

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

**COMPARAÇÃO DE MODELOS ADITIVO, ADITIVO-DOMINANTE E DE
MACHINE LEARNING (LightGBM) NA PREDIÇÃO DE HÍBRIDOS DE
MILHO**

Jorge Otavio Domingos Missima

Biólogo

2023

UNIVERSIDADE ESTADUAL PAULISTA – UNESP

CÂMPUS DE JABOTICABAL

**COMPARAÇÃO DE MODELOS ADITIVO, ADITIVO-DOMINANTE E DE
MACHINE LEARNING (LightGBM) NA PREDIÇÃO DE HÍBRIDOS DE
MILHO**

Jorge Otavio Domingos Missima

Orientador: Prof. Dr. Dilermando Perecin

Coorientador: Dr. Ivan Schuster

Dissertação apresentada à Faculdade de Ciências Agrárias e Veterinárias – UNESP, Câmpus de Jaboticabal, como parte das exigências para a obtenção do título de Mestre em Agronomia (Genética e Melhoramento de Plantas)

M678c Missima, Jorge Otavio Domingos
Comparação de modelos aditivo, aditivo-dominante e de machine learning (LightGBM) na predição de híbridos de milho / Jorge Otavio Domingos Missima. -- Jaboticabal, 2023
27 p. : il., tabs.

Dissertação (mestrado) - Universidade Estadual Paulista (Unesp), Faculdade de Ciências Agrárias e Veterinárias, Jaboticabal
Orientador: Dilermando Perecin
Coorientador: Ivan Schuster

1. Melhoramento genético. 2. Estatística. 3. Inteligência artificial. I. Título.

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

Essa ficha não pode ser modificada.

CERTIFICADO DE APROVAÇÃO

**TÍTULO DA DISSERTAÇÃO: COMPARAÇÃO DE MODELOS ADITIVO, ADITIVO-DOMINANTE E DE
MACHINE LEARNING (LightGBM) NA PREDIÇÃO DE HÍBRIDOS DE MILHO**

AUTOR: JORGE OTAVIO DOMINGOS MISSIMA**ORIENTADOR: DILERMANDO PERECIN****COORDENADOR: IVAN SCHUSTER**

Aprovado como parte das exigências para obtenção do Título de Mestre em Agronomia (Genética e Melhoramento de Plantas), pela Comissão Examinadora:

Pesquisador Dr. IVAN SCHUSTER (Participação Virtual)
Longping HighTech Pesquisa e Desenvolvimento / Cravinhos/SP

Documento assinado digitalmente
gov.br IVAN SCHUSTER
Data: 06/08/2023 17:04:21-0300
Verifique em <https://validar.iti.gov.br>

Prof. Dr. CARLOS ALBERTO SCAPIM (Participação Virtual)
Departamento de Agronomia / Universidade Estadual de Maringá - Maringá/SP

Documento assinado digitalmente
gov.br CARLOS ALBERTO SCAPIM
Data: 15/08/2023 12:08:50-0300
Verifique em <https://validar.iti.gov.br>

Prof. Dr. GUSTAVO VITTI MÔRO (Participação Virtual)
Departamento de Ciências da Produção Agrícola / FCAV UNESP Jaboticabal

Documento assinado digitalmente
gov.br GUSTAVO VITTI MORO
Data: 07/08/2023 07:52:45-0300
Verifique em <https://validar.iti.gov.br>

Jaboticabal, 10 de julho de 2023

DADOS CURRICULARES DO AUTOR

Jorge Otavio Domingos Missima – filho de Jorge Missima e Solange Aparecida Domingos Missima, nasceu em 11 de abril de 1992 na cidade de Ribeirão Preto, São Paulo. Ingressou em 2010 no curso de Tecnologia em Biocombustíveis na Faculdade de Tecnologia de Jaboticabal - FATEC, obtendo em dezembro de 2013 o título de Tecnólogo em Biocombustíveis; Em maio de 2016, iniciou atividade na empresa Dow Agrosiences como Técnico de Pesquisa e Desenvolvimento; em 2017 ingressou no curso de Ciências Biológicas no Centro Universitário Barão de Mauá. Em julho de 2021 iniciou seu trabalho na empresa LongPing High-Tech Ltda como Pesquisador em Biotecnologia na área de análise de dados moleculares. Em março de 2021, iniciou o curso de Mestrado em Agronomia (Genética e Melhoramento de Plantas) na Universidade Estadual Paulista “Júlio de Mesquita Filho” sob orientação do Prof. Dr. Dilermando Perecin.

“N3o existe um caminho para a felicidade. A felicidade 3 o caminho.”

Thich Nhat Hanh

AGRADECIMENTOS

Ao Professor Dr. Dilermando Perecin, pela sua orientação, paciência e dedicação em me guiar durante todo o tempo de mestrado.

Ao Dr. Ivan Schuster, pela sua coorientação, incentivos, ensinamentos e conselhos durante toda essa jornada acadêmica e profissional.

Aos membros da banca de defesa, pela participação, contribuição e aprendizado enriquecedor.

A UNESP - campus de Jaboticabal, e ao Programa de pós-graduação em genética e melhoramento de plantas pela experiência e aprendizado.

A LongPing High-Tech por todo o subsídio, e pela valiosa oportunidade de desenvolver esse trabalho.

Aos meus amigos do Laboratório de Pesquisa e Desenvolvimento de Cravinhos por todo o companheirismo e parceria.

Aos meus pais, Solange e Jorge, e minha irmã Júlia, cujo amor, encorajamento e sacrifícios sempre foram meu alicerce.

A minha sobrinha Helena e irmã Maria Victória por serem uma fonte inspiradora de pureza e amor em minha vida.

As minhas avós Fumi (In Memoria) e Helena (In Memoria), pela benção, profundo amor e carinho, que foram muito importantes para me incentivar todos os dias.

A minha namorada Fernanda pelo amor, incentivo e apoio incondicional durante o desenvolvimento desse trabalho. Sua presença em minha vida é um verdadeiro presente, e sou muito grato por tê-la ao meu lado.

SUMÁRIO

RESUMO	ii
ABSTRACT	iii
LISTA DE TABELAS.....	IV
LISTA DE FIGURAS	V
1 INTRODUÇÃO	1
2 REVISÃO DE LITERATURA	2
2.1 Melhoramento genético do milho	2
2.2 Marcadores moleculares no melhoramento	3
2.3 Seleção genômica ampla.....	3
2.4 Modelos de Seleção genômica ampla	5
3 MATERIAL E MÉTODOS.....	7
3.1 Condução experimental e dados fenotípicos	7
3.2 Herdabilidade.....	7
3.3 Dados de marcadores moleculares	7
3.4 Modelos estatísticos	8
3.5 Avaliação e comparação dos modelos	9
4 RESULTADOS E DISCUSSÃO	11
4.1 Resultados.....	11
4.2 Discussão	17
5 CONCLUSÃO.....	22
6 REFERÊNCIA	23

COMPARAÇÃO DE MODELOS ADITIVO, ADITIVO-DOMINANTE E DE MACHINE LEARNING (LightGBM) NA PREDIÇÃO DE HÍBRIDOS DE MILHO

RESUMO - A escolha do método estatístico para predição dos valores genéticos genômicos em estudos de seleção genômica ampla (GWS) é fundamental para obtenção de elevadas acurácias preditivas. O objetivo deste estudo foi avaliar e comparar a performance de modelos paramétricos contendo efeito Aditivo (rrBLUP) e Aditivo-Dominante (BL), além de um modelo não paramétrico baseado em Machine Learning (LightGBM). Tais modelos foram aplicados na seleção genômica de híbridos de milho avaliados em três locais (Jataí-GO, Rolândia-PR e Sorriso-MT) para duas características importantes, produtividade e ciclo. Através dos resultados foi demonstrado que o modelo BL apresenta ótima estabilidade em termos de capacidade preditiva, porém sua performance computacional no treinamento dos modelos é consideravelmente inferior aos métodos rrBLUP e LightGBM. Caso o tempo computacional seja um gargalo para o desenvolvimento do estudo de seleção genômica, o modelo LightGBM que apresenta alta eficiência computacional, pode ser utilizado, entretanto essa utilização pode implicar em perda significativa de acurácia preditiva. Além disso, foi observado que a herdabilidade das características afeta a acurácia de predição dos modelos, e que os efeitos genéticos provenientes de herdabilidades menores podem ser capturados de maneira mais eficiente com o uso de modelos que incorporem efeitos aditivo-dominantes ou modelos de Machine learning.

Palavras-chave: acurácia, aditivo-dominante, herdabilidade, LightGBM, machine learning, seleção genômica ampla.

COMPARISON OF ADDITIVE, ADDITIVE-DOMINANT AND MACHINE LEARNING (LightGBM) MODELS IN CORN HYBRID PREDICTION

RESUMO - The choice of statistical method for predicting genomic breeding values in genomic wide selection (GWS) is essential to obtain high prediction accuracy. The aim of this study was to evaluate and compare the performance of parametric models containing Additive (rrBLUP) and Additive-Dominant (BL) effects, as well as a non-parametric model based on Machine Learning (LightGBM). Such models were applied in the genomic selection of corn hybrids evaluated in three locations (Jataí-GO, Rolândia-PR and Sorriso-MT) for two important traits, yield and maturity. Through the results it was demonstrated that the BL model presents excellent stability in terms of predictive ability, however its computational performance in the model training is considerably inferior to the rrBLUP and LightGBM methods. If computational time is a bottleneck for the development of the genomic selection study, the LightGBM model, which has high computational efficiency, can be used, however this use may imply a significant loss of predictive accuracy. Furthermore, it was observed that the heritability of the traits affects the prediction accuracy of the models, and the genetic effects from lower heritabilities can be captured more efficiently using models that incorporate additive-dominant effects or machine learning models.

Keywords: accuracy, additive-dominant, heritability, LightGBM, machine learning, genomic wide selection.

LISTA DE TABELAS

Tabela 1. Estatísticas descritivas e estimativa de herdabilidade	11
Tabela 2. Média das acurácias de predição obtidas a partir da validação cruzada	13
Tabela 3. Porcentagens de híbridos que ficaram acima dos 10% mais produtivos para PG, e acima dos 10% mais precoces para UG tanto no resultado de campo como na predição	12

LISTA DE FIGURAS

Figura 1 . Esquema de aplicação da seleção genômica ampla em um programa de melhoramento genético	4
Figura 2. Esquema do processo de validação cruzada $K = 3$	9
Figura 3. Comparação de acurácia de predição de produtividade e umidade de grãos para os modelos BL, LightGBM e rrBLUP nos três locais de avaliação (Jataí-GO, Rolândia-PR e Sorriso-MT)	12
Figura 4. Gráfico de dispersão indicando a relação entre os fenótipos preditos e observados para os três modelos (BL, LightGBM e rrBLUP)	13
Figura 5. Comparação do tempo computacional gasto (segundos) pelos três métodos avaliados (BL, rrBLUP e LightGBM) no treinamento dos modelos	15
Figura 6. Comparação do tempo computacional gasto (segundos) pelos três métodos avaliados (BL, LightGBM e rrBLUP) na predição dos VGG	16

1. INTRODUÇÃO

O alcance de altas produtividades na cultura do Milho (*Zea Mays* L.) é dependente do desenvolvimento de híbridos, sendo esse o principal produto explorado nos programas de melhoramento dessa cultura (Galli et al., 2018). A obtenção e avaliação de híbridos são atividades que demandam bastante tempo e recursos pois envolvem numerosos cruzamentos para identificar as melhores combinações de linhagens parentais (Schrag et al., 2009). A partir de poucas linhagens é possível obter uma vasta quantidade de híbridos, o que torna o processo de avaliação a campo de todos os híbridos possíveis oneroso e impraticável (Technow et al., 2014; de Oliveira et al., 2020).

Nesse sentido, a crescente disponibilidade de marcadores moleculares e a redução no custo de genotipagem tem tornado a metodologia de seleção genômica ampla, do inglês “Genome-Wide-Selection” – GWS (Meuwissen et al., 2001) uma ferramenta alternativa promissora para os programas de melhoramento (de Oliveira et al., 2020; Zhao et al., 2015). A partir dessa ferramenta pode-se prever a performance fenotípica de materiais somente genotipados sem a necessidade de testá-los no campo, aumentando consideravelmente a eficiência de seleção e ganho genético (de Oliveira et al., 2020; Zhang et al., 2022). A GWS consiste na estimação simultânea dos efeitos genéticos de marcadores dispersos em todo o genoma, com o objetivo de capturar os efeitos de todos os locos, de pequenos à grandes efeitos, e assim explicar grande parte da variância genética de um caráter quantitativo (Meuwissen et al., 2001; Cantelmo et al., 2017).

A seleção genômica apresenta desafios estatísticos e computacionais que devem ser considerados na sua implementação, dentre eles estão: o grande número de marcadores moleculares e a colinearidade entre eles, alta dimensionalidade dos dados e a complexidade da característica quantitativa. (Crossa et al., 2017; Pungpapong et al., 2012). Para superar esses desafios, diversos métodos têm sido propostos em estudos de seleção genômica, incluindo modelos paramétricos, não-paramétricos e mais recentemente modelos de Machine Learning (Crossa et al., 2017). Esses métodos diferem na suposição sobre o modelo genético associado ao caráter quantitativo, sendo que essas diferenças podem refletir diretamente sobre a acurácia de predição e consequentemente na identificação dos indivíduos de fato superiores

(Wang et al., 2018; Andrade et al., 2019). Diante desse cenário, o objetivo desse trabalho foi comparar a performance e acurácia de predição de dois métodos paramétricos lineares (BL e rrBLUP) e um método não paramétrico de machine learning (LightGBM) na predição genômica de híbridos de milho para duas características importantes, produtividade e ciclo.

2. REVISÃO DE LITERATURA

2.1 Melhoramento genético do milho

O milho (*Zea mays* L.) é uma das culturas mais importantes produzidas no mundo, sendo utilizado como alimento humano e para ração animal, devido principalmente às suas qualidades nutricionais (Costa, 2020). O Brasil se destaca como um dos maiores produtores dessa commodity agrícola, ocupando a terceira colocação, depois dos Estados Unidos e da China (ABIMILHO, 2019). A importância socioeconômica do milho foi um dos grandes motivos que levaram ao interesse dos pesquisadores por seu melhoramento genético visando o aumento de produtividade e adaptação em diferentes áreas de cultivo (Costa, 2020).

Os programas de melhoramento genético de milho foram impulsionados, no início do século XX, com o desenvolvimento de linhas puras ou linhagens, provenientes do processo clássico de autofecundação das plantas por várias gerações, e com a descoberta do vigor híbrido (Guimarães, 2007). O emprego de linhagens em programas de melhoramento está associado justamente à busca do vigor híbrido, fenômeno que proporciona grande produtividade em plantas F1 provenientes do cruzamento de linhagens parentais que exibem alta divergência entre si (Paterniani et al., 2008).

Com o advento da tecnologia de Duplos Haploides foi possível aumentar rapidamente o número de linhagens obtidas dentro dos programas de melhoramento. Por consequência desse aumento, o número potencial de híbridos a serem testados ao final de cada ciclo de melhoramento aumentou exponencialmente, tornando o processo de avaliação oneroso e impraticável (Technow et al., 2014; de Oliveira et al., 2020). Na prática, apenas uma pequena parcela de todas as possíveis combinações de linhagens parentais pode ser avaliada no campo experimental (Schrag et al., 2010).

2.2 Marcadores moleculares nos programas de melhoramento

Com o surgimento dos marcadores moleculares foi possível não só o estudo do genoma de várias espécies, mas também a identificação de regiões genômicas associadas a expressão de determinados caracteres (Pires, 2017). No melhoramento, as informações moleculares podem contribuir de maneira relevante para obtenção de maiores ganhos genéticos do que quando se utiliza apenas os dados fenotípicos (Meuwissen; Hayes; Goddard, 2001). Os marcadores moleculares têm sido utilizados nos programas de melhoramento de milho nas últimas décadas principalmente em programas de retrocruzamento, para diminuir o número de gerações para recuperação do genoma recorrente, e para seleção assistida de algumas características de herança simples. Mais recentemente, com o avanço das tecnologias de genotipagem e o surgimento dos marcadores do tipo SNP (do inglês, single nucleotide polymorphism - SNP) , milhares de marcadores podem ser genotipados em cada linhagem de milho, permitindo a aplicação de estudos de seleção genômica nessa cultura (Bernardo, 2008).

Os marcadores SNPs são caracterizados pela diferença em uma única base entre dois cromossomos homólogos ou entre dois indivíduos da mesma espécie, são marcadores codominantes e geralmente bialélicos, assumindo apenas duas formas alélicas (Júnior, 2013). As principais vantagens desse tipo de marcador são a ampla distribuição no genoma e a possibilidade de genotipar painéis de marcadores densos em alta escala, esses fatores promovem seu uso na seleção genômica ampla dentro dos programas de melhoramento genético (Meuwissen et al., 2001).

2.3 Seleção genômica ampla (GWS)

A utilização massiva de marcadores moleculares em técnicas de seleção baseada no genótipo foi proposta, inicialmente, por Meuwissen, Hayes e Goddard (2001), que definiram GWS como a seleção baseada em marcadores sem a necessidade de testes de significância em nível de marca ou identificação de um subconjunto de marcadores associados com a característica de estudo (Pires, 2017).

A seleção genômica consiste na estimação simultânea dos efeitos genéticos de marcadores dispersos em todo o genoma, sob o pressuposto de que a maioria dos

alelos de interesse esteja associado a esses marcadores (em desequilíbrio de ligação) e possa explicar grande parte da variância genética (Resende et al., 2014). Na obtenção dos modelos de seleção genômica, geralmente usa-se uma população de referência que pode ser dividida em duas partes: a primeira é conhecida como população de treinamento ou estimação, onde são utilizados os dados fenotípicos e genotípicos para estimar os efeitos genéticos de cada marcador molecular, a segunda é chamada de população de validação, onde os dados genotípicos são usados para predição dos valores genéticos genômicos (VGG) e os dados fenotípicos utilizados apenas na validação do modelo (Goddard & Hayes, 2009). Posteriormente, o modelo construído poderá ser aplicado na predição de fenótipos em populações de seleção para uma determinada característica de interesse, sendo que essas populações precisam ser apenas genotipadas (Goddard & Hayes, 2009)

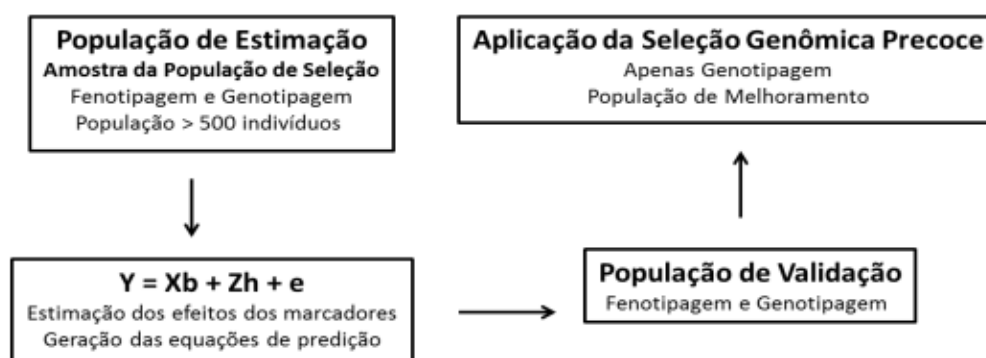


Figura 1. Esquema de aplicação da seleção genômica ampla em um programa de melhoramento genético (Resende Júnior, 2010).

A predição do desempenho de híbridos de milho sem a necessidade da realização dos cruzamentos a campo, possibilita a avaliação de centenas de milhares de híbridos de forma virtual, pelas suas predições, e selecionar apenas aqueles com as melhores predições para serem cruzados e avaliados a campo. Com isso, a tendência é que sejam avaliados a campo mais híbridos com alto desempenho, aumentando a chance de obtenção de híbridos mais produtivos em combinações que de outra forma não teriam sido avaliadas, contribuindo assim para o ganho genético (Windhausen et al., 2012).

A utilidade do modelo aplicado na seleção genômica depende da acurácia de predição, que é obtida através da correlação entre os valores preditos e observados na população de validação (Kaler et al., 2022). Inúmeros são os fatores que interferem na acurácia de predição de um modelo de seleção genômica, dentre eles podemos destacar o tamanho da população de treinamento, densidade de marcadores utilizada, herdabilidade da característica, número de QTLs envolvidos e o modelo estatístico utilizado (Grattapaglia & Resende, 2011).

2.4 Modelos de Seleção Genômica Ampla

A seleção genômica pode ser realizada utilizando diferentes abordagens estatísticas, que vão desde métodos de regressão linear, regressão não linear e bayesianos, até métodos de aprendizado de máquina (do inglês, Machine Learning - ML) (Mendes, 2015). De forma geral, os métodos estatísticos proporcionam diferentes suposições sobre o número e distribuição dos efeitos dos QTLs, apresentando diferentes abordagens sobre o modelo genético associado ao caráter estudado (Romero, 2017). O método ideal para a GWS deve acomodar a arquitetura genética da característica em estudo, além de considerar problemas de multicolinearidade e dimensionalidade dos marcadores utilizados no modelo (Mendes, 2015).

Um dos métodos mais utilizados na seleção genômica é o Ridge Regression–Best Linear Unbiased Prediction (rrBLUP), que possui pressuposição mais simples e assume que os marcadores explicam a mesma proporção da variância genética, geralmente sua aplicação é adotada utilizando apenas efeitos aditivos (Wang et al., 2018; Endelman, 2011). Outro modelo linear paramétrico que tem sido aplicado na seleção genômica é o Bayesian Lasso (BL). Essa metodologia permite ajustar um modelo de regressão linear de alta dimensão, incluindo marcadores moleculares densos, informação de pedigree e outras covariáveis (Pérez et al., 2010). O BL consegue incorporar efeitos aditivos e de dominância na sua implementação (Wellman & Bennewitz, 2012), o que pode aumentar a acurácia de predição, especialmente para espécies que apresentam heterose, como é o caso de híbridos (Ishimori et al., 2020).

Métodos não-paramétricos de ML têm sido aplicados com bastante sucesso na seleção genômica (Wang et al., 2018). Esses métodos são capazes de aprender automaticamente padrões de dados e otimizar parâmetros do modelo (Ma et al., 2014). Diferentemente dos métodos convencionais de seleção genômica, as metodologias baseadas em inteligência artificial não requerem suposições a priori sobre a relação entre as entradas (marcadores) e a saída (valores fenotípicos), permitindo grande flexibilidade para lidar com diferentes arquiteturas genéticas, incluindo efeitos aditivos, dominantes e epistáticos (Yan et al., 2021). Outra vantagem dos algoritmos de ML é a ótima capacidade em lidar com grandes conjuntos de dados através da seleção automática de variáveis (Crossa et al., 2017; Yan et al., 2021).

Um dos algoritmos de ML que vem sendo explorado com resultados bem promissores em estudos de seleção genômica é o Gradiente Boosting (GB) (Ghafouri-Kesbi et al., 2017; Abdollahi-Arpanahi et al., 2020). O GB é um método de aprendizado por agrupamento que consiste em reunir vários modelos fracos para estabelecer um modelo mais robusto e complexo (Montesinos-López et al., 2022). Uma das variantes do GB – LightGBM (Light Gradient Boosting Machine) tem demonstrado performance superior em lidar com grandes conjuntos de dados, atuando com alta velocidade de treinamento e acurácia preditiva (Yan et al., 2021). O LightGBM é uma estrutura de código aberto que foi desenvolvido pela Microsoft e baseia-se em algoritmos de árvore de decisão. Segundo Yan et al., (2021) esse modelo apresenta ótima performance em termos de precisão, estabilidade e eficiência computacional para estudos de GWS. Apesar do potencial preditivo e computacional desse modelo, há poucos estudos sobre a performance desse algoritmo na seleção genômica de plantas, especialmente para características quantitativas em híbridos de milho.

Para híbridos é esperado que o efeito de desvio de dominância seja um componente importante para o desempenho do modelo preditivo. Nesse sentido a comparação da acurácia entre modelos aditivos, aditivo-dominantes e de machine learning pode ser fundamental para definir a melhor metodologia para o estudo de seleção genômica ampla. A escolha do melhor modelo de predição para cada situação particular, promoverá maior eficiência nos programas de melhoramento, e economia de recursos pela decisão acertada dos híbridos realmente superiores (Pereira, 2015).

3. MATERIAL E MÉTODOS

3.1 Condução experimental e Dados fenotípicos

Nesse estudo foram avaliados 11.197 híbridos simples provenientes de cruzamentos entre linhagens endogâmicas de diferentes grupos heteróticos do banco de germoplasma da empresa LongPing High-Tech, Brasil. Os híbridos foram avaliados na safrinha de 2022 em três localidades diferentes, seguindo as seguintes quantidades: 3.985 híbridos em Jataí-GO, 1.811 em Rolândia-PR e 5.401 em Sorriso-MT. O delineamento experimental foi de Blocos Aumentados de Federer, com 32 tratamentos em cada bloco. Cada bloco continha 4 testemunhas, que se repetiram em todos os blocos. As parcelas experimentais continham 4 linhas de 5m, com espaçamento de 50cm entre as linhas e três plantas por metro linear, correspondendo a densidade de 60 mil plantas por ha. Foram avaliadas as características Produtividade (PG) e Umidade (UG) de grãos na colheita. Como todas as parcelas foram colhidas na mesma data, a umidade de colheita está relacionada ao ciclo de maturidade dos híbridos.

3.2 Herdabilidade

As estimativas de herdabilidade para as duas características (PG e UG) foram obtidas a partir de uma subpopulação dos híbridos avaliados em cada local, sendo que essa subpopulação foi composta por 2 repetições de cada híbrido. A herdabilidade no sentido amplo foi estimada através dos estimadores dos quadrados médios da análise de variância, utilizando-se o programa estatístico GENES (Cruz, 2013).

3.3 Dados de marcadores moleculares

A extração de DNA das linhagens parentais (5682 Linhagens) foi realizada a partir de tecidos foliares, utilizando-se o Kit Fast ID Genomic DNA Extraction® (GENETIC ID., Fairfield, IA, EUA), de acordo com as instruções do fabricante. A genotipagem das amostras foi feita na plataforma Ion GeneStudio S5 Prime (Thermo Fisher Scientific., Waltham, MA, EUA) seguindo as instruções do fabricante. As linhagens parentais foram genotipadas com painel de 5076 marcadores do tipo SNP.

Os genótipos dos híbridos foram obtidos pela concatenação dos genótipos dos parentais de cada híbrido.

3.4 Modelos estatísticos

Três modelos foram utilizados para estimar os efeitos dos marcadores e prever a performance fenotípica dos híbridos:

rrBLUP (Ridge Regression-Best Linear Unbiased Prediction – Modelo Aditivo)

O modelo rr-BLUP (Meuwissen et al., 2001) calcula os efeitos dos marcadores por regressão:

$$y = Xb + Zm + e$$

em que, y é o vetor de observações fenotípicas; b é o vetor de efeitos fixos; m é o vetor de efeitos aleatórios das marcas; e é o vetor de resíduos aleatórios; X e Z são as matrizes de incidência para b e m respectivamente. O modelo rrBLUP utilizado nesse trabalho considera apenas os efeitos aditivos na predição dos VGG. Assim, a matriz Z foi codificada por valores -1, 0 e 1, sendo -1 e 1 os homozigotos e 0 os heterozigotos. Esse modelo foi implementado através do pacote R “rrBLUP” (Endelman, 2011).

BL (Bayesian Lasso - Modelo Aditivo-Dominante)

O modelo bayesiano pode ser descrito da seguinte forma:

$$y = 1_n \mu + Z\mu + \varepsilon$$

onde y é o vetor de observações, Z é a matriz de incidência para efeitos aleatórios, e μ é o vetor de efeitos aleatórios (Liu et al., 2018). O modelo BL considera efeitos aditivos e de dominância na predição dos VGG. Assim, a matriz Z foi codificada

em -1, 0 e 1 para contemplar os efeitos aditivos, sendo os valores -1 e 1 os homozigotos e 0 os heterozigotos. Para efeitos de dominância, os valores foram 1 para heterozigotos e 0 para os homozigotos. Esse modelo foi implementado através do pacote R “BLR” (De los Campos et al., 2013).

LightGBM (Light Gradient Boosting Machine)

A construção do modelo LightGBM foi realizada através do pacote `lightgbm` versão 3.3.4 utilizando linguagem R (Shi et al., 2022). A codificação dos genótipos para esse modelo foi realizada através dos valores -1, 0 e 1, sendo -1 e 1 os homozigotos e 0 os heterozigotos.

3.5 Avaliação e comparação dos modelos

A capacidade preditiva dos três modelos foi avaliada utilizando o procedimento de validação cruzada 3-fold (Montesinos-López et al., 2022). Nesse procedimento, o conjunto de dados fenotípicos de cada local avaliado foi dividido aleatoriamente em 3 grupos de tamanhos similares, com um grupo representando a população de validação e dois grupos representando a população de treinamento (Figura 2). Esse processo foi repetido três vezes, de forma que cada grupo fosse uma vez o grupo de validação. Na prática, foram gerados 27 modelos de predição (3 repetições x 3 locais de avaliação x 3 métodos).

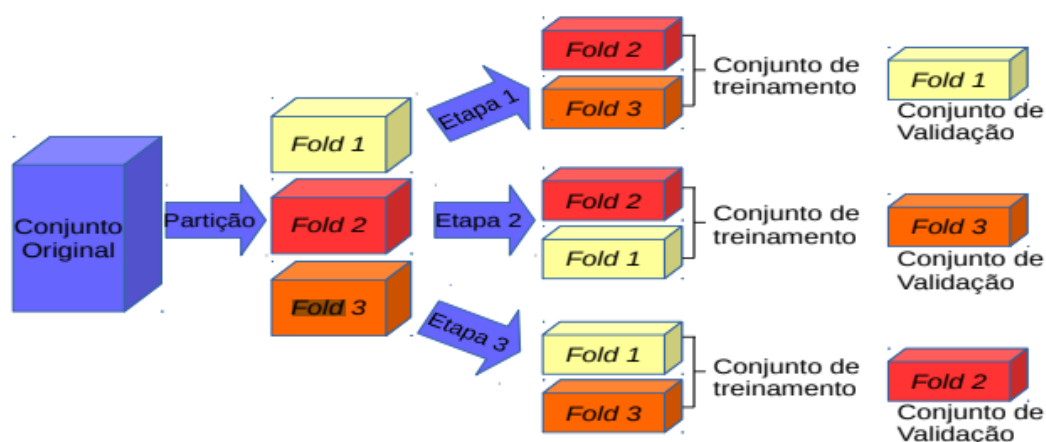


Figura 2. Esquema do processo de validação cruzada K = 3 (Clebiano da Costa Sá, 2018).

A acurácia de predição de cada modelo foi mensurada pela média da correlação simples (correlação de Pearson) entre os fenótipos observados e preditos ($y, \hat{y}$) obtidos a partir das três validações cruzadas para os três locais avaliados (Jataí-GO, Rolândia-PR e Sorriso-MT). Para determinar se existiam diferenças estatísticas entre as médias de acurácia preditiva dos modelos avaliados foram realizadas análises de variância e, em seguida, o teste de Tukey a 5% de probabilidade, considerando cada local e característica avaliada.

O segundo parâmetro utilizado na avaliação da capacidade preditiva dos modelos foi o T10Recall. Esse parâmetro indica a proporção dos híbridos que ficaram concomitantemente entre os 10% das melhores predições e 10% dos melhores resultados na avaliação de campo. As estimativas desse parâmetro foram obtidas pela média das porcentagens calculadas a partir das três validações cruzadas para os três locais avaliados.

A performance computacional dos modelos foi avaliada através do tempo gasto (segundos) para estimar os efeitos dos marcadores para as duas características avaliadas. Todos os modelos foram gerados a partir de um computador Intel Core i7-8550, contendo @1.80GHz e 8GB de RAM.

4. RESULTADOS E DISCUSSÕES

4.1 Resultados

A estatística descritiva e herdabilidade das duas características avaliadas (PG e UG) para os três locais (Jataí-GO, Rolândia-PR e Sorriso-MT) estão descritos na Tabela 1. Os valores de herdabilidade para PG variaram entre 31.23% e 73.39%, e para UG entre 63.37% e 87.15%. O local que apresentou as maiores estimativas de herdabilidade foi Rolândia, tendo 73.39% para PG e 87.15% para UG. Jataí-GO e Sorriso-MT apresentaram herdabilidades inferiores a 50% para PG (31.23% e 34.62% respectivamente), e foram equivalentes para UG (64.89% e 63.37 respectivamente).

Tabela 1. Estatísticas descritivas e estimativa de herdabilidade para as características de Produtividade (PG) e Umidade (UG) nos três locais de avaliação.

	Jataí-GO		Rolândia-PR		Sorriso-MT	
Estatística Descritiva	PG	UG	PG	UG	PG	UG
Média	9.65	21.19	8.10	24.40	11.81	20.54
Variância	1.34	3.39	2.36	9.17	2.52	4.54
Mínimo	3.20	13.40	2.72	14.10	5.45	10.20
Máximo	16.41	30.30	12.57	38.90	17.24	29.20
Herdabilidade (%)	31.23	64.89	73.39	87.15	34.62	63.37

A Figura 3 apresenta os resultados obtidos a partir da validação cruzada (3-fold) para as duas características em cada ambiente de avaliação. Cada gráfico boxplot ilustra os valores de correlação entre os fenótipos preditos e observados em cada um dos três modelos avaliados (BL, LightGBM e rrBLUP). A capacidade preditiva de cada modelo foi avaliada através da média das acurácias de predição obtidas a partir do processo de validação cruzada.

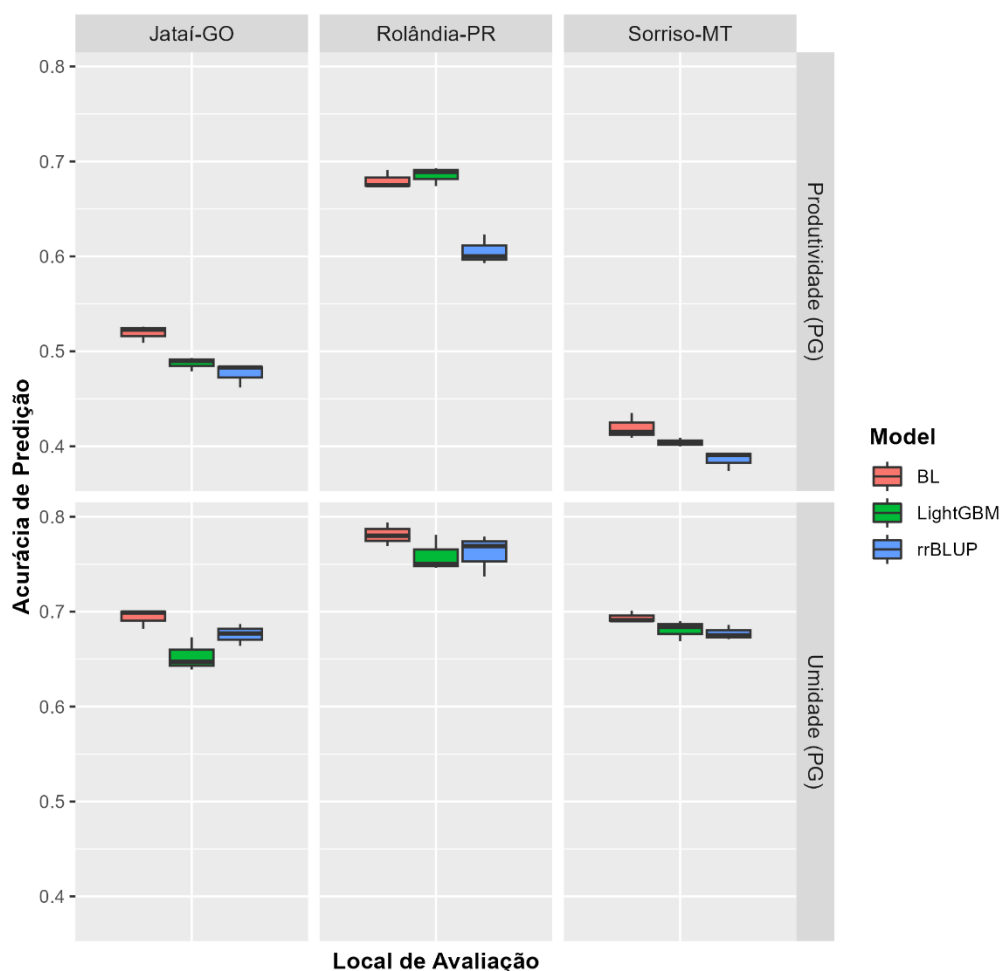


Figura 3. Comparação da acurácia de predição de PG e UG para os modelos BL, LightGBM e rrBLUP nos três locais de avaliação (Jataí-GO, Rolândia-PR e Sorriso-MT). Cada gráfico representa a acurácia preditiva utilizando o método de validação cruzada 3-fold (CV).

As médias de acurácia de predição dos modelos BL, LightGBM e rrBLUP para as duas características avaliadas nos três locais está descrita na Tabela 2. As médias de acurácia preditiva para os locais avaliados variaram de 0.404 a 0.685 para PG, e 0.653 a 0.781 para UG, sendo que as maiores médias foram obtidas para Rolândia-PR.

Para Jataí-GO, o modelo BL apresentou a maior acurácia de predição para as duas características avaliadas, tendo valores de 0.514 para PG e 0.694 para UG. Essas acurácias foram significativamente maiores que os valores de acurácia dos outros dois métodos, pelo teste de Tukey ao nível de 5% de probabilidade. Ainda nesse local, foi observado que o modelo LightGBM e rrBLUP tiveram valores de acurácia preditiva equivalentes, tanto para PG quanto para UG. Para Rolândia-PR, os modelos BL e LightGBM tiveram as maiores acurácias preditivas de PG (0.680 e 0.685

respectivamente), e diferiram estatisticamente do método rrBLUP. Para UG, os modelos tiveram performance preditiva equivalente. Para Sorriso-MT, não houve diferença significativa entre os modelos avaliados para UG, entretanto para PG, os modelos BL e LightGBM apresentaram os maiores valores de acurácia preditiva, sendo 0.420 e 0.404 respectivamente. Em relação às características avaliadas, foi observado que a acurácia de predição para UG foi maior que a acurácia de predição para PG nos três modelos, e para os três locais avaliados.

Tabela 2. Média das acurácias de predição obtidas a partir da validação cruzada K=3 para os três modelos desse estudo (BL, LightGBM e rrBLUP) nos três locais de avaliação. Médias seguidas pela mesma letra na coluna não diferem estatisticamente a nível de 5% de probabilidade pelo teste de Tukey.

Modelos	Jataí-GO		Rolândia-PR		Sorriso-MT	
	PG	UG	PG	UG	PG	UG
BL	0.519 ^a	0.694 ^a	0.680 ^a	0.781 ^a	0.420 ^a	0.694 ^a
LightGBM	0.487 ^b	0.653 ^b	0.685 ^a	0.759 ^a	0.404 ^{ab}	0.681 ^a
rrBLUP	0.477 ^b	0.676 ^{ab}	0.605 ^b	0.762 ^a	0.386 ^b	0.677 ^a

A relação entre os valores fenotípicos preditos e observados para os três modelos avaliados em cada local está apresentado na Figura 4.

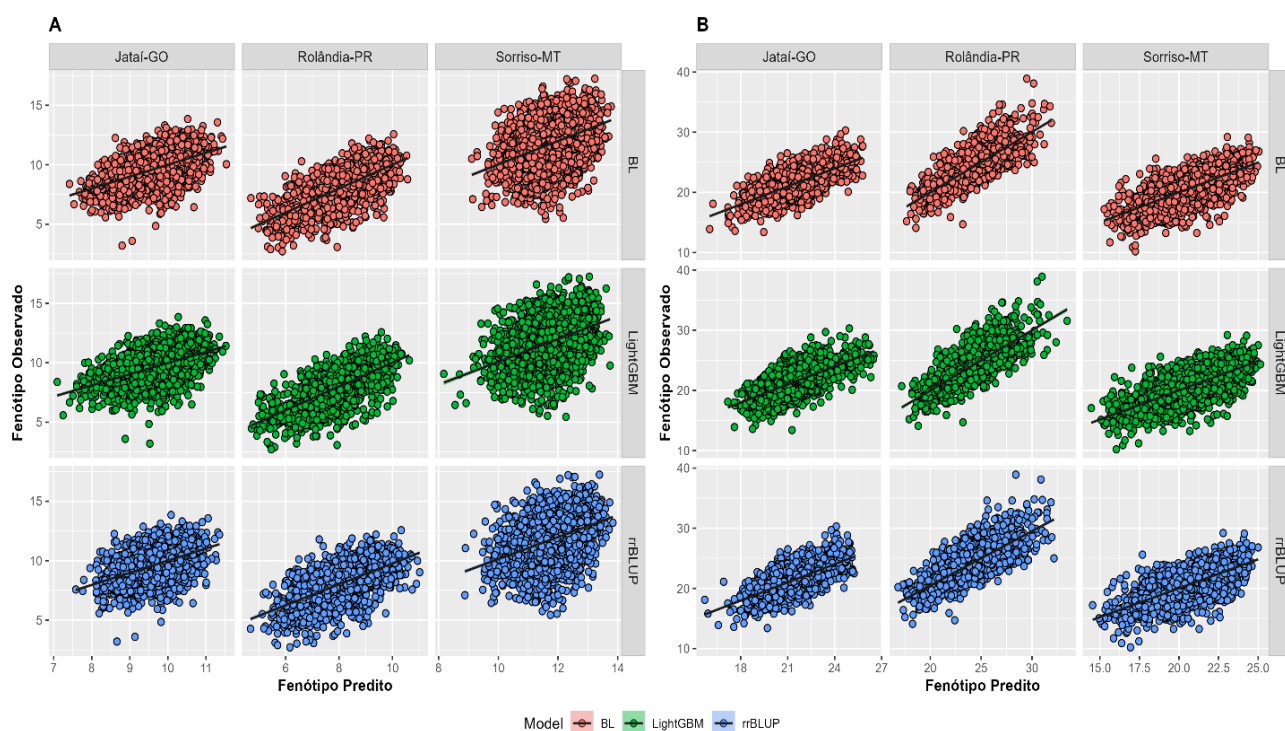


Figura 4. Gráficos de dispersão mostrando a relação entre os valores fenotípicos preditos e observados para os três modelos (BL, LightGBM e rrBLUP). **A.** Relação para Produtividade (PG). **B.** Relação para Umidade (UG).

O parâmetro T10Recall indica a porcentagem de correspondência entre os dados obtidos na predição e os dados observados a campo para os 10% dos híbridos mais produtivos para PG, e os 10% mais precoces para UG.

Para esse parâmetro foi observado que o modelo BL apresentou as maiores porcentagens de PG nos três locais avaliados, sendo 34%, 38% e 28% para Jataí-GO, Rolândia-PR e Sorriso-MT respectivamente, sendo que nesse último local sua porcentagem foi igual à obtida pelo modelo LightGBM. Para UG, o modelo BL teve a maior porcentagem para Jataí-GO (46%) e Rolândia-PR (68%). O modelo LightGBM apresentou as menores porcentagens de PG e UG para Jataí-GO (32% e 40% respectivamente) e para Rolândia-PR (36% e 62% respectivamente), porém esse modelo teve os maiores valores de T10Recall para Sorriso-MT, sendo 28% para PG e 43% para UG.

O modelo rrBLUP apresentou 33% de correspondência para PG e 45% para UG em Jataí-GO, e teve os menores valores de PG e UG para Sorriso-MT (25% e 39% respectivamente). Para Rolândia-PR, o modelo rrBLUP apresentou a menor porcentagem para PG (36%), e foi o modelo com a maior porcentagem obtida para

UG, juntamente com o modelo BL (66%).

Tabela 3. Porcentagens de híbridos que ficaram acima dos 10% mais produtivos para PG, e acima dos 10% mais precoces para UG, tanto no resultado de campo (Observado) como na predição (Predito). Quanto maior a porcentagem maior a correspondência entre os dados Preditos e Observados.

Modelos	Jataí-GO		Rolândia-PR		Sorriso-MT	
	PG	UG	PG	UG	PG	UG
BL	34%	46%	38%	68%	28%	41%
LightGBM	32%	40%	36%	62%	28%	43%
rrBLUP	33%	45%	37%	66%	25%	39%

Na Figura 5 está descrito o tempo computacional (segundos) gasto no treinamento dos modelos para as duas características (PG e UG) em cada local de avaliação. Os valores foram obtidos a partir da média do tempo gasto em cada processo de validação cruzada (3-fold).

O tempo computacional do modelo BL foi consideravelmente maior que os outros dois métodos em todos os locais avaliados. Esse modelo levou 8.214, 3.186 e 9.282 segundos para o treinamento dos modelos de PG e UG para Jataí-GO, Rolândia-PR e Sorriso-MT respectivamente. O modelo de Machine Learning (LightGBM) apresentou o menor tempo computacional para todos os locais avaliados, sendo 12,42, 10,04 e 17,64 segundos para Jataí-GO, Rolândia-PR e Sorriso-MT respectivamente. O desempenho computacional do modelo rrBLUP no treinamento dos modelos foi intermediário aos modelos BL e LightGBM, apresentando valores de 319,53, 298,19 e 362,41 segundos para Jataí-GO, Rolândia-PR e Sorriso-MT respectivamente.

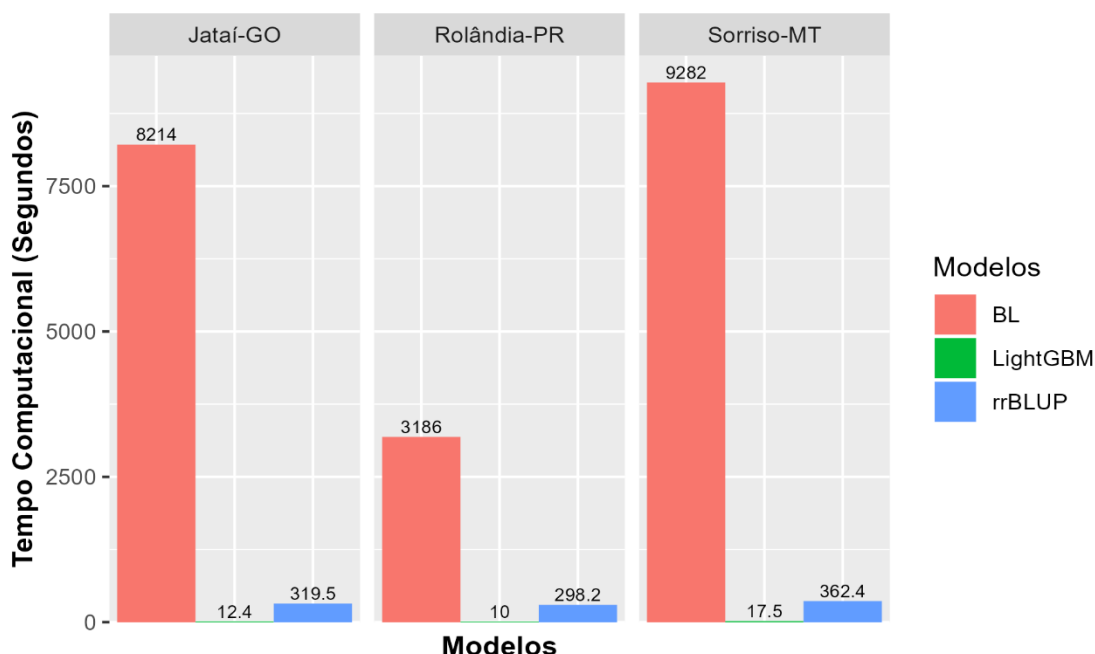


Figura 5. Comparação do tempo computacional gasto (segundos) pelos três métodos avaliados (BL, rrBLUP e LightGBM) no treinamento dos modelos para as duas características (UG e PG).

O tempo computacional gasto na predição dos VGG, e não somente na fase de treinamento dos modelos, também é um fator bastante importante na escolha do método a ser utilizado na seleção genômica. Isso se deve ao fato dessa etapa ser constantemente realizada na rotina dos programas de melhoramento genético de milho para seleção precoce das melhores combinações híbridas. Nesse sentido, foi realizado a estimativa do tempo gasto (segundos) para obtenção dos valores fenotípicos preditos a partir dos modelos construídos para PG e UG na validação cruzada de Jataí-GO.

Nessa estimativa o tempo gasto na predição dos valores fenotípicos variou de 1,04 a 53,48 segundos. O maior tempo computacional para predição dos VGG foi do modelo rrBLUP (53,48) e o menor tempo foi obtido para o modelo LightGBM (1,04), o modelo BL foi intermediário aos dois modelos (22,40 segundos). O tempo estimado para predição dos VGG em cada modelo está apresentado na Figura 6.

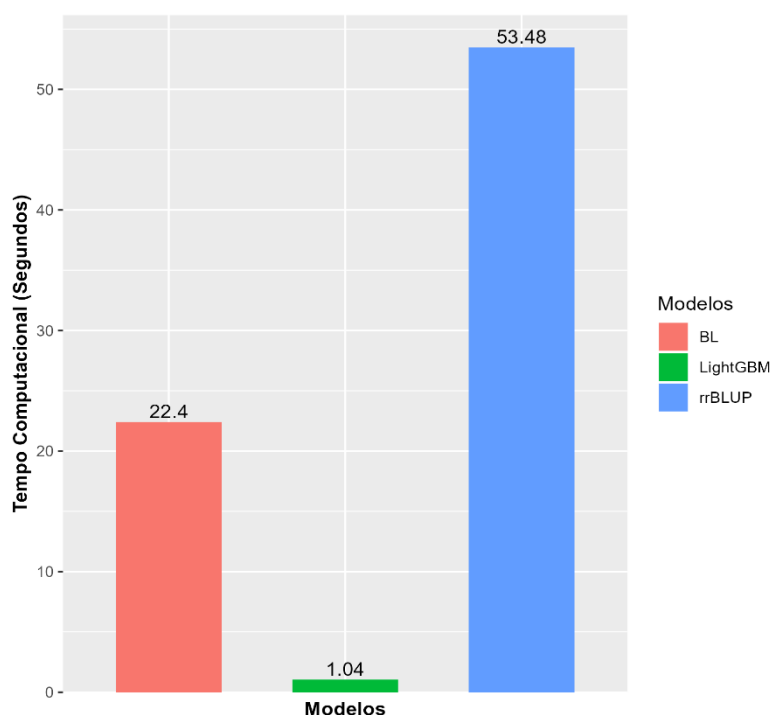


Figura 6. Comparação do tempo computacional gasto (segundos) pelos três métodos avaliados (BL, LightGBM e rrBLUP) na predição dos VGG para as duas características (UG e PG).

4.2 Discussão

No presente estudo foi realizada a comparação da performance de predição genômica de dois métodos paramétricos lineares (BL e rrBLUP) e um não-paramétrico de Machine learning (LightGBM). A análise comparativa foi realizada utilizando a abordagem de seleção genômica em híbridos de milho provenientes de cruzamentos entre linhagens endogâmicas da empresa LongPing High-Tech nas localidades de Jataí-GO, Rolândia-PR e Sorriso-MT na Safrinha de 2022. Duas importantes características da cultura do milho foram avaliadas nesse estudo: Produtividade (PG) e Umidade (UG) de grãos. Vale ressaltar que a umidade na colheita é utilizada como estimador de maturidade fisiológica (Sala et al., 2007), uma vez que todas as parcelas foram colhidas na mesma data.

Os critérios utilizados na comparação dos métodos foram: Acurácia de predição, definida como a correlação entre os dados fenotípicos observados e preditos; T10Recall, que indica a proporção de híbridos da população de validação

que ficaram entre os 10% mais produtivos para PG e 10% mais precoces para UG, considerando tanto a predição como a avaliação de campo; além da performance computacional no treinamento dos modelos de PG e UG.

As acurácias de predição obtidas a partir dos modelos BL, LightGBM e rrBLUP para as duas características avaliadas nos três locais (Jataí-GO, Rolândia-PR e Sorriso-MT) foram médias-altas, variando de 0.386 a 0.680 para PG, e 0.653 a 0.781 para UG (Tabela 2). Resultados similares foram encontrados por Terraillon et al. 2023, que utilizando dados de híbridos de milho avaliados dentro do mesmo ano obtiveram acurácias preditivas que variaram de 0.300 a 0.710 para PG e 0.490 a 0.830 para UG.

As maiores estimativas de acurácia preditiva de PG e UG foram observadas para Rolândia-PR. Sorriso-MT e Jataí-GO tiveram as menores médias de acurácia para PG e UG respectivamente. De forma geral, para as duas características avaliadas nos três locais, foi observado que a partir das maiores herdabilidades as acurácias preditivas também foram mais elevadas. Esse comportamento é esperado, e está de acordo com o trabalho reportado por Liu et al., 2018, no qual trabalhando com seleção genômica para produção de grãos em milho verificaram que a herdabilidade tem correlação positiva com a acurácia de predição. Também em concordância com os resultados aqui encontrados, Hayes et al. 2010 mostraram que aumentando a herdabilidade de 10% para 90%, a acurácia preditiva também foi aumentada de 0.35 para 0.75.

Considerando os três locais avaliados foi observado que as maiores acurácias de predição para PG foram obtidas pelo modelo aditivo-dominante (BL), e as menores acurácias foram obtidas pelo modelo aditivo (rrBLUP). A amplitude média entre esses dois modelos em termos de capacidade preditiva para PG foi de 0.050.

Para UG, o modelo BL também teve as maiores acurácias preditivas comparado ao modelo aditivo (rrBLUP), porém a diferença não foi estatisticamente significativa. Esses resultados indicam que o efeito genético de dominância considerado no modelo BL contribuiu de forma mais efetiva na predição de PG. Resultados bem similares foram encontrados no estudo realizado por Ferrão et al. (2020), que ao comparar modelos aditivos e aditivo-dominantes na seleção genômica de híbridos de milho verificaram acurácias de predição maiores a partir dos modelos com incorporação do efeito de dominância, especialmente para a característica de

produtividade de grãos.

Devido ao efeito da heterose em híbridos de milho, a incorporação de efeitos não-aditivos nos modelos de predição genômica pode contribuir consideravelmente para o aumento da acurácia de predição (Fritsche-Neto et al., 2021), mas essa performance superior também depende da arquitetura genética e complexidade da característica de interesse. Isso foi demonstrado por Bernardo (1996), que ao avaliar a acurácia de predição para diferentes caracteres na cultura do milho verificou que a inclusão do efeito de dominância na modelagem não foi consistente para todas as características estudadas.

A influência dos efeitos genéticos de cada modelo na predição genômica está relacionada com a herdabilidade da característica (Zhang et al., 2017). Neste estudo, os valores de herdabilidade de UG foram maiores do que os valores de PG, para todos os locais avaliados. Isso indica que a proporção do efeito de aditividade na variância fenotípica de UG é maior do que em PG. Essa proporção pode explicar as menores diferenças encontradas entre as acurácias preditivas dos modelos aditivo (rrBLUP) e aditivo-dominante (BL) para UG, porém maiores diferenças para PG.

Outra abordagem que tem sido aplicada em estudos de seleção genômica é baseada em Machine Learning (ML). Os algoritmos dessa abordagem são capazes de capturar efeitos não-aditivos sem precisar de nenhuma pressuposição sobre a arquitetura genética da característica avaliada. Isso é resultado do seu alto poder em captar padrões complexos e não lineares (Abdollahi-Arpanahi et al., 2020; Wang et al., 2018). Nesse estudo, foi avaliado e comparado a performance de predição genômica do método de ML (LightGBM) em relação aos dois métodos estatísticos convencionais BL e rrBLUP.

Em termos de acurácia de predição, o modelo LightGBM foi equivalente aos modelos BL e rrBLUP na predição de UG para Rolândia-PR (0.759) e Sorriso-MT (0.681). Houve diferença estatística apenas para Jataí-GO (0.653), onde o LightGBM teve performance preditiva equivalente ao rrBLUP, porém foi inferior à performance do modelo BL (Tabela 2). Para PG, sua acurácia de predição foi equivalente ao modelo BL para Rolândia-PR (0.685) e Sorriso-MT (0.404), diferindo-se apenas na predição de Jataí-GO (0.487). Nesse local sua performance foi similar ao modelo aditivo rrBLUP. Esses resultados confirmam que para características

contendo ação gênica complexa, como é o caso de PG, os modelos de ML podem desempenhar de forma equivalente ou até melhor que os métodos estatísticos convencionais (Abdollahi-Arpanahi et al., 2020). Em concordância com os resultados aqui encontrados, Perez et al. (2022) também observaram a superioridade de modelos de ML quando comparados à modelos lineares aditivos para predição genômica de características de menor herdabilidade.

Outro parâmetro bastante útil na avaliação da capacidade preditiva dos modelos de predição genômica é o T10Recall, que indica a porcentagem de correspondência entre os dados obtidos na predição e os dados observados a campo para os 10% dos híbridos mais produtivos para PG, e os 10% mais precoces para UG. Na prática, através desse parâmetro conseguimos estimar para os modelos avaliados, o índice de acerto daqueles híbridos de fato superiores no campo.

Para esse parâmetro os resultados indicaram que, de forma geral o modelo LightGBM foi inferior aos modelos paramétricos (BL e rrBLUP) para Jataí-GO e Rolândia-PR, porém esse modelo foi superior para Sorriso-MT (Tabela 3). Os modelos BL e rrBLUP tiveram porcentagens bem similares para Jataí-GO e Rolândia-PR, em ambas características (PG e UG). Diante do exposto, percebe-se que não há uma clara associação entre os resultados obtidos para acurácia de predição e T10Recall, mas vale ressaltar a estabilidade de desempenho do modelo BL frente aos outros dois métodos para esses dois parâmetros.

Em termos de eficiência computacional, o modelo LightGBM se destacou como o mais eficiente e rápido para os três locais avaliados (Jataí-GO, Rolândia-PR e Sorriso-MT), e em ambas características (PG e UG). Para o local de maior quantidade de dados de treinamento (Sorriso-MT), esse modelo levou menos de 20 segundos para estimar os efeitos dos marcadores das duas características avaliadas (Figura 5). Essa excelente performance computacional do LightGBM era esperada, e é justificada pela forma que as árvores são construídas nesse algoritmo, que diferentemente de outros algoritmos de gradient boosting é baseada em folhas ao invés de uma estrutura baseada em níveis (Shi et al., 2022).

Dentre os dois modelos paramétricos, o modelo aditivo (rrBLUP) apresentou a melhor performance computacional para o treinamento dos modelos de PG e UG, sendo consideravelmente mais rápido que o modelo aditivo-dominante (BL).

Considerando o local de maior quantidade de dados de treinamento (Sorriso-MT), a diferença de tempo entre os dois modelos foi maior que 8.000 segundos (Figura 5). A baixa performance computacional do modelo BL está associada ao método utilizado nas inferências estatísticas, que nesse caso é baseado em um algoritmo de Monte Carlo via Cadeias de Markov (MCMC) (Yang et al., 2020).

Apesar do BL ter sido amplamente superado pelas outras duas metodologias (LightGBM e rrBLUP) em termos de performance computacional para o treinamento dos modelos de PG e UG, essa inferioridade é atenuada com a maior rapidez na execução das estimativas dos VGG. Uma vez que a etapa de treinamento e validação dos modelos é finalizada, o tempo gasto na predição dos VGG de novas populações de seleção se torna o maior gargalo. Nesse cenário, o modelo BL se torna o segundo mais rápido, levando cerca de 31 segundos a menos que o modelo rrBLUP e 21 segundos a mais que o modelo LightGBM (Figura 6). Como em projetos de predição genômica nos programas de melhoramento genético pode ser necessário a predição de alguns milhões de híbridos, o tempo computacional é um parâmetro importante na escolha dos modelos de predição.

5. CONCLUSÃO

O modelo BL apresentou maior estabilidade em termos de capacidade preditiva. Se o tempo computacional não for um fator limitante, esse é o método mais indicado para obter maiores acurácias de predição para situações similares as encontradas nesse estudo. Diante de um cenário com alta quantidade de dados para treinamento dos modelos, o método aditivo (rrBLUP) e de Machine Learning (LightGBM) podem ser usados para atingir uma melhor performance computacional, entretanto essa utilização pode implicar em perda significativa de acurácia preditiva.

A herdabilidade das características tem influência na acurácia de predição dos modelos. Acurácias mais elevadas são obtidas com herdabilidades maiores. Os efeitos genéticos decorrentes de herdabilidades menores podem ser melhor capturados a partir de modelos aditivo-dominante e de Machine Learning.

6. REFERÊNCIAS

- Abdollahi-Arpanahi R, Gianola D, Peñagaricano F (2020) Deep learning versus parametric and ensemble methods for genomic prediction of complex phenotypes. **Genet Sel Evol**, 52, 12.
- Abimilho. Associação brasileira das indústrias do milho. **Estatísticas - 6º Acompanhamento da Safra de Milho - 2018/19**. São Paulo. 2019. Disponível em: <<http://www.abimilho.com.br/estatisticas/acompanhamento-da-safra>> Acesso em: 02 dez. 2020.
- Andrade L, Sousa M, Oliveira E, Resende M, Azevedo C (2019) Cassava yield traits predicted by genomic selection methods. **PLoS One**, 14:e0224920.
- Bernardo R (1996) Testcross additive and dominance effects in best linear unbiased prediction of maize single-cross performance. **Theoretical and Applied Genetics**, Berlin, v. 93, n. 7, p. 1098-1102
- Bernardo R (2008) Molecular markers, and selection for complex traits in plants: learning from the last 20 years. **Crop Science**, v. 48, n. 5, p. 1649-1664.
- Cantelmo NF, Von Pinho RG, Balestre M (2017) Genome-wide prediction for maize single-cross hybrids using the GBLUP model and validation in different crop seasons. **Mol Breeding** 37, 51.
- Costa NV (2020) **Seleção de linhagens de milho a partir do esquema de topcrosses com testadores recíprocos intergrupos**. 77 f. Dissertação (Mestrado em Genética e Melhoramento de Plantas). Universidade Federal de Goiás, Goiânia.
- Crossa J, Perez-Rodriguez P, Cuevas J, Montesinos-Lopez O, Jarquin D, de Los Campos G, Burgueno J, Gonzalez-Camacho JM, Perez-Elizalde S, Beyene Y, Dreisigacker S, Singh R, Zhang X, Gowda M, Roorkiwal M, Rutkoski J, Varshney RK (2017) Genomic Selection in Plant Breeding: Methods, Models, and perspectives. **Trends Plant Sci**, 22, 961-975.
- Cruz CD (2013) GENES, A software package for analysis in experimental statistics and quantitative genetics. **Acta Scientiarum**. v.35, n.3, p.271-276.
- De los Campos G, Pérez P, Vazquez AI, Crossa J (2013) Genome-enabled prediction using the BLR (Bayesian linear regression) R-package. In *Methods in Molecular Biology*, **Springer: Berlin/Heidelberg**, Germany, 299–320.
- De Oliveira AA, Resende MFR, Ferrão LFV, Amadeu RR, Guimarães LJM, Guimarães CT, Pastina MM, Margarido GRA (2020) Genomic prediction applied to multiple traits and environments in second season maize hybrids. **Heredity** 125, 60–72.

Endelman JB (2011) Ridge regression and other kernels for genomic selection with R package rrBLUP. **Plant Genome**, 4, 250–255.

Ferrão LFV, Marinho CD, Munoz PR, Resende Jr MFR (2020) Improvement of predictive ability in maize hybrids by including dominance effects and marker × environment models. **Crop Sci**.

Fritsche-Neto R, Galli G, Borges KLR, Costa-Neto G, Alves FC, Sabadin F, Lyra DH, Morais PPP, Braatz de Andrade LR, Granato I, Crossa J (2021) Optimizing Genomic-Enabled Prediction in Small-Scale Maize Hybrid Breeding Programs: A Roadmap Review. **Front. Plant Sci**, 12, 658267.

Galli G, Lyra DH, Alves FC, Granato ÍSC, e Sousa MB, Fritsche-Neto R (2018) Impact of phenotypic correction method and missing phenotypic data on genomic prediction of maize hybrids. **Crop Sci**, 58, 1481–1491.

Ghafouri-Kesbi F, Rahimi-Mianji G, Honarvar M, Nejati-Javaremi A (2017) 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, 229–236.

Grattapaglia D, Resende MDV (2011) Genomic selection in forest tree breeding. **Tree Genetics & Genomes**, v. 7, p. 241-255.

Goddard ME, Hayes BJ (2009) Mapping genes for complex traits in domestic animals and their use in breeding programs. **Nat Rev Genet**, 10, 381-91.

Guimarães PS, Paterninani RRL, Souza AP, Laborda PR, Oliveira KM (2007) Correlação da heterose de híbridos de milho com divergência genética entre linhagens. **Pesquisa Agropecuária Brasileira**, v. 42, n. 6, p. 811- 816.

Hayes BJ, Daetwyler HD, Bowman P, Moser G, Tier B, Crump R, Khatkar M, Raadsma HW, Goddard ME (2010) Accuracy of genomic selection: comparing theory and results. In **Proceedings of the 18th Conference of the Association for the Advancement of Animal Breeding and Genetics**. Barossa Valley, Australia.

Ishimori M, Hattori T, Yamazaki K, Takanashi H, Fujimoto M, Kajiya-Kanegae H, Yoneda J, Tokunaga T, Fujiwara T, Tsutsumi N, Iwata H (2020) Impacts of dominance effects on genomic prediction of sorghum hybrid performance. **Breed Sci**, 70(5), 605-616.

Júnior OPM (2013) **Seleção genômica ampla (GWS) aplicada ao melhoramento populacional**. 38 f. Dissertação (Mestrado em Genética e Melhoramento de Plantas) – Universidade Federal de Goiás, Goiânia – GO.

Kaler AS, Purcell LC, Beissinger T, Gillman JD (2022) Genomic prediction models for traits differing in heritability for soybean, rice, and maize. **BMC Plant Biol**, 22, 87.

Liu X, Wang H, Wang H, Guo Z, Xu X, Liu J, Wang S, Li WX, Zou C, Prasanna BM, Olsen MS, Huang C, Xu Y (2018) Factors affecting genomic selection revealed by empirical evidence in maize. **Crop J**, 6, 341–352.

Ma C, Zhang HH, Wang X (2014) Machine learning for Big Data analytics in plants. **Trends Plant Sci**, 19(12), 798-808.

Mendes MP (2015) **Predição genômica de híbridos simples de milho**, 106 f. Dissertação (Doutorado em Genética e Melhoramento de Plantas) – Universidade de São Paulo, Piracicaba – SP.

Meuwissen TH, Hayes BJ, Goddard ME (2001) Prediction of total genetic value using genome-wide dense marker maps. **Genetics**, 157: 1819–1829.

Montesinos-López OA, Gonzalez HN, Montesinos-López A, Daza-Torres M, Lillemo M, Montesinos-López JC, Crossa J (2022) Comparing gradient boosting machine and Bayesian threshold BLUP for genome-based prediction of categorical traits in wheat breeding. **Plant Genome**.

Montesinos-López OA, Montesinos-López A, Crossa J (2022) Multivariate Statistical Machine Learning Methods for Genomic Prediction. **International Publishing: Cham, Switzerland**, 115-117

Paterniani MEAGZ, Guimarães OS, Luders RR, Gallo PB, Souza AP, Laborda PR, Oliveira KM (2008) Capacidade combinatória, divergência genética entre linhagens de milho e correlação com heterose. **Bragantia**, Campinas, v. 67, n. 3, p.639-648.

Pereira HD (2015) **Predição genômica ampla de híbridos simples de milho considerando modelo não aditivo**, 79 f. Dissertação (Doutorado em Genética e Melhoramento de Plantas) – Universidade Federal de Lavras, Lavras – MG.

Perez BC, Bink MCAM, Svenson KL, Churchill GA, Calus MPL (2022) Prediction performance of linear models and gradient boosting machine on complex phenotypes in outbred mice. **G3 (Bethesda)**, 4, 12(4):jkac039.

Pérez P, de los Campos G, Crossa J, Gianola D (2010) Genomic-enabled prediction based on molecular markers and pedigree using the Bayesian linear regression package in R. **plant. Genome**, 3, 106–116.

Pires LPM (2017) **Predição genômica do desempenho de híbridos de milho considerando a interação genótipos por ambientes**. 76 f. Tese (Doutorado em Agronomia/Fitotecnia). Universidade Federal de Lavras, Lavras.

Pungpapong V, Muir WM, Li X, Zhang D, Zhang M (2012) A fast and efficient approach for genomic selection with high-density markers. **G3: Genes Genomes Genetics**, 2, 1179-1184.

R Core Team (2013) A language and environment for statistical computing, Vienna, Austria: **Foundation for Statistical Computing**.

Resende Júnior MFR (2010) **Seleção genômica ampla no melhoramento vegetal**. 2010. 67 f. Dissertação (Mestrado em Genética e Melhoramento de Plantas) – Universidade Federal de Viçosa, Viçosa – MG.

Romero Júnior MFR (2017) **Seleção genômica usando genotipagem de baixa saturação no melhoramento genético do cafeeiro**. 67 f. Dissertação (Doutorado em Genética e Melhoramento de Plantas) – Universidade Federal de Viçosa, Viçosa – MG.

Sá CC (2018) **Métodos de validação tradicional e temporal aplicados à avaliação de classificadores de RNAs codificantes e não codificantes**. 102 f. Dissertação (Mestrado em Bioinformática) – Instituto de Matemática e Estatística, São Paulo – SP.

Sala RG, Andrade FH, Westgate ME (2007) Maize kernel moisture at physiological maturity as affected by the source–sink relationship during grain filling. **Crop Science**, 47, 711–714

Schrag TA, Frisch M, Dhillon B, Melchinger AE (2009) Marker-based prediction of hybrid performance in maize single-crosses involving doubled haploids. **Maydica**, 54, 353–362.

Schrag TA, Möring J, Melchinger AE, Kusterer B, Dhillon BS, Piepho HP, Frisch M (2010) Prediction of hybrid performance in maize using molecular markers and joint analyses of hybrids and parental inbreds. **Theoretical and Applied Genetics**, 120:451-461.

Shi Y, Ke G, Soukhavong D, Lamb J, Men, Q, Finley T, Wang T, Chen W, Ma W, Ye Q, Liu T, Titov N, Yan Y, Microsoft Corporation, Dropbox Inc, Loden J, Daeschler D, Rodola G, Ferreira A, Lemire D, Zverovich V, IBM Corporation, Cortes D (2022) Lightgbm: Light Gradient Boosting Machine. R package version 3.3.4.

Technow F, Schrag TA, Schipprack W, Bauer E, Simianer H, Melchinger AE (2014) Genome properties and prospects of genomic prediction of hybrid performance in a breeding program of maize. **Genetics**, 197,1343–55.

Terraillon J, Roeber FK, Flachenecker C, Frisch M (2023) Training set designs for prediction of yield and moisture of maize test cross hybrids with unreplicated trials. **Front. Plant Sci.** 14:1080087.

Wang X, Xu Y, Hu Z, Xu C (2018) Genomic selection methods for crop improvement: Status and prospects. **The Crop Journal**, 6, 330-340.

Wellmann R, Bennewitz J (2012) Bayesian models with dominance effects for genomic evaluation of quantitative traits. **Genet. Res. (Camb)**, 94, 21–37.

Windhausen VS, Atlin GN, Hickey JM, Crossa J, Jannink JL, Sorrells ME, Raman B, Cairns JE, Tarekegne A, Semagn K, Beyene Y, Grudloyma P, Technow F, Riedelsheimer C, Melchinger AE (2012) Effectiveness of genomic prediction of maize hybrid performance in different breeding populations and environments. **G3 Genes Genomes Genetics**, 2, 1427–1436.

Yan J, Xu Y, Cheng Q, Jiang S, Wang Q, Xiao Y, Ma C, Yan J, Wang X (2021) LightGBM: accelerated genomically designed crop breeding through ensemble learning. **Genome Biol** 22, 271.

Yang Y, Wainwright MJ, Jordan MI (2016) On the computational complexity of high-dimensional Bayesian variable selection. **The Annals of Statistics** 44 (6), 2497–2532.

Zhang A, Pérez-Rodríguez P, San Vicente F, Palacios-Rojas N, Dhliwayo T, Liu Y, Cui Z, Guan Y, Wang H, Zheng H, Olsen M (2022) Genomic prediction of the performance of hybrids and the combining abilities for line by tester trials in maize. **Crop J**, 10, 109–116.

Zhang A, Wang H, Beyene Y, Semagn K, Liu Y, Cao S, Cui Z, Ruan Y, Burgueño, J, San Vicente F, Olsen M, Prasanna BM, Crossa J, Yu H, Zhang X (2017) Effect of Trait Heritability, Training Population Size and Marker Density on Genomic Prediction Accuracy Estimation in 22 bi-parental Tropical Maize Populations. **Front Plant Sci**, 8, 8:1916.

Zhao Y, Mette MF, Reif JC (2015) Genomic selection in hybrid breeding. **Plant Breed**. 134, 1–10.