. In the context of MET, predicting the genetic value of individuals not observed in specific environments has led to the development of several models, including interaction model of genotypes by environments, environmental covariates, and additive and dominance effects [9,13–16].

Recently, machine learning methodologies have gained attention due to their ability to recognize complex interaction structures in data sets[17]. Machine learning algorithms approximate the mapping function linking input variables (e.g., phenotypic trait) to output variables from the training datasets without making a priori assumptions about data distribution or the genotype-phenotype relationship [18]. This flexibility allows these methods to capture more complex genetic architectures in prediction models.

Among the machine learning methodologies used in genomic prediction, decision trees and their refinements (such as bootstrap aggregation (bagging), random forest, and boosting) stand out, as they stratify the predictor space into many sub-regions [19,20]. These refinements aim to build more accurate prediction models; for example, bagging (Bag) reduces the variance observed in decision trees, random forest (RF) improves accuracy by avoiding high tree correlation [21], and boosting (Boost) builds trees sequentially using information from previously built trees[22].

Studies using simulated and real data have concluded that tree-based machine learning tools can serve as an alternative to traditional techniques for genomic prediction[23–26]. For instance, Sousa et al. [27], evaluating genomic prediction for

resistance to rust disease in *Coffea arabica*, observed that Bag, RF, and Boost showed superior predictive abilities to Generalized Bayesian Linear Regression. Westhues et al.[28], using genomic and environmental variables, found that machine learning models can provide similar or slightly superior predictive abilities to GBLUP models for traits strongly influenced by environmental factors. Despite the potential of tree-based machine learning, there are still few studies that have evaluated MET data, and these methodologies could prove beneficial in such cases.

Therefore, our study aimed to evaluate and to compare the efficiency of statistical methodologies (GBLUP) and machine learning (Bag, RF, and Boost) for genomic prediction for single cross hybrid evaluated for drought tolerance traits in MET. We considered two different prediction scenarios, mimicking two situations that plant breeders may encounter: (i) predicting the performance of newly developed single hybrids for which there are no existing phenotypic records; (ii) predicting the performance of single cross hybrids in sparse design trials, where some hybrids were evaluated in some environments, but not in others.

3.3. Material and methods

3.3.1. Phenotypic data

The data are composed of 265 single cross hybrids from the maize breeding program of Embrapa Maize and Sorghum evaluated in eight combinations of trials/locations/years under irrigated trials (WW) and water stress (WS) conditions at two locations in Brazil (Janaúba – Minas Gerais and Teresina – Piauí) over two years (2010 and 2011). The hybrids were obtained from crosses between 188 inbred lines and two testers. The inbred lines belong to heterotic groups: dent (85 inbred lines), flint (86 inbred lines), and an additional group, referred to as group C (17 inbred lines), which is unrelated to the dent and flint origins. The two testers are inbred lines belonging to the flint (L3) and dent (L228-3) groups. Among the inbred lines, 120 were crossed with both testers, 52 were crossed with the L228-3 tester only, and 16 lines were crossed with the L3 tester only. As demonstrated by Silva et al. (2020)[29], there are subgroups within each heterotic group from the maize diversity panel from Embrapa's Maize and Sorghum breeding program. Once these groups were not well defined genetically, we assume that there is the same effect of allelic substitution in

both groups. More details on the experimental design and procedures can be found in Dias et al.[13,30].

The experiment originally included 308 entries, but hybrids that were not present in all environments were also removed to evaluate the genomic prediction within each environment, resulting in a total of 265 hybrids for analysis. Each trial consisted of 308 maize single cross hybrids, randomly divided into six sets: sets 1-3 for crosses with L3 (61, 61, and 14 hybrids each), and sets 4–6 for crosses with L228-3 (80, 77, and 15 hybrids each). Four checks (commercial maize cultivars) were included in each set, and the experiment was designed in completely randomized blocks. Between trials, hybrids within each set remained the same, but hybrids and checks were randomly allocated into groups of plots within each set. This allocation varied between replicates of sets and between trials. The WS trials had three replications, except for the set containing 15 hybrids and the trials evaluated in 2010, which had two replications. All WW trials, except for the trial in 2011, had two replicates.

Two agronomic traits related to drought tolerance were analyzed: grain yield (GY) and female flowering time (FFT). GY was determined by weighing all grains in each plot, adjusted for 13% grain moisture, and converted to tons per hectare (t/ha), accounting for differences in plot sizes across trials. FFT was measured as the number of days from sowing until the stigmas appeared in 50% of the plants. A summary of means, standard deviations, and ranges of both evaluated traits are available in Table 1.

To conduct the analyses, hybrids considered as outliers were removed (i.e., hybrids that presented phenotypic values greater than 1.5 × interquartile range above the third quartile or below the first quartile) for the GY and FFT traits. The variations in predictive abilities among hybrids of T2, T1, and T0 are widely recognized[31]. However, the primary aim of our study was to compare different prediction methodologies in MET assays. In this study, there were 240 T2 hybrids and 68 T1 hybrids, with T2 hybrids had both parents evaluated in different hybrid combinations, while hybrids being single-cross hybrids sharing one parent with the tested hybrids. Given the realistic nature of our scenario, we have a limited and imbalanced distribution of these hybrid groups, making a fair comparison challenging. Consequently, we opted to construct a training set comprising T2 and T0 hybrids.

Table 1. Summary of means (Mean), standard deviations (SD), minimum (Min), and maximum (Max) for grain yield (GY) and female flowering time (FFT) obtained under irrigated (WW) and water stress (WS) conditions, evaluated in the years 2010 and 2011, at the locations of Janaúba (J) and Teresina (T).

		WS				**WW**			
		2010		2011		2010		2011	
		J	T	J	T	J	T	J	T
GY	*Mean*	3.49	2.86	3.34	4.93	6.59	6.63	6.98	5.56
	SD	1.34	1.65	1.22	1.63	1.63	1.70	1.68	1.92
	Min	0.24	0.12	0.17	0.30	0.98	2.02	2.20	1.05
	Max	8.36	9.57	6.82	9.42	11.55	12.49	10.86	10.24
FFT	*Mean*	64.20	54.76	70.59	53.50	64.62	54.16	68.36	52.40
	SD	2.61	5.30	3.25	2.54	3.07	2.77	2.19	2.10
	Min	59.00	50.00	62.00	48.00	58.00	49.00	63.00	46.00
	Max	73.00	69.00	88.00	64.00	74.00	64.00	75.00	59.00

3.3.2. Statistical analysis of phenotypic data

To correct the phenotypic values for experimental design effects, each trial (WW and WS) and environment were analyzed independently to obtain the Best Linear Unbiased Estimator (eBLUEs) for each hybrid, for the two traits evaluated. The estimates were obtained based on the following model:

$$\boldsymbol{y} = \mathbf{1}\mu + \boldsymbol{X_1 r} + \boldsymbol{X_2 s} + \boldsymbol{X_3 h} + \boldsymbol{e} \quad (1)$$

where $\boldsymbol{y}$ $(n \times 1)$ is the phenotype vector for f replicates, t sets of p hybrids, and n is the number of observations; μ is the mean; $\boldsymbol{r}$ $(f \times 1)$ is the fixed effect vector of the replicates; $\boldsymbol{s}$ $(t \times 1)$ is the fixed effect vector of the sets; $\boldsymbol{h}$ $(p \times 1)$ is the fixed effect vector of the hybrids; and $\boldsymbol{e}$ $(k \times 1)$ is the residue vector, with $\boldsymbol{e} \sim , MVN(0, \boldsymbol{I}\sigma_e^2)$, where $\boldsymbol{I}$ is an identity matrix of corresponding order, and σ_e^2 the residual variance. $\boldsymbol{X_1}$ $(k \times f)$, $\boldsymbol{X_2}$ $(k \times t)$ e $\boldsymbol{X_3}$ $(k \times p)$ represents incidence matrices for their respective effects. The eBLUES of each environment were used in further analyses.

3.3.3. Genotypic data

A total of 57,294 Single Nucleotide Polymorphisms (SNPs) markers were obtained from 188 inbred lines, and two testers used as parents of the 265 single cross hybrids. The genotyping by sequencing (GBS) strategy are detailed in Dias et al.[13]. For

the quality control, SNPs were discarded if: the minor allele frequency was smaller than 5%, more than 20% of missing genotypes were found, and/or there were more than 5% of heterozygous genotypes. After filtering, missing data were imputed using NPUTE. Then, for each SNP, the genotypes of the hybrids were inferred based on the genotype of their parents (inbred line and tester). The number of SNPs per chromosome ranged from 3121 (chromosome 10) to 7705 (chromosome 1), totalizing 47,127 markers.

3.3.4. Genomic relationship matrix

The additive and dominance genomic relationship matrices were constructed [32] based on information from the SNPs using the package AGHmatrix[33], following VanRaden[34] and Vitezica et al., respectively.

3.3.5. Genomic Prediction

Genomic predictions were performed using the Genomic Best Line Unbiased Prediction (GBLUP) method using the package AsReml v. 4[35]. Two groups were considered: the first group comprised four environments under WW conditions, and the second included four environments under WS conditions. The linear model is described below:

$$\bar{\boldsymbol{y}} = \mu\mathbf{1} + \boldsymbol{Xb} + \boldsymbol{Z_1 u_a} + \boldsymbol{Z_2 u_d} + \boldsymbol{e} \quad (2)$$

where $\bar{\boldsymbol{y}}$ $(pq \times 1)$ is the vector of eBLUES previously estimated for each environment with p hybrids and q environments; μ is the mean; $\boldsymbol{b}$ $(q \times 1)$ is the vector of environmental effects (fixed); $\boldsymbol{u_a}$ $(pq \times 1)$ is the vector of individual additive genetic values nested within environments (random), with $\boldsymbol{u_a} \sim MVN(0, [\boldsymbol{I_q}\sigma^2_{u_a} + \rho_a(\boldsymbol{J_q} - \boldsymbol{I_q})] \otimes \boldsymbol{A})$, where $\boldsymbol{A}$ is the genomic relationship matrix between individuals for additive effects, ρ_a is the additive genetic correlation coefficient between environments, $\boldsymbol{I_q}$ $(q \times q)$ is an identity matrix, $\boldsymbol{J_q}$ $(q \times q)$ is a matrix of ones, and $\otimes$ denotes the Kronecker product; $\boldsymbol{u_d}$ $(pq \times 1)$ is the vector of individual dominance genetic values nested within environments (random), with $\boldsymbol{u_d} \sim MVN(0, [\boldsymbol{I_q}\sigma^2_{u_d} + \rho_d(\boldsymbol{J_q} - \boldsymbol{I_q})] \otimes \boldsymbol{D})$, where $\boldsymbol{D}$ is the genomic relationship matrix between individuals

for dominance effects, $\rho_{\boldsymbol{d}}$ is the dominance correlation coefficient between environments; $\boldsymbol{e}$ $(pq \times 1)$ is the random residuals vector with $\boldsymbol{e} \sim MVN(0, \boldsymbol{I}\sigma_e^2)$. The capital letters $\boldsymbol{X}$ $(pq \times q)$, $\boldsymbol{Z_1}(pq \times pq)$ and $\boldsymbol{Z_2}$ $(pq \times pq)$ represent the incidence matrices for their respective effects, $\mathbf{1}$ $(pq \times 1)$ is a vector of ones. The (co)variance components were obtained using the residual maximum likelihood method (REML)[36].

Two alternative models were also used. The first for genomic prediction retained only additive effects by removing $\boldsymbol{u_d}$ from Equation (2). The second model was used to estimate the genetic parameters within each environment separately.

The significance of random effects was tested using the Likelihood Ratio Test (LRT) [37], given by:

$$LRT = 2 * (LogL_c - LogL_r) \tag{3}$$

where $LogL_c$ is the logarithm of the likelihood function of the complete model (with all effects included), and $LogL_r$ is the logarithm of the restricted likelihood function of the reduced model (without the effect under test). Effect significance was tested by LRT using the chi-square (χ^2) probability density function with a degree of freedom and significance level of 5% [38].

The narrow-sense heritability (h^2), the proportion of variance explained by dominance effects (d^2), and the broad-sense heritability (H^2) for each trait were estimated following Falconer and Mackay,1996 [39].

3.3.6. Machine learning

Similar to the previous topic, the trials were divided between WW and WS conditions, and the potential of regression trees (RT) was explored using the following three algorithms: bagging, random forest, and boosting[22]. Bagging (Bag) is a methodology that aims to reduce the RT variance [22]. In other words, it consists of obtaining D samples with available sampling replacement, thus obtaining D models $\hat{f}^1(x), \hat{f}^2(x), \ldots, \hat{f}^D(x)$, and finally use the generated models to obtain an average, given by:

$$\hat{f}_{medio}(x) = \frac{1}{D}\sum_{d=1}^{D} \hat{f}^d(x) \tag{4}$$

This decreases the variability obtained in the decision trees. The number of trees used in Bag is not a parameter that will result in overfitting of the model. In practice, a number of trees is used until the error has stabilized[22]. The number of trees sampled for Bag was set at 500 trees.

Random forest (RF) was proposed by HO[40] and it is an improvement of Bag to avoid the high correlation of the trees and to improve the accuracy in the selection of individuals. RF changes only the number of predictor variables used in each split. That is, each time a split in a tree is considered, a random sample of m variables is chosen as candidates from the complete set of p variables. Hastie et al.[21] suggest that the number of predictor variables used in each partition is equal to $m = \frac{p}{3}$ for regression trees. The number of trees for the RF was set at 500.

Boosting uses RT by adjusting the residual of an initial model. The residual is updated with each tree that grows sequentially from the previous tree's residual, and the response variable involves a combination of a large number of trees, such that:

$$\hat{f}(x) = \sum_{b=1}^{B} \lambda \, \hat{f}^{b}(x) \tag{5}$$

The function $\hat{f}(.)$ refers to the final tree combined with sequentially adjusted trees, and λ is the shrinkage parameter that controls the learning rate of the method. Furthermore, this method needs to be adjusted with several splits in each of the trees. This parameter controls the complexity of the Boost and is known as the depth. For Boosting, the number of trees sampled was 250, with a learning rate of 0.1 and a depth of 3.

To perform hybrid prediction for each environment based on MET dataset, we propose the incorporation of location and year information in which the experiments were carried out as factors in the data input file together with SNPs markers as predictors in machine learning methodologies. As a response variable, the eBLUEs previously estimated by Equation (1) were used.

For the construction of the bagging and random forest models, the *randomForest* function from the package randomForest[41] was used. Finally, the package's gbm function *gbm* [42] was used for boosting. All analyzes were implemented in the software R[43].

3.3.7. Model validation

Genomic predictions were carried out following Burgueño et al.[16], considering two different prediction problems, CV1 and CV2, which simulate two possible scenarios a breeder can face. In CV1, the ability of the algorithms to predict the performance of hybrids that have not yet been evaluated in any field trial was evaluated. Thus, predictions derived from the CV1 scenario are entirely based on phenotypic and genotypic records from other related hybrids. In CV2, the ability of the algorithms to predict the performance of hybrids using data collected in other environments was evaluated. It simulates the prediction problem found in incomplete MET trials. Here, information from related individuals is used, and the prediction can benefit from genetic relationships between hybrids and correlations between environments. Within the CV2 scenario, two different situations of data imbalance were evaluated. In the first, called CV2 (50%), the tested hybrids were not present in half of the environments, while in the second, called CV2 (25%), the tested hybrids were not present in only 25% of the environments. Table 2 provides a hypothetical representation of this CV1, CV2 (50%), and CV2 (25%) validation scheme.

Table 2. Representation of the three scenarios (CV1, CV2-50% and CV2-25%) for four hybrids (Hybrid 1-4) and four environments (J10, J11, T10, T11).

CV1				
	J10	**J11**	**T10**	**T11**
Hybrid 1	H_{11}	H_{12}	H_{13}	H_{14}
Hybrid 2	NA	NA	NA	NA
Hybrid 3	H_{31}	H_{32}	H_{33}	H_{34}
Hybrid 4	H_{41}	H_{42}	H_{43}	H_{44}
CV2 (50%)				
	J10	**J11**	**T10**	**T11**
Hybrid 1	H_{11}	H_{12}	H_{13}	H_{14}
Hybrid 2	NA	H_{22}	H_{23}	NA
Hybrid 3	H_{31}	H_{32}	H_{33}	H_{34}
Hybrid 4	H_{41}	NA	NA	H_{44}
CV2 (25%)				
	J10	**J11**	**T10**	**T11**
Hybrid 1	H_{11}	H_{12}	H_{13}	NA
Hybrid 2	NA	H_{22}	H_{23}	H_{24}
Hybrid 3	H_{31}	H_{32}	NA	H_{34}

Hybrid 4	H_{41}	NA	H_{43}	H_{44}

Hybrids with phenotypes not observed in the scenario are indicated with NA (not evaluated); Hybrids with observed phenotypes are named as H_{ij} for i,j = 1, 2, 3 e 4.

To separate the training and validation sets, the k-folds procedure was used, considering $k = 5$. The set of 265 hybrids was divided into five groups, with 80% of the hybrids considered as the training population, and the remaining 20% hybrids considered as the validation population. The hybrids were separated into sets proportionally containing all the crosses performed (*Dent* × *Dent*, *Dent* × *Flint*, *Flint* × *Flint*, C × *Dent*, C × *Flint*). The cross-validation process was performed separately for each trait, condition (WS or WW) and scenario (CV1, CV2-50% and CV2-25%) and was repeated five times to assess the predictive ability of the analyses.

The predictive ability within each environment for the conditions (WS and WW) was estimated by the Pearson correlation coefficient[44] between the corrected phenotypic values (eBLUES) of Equation (1) for each environment and the GEBVs predicted by each fitted method.

3.4. Results

3.4.1. Variance components and estimation of genetic parameters

Estimates of variance components and genetic parameters for GY and FFT under WW and WS conditions, obtained for the joint analysis with the four environments and analyses within each environment, are shown in Table 3. For the joint analysis, the heritability estimates for GY and FFT were slightly different from those obtained by Dias et al.[13] using the same material, since here, a different statistical model was used to estimate the genetic parameters, and hybrids that were not present in all environments were removed.

The additive variance found for GY and FFT was greater than the variance due to dominance effects, in both WS and WW conditions. For GY, the variances due to dominance effects represented about 33.3% and 31.0% of the genetic variance in WS and WW conditions, respectively. Lower broad-sense heritability was observed for this trait in WW (0.42) when compared to the WS condition (0.53). As for FFT, the variances due to dominance effects represented about 19.9% and 20.8% of the genetic variance in WS and WW conditions, respectively, and the broad-sense heritabilities for FFT were greater in WW conditions (0.71) than in conditions WS (0.56).

For GY, the narrow-sense heritabilities within environments ranged from 0.30 (T11) to 0.38 (J10) under WS conditions and from 0.20 (T11) to 0.57 (J10) under WW conditions. The proportion of genetic variance explained by dominance deviations ranged from 0.04 (T10) to 0.43 (T11) under WS conditions and from 0.10 (T11) to 0.29 (J11) under WW conditions. The broad-sense heritabilities were lower for the experimental tests that had a lower number of repetitions (2010 under WS conditions, and 2011 under WW conditions).

For FFT, the narrow-sense heritabilities ranged from 0.09 (T10) to 0.80 (J10) under WS conditions and from 0.49 (T10) to 0.74 (J10) under WW conditions. The proportion of genetic variance explained by dominance deviations ranged from 0.01 (T10) to 0.26 (T11) under WS conditions and from 0.07 (J11) to 0.23 (T11) under WW conditions. The broad-sense heritabilities were higher for J10 (0.89 and 0.88) under WS and WW conditions.

The Equation (2) is an implicit model to perform MET analyses[45] and provide genetic correlations for additive and dominance effects across environments. This model, reflects on the levels of genotypes-by-environment interaction. Particularly, for GY, the environmental correlations were 0.35 and 0.24 for WS and WW conditions, respectively, indicating an inconsistent ranking of hybrids across environments. As for FFT, the lowest correlation was observed for dominance effects.

3.4.2. Efficiency of prediction methodologies in multi-environment trials

Figures 1 and 2 show the predictive abilities observed in the three scenarios (CV1, CV2-50%, and CV2-25%) for each of the five compared methods: GBLUP additive model (GBLUP-A), GBLUP additive-dominant model (GBLUP-AD), bagging (Bag), random forest (RF) and boosting (Boost). The numerical results of these figures are presented in Supplementary Tables 1, 2, and 3.

Table 3. Estimates of variance components and genetic parameters for grain yield (GY) and female flowering time (FFT) were obtained considering the joint analysis for the four evaluated environments and analyses within each environment, for the irrigated (WW) and water stress (WS) conditions.

				WS				**WW**			
		WS	**WW**	2010		2011		2010		2011	
				Janaúba	Teresina	Janaúba	Teresina	Janaúba	Teresina	Janaúba	Teresina
GY	$\hat{\sigma}^2_{u_a}$	0.60*	1.00*	0.64*	0.88*	0.35*	0.50*	1.47*	1.15*	1.59*	0.96*
	$\hat{\sigma}^2_{u_d}$	0.30*	0.45*	0.39*	0.11^{ns}	0.34*	0.71*	0.56*	0.61*	1.42*	0.45^{ns}
	$\hat{\sigma}^2_e$	0.81	1.98	0.66	1.61	0.33	0.44	0.53	1.02	1.83	3.30
	ρ_a	0.35	0.24	-	-	-	-	-	-	-	-
	ρ_d	0.83	0.85	-	-	-	-	-	-	-	-
	h^2	0.35	0.29	0.38	0.34	0.34	0.30	0.57	0.41	0.33	0.20
	d^2	0.17	0.13	0.23	0.04	0.33	0.43	0.22	0.22	0.29	0.10
	H^2	0.53	0.42	0.61	0.38	0.68	0.73	0.79	0.63	0.62	0.30
FFT	$\hat{\sigma}^2_{u_a}$	4.66*	3.13*	6.86*	1.54*	6.85*	2.04*	4.24*	2.93*	3.42*	2.38*
	$\hat{\sigma}^2_{u_d}$	1.16*	0.82*	0.75*	0.12^{ns}	1.88*	1.20*	0.83*	1.15*	0.45^{ns}	1.03*
	$\hat{\sigma}^2_e$	4.54	1,64	0.97	14.73	1.73	1.40	0.69	1.88	2.70	1.04
	ρ_a	0.64	0.82	-	-	-	-	-	-	-	-
	ρ_d	0.43	0.37	-	-	-	-	-	-	-	-
	h^2	0.45	0.56	0.80	0.09	0.65	0.44	0.74	0.49	0.52	0.54
	d^2	0.11	0.15	0.09	0.01	0.18	0.26	0.14	0.19	0.07	0.23
	H^2	0.56	0.71	0.89	0.10	0.83	0.70	0.88	0.68	0.59	0.77

$\hat{\sigma}^2_{u_a}$: additive genetic variance; $\hat{\sigma}^2_{u_d}$: dominance genetic variance; $\hat{\sigma}^2_e$: residual variance; $\rho_{a:}$ additive genetic correlation coefficient between environments, ρ_d: dominance genetic correlation coefficient between environments; h^2*:* narrow sense heritability; d^2: proportion of the variance explained by the dominance effect and H^2*:* broad sense heritability. ^{ns} and *, not significant and significant at 5% probability of error

For the GBLUP models, the predictive abilities were lower for the CV1 scenario, where the predictions were based only on phenotypic and genotypic records of other related hybrids. However, the predictive ability increased when the predicted hybrid was present in some environments, being intermediate in CV2-50% and higher in CV2-25%. A more notable improvement in predictive abilities was observed when transitioning from CV1 to CV2-50%. Specifically, for GBLUP-A, the average increase was about 107%, 19%, 18%, and 19% for GY-WS, GY-WW, FFT-WS, and FFT-WW, respectively. As for GBLUP-AD, the average increase was 19%, 7%, 12%, and 15% for GY-WS, GY-WW, FFT-WS, and FFT-WW, respectively (Table Sup. 4). For GY, considering only CV1 scenario, the average predictive abilities were higher in WW conditions. For FFT, mean predictive abilities for WW conditions were higher than for WS for all scenarios (Table Sup. 4).

The environments with the highest broad-sense heritabilities also exhibited the highest average predictive abilities across all evaluated methodologies (Table 3, Figures 1 and 2). The GBLUP-AD model demonstrated superior predictive abilities for the GY and FFT traits in almost all environments and scenarios (CV1, CV2-50%, and CV2-25%). Conversely, the GBLUP-A performed equally or better than the GBLUP-AD model for FFT in environments where the dominance effect was not significant (T10-WS and J11-WW). However, for GY, in environments where the proportion of dominance genetic variance was close to or greater than the additive variance (J11 and T11 under WS), GBLUP-A exhibited lower predictive ability.

Unlike GBLUP, machine learning methodologies did not show a consistent pattern of increased predictive ability with the presence of phenotypic records of the hybrid to be predicted in correlated environments. Overall, an increase in predictive abilities for GY was observed when the scenario changed from CV1 to CV2, under WS conditions, which was not observed in many environments under WW conditions. For example, for RF, the average predictive abilities in the environments increased by approximately 51% for GY in WS conditions, while a decrease of about 10% was observed when the prediction changed from CV1 to CV2-50% for the same trait in WW conditions (Table Sup. 4). As for FFT, Bag drew the most attention, presenting a large standard error in predictive abilities with statistically similar results for CV1, CV2-50% and CV2-25% in WS and WW conditions, in almost all environments.

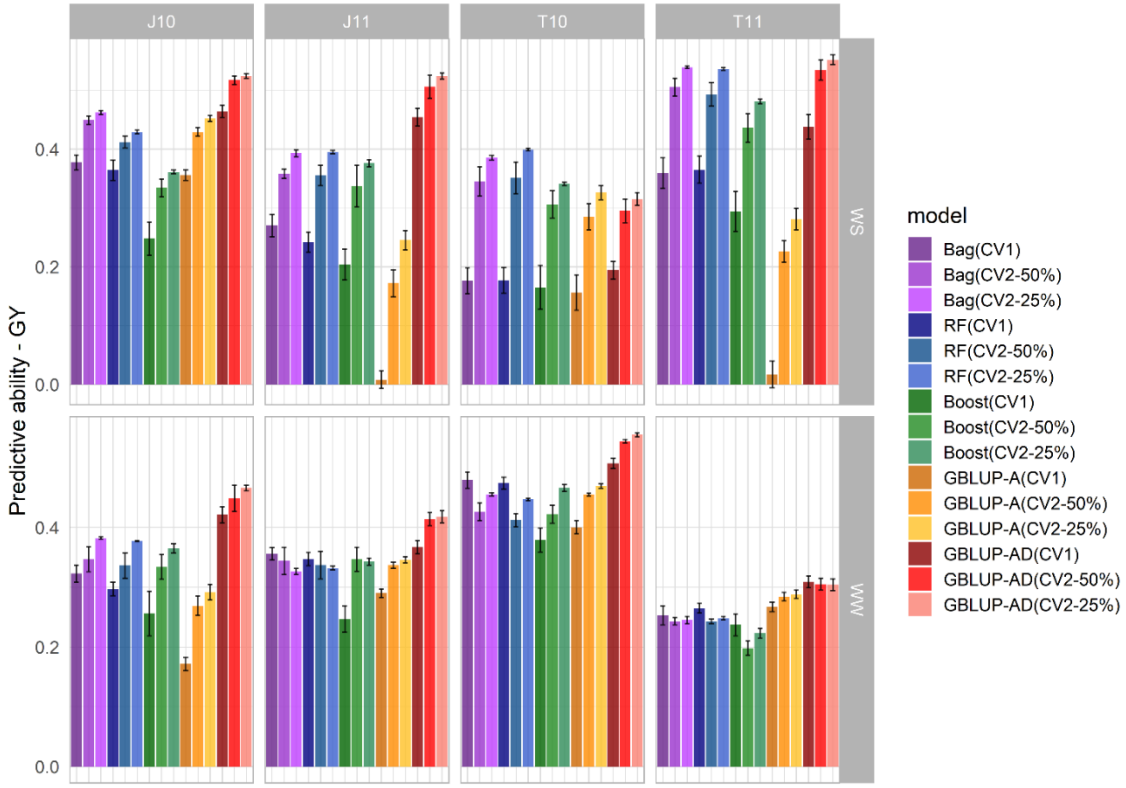

Figure 1. Mean predictive abilities and their respective standard errors for grain yield (GY), evaluated in the environments of Janaúba/2010 (J10), Janaúba/2011 (J11), Teresina/2010 (T10) and Teresina/2011 (T11)), under the CV1, CV2 (50%) and CV2 (25%) scenarios, considering irrigated (WW) and water stress (WS) conditions. The evaluated methodologies include bagging (Bag), random forest (RF), boosting (Boost), GBLUP additive model (GBLUP-A), and GBLUP additive-dominant model (GBLUP-AD).

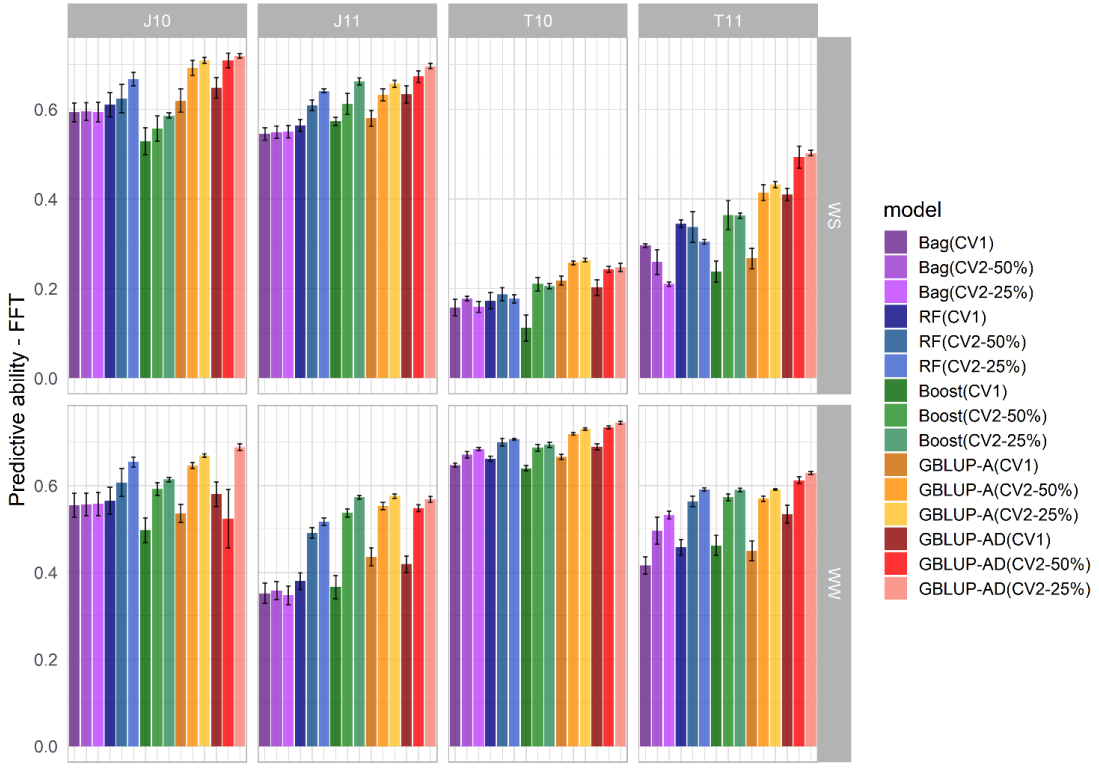

Figure 2. Mean predictive abilities and their respective standard errors for female flowering time (FFT), evaluated in the environments of Janaúba/2010 (J10), Janaúba/2011 (J11), Teresina/2010 (T10) and Teresina/2011 (T11), under CV1, CV2 (50%) and CV2 (25%) scenarios, considering irrigated (WW) and water stress (WS) conditions. The evaluated methodologies include bagging (Bag), random forest (RF), boosting (Boost), GBLUP additive model (GBLUP-A), and GBLUP additive-dominant model (GBLUP-AD).

In environments where dominance effects accounted for a large portion of the genetic variance, the Bag, RF and Boost methodologies showed intermediate results between the two GBLUP models. Among the evaluated machine learning methodologies, Bag and RF produced very similar predictions and Boost presented the lowest predictive ability for GY. As for FFT, machine learning did not show improvement in predictive abilities when compared to GBLUP.

3.5. Discussion

We employed both statistical and machine learning methods to evaluate the efficiency of genomic predictions for GY and FFT. These models allow us to leverage information about relationships between hybrids and correlated environments. Three scenarios, CV1, CV2-50% and CV2-25%, were used, each characterized by different degrees of data imbalance. Predicting the performance of new hybrids that have not been evaluated in the field (CV1) is more challenging compared to predicting hybrids that have been evaluated in an unbalanced manner across correlated environments (CV2-50% and CV2-25%)[15,16].

Among the evaluated methodologies, GBLUP stands out as it allows to estimate relationships between individuals based on molecular markers for additive and dominant effects[32,34]. This methodology also allows to incorporate variance-covariance structures to handle correlated environments and unbalanced data.[14,46]. As expected, GBLUP showed the highest predictive abilities in the CV2 validation schemes (CV2-50% and CV2-25%) as the predictions benefited from phenotypic records of the same hybrids in other correlated environments. Similar results have been reported in previous studies using wheat inbred lines and maize single cross hybrids [16,14,15,13].

GBLUP-AD demonstrated greater efficiency to predict hybrids for traits and environments in which the effects due to dominance deviations were significant. As a substantial part of the genetic variance in maize hybrids for GY is attributed to dominance effects[47], incorporating theses effects in the model significantly improved the prediction of this trait. However, when the dominance effect accounts for a small portion of the genetic variance (as observed for FFT), additive and additive-dominant models tend to show minor improvements in the predictions of the estimated total genetic effects [13,48]. This emphasizes the importance of understand the genetic bases

of hybrids which can be decomposed into general combining ability (GCA) and specific combining ability (SCA). For GCA, the additive variance is the major component of the variance[49] and, SCA is largely dependent on genes with dominance or epistatic effects[50]. In this context, gaining a comprehension of the relative magnitudes of GCA and SCA variations is instrumental in guiding the formulation of optimal strategies for hybrid breeding programs, as highlighted in previous research[51].

One possible reason why tree-based learning models did not benefit from the presence of phenotyped hybrids in correlated environments is related to their shared concept of recursive division[52]. These models aim to find decision rules that naturally partition the prediction space to provide an informative and robust hierarchical model[21,53]. In this context, it is conceivable that the location/year variables played a crucial role in most of the tree construction scenarios, leading to the split of the same hybrids at the beginning of branching and making it difficult for them to be grouped at the last nodes, which are responsible for the predictive response.

Among the evaluated methodologies, Bag showed, in most environments, the same prediction pattern for both scenarios where the predicted hybrid was absent in all environments and where it was absent only in some environments. Bag is a variation of the decision tree that uses resampling of the original data in subsamples (bootstrap) according to the determined number of trees. This process may lead to high correlation between the generated trees, resulting in the same variable consistently being the most important one[21]. As consequence, the same hybrids in different environments may end up at very distant nodes in the construction of the tree. The RF further reduces the dependence between the decision trees by randomly selecting part of the original data and variables to build the trees[22,52]. This allows different variables to have a chance to be at the top in their construction. On the other hand, boost sequentially combines different predictors, fitting a new tree to the residuals of the previous model using a specified loss function (e.g. mean squared error for regression)[54]. It incorporates automatic indirect selection of markers and is generally recommended for regression analysis[17,20]. It is important to note that these variations of decision trees mentioned above are typically used when obtaining accurate predictions is more important than the biological interpretation of the model itself[53].

One of the advantages of using machine learning methodologies is that they do not require the specification of the trait inheritance, having as an initial hypothesis the

possibility of capturing non-additive effects in the genome, which are often not specified in traditional statistical methods[25]. Possibly tree-based machine learning methodologies managed to capture part of the dominance effect, presenting better results than GBLUP-A in environments where dominance represented a large part of the genetic variance. These methodologies are competitive with statistical models and tend to outperform them when applied to large data sets. However, when applied to small training sets, machine learning is probably not able to capture all non-additive information and linear models may perform better[55].

Tree-based machine learning methodologies (Bag, RF, and Boost) are considered promising for genomic prediction, especially for traits controlled by a few quantitative trait locus (QTL), capturing non-additive effects in their models[25]. Among the machine learning methodologies evaluated in this study, Boost is considered the most sensitive methodology for the variation of traits and environments[25,26,56]. In these studies, carried out with simulated data, the lower the heritability and the greater the number of QTL that control the trait, the lower the Boost prediction efficiency. This fact may be related to the lower values of predictive ability for GY when compared to other methodologies, since this is a quantitative trait with low heritability, greatly influenced by the environment[57,58].

A possible explanation for the lower values of predictive ability of machine learning methodologies when compared to GBLUP in FFT prediction is that, even though it is a complex trait[59,60], additive genetic variance contributed with a large proportion of this trait. If the relationship between trait and response is close to a linear model, then a linear approach will probably work well and be superior to a method like a regression tree that does not explore this linear structure[22,61].

In the context of a hybrid breeding program, identifying high-performance genotypes is essential[62]. However, extensive field trials are needed to identify the best hybrids in the target environment. These trials require resources in terms of people, equipment, and area to carry out the phenotyping of the plants. Furthermore, most crosses are discarded after evaluation due to poor overall performance[63]. Associated with this fact is the trend of stagnating or rising costs of field evaluations, unlike genotyping which tends to decrease [64]. In this sense, the use of genomic information in multiple-environment trials is an alternative to traditional field trials, as it allows to reduce the phenotyping costs and to increase the number of hybrid combinations

evaluated when performing the prediction of the genetic value of individuals that were not evaluated in the field [11,65].

Based on the results of this study, a practical application of cost reduction and the efficiency of genomic selection in a breeding program can be demonstraded[10]. Assuming the cost of an experimental maize trial is $17.00 per plot[66] and the cost of sequencing genotyping (GBS) standard is $25.38 per parental line[67], the total budget needed for a breeding program would be $108,120.00 for phenotyping the 265 single cross hybrids in eight experimental trials using three replications, and $5,076.00 for genotyping the 188 parental lines and the two testers. Using the GBLUP-AD model, it was shown that the predictive abilities for the untested single hybrids averaged greater than 0.40 for GY under WW conditions, with an imbalance of 25% of hybrids randomly missing in each environment. Consequently, the cost reduction for the breeding program would be approximately 25% or $27,030.00 compared to phenotyping alone, which would cover the costs of genotyping the lines ($5,076.00).

The models proposed in this study can be applied to other crops that use hybrids to explore heterosis, and they can be expanded to include environmental variables to predict non-evaluated environments. In addition, this study demonstrates the application of machine learning methodologies in tests of multiple environments for predicting of hybrids, comparing their performance with GBLUP with non-additive effects, thus highlighting the potential of both methodologies.

3.6. Conclusion

Genomic prediction for untested maize single cross hybrids using both statistical and machine learning approaches were applied in MET assays. The results demonstrate that both methodologies are efficient and can be valuable tools in maize breeding programs to accurately predict the performance of hybrids in specific environments. The choice of the best methodology depends on the specific case. To optimize the predictive ability of GBLUP, it is crucial to accurately model the variance components. On the other hand, machine learning methodologies have the advantage of capturing non-additive effects without making any assumptions at the outset of the model. We found that predicting the performance of new hybrids that were not evaluated in any field trials was more challenging than predicting hybrids in unbalanced trials. However, regardless of the methodology used, environments with the lowest

heritability showed the lowest predictive abilities, underscoring the importance of conducting well-designed and properly replicated experiments.

References

1. Hossain, F. *et al.* Molecular breeding for increasing nutrition quality in maize: recent progress. in *Molecular breeding in wheat, maize and sorghum: strategies for improving abiotic stress tolerance and yield* 360–379 (CABI, 2021). doi:10.1079/9781789245431.0021.
2. Hossain, F. *et al.* Maize Breeding. in *Fundamentals of Field Crop Breeding* 221–258 (Springer Nature Singapore, 2022). doi:10.1007/978-981-16-9257-4_4.
3. Lobell, D. B. *et al.* Greater Sensitivity to Drought Accompanies Maize Yield Increase in the U.S. Midwest. *Science (80-.).* **344**, 516–519 (2014).
4. ONU. World Population Prospects 2022. https://population.un.org/wpp/Graphs/Probabilistic/POP/TOT/900 (2022).
5. Cruz, C. D., Regazzi, A. J. & Carneiro, P. C. S. Modelos biométricos aplicados ao melhoramento. *UFV, Viçosa* (2012).
6. Malosetti, M., Ribaut, J.-M. & van Eeuwijk, F. A. The statistical analysis of multi-environment data: modeling genotype-by-environment interaction and its genetic basis. *Front. Physiol.* **4**, (2013).
7. Crossa, J. Statistical Analyses of Multilocation Trials. in 55–85 (1990). doi:10.1016/S0065-2113(08)60818-4.
8. Burgueño, J., Crossa, J., Cotes, J. M., Vicente, F. S. & Das, B. Prediction Assessment of Linear Mixed Models for Multienvironment Trials. *Crop Sci.* **51**, 944–954 (2011).
9. Jarquin, D. *et al.* Genomic Prediction Enhanced Sparse Testing for Multi-environment Trials. *G3 Genes|Genomes|Genetics* **10**, 2725–2739 (2020).
10. Krause, M. D. *et al.* Boosting predictive ability of tropical maize hybrids via genotype-by-environment interaction under multivariate GBLUP models. *Crop Sci.* **60**, 3049–3065 (2020).
11. Bernardo, R. Prediction of maize single-cross performance using RFLPs and information from related hybrids. *Crop Sci.* **34**, 20–25 (1994).
12. Meuwissen, T. H. E., Hayes, B. J. & Goddard, M. Prediction of total genetic

value using genome-wide dense marker maps. *Genetics* **157**, 1819–1829 (2001).

13. Dias, K. O. D. G. *et al.* Improving accuracies of genomic predictions for drought tolerance in maize by joint modeling of additive and dominance effects in multi-environment trials. *Heredity (Edinb).* **121**, 24–37 (2018).
14. Jarquin, D. *et al.* Increasing Genomic-Enabled Prediction Accuracy by Modeling Genotype × Environment Interactions in Kansas Wheat. *Plant Genome* **10**, (2017).
15. Jarquin, D. *et al.* A reaction norm model for genomic selection using high-dimensional genomic and environmental data. *Theor. Appl. Genet.* **127**, 595–607 (2014).
16. Burgueño, J., de los Campos, G., Weigel, K. & Crossa, J. Genomic Prediction of Breeding Values when Modeling Genotype × Environment Interaction using Pedigree and Dense Molecular Markers. *Crop Sci.* **52**, 707–719 (2012).
17. González-Recio, O., Rosa, G. J. M. & Gianola, D. Machine learning methods and predictive ability metrics for genome-wide prediction of complex traits. *Livest. Sci.* **166**, 217–231 (2014).
18. Zhou, Z.-H. *Machine learning*. (Springer Nature, 2021).
19. Jannink, J.-L. J.-L., Lorenz, A. J. & Iwata, H. Genomic selection in plant breeding: from theory to practice. *Brief. Funct. Genomics* **9**, 166–177 (2010).
20. Ogutu, J. O., Piepho, H.-P. & Schulz-Streeck, T. A comparison of random forests, boosting and support vector machines for genomic selection. *BMC Proc.* **5**, S11 (2011).
21. Hastie, T., Tibshirani, R., Friedman, J., Cruz, C. D. & Nascimento, M. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. *Viçosa, MG: Editora UFV* (Springer-Verlag New York, 2009).
22. Gareth, J., Daniela, W., Trevor, H. & Robert, T. *An introduction to statistical learning: with applications in R*. (Spinger, 2013).
23. Sarkar, R. K., Rao, A. R., Meher, P. K., Nepolean, T. & Mohapatra, T. Evaluation of random forest regression for prediction of breeding value from genomewide SNPs. *J. Genet.* **94**, 187–192 (2015).
24. Farooq, M., van Dijk, A. D. J., Nijveen, H., Mansoor, S. & de Ridder, D. Genomic prediction in plants: opportunities for ensemble machine learning

based approaches. *F1000Research* **11**, 802 (2022).

25. Barbosa, I. de P. *et al.* Genome-enabled prediction through machine learning methods considering different levels of trait complexity. *Crop Sci.* **61**, 1890–1902 (2021).
26. Costa, W. G. da *et al.* Genomic prediction through machine learning and neural networks for traits with epistasis. *Comput. Struct. Biotechnol. J.* **20**, 5490–5499 (2022).
27. Sousa, I. C. de *et al.* Genomic prediction of leaf rust resistance to Arabica coffee using machine learning algorithms. *Sci. Agric.* **78**, (2021).
28. Westhues, C. C. *et al.* Prediction of Maize Phenotypic Traits With Genomic and Environmental Predictors Using Gradient Boosting Frameworks. *Front. Plant Sci.* **12**, (2021).
29. Silva, K. J. *et al.* High-density SNP-based genetic diversity and heterotic patterns of tropical maize breeding lines. *Crop Sci.* **60**, 779–787 (2020).
30. Dias, K. O. D. G. *et al.* Estimating Genotype × Environment Interaction for and Genetic Correlations among Drought Tolerance Traits in Maize via Factor Analytic Multiplicative Mixed Models. *Crop Sci.* **58**, 72–83 (2018).
31. Technow, F. *et al.* Genome Properties and Prospects of Genomic Prediction of Hybrid Performance in a Breeding Program of Maize. *Genetics* **197**, 1343–1355 (2014).
32. Vitezica, Z. G., Varona, L. & Legarra, A. On the additive and dominant variance and covariance of individuals within the genomic selection scope. *Genetics* **195**, 1223–1230 (2013).
33. Amadeu, R. R. *et al.* AGHmatrix: R Package to Construct Relationship Matrices for Autotetraploid and Diploid Species: A Blueberry Example. *Plant Genome* **9**, (2016).
34. VanRaden, P. M. Efficient methods to compute genomic predictions. *J. Dairy Sci.* **91**, 4414–4423 (2008).
35. Gilmour, A. R., Gogel, B. J., Cullis, B. R., Welham, S. J. & Thompson, R. ASReml User Guide Release 4.2 Functional Specification. *VSN Int. Ltd* (2021).
36. Corbeil, R. R. & Searle, S. R. Restricted Maximum Likelihood (REML) Estimation of Variance Components in the Mixed Model. *Technometrics* **18**,

31 (1976).

37. Wilks, S. S. The Large-Sample Distribution of the Likelihood Ratio for Testing Composite Hypotheses. *Ann. Math. Stat.* **9**, 60–62 (1938).
38. Dobson, A. & Barnett, A. *An Introduction to Generalized Linear Models*. (Chapman and Hall/CRC, 2008). doi:10.1201/9780367807849.
39. Falconer, D. S. & Mackay, T. F. C. Introduction to quantitative genetics. Essex. *UK Longman Gr.* (1996).
40. Ho, T. K. Random decision forests. in *Proceedings of 3rd international conference on document analysis and recognition* vol. 1 278–282 (IEEE, 1995).
41. Breiman, L. Random forests. *Mach. Learn.* **45**, 5–32 (2001).
42. Greenwell, B., Boehmke, B., Cunningham, J. & GBM, D. gbm: Generalized boosted regression models. R package version 2.1. 5. *Website https//cran. r-project. org/package= gbm [accessed 12 January 2020]* (2019).
43. R Core Team. R: A language and environment for statistical computing. at (2021).
44. Resende, M. D. V. de, Silva, F. F. e & Azevedo, C. F. Estatística matemática, biométrica e computacional: Modelos mistos, multivariados, categóricos e generalizados (REML/BLUP), inferência bayesiana, regressão aleatória, seleção genômica, QTL-GWAS, estatística espacial e temporal, competição, sobrevivência. *Viçosa Ed. UFV* 1–881 (2014).
45. Gezan, S. A., de Carvalho, M. P. & Sherrill, J. Statistical methods to explore genotype-by-environment interaction for loblolly pine clonal trials. *Tree Genet. Genomes* **13**, 1 (2017).
46. Krause, M. D. *et al.* Boosting predictive ability of tropical maize hybrids via genotype-by-environment interaction under multivariate GBLUP models. *Crop Sci.* **60**, 3049–3065 (2020).
47. Fernandes, S. B., Dias, K. O. G., Ferreira, D. F. & Brown, P. J. Efficiency of multi-trait, indirect, and trait-assisted genomic selection for improvement of biomass sorghum. *Theor. Appl. Genet.* **131**, 747–755 (2018).
48. Nishio, M. & Satoh, M. Including Dominance Effects in the Genomic BLUP Method for Genomic Evaluation. *PLoS One* **9**, e85792 (2014).
49. Reif, J. C., Gumpert, F.-M., Fischer, S. & Melchinger, A. E. Impact of

Interpopulation Divergence on Additive and Dominance Variance in Hybrid Populations. *Genetics* **176**, 1931–1934 (2007).

50. Sprague, G. F. & Tatum, L. A. General vs. specific combining ability in single crosses of corn. *J. Am. Soc. Agron.* (1942).
51. Giraud, H. *et al.* Reciprocal Genetics: Identifying QTL for General and Specific Combining Abilities in Hybrids Between Multiparental Populations from Two Maize (Zea mays L.) Heterotic Groups. *Genetics* **207**, 1167–1180 (2017).
52. Hofmarcher, P. & Grün, B. *Macroeconomic Forecasting in the Era of Big Data*. vol. 52 (Springer International Publishing, 2020).
53. Myles, A. J., Feudale, R. N., Liu, Y., Woody, N. A. & Brown, S. D. An introduction to decision tree modeling. *J. Chemom.* **18**, 275–285 (2004).
54. Friedman, J. H. Greedy function approximation: A gradient boosting machine. *Ann. Stat.* **29**, (2001).
55. Westhues, C. C., Simianer, H. & Beissinger, T. M. learnMET: an R package to apply machine learning methods for genomic prediction using multi-environment trial data. *G3 Genes|Genomes|Genetics* **12**, (2022).
56. Ghafouri-Kesbi, F., Rahimi-Mianji, G., Honarvar, M. & Nejati-Javaremi, A. Predictive ability of Random Forests, Boosting, Support Vector Machines and Genomic Best Linear Unbiased Prediction in different scenarios of genomic evaluation. *Anim. Prod. Sci.* **57**, 229 (2017).
57. Zhang, X. *et al.* Genetic architecture of maize yield traits dissected by QTL mapping and GWAS in maize. *Crop J.* **10**, 436–446 (2022).
58. Zhang, X. *et al.* A combination of linkage mapping and GWAS brings new elements on the genetic basis of yield-related traits in maize across multiple environments. *Theor. Appl. Genet.* **133**, 2881–2895 (2020).
59. Steinhoff, J. *et al.* Detection of QTL for flowering time in multiple families of elite maize. *Theor. Appl. Genet.* **125**, 1539–1551 (2012).
60. Buckler, E. S. *et al.* The Genetic Architecture of Maize Flowering Time. *Science (80-.).* **325**, 714–718 (2009).
61. Abdollahi-Arpanahi, R., Gianola, D. & Peñagaricano, F. Deep learning versus parametric and ensemble methods for genomic prediction of complex phenotypes. *Genet. Sel. Evol.* **52**, 12 (2020).

62. Technow, F., Riedelsheimer, C., Schrag, T. A. & Melchinger, A. E. Genomic prediction of hybrid performance in maize with models incorporating dominance and population specific marker effects. *Theor. Appl. Genet.* **125**, 1181–1194 (2012).
63. Windhausen, V. S. *et al.* Effectiveness of genomic prediction of maize hybrid performance in different breeding populations and environments. *G3 Genes| Genomes| Genet.* **2**, 1427–1436 (2012).
64. Krchov, L.-M. & Bernardo, R. Relative Efficiency of Genomewide Selection for Testcross Performance of Doubled Haploid Lines in a Maize Breeding Program. *Crop Sci.* **55**, 2091–2099 (2015).
65. Massman, J. M., Gordillo, A., Lorenzana, R. E. & Bernardo, R. Genomewide predictions from maize single-cross data. *Theor. Appl. Genet.* **126**, 13–22 (2013).
66. Tech Services. Pricing brochure TSI 2023 test sites. *Bluffton IN:TechServices* https://techservicespro.com/test-locations/ (2023).
67. University of Minnesota. Genotyping-by-sequencing (Pricing). *Genomics Center* https://genomics.umn.edu/service/standard-genotyping-sequencing (2023).

3.7. Supporting information

Table S1. Predictive abilities analyzed within each environment and their respective standard errors for grain yield (GY) and female flowering time (FFT), for the four methodologies using CV1 scenario (tested hybrids were not evaluated in any environment) for the irrigated (WW) and water stress (WS) conditions.

CV1									
		WS				**WW**			
Trait	Methodology	2010		2011		2010		2011	
		Janaúba	Teresina	Janaúba	Teresina	Janaúba	Teresina	Janaúba	Teresina
GY	*Bagging*	0.377±0.013	0.176±0.022	0.270±0.019	0.359±0.026	0.323±0.014	0.479±0.014	0.356±0.010	0.253±0.016
	random forest	0.364±0.017	0.177±0.022	0.242±0.017	0.365±0.023	0.297±0.011	0.474±0.010	0.347±0.011	0.265±0.008
	Boosting	0.248±0.028	0.165±0.037	0.204±0.026	0.294±0.034	0.256±0.037	0.379±0.020	0.247±0.022	0.237±0.018
	GBLUP-A	0.356±0.009	0.156±0.030	0.008±0.015	0.017±0.023	0.172±0.011	0.400±0.011	0.290-0.007	0.267±0.008
	GBLUP-AD	0.464±0.010	0.194±0.015	0.454±0.015	0.438±0.021	0.421±0.013	0.507±0.008	0.367±0.011	0.309±0.010
FFT	*Bagging*	0.593±0.021	0.158±0.018	0.545±0.014	0.296±0.004	0.555±0.027	0.647±0.005	0.352±0.023	0.416±0.020
	random forest	0.610±0.027	0.173±0.018	0.564±0.013	0.345±0.008	0.565±0.031	0.662±0.006	0.380±0.019	0.458±0.018
	Boosting	0.529±0.030	0.112±0.029	0.573±0.009	0.238±0.023	0.497±0.028	0.640±0.006	0.366±0.027	0.462±0.023
	GBLUP-A	0.619±0.026	0.218±0.010	0.580±0.018	0.267±0.023	0.536±0.021	0.666±0.006	0.436±0.020	0.450±0.023
	GBLUP-AD	0.648±0.023	0.202±0.017	0.633±0.019	0.410±0.014	0.580±0.028	0.689±0.007	0.419±0.019	0.534±0.021

Table S2. Predictive abilities analyzed within each environment and their respective standard errors for grain yield (GY) and female flowering time (FFT), for the four methodologies using CV2-50% scenario (tested hybrids were not evaluated at 50% environments), for irrigated (WW) and water stress (WS) conditions.

CV2 (50%)									
		WS				**WW**			
Trait	Methodology	2010		2011		2010		2011	
		Janaúba	Teresina	Janaúba	Teresina	Janaúba	Teresina	Janaúba	Teresina
GY	*bagging*	0.449±0.007	0.345±0.025	0.358±0.008	0.505±0.015	0.347±0.021	0.426±0.015	0.344±0.023	0.243±0.006
	random forest	0.412±0.010	0.351±0.027	0.355±0.017	0.493±0.020	0.336±0.021	0.412±0.011	0.337±0.023	0.243±0.004
	boosting	0.334±0.015	0.306±0.023	0.337±0.035	0.436±0.024	0.334±0.021	0.422±0.015	0.347±0.020	0.198±0.012
	GBLUP-A	0.429±0.007	0.285±0.022	0.172±0.023	0.226±0.018	0.269±0.016	0.455±0.002	0.337±0.005	0.284±0.007
	GBLUP-AD	0.517±0.007	0.295±0.020	0.506±0.020	0.534±0.017	0.449±0.022	0.544±0.003	0.414±0.011	0.305±0.010
FFT	*bagging*	0.595±0.020	0.178±0.005	0.549±0.014	0.259±0.028	0.556±0.026	0.671±0.007	0.358±0.021	0.496±0.031
	random forest	0.624±0.032	0.187±0.015	0.609±0.012	0.337±0.034	0.607±0.032	0.700±0.009	0.491±0.012	0.564±0.012
	boosting	0.557±0.028	0.210±0.015	0.612±0.024	0.364±0.032	0.592±0.015	0.687±0.008	0.537±0.009	0.573±0.008
	GBLUP-A	0.692±0.017	0.257±0.004	0.632±0.013	0.414±0.018	0.646±0.007	0.719±0.003	0.553±0.008	0.570±0.006
	GBLUP-AD	0.709±0.016	0.243±0.007	0.673±0.013	0.493±0.025	0.670±0.010	0.734±0.003	0.548±0.008	0.612±0.008

Table S3. Predictive abilities analyzed within each environment and their respective standard errors for grain yield (GY) and female flowering time (FFT), for the four methodologies using CV2-25% scenario (tested hybrids were not evaluated at 25% environments), for irrigated (WW) and water stress (WS) conditions.

CV2 (25%)									
		WS				**WW**			
Trait	Methodology	2010		2011		2010		2011	
		Janaúba	Teresina	Janaúba	Teresina	Janaúba	Teresina	Janaúba	Teresina
GY	*bagging*	0.462±0.003	0.385±0.004	0.393±0.006	0.539±0.002	0.382±0.002	0.455±0.003	0.326±0.005	0.245±0.006
	random forest	0.429±0.003	0.399±0.002	0.395±0.003	0.536±0.002	0.377±0.001	0.447±0.002	0.332±0.003	0.248±0.003
	boosting	0.361±0.003	0.341±0.003	0.376±0.006	0.481±0.004	0.365±0.008	0.466±0.006	0.343±0.006	0.223±0.008
	GBLUP-A	0.452±0.005	0.326±0.012	0.245±0.016	0.281±0.018	0.292±0.013	0.469±0.004	0.346±0.005	0.288±0.007
	GBLUP-AD	0.524±0.004	0.315±0.011	0.524±0.005	0.552±0.008	0.466±0.005	0.555±0.003	0.418±0.010	0.304±0.010
FFT	*bagging*	0.594±0.022	0.159±0.012	0.550±0.014	0.210±0.005	0.558±0.027	0.684±0.004	0.347±0.021	0.532±0.009
	random forest	0.667±0.015	0.177±0.009	0.641±0.004	0.304±0.005	0.654±0.012	0.707±0.002	0.517±0.008	0.591±0.004
	boosting	0.586±0.006	0.205±0.006	0.662±0.008	0.363±0.006	0.614±0.005	0.694±0.006	0.573±0.004	0.590±0.004
	GBLUP-A	0.709±0.007	0.263±0.004	0.657±0.008	0.432±0.007	0.669±0.004	0.730±0.003	0.576±0.005	0.591±0.002
	GBLUP-AD	0.719±0.005	0.247±0.009	0.696±0.006	0.502±0.006	0.688±0.008	0.745±0.003	0.568±0.007	0.629±0.003

Table S4. Mean predictive abilities and their respective standard errors for grain yield (GY) and female flowering time (FFT), for the four methodologies, using CV1, CV2 (50%) and CV2 (25%) scenarios, for the irrigated (WW) and water stress (WS) conditions.

Trait	Methodology	**WS**			**WW**		
		CV1	CV2 (50%)	CV2 (25%)	CV1	CV2 (50%)	CV2 (25%)
GY	*bagging*	0.296±0.046	0.414±0.038	0.445±0.036	0.311±0.026	0.340±0.037	0.352±0.044
	random forest	0.287±0.047	0.403±0.033	0.440±0.033	0.346±0.046	0.332±0.035	0.351±0.042
	boosting	0.228±0.028	0.353±0.028	0.390±0.031	0.280±0.033	0.325±0.047	0.349±0.050
	GBLUP-A	0.134±0.081	0.278±0.055	0.326±0.045	0.282±0.047	0.336±0.042	0.349±0.042
	GBLUP-AD	0.388±0.065	0.463±0.056	0.479±0.055	0.401±0.042	0.428±0.049	0.436±0.052
FFT	*bagging*	0.398±0.103	0.395±0.104	0.378±0.113	0.493±0.067	0.520±0.065	0.530±0.070
	random forest	0.423±0.101	0.439±0.107	0.447±0.122	0.516±0.062	0.591±0.044	0.617±0.041
	boosting	0.363±0.112	0.436±0.092	0.454±0.104	0.491±0.057	0.597±0.032	0.618±0.027
	GBLUP-A	0.421±0.104	0.499±0.100	0.515±0.103	0.522±0.053	0.622±0.038	0.642±0.036
	GBLUP-AD	0.4730.106	0.530±0.107	0.541±0.109	0.556±0.056	0.641±0.040	0.658±0.038

4. CONCLUSÕES GERAIS

Considerando diferentes cenários de predição, este estudo demonstrou que o GBLUP, especialmente quando incorporam efeitos não aditivos, tendem a apresentar maiores capacidades preditivas em comparação com as metodologias de aprendizado de máquina avaliadas. Em particular, o modelo GBLUP-AD se destacou ao demonstrar maior eficiência na predição de híbridos para características e ambientes nos quais os efeitos de dominância foram significativos.

Em suma, os resultados deste estudo oferecem *insights* valiosos para aprimorar os métodos de predição genômica em híbridos de milho e fornecem orientações práticas para otimizar os programas de melhoramento genético, destacando o potencial de integração entre abordagens estatísticas e de aprendizado de máquina para enfrentar os desafios emergentes na agricultura moderna.