PATRICK WÖHRLE GUIMARÃES

REDUÇÃO DE DIMENSIONALIDADE EM PREDIÇÕES BIOMÉTRICAS

Dissertação apresentada à Universidade Federal de Viçosa, como parte das exigências do Programa de Pós-Graduação em Ciência da Computação, para obtenção do título de *Magister Scientiae*.

Orientador: Alcione de Paiva Oliveira

VIÇOSA - MINAS GERAIS
2024

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

T
G963r
2024

Guimarães, Patrick Wöhrle, 1975-
Redução de dimensionalidade em predições biométricas / Patrick Wöhrle Guimarães. – Viçosa, MG, 2024.
54 f.: il. (algumas color.).

Orientador: Alcione de Paiva Oliveira.
Dissertação (mestrado) - Universidade Federal de Viçosa, Departamento de Informática, 2024.
Inclui bibliografia.
DOI: https://doi.org/10.47328/ufvbbt.2024.215
Modo de acesso: World Wide Web.

1. Redução de dimensão (Estatística). 2. Polimorfismo de Nucleotídeo Único. 3. Predição. 4. Biometria. I. Oliveira, Alcione de Paiva, 1962-. II. Universidade Federal de Viçosa. Departamento de Informática. Programa de Pós-Graduação em Ciência da Computação. III. Título.

CCD 22. ed. 519.5

Bibliotecário(a) responsável: Renata de Fátima Alves CRB-6/2875

PATRICK WOHRLE GUIMARÃES

REDUÇÃO DE DIMENSIONALIDADE EM PREDIÇÕES BIOMÉTRICAS

Dissertação apresentada à Universidade Federal de Viçosa, como parte das exigências do Programa de Pós-Graduação em Ciência da Computação, para obtenção do título de *Magister Scientiae*.

APROVADA: 21 de março de 2024.

Assentimento:

Patrick Wohrle Guimarães
Autor

Alcione de Paiva Oliveira
Orientador

Dedico este trabalho aos bravos pesquisadores brasileiros que fazem ciência num país desigual e formam gerações imbuídas desse mesmo sonho.

AGRADECIMENTOS

Meu agradecimento inicial é voltado a Universidade Federal de Viçosa (UFV) que persiste em fazer ciência, pesquisa e extensão muitas vezes num cenário adverso de recursos e num período que muitas vidas foram perdidas (pandemia de SARS-COV-2). Mais especificamente, agradeço aos professores e funcionários dos Programas de Pós-Graduação em Ciência da Computação e em Estatística da UFV aonde realizei os meus créditos para concluir o mestrado. Agradeço também aos demais funcionários, da UFV e terceirizados, por toda estrutura oferecida e pela excelência no serviço prestado de maneira gratuita numa universidade pública.

Tive muitas experiências enriquecedoras e angustiantes nas disciplinas cursadas (como aluno regular e/ou ouvinte) e só foi possível superar cada uma graças aos colegas de curso e gostaria de poder agradecer a generosidade de cada um nessa parceria. A maior lembrança do mestrado que vou levar para a vida será desses momentos e na impossibilidade de agradecer a cada um, deixo meu muito obrigado representando todos os alunos nas figuras dos meus colegas (três - 03 - Ângelo, Lucas e Leonardo) de 2020.02 de Pós-Graduação em Ciência da Computação da UFV.

Agradeço ao meu orientador professor Alcione de Paiva Oliveira que ao longo de todo o meu mestrado sempre teve uma atitude positiva, uma palavra amiga e sempre foi parte da solução (e não do problema). Agradeço também ao professor Cosme Damião Cruz que com sua simplicidade, otimismo e excelência; consegue fazer com que a Biometria seja um problema de pesquisa estimulante mesmo para um estudante que entende muito pouco de Genética Quantitativa.

Agradeço ao professor Ricardo dos Santos Ferreira do Programas de Pós-Graduação em Ciência da Computação da UFV pela participação em minha banca de defesa e mais especificamente por me permitir lapidar meu problema de pesquisa cursando uma disciplina sua (em Aprendizado de Máquina). Agradeço também ao professor Ricardo Costa Pinto e Santos do Instituto Sudeste de Minas Gerais (IFSEMG) campus de Juiz de Fora pelos encontros (virtuais) que me geraram ideias inclusas nesse projeto e os demais correlatos que estou envolvido.

Por fim, um agradecimento especial aos familiares. A família da minha esposa que me acolheu em Viçosa-MG para a realização do mestrado (ao meu sogro e a minha sogra Rita), a minha mãe (Selma) que acompanha minha aventura pela área de educação com seu otimismo (e novidades diárias) e *"in memorian"* a minha avó (Nieves) que não resistiu a pandemia mas sempre foi uma entusiasta das minhas escolhas acadêmicas. Um agradecimento mais do que especial a minha esposa Patrícia que sempre me incentivou nos momentos difíceis e foi fundamental para a conclusão

do curso.

Fiz essa jornada (mestrado) uma vez em 2002 e agora repito em 2024 de novo. Será a última?

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

"Your compile timed out."
(Overleaf)

RESUMO

WÖHRLE-GUIMARÃES, Patrick, M.Sc., Universidade Federal de Viçosa, março de 2024. **Redução de Dimensionalidade em Predições Biométricas**. Orientador: Alcione de Paiva Oliveira.

A computação aplicada tem um papel crucial no Melhoramento Genético seja no sequenciamento e análise de genoma, na identificação de marcadores moleculares ou no aprimoramento de técnicas de inteligência computacional. Um dos domínios do Melhoramento Genético que tem feito extenso uso de dados reais e simulação é a Biometria. Um dos focos da Biometria é a seleção de modelos eficientes para o processo de Seleção Genômica Ampla (GWS) utilizando marcadores moleculares (SNPs) e esse procedimento apresenta vários desafios. Esse estudo buscou fornecer respostas a dois (02) desses desafios: a) comparar eficiência de diferentes técnicas de Predição em Biometria para identificação de marcadores moleculares (SNPs) relevantes para controle genético de características simuladas; b) estabelecer procedimentos de Redução de Dimensionalidade (ou *Feature Selection*) de conjunto de marcadores moleculares (SNPs) para fins de Predição de características complexas simuladas, considerando interação gênica, dominância e herdabilidade diferenciada. Para elucidar tais problemas foram gerados dados por simulação (via programa Genes) do tipo marcadores moleculares (*Single Nucleotide Polymorphisms - SNPs*) e esse conjunto de dados reflete os principais problemas encontrados nesta linha de investigação (alta dimensionalidade, não-linearidade e multicolinearidade). A comparação da eficiência foi feita através da avaliação da acurácia seletiva de nove (09) modelos de Predição que seguem diferentes paradigmas no contexto da Biometria. Como principais resultados, pode-se destacar o aumento da eficiência preditiva à medida que o ruído dos dados diminui, a superioridade do paradigma da árvore (para baixos níveis de ruído, BOO) e a eficiência do paradigma da rede neural (para altos níveis de ruído, RBF). O segundo desafio teve como ponto de partida o fato de que Redução de Dimensionalidade (RD) tem se tornado uma ferramenta fundamental para melhorar a eficiência de modelos de Predição e mais especificamente ainda a classe de modelos do tipo *Feature Selection (FS)*. Visando contribuir com esse processo, este estudo fez uso três (03) técnicas de *Feature Selection* (*Bagging*, Sonda e *Stepwiswe*) aplicadas a um modelo do tipo (*Random Forest - RF*). Os resultados obtidos sinalizam melhorias na acurácia seletiva (R^2) entre 10% e 28% e uma melhor adequação de dois (02) modelos (*Bagging* e *Stepwiswe*) em detrimento a um modelo (Sonda), em captar de maneira mais adequada as situações que envolvem o controle gênico de uma característica. Por fim,a medida que o ruído diminui, a acurácia seletiva aumenta (isso vale para modelos com e sem RD) e as

taxas de crescimento da acurácia seletiva se tornam decrescentes.

Palavras-chave: Modelos de Predição. Redução de Dimensionalidade. Feature Selection. SNPs. Biometria.

ABSTRACT

WÖHRLE-GUIMARÃES, Patrick, M.Sc., Universidade Federal de Viçosa, March, 2024. **Dimensionality Reduction in Biometric Predictions**. Advisor: Alcione de Paiva Oliveira.

Applied computing plays a crucial role in Genetic Improvement, whether in genome sequencing and analysis, the identification of molecular markers or the improvement of computational intelligence techniques. One of the areas of breeding that has made extensive use of real data and simulation is biometrics. One of the focuses of Biometrics is the selection of efficient models for the Genome-Wide Selection (GWS) process using molecular markers (SNPs) and this procedure presents several challenges. This study sought to provide answers to two (02) of these challenges: a) to compare the efficiency of different Biometrics Prediction techniques for identifying relevant molecular markers (SNPs) for genetic control of simulated traits; b) to establish procedures for Dimensionality Reduction (or Feature Selection) of a set of molecular markers (SNPs) for the purposes of Prediction of complex simulated traits, considering gene interaction, dominance and differentiated heritability. In order to elucidate these problems, simulation data was generated (via the Genes program) for molecular markers (Single Nucleotide Polymorphisms - SNPs) and this data set reflects the main problems encountered in this line of research (high dimensionality, non-linearity and multicollinearity). Efficiency was compared by evaluating the selective accuracy of nine (09) prediction models that follow different paradigms in the context of Biometrics. The main results were the increase in predictive efficiency as data noise decreased, the superiority of the tree paradigm (for low levels of noise, BOO) and the efficiency of the neural network paradigm (for high levels of noise, RBF). The second challenge was based on the fact that Dimensionality Reduction (DR) has become a fundamental tool for improving the efficiency of prediction models, and more specifically the class of Feature Selection (FS) models. In order to contribute to this process, this study used three (03) Feature Selection techniques (Bagging, *Sonda* and Stepwiswe) applied to a Random Forest (RF) model. The results obtained indicate improvements in selective accuracy (R^2) between 10% and 28% and a better suitability of two (02) models (Bagging and Stepwiswe) to the detriment of one model (*Sonda*), in more adequately capturing situations involving the genetic control of a trait. Finally, as noise decreases, selective accuracy increases (this is true for models with and without RD) and selective accuracy growth rates become decreasing.

Keywords: Prediction Models. Dimensionality Reduction. Feature Selection. SNPs. Biometric.

LISTA DE FIGURAS

LISTA DE TABELAS

SUMÁRIO

Capítulo 1

Introdução

A busca de eficiência em modelos de Predição e Classificação é um ponto recorrente em várias áreas de conhecimento e mais especificamente em Biometria. A Biometria que se faz referência nesse trabalho é uma subárea do melhoramento genético que utiliza dados oriundos de fenômenos biológicos e os interpreta levando em conta conceitos teóricos da Genética Quantitativa, com o intuito de aperfeiçoar modelos (procedimentos estatísticos) e desenvolver arquiteturas (inteligência computacional).

Para entender a busca da eficiência em Biometria é importante discutir alguns pontos iniciais. O primeiro ponto é que os modelos biométricos geralmente fazem uso de marcadores moleculares do tipo SNP (*Single Nucleotide Polymorphisms*), sendo que tais marcadores cobrem amplamente o genoma, permitindo a incorporação de informações moleculares no processo de Seleção Genômica Ampla (GWS) e na predição (ou classificação) do mérito genético dos indivíduos [1].

O segundo ponto é que trabalhar com dados do tipo marcador molecular (SNP) gera alguns desafios estatísticos. Dentre esses desafios, os mais recorrentes são: dimensionalidade dos dados, multicolinearidade e não linearidade (interação gênica ou epistasia). O problema de dimensionalidade dos dados está associado ao fato de que ao mapear todo genoma do indivíduo ou população de interesse, o número de marcadores moleculares (SNPs) seja quase sempre muito maior que o número de indivíduos [2]. A questão da multicolinearidade entre marcadores moleculares (SNPs) está associada ao espaçamento que existe entre cada marca e sua localização no cromossomo em si. Ambos os fatos geram existência de alta correlação entre os marcadores moleculares.

O último ponto envolve a origem dos dados considerados (fenômeno biológico) e as leis que regem tais relações (genética quantitativa). Um ajuste de um modelo eficiente deve ser capaz de reproduzir princípios biológicos (por exemplo, a via biossintética) ou mesmo princípios da Genética Quantitativa (por exemplo, efeitos de dominância e epistasia). Mais especificamente, o ajuste de qualquer modelo deve ser feito levando em conta uma equação fundamental da genética quantitativa que estabelece que a manifestação de uma caraterística observável (Fenótipo - P) resulta da soma de dois efeitos (Genótipo e ruído ambiental, $G + E$), e este é outro problema que

testa a eficiência dos modelos [3][1].

Em termos práticos esse último ponto ressalta que entre modelos que seguem paradigmas diferentes há alguns que se ajustam melhor aos fenômenos inerentes a Biometria, mas o ruído (variação da herdabilidade em diferentes conjuntos de dados) pode tornar isso menos aparente e deve ser levado em conta. Em síntese, melhorar a eficiência de modelos biométricos pode ser uma questão associada ao processo de fenotipagem, a superação de desafios estatísticos (dimensionalidade dos dados, multicolinearidade e não linearidade) ou a incorporação de princípios de genética quantitativa (e/ou biológicos) na seleção dos modelos.

Uma maneira de melhorar a eficiência de modelos biométricos de Predição (ou Classificação) é fazer uso de métodos de Redução de Dimensionalidade [4, 5]. Mais especificamente ainda no contexto da Biometria, fazer uso de técnicas do tipo *Feature Selection* por causa da sua carecterística principal. Os métodos do tipo *Feature Selection* fazem uma seleção de um subconjunto de variáveis do total do conjunto de dados original sem alterar a sua representação inicial e por esse motivo são os métodos de Redução de Dimensionalidade (RD) utilizados em estudos que fazem uso de marcadores moleculares (SNPs).

A literatura sobre *Feature Selection* tem crescido de maneira exponencial nas publicações da IEEE e da ACM no período de 1990 a 2020 [6][2]. Essa literatura exalta a importância de se utilizar um método de Redução de Dimensionalidade do tipo *Feature Selection* e alguns benefícios desse processo [7, 8, 9, 10]: a) a redução do número de variáveis irrelevantes, b) o aumento da precisão da previsão. c) a redução do sobreajuste, d) a melhoria da interpretabilidade do modelo, e) a criação de algoritmos mais rápidos e econômicos, e por fim f) o reforço a estabilidade dos modelos.

1.1 Objetivos Gerais e Específicos

O objetivo geral deste trabalho é comparar a eficiência de modelos de Predição e métodos de Redução de Dimensionalidade (RD) no contexto da Biometria e mais especificamente no contexto de marcadores moleculares (SNPs) gerados por processo de simulação. Os objetivos específicos são:

1. Comparar eficiência de diferentes técnicas de Predição em Biometria para identificação de marcadores moleculares (SNPs) relevantes para controle genético de características simuladas;

[1] Esta relação pode ser apresentada em termos de variância $(VP = VG + EV)$ e a hereditariedade pode ser definida a partir desta equação como: $h^2 = VG/VP$ (quanto da variância total é variância genética).

[2] Usando como métrica a contagem de publicações por ano dos artigos relacionados a palavra-chave *"Feature Selection"*.

2. Estabelecer procedimentos de Redução de Dimensionalidade (ou *Feature Selection*) de conjunto de marcadores moleculares (SNPs) para fins de Predição de características complexas simuladas, considerando interação gênica, dominância e herdabilidade diferenciada;

3. Confrontar a eficiência de modelos de Predição utilizando tanto conjuntos de dados originais quanto aqueles submetidos à Redução de Dimensionalidade (RD).

1.2 O problema e sua importância

Esse estudo buscou desenvolver uma aplicação de computação em Biometria e com esse intuito fez uso de maneira sequencial de Modelos de Predição, e posteriormente, de técnicas de Redução de Dimensionalidade (ou mais especificamente *Feature Selection*).

Desde a proposta inicial da Seleção Genômica (SG) diversos modelos estatísticos e de Aprendizado de Máquina são propostos para a tarefa de predizer valores genéticos. A gama de modelos propostos é vasta, entretanto, nenhum modelo, estatístico ou de Aprendizado de Máquina, oferece uma solução ótima para todos os casos [11, 12].

Nesse contexto, [13] avaliaram doze (12) modelos de Seleção Genômica (SG) e não conseguiram identificar um modelo que se destacasse em todos os cenários. A literatura ressalta a dificuldade de generalizar sobre uma única técnica capaz de resolver todos os problemas de Predição na área genômica e por isso para se determinar o melhor modelo deve-se testar diferentes metodologias no mesmo conjunto de dados.

Como se trata de um problema complexo, uma possível solução é alimentar a literatura científica com uma série de informações, com cenários bem diferenciados, para que, com o acúmulo de informação, possa ser estabelecido um padrão de recomendação de procedimentos mais eficazes. Nesta tarefa é necessário pesquisas que abordem diferentes metodologias de Predição, diferentes particularidade das variáveis respostas (controle genético, ruído ambiental, distribuição dos dados) e diferentes particularidades das variáveis preditoras (tipos, quantidade, densidade, efeitos genéticos aditivos e não aditivos, dentre outros).

Por isso, esse estudo pode ser considerado agregativo no contexto da informação. Quanto aos métodos, foram empregados nove (09) modelos de Predição com diferentes conceitos ou paradigmas e que permitissem captar efeitos lineares e não-lineares dado o conjunto de dados gerados por simulação. Quanto às variáveis respostas, foram apresentadas respostas de características representativas de situações de alto, médio e baixo resíduo ambiental, por meio da escolha arbitrária de herdabilidades. Quando às variáveis preditoras, uma preocupação foi exemplificar a Predição usando marcadores codominantes, em genomas de boa densidade e, principalmente, com

ação gênica do tipo epistático.

Em relação aos métodos de Redução de Dimensionalidade e mais especificamente as técnicas de *Feature Selection*, foram selecionadas técnicas para captar efeitos lineares e não-lineares no processo de seleção e exclusão de variáveis. Os métodos também foram utilizados numa estratificação de eliminação de variáveis num contexto de proporção fixa *versus* variável e o intuito foi determinar a relevância de se estabelecer um % ótimo de descarte da matriz de *Features*[3].

Em síntese, vislumbra-se que os resultados obtidos com esse estudo se somem a tantos outros para que, em estudos de metanálise, alguma generalização possa ser estabelecida no futuro.

1.3 Organização dos Capítulos

Esta dissertação foi elaborada no formato de "artigos científicos" conforme o modelo estabelecido pelo Conselho Técnico de Pós Graduação da Universidade Federal de Viçosa [14]. Nesse padrão, esse trabalho foi dividido em quatro (04) partes: a presente introdução geral (Capítulo 1), dois (02) artigos científicos publicados e/ou em fase de publicação (Capítulos 2 e 3) e as conclusões gerais da pesquisa que englobam pontos fundamentais ao longo do estudo (Capítulo 4).

Os artigos científicos foram desenvolvidos para estudar um problema recorrente em Computação (Redução de Dimensionalidade), mas aplicado a área da Biometria. O primeiro artigo é apresentado no Capítulo 2 e é intitulado como *"Predcitions in Biometric Models"*[4]. Esse artigo apresenta o conjunto de dados que foi utilizado ao longo da dissertação pela primeira vez (SNPs) e que foi gerado por simulação, introduz alguns conceitos fundamentais para analisar os resultados do trabalho e do artigo posterior (por exemplo: ruído, epistasia, via biossintética, etc...), lista de maneira sucinta nove (09) modelos de Predição e mostra como a seleção do modelo mais eficiente esta relacionada com a natureza do problema.

No Capítulo 3 é apresentado o artigo intitulado "*Feature Selection* em Biometria". Esse artigo faz uso do conhecimento gerado no artigo anterior e seleciona um modelo de Predição (*Random Forest - RF*) para testar modelos de Redução de Dimensionalidade (ou mais especificamente, *Feature Selection*) com proporções fixas (*Bagging* e Sonda) e variável (*Stepwise*). Por fim, no Capítulo 4, são apresentadas as considerações finais desse estudo.

[3]Esse último ponto foi implementado de uma maneira exploratória e não num contexto de otimização.

[4]O artigo foi feito em parceria com os pesquisadores Alcione de Paiva Oliveira e Cosme Damião Cruz e aceito para publicação (em fase final de edição) na *Acta Scientiarum. Agronomy.*

Referências

[1] Weverton Gomes da Costa, Maurício de Oliveira Celeri, Ivan de Paiva Barbosa, Gabi Nunes Silva, Camila Ferreira Azevedo, Aluizio Borem, Moysés Nascimento, and Cosme Damião Cruz. Genomic prediction through machine learning and neural networks for traits with epistasis. *Computational and Structural Biotechnology Journal*, 20:5490–5499, January 2022. ISSN 2001-0370. doi: 10.1016/j.csbj.2022.09.029. URL https://www.sciencedirect.com/science/article/pii/S2001037022004342.

[2] Cosme Damião Cruz, Caio Césio Salgado, and Leonardo Lopes Bhering. *Genômica Aplicada*. Suprema, Visconde do Rio Branco, 2013.

[3] Shizhong Xu. *Quantitative Genetics*. Springer International Publishing, Cham, 2022. ISBN 978-3-030-83939-0 978-3-030-83940-6. doi: 10.1007/978-3-030-83940-6. URL https://link.springer.com/10.1007/978-3-030-83940-6.

[4] Gabriel Feresin Pantalião, Marcelo Narciso, Cléber Guimarães, Adriano Castro, José Manoel Colombari, Flavio Breseghello, Luana Rodrigues, Rosana Pereira Vianello, Tereza Oliveira Borba, and Claudio Brondani. Genome wide association study (GWAS) for grain yield in rice cultivated under water deficit. *Genetica*, 144(6):651–664, December 2016. ISSN 0016-6707, 1573-6857. doi: 10.1007/s10709-016-9932-z. URL http://link.springer.com/10.1007/s10709-016-9932-z.

[5] Gareth James, Daniela Witten, Trevor Hastie, and Robert Tibshirani. *An introduction to statistical learning: with applications in R*. Springer texts in statistics. Springer, New York NY, second edition edition, 2021. ISBN 978-1-07-161417-4 978-1-07-161420-4. doi: 10.1007/978-1-0716-1418-1.

[6] Dipti Theng and Kishor K. Bhoyar. Feature selection techniques for machine learning: a survey of more than two decades of research. *Knowledge and Information Systems*, December 2023. ISSN 0219-1377, 0219-3116. doi: 10.1007/s10115-023-02010-5. URL https://link.springer.com/10.1007/s10115-023-02010-5.

[7] Isabelle Guyon and André Elisseeff. An introduction to variable and feature selection. *Journal of Machine Learning Research*, 3(7-8):1157–1182, 2003. ISSN 1533-7928(Electronic),1532-4435(Print). doi: 10.1162/153244303322753616. Place: US Publisher: MIT Press.

[8] Girish Chandrashekar and Ferat Sahin. A survey on feature selection methods. *Computers & Electrical Engineering*, 40(1):16–28, January 2014. ISSN 00457906.

doi: 10.1016/j.compeleceng.2013.11.024. URL https://linkinghub.elsevier.com/retrieve/pii/S0045790613003066.

[9] Yun Li, Tao Li, and Huan Liu. Recent advances in feature selection and its applications. *Knowledge and Information Systems*, 53(3):551–577, December 2017. ISSN 0219-1377, 0219-3116. doi: 10.1007/s10115-017-1059-8. URL http://link.springer.com/10.1007/s10115-017-1059-8.

[10] Jundong Li, Kewei Cheng, Suhang Wang, Fred Morstatter, Robert P. Trevino, Jiliang Tang, and Huan Liu. Feature Selection: A Data Perspective. *ACM Computing Surveys*, 50(6):1–45, November 2018. ISSN 0360-0300, 1557-7341. doi: 10.1145/3136625. URL https://dl.acm.org/doi/10.1145/3136625.

[11] J Crossa, P Pérez, J Hickey, J Burgueño, L Ornella, J Cerón-Rojas, X Zhang, S Dreisigacker, R Babu, Y Li, D Bonnett, and K Mathews. Genomic prediction in CIMMYT maize and wheat breeding programs. *Heredity*, 112(1):48–60, January 2014. ISSN 0018-067X, 1365-2540. doi: 10.1038/hdy.2013.16. URL https://www.nature.com/articles/hdy201316.

[12] Philipp E. Bayer, Jakob Petereit, Monica Furaste Danilevicz, Robyn Anderson, Jacqueline Batley, and David Edwards. The application of pangenomics and machine learning in genomic selection in plants. *The Plant Genome*, 14(3):e20112, November 2021. ISSN 1940-3372, 1940-3372. doi: 10.1002/tpg2.20112. URL https://acsess.onlinelibrary.wiley.com/doi/10.1002/tpg2.20112.

[13] Christina B Azodi, Emily Bolger, Andrew McCarren, Mark Roantree, Gustavo De Los Campos, and Shin-Han Shiu. Benchmarking Parametric and Machine Learning Models for Genomic Prediction of Complex Traits. *G3 Genes|Genomes|Genetics*, 9(11):3691–3702, November 2019. ISSN 2160-1836. doi: 10.1534/g3.119.400498. URL https://academic.oup.com/g3journal/article/9/11/3691/6026746.

[14] UFV, Conselho Técnico de Pós Graduação da Universidade Federal de Viçosa. Normas de redação de teses e dissertações. https://ppg.ufv.br/wp-content/uploads/2012/08/Normas-gerais-de-Teses-e-Dissertac%CC%A7o%CC%83es-12.pdf, 2019. Accessed: 11/01/2023.

Chapter 2

Predictions in Biometric Models

One of the domains of genetic enhancement that has extensively employed both simulation and authentic data is Biometrics. Selecting efficient models for the Genome-Wide Selection (GWS) process using molecular markers (SNPs) presents several challenges. Among these challenges is the effective identification of the optimal model for fitting a given dataset. To contribute to this endeavor, this paper's primary objective is to assess the predictive accuracy of nine (9) distinct models, each following different paradigms within the realm of Biometrics. The data employed in this study were generated through simulation, encompassing the primary issues encountered in this field of research, including high dimensionality, nonlinearity, and multicollinearity. As the primary findings, notable observations include the enhancement of predictive efficiency as data noise decreases, the predominance of the tree paradigm (for low noise levels, BOO), and the efficacy of the neural network paradigm (for high noise levels, RBF).

2.1 Introduction

Data analysis plays a pivotal role in decision-making and knowledge extraction across diverse research domains. The quest for identifying regular patterns and frameworks to address challenges is a burgeoning demand in computer science and fields that extensively employ this discipline. One such field that has embraced this approach is Biometrics.

Biometrics represents the forefront of genetic breeding, involving the analysis, processing, and interpretation of biological phenomena through data and theoretical concepts from Quantitative Genetics. It focuses on refining models (statistical procedures) and architectures (computational intelligence). An effective biometric analysis optimizes physical, financial, and human resources.

Efficiency in Biometrics often hinges on determining the most suitable model for fitting a given dataset [1]. Specifically, this is accomplished through predictive models rooted in various paradigms for data processing. Biometricians typically employ prediction models that utilize molecular markers of the SNP type (Single Nucleotide

Polymorphisms). These markers span the genome, enabling the incorporation of molecular information in predicting the genetic merit of individuals [2].

Leveraging molecular markers in Genome-Wide Selection (GWS) via prediction models presents several statistical challenges, with three primary ones being data dimensionality, multicollinearity, and nonlinearity (gene interaction or epistasis). Additionally, the manifestation of an observable characteristic (Phenotype - F) results from the interplay of two factors (Genotype and environmental noise, $G + E$), further testing the prediction models' efficiency (i.e., how to effectively fit prediction models, even with high noise in the dataset) [1].

Beyond statistical issues and environmental noise, biometricians face other challenges. One such challenge is selecting prediction models that can reflect biological principles (e.g., the biosynthetic pathway) or principles of Quantitative Genetics (e.g., dominance and epistasis effects).

Biometric models serve a vital purpose: selecting the most suitable individuals for breeding, thereby ensuring the efficiency of the selection process. Therefore, chosen prediction models must aptly classify the individuals under consideration, achieved through the selection of an appropriate metric.

Two measures to assess predictive techniques' effectiveness are selective accuracy (indicating the model's efficiency in ranking individuals correctly for selection) and predictive accuracy (estimated through mean square error, representing the model's efficiency in predicting values close to the true values). Both metrics can be used without compromising the analysis's inference or generalization, with the choice dependent on the study's primary objective.

Prediction studies, or model fitting, hold significant importance across various scientific fields. They seek to establish functional relationships between a response variable Y and a set of predictor variables X, enabling the prediction of the response variable using new information from the explanatory variables or understanding the cause-and-effect relationships unveiled by the model.

In genetic improvement, prediction has distinctive features that deserve mention. While it also aims to establish relationships between explanatory and response variables, it emphasizes the need not only for a robust predictor of Y (a directly observed and measured variable, termed phenotypic value, F) but also one that approximates the true genotypic value (G) upon which selection processes are based.

This situation can be exemplified by considering the prediction of complex and economically valuable traits, such as grain or fruit production, based on auxiliary variables like plant height, flowering, thatch width, and others. In such cases, the goal is to identify predictors that not only exhibit a good statistical model fit but also

[1] This relationship can be presented in terms of variance ($VF = VG + EV$) and heritability can be defined from this equation as: $h^2 = VG/VF$ (how much of the total variance is genetic variance).

result in more substantial gains through indirect selection. This involves considering information about genetic correlations and the heritability of traits when selecting the most suitable auxiliary variable. While various factors influence the choice of an effective explanatory trait, a general guideline favors traits with higher heritability and stronger correlations with the main variable. Such guidance can stem from theoretical foundations, empirical findings, or simulation studies.

Currently, prediction studies contemplate the utilization of additional information beyond what is routinely measured in agronomic trials. Spectral data, including infrared and imaging data, and particularly molecular information, have gained prominence. Molecular information is particularly valuable as it directly relates to loci controlling quantitative traits (QTLs) and aligns with Mendelian and quantitative genetics principles.

Molecular information, while still relatively costly for extensive use in many agricultural species, is of great interest because it provides genotypic values (referred to as genomic values in this case) that hold the potential for more substantial genetic gains. The availability of these markers, particularly SNP markers, sparks interest and encourages biometricians to employ more parameterized models. These models take into account allelic dose effects manifested in the markers, dominance effects that defy explanation through the recognition of homozygous and heterozygous forms, and especially the effects of epistatic interactions arising from the joint action of two or more markers influencing the expression of traits.

Consequently, predicting quantitative traits based on molecular information poses challenges in both genetic and statistical contexts, highlighting issues related to dimensionality and multicollinearity in prediction. It is essential to bear in mind that several factors will influence the selection of the best model across various agricultural species. The pursuit of a general rule of thumb for modeling guidance becomes pivotal, aiming to streamline efforts and concentrate resources on models that align best with prediction objectives while capturing the genetic nuances of the phenomena under scrutiny.

Also underscore the significance of comparing algorithms across a wide range of scenarios, considering the advancements in computing speeds, the evolution of graphics processing units (GPUs), and progress in backpropagation learning algorithms [3]. To delve into some of the mentioned issues, our primary goal is to evaluate the selective accuracy of nine (09) predictive models, each grounded in various paradigms applied within the field of Biometrics. The data have been generated via simulation and reflect the principal challenges outlined in this brief introduction, encompassing dimensionality reduction, multicollinearity, and nonlinearity. Additionally, we present related considerations in model fitting, such as sample size and data representativeness.

2.2 Material and Methods

2.2.1 Prediction and selected models

The prediction models chosen in this study adhere to the "algorithm modeling culture," with a primary focus on developing suitable algorithms for individual selection [4]. To achieve this objective, nine (09) models were specifically selected: a) Ridge Regression Best Linear Unbiased Predictor (RR-Blup), b) Multi-Layer Perceptron (MLP), c) Radial Basis Function Network (RBF), d) Decision Tree (DT), e) Bootstrap Aggregating (BAG), f) Random Forest (RF), g) Boosting (BOO), and h) Multivariate Adaptive Spline (linear and interacting MARS, MARS1 and MARS2)[2].

These models can be classified into three primary paradigms: a) Statistical - this paradigm relies on fundamental information regarding means, variances, covariances, and distributions; b-c) Computational Intelligence / Neural Network - in this paradigm, models are developed by repeatedly presenting all available data and incorporating learning rules to achieve a final model with maximum accuracy; d-h) Machine Learning - this paradigm involves the partitioning of predictor spaces and their organization in a hierarchical tree diagram structure, typically for model fitting or classification purposes.

The objective of employing models from diverse paradigms is to assess how each one addresses the challenges posed by the dataset (including issues like nonlinearity, multicollinearity, and dimensionality), accounts for the existing noise (which exhibits differentiated heritability), and accommodates the specific challenges associated with Genome-Wide Selection (GWS).

The Ridge Regression Best Linear Unbiased Predictor (RR-BLUP) is a model that employs a statistical approach to estimate genetic values in genetic improvement studies. RR-BLUP extends the linear regression model, making it a parametric method. It introduces a penalty term called "Ridge" to manage the impact of multicollinearity and enhance stability in the estimates, particularly in scenarios where the number of explanatory variables exceeds the number of observations [5, 6]. The shrinkage parameter, denoted as λ (penalty parameter), is defined as follows:

$$\lambda = \sigma_e^2 / \left(\sigma_g^2 / n_q\right) \tag{2.1}$$

Where: σ_e^2 is the error variance; σ_g^2 is the additive variance of the character; and n_q is the number of QTLs. The RR-Blup model has the following formulation [7]:

[2]This item provides a brief overview of the nine models used without delving into the specifics of mathematics, parameter setting, or algorithm tuning. Additionally, it presents some fundamental issues to be considered and suggestions for references for the deepening of each model.

$$y = Wb + Xm + e \tag{2.2}$$

Where: y is the vector of phenotypic observations (matrix of dimension $1000x1$); b is the vector of fixed effects (general mean) with incidence matrix W; m is the vector of random effects of markers with incidence matrix X (X is the incidence matrix composed of the values 1, 0, and -1 according to the number of alleles of the marker of the genotypes AA, Aa, and aa, respectively) with $m \sim N(0, I\sigma_g^2)$ and e refers to the vector of random errors with $e \sim N(0, I\sigma_e^2)$[3].

The prediction equations were formulated with the assumption that all loci contribute equally to the genetic variation, hence sharing a common σ_g^2. Consequently, the genetic variation attributed to each locus can be expressed as (σ_g^2/n), where σ_g^2 represents the total genetic variation, and n is the count of markers utilized. In this study, the RR-BLUP model employed the simplest version, considering only the additive effects of the predictor variables.

The Multi-layer Perceptron (MLP) model comprises artificial neurons organized into layers, including an input layer, hidden layers, and an output layer. MLP relies on five (05) fundamental concepts [8, 9]: artificial neuron, layers, synaptic weight, activation function (which introduces nonlinearity into the model), backpropagation, and supervised learning.

Each neuron conducts a linear combination of inputs, weighted by synaptic weights, adds a bias, and employs a nonlinear activation function. Backpropagation is used to fine-tune the synaptic weights through supervised learning, minimizing the discrepancy between predicted and expected outcomes.

The equation that summarizes the MLP model, considering two hidden layers, can be defined as follows:

$$y = f(W^2 * f(W^1 * X + b_1) + b_2) \tag{2.3}$$

Where: y represents the output of the MLP, which corresponds to the predicted genetic value for each F2 individual included in the training or validation dataset; X is an incidence matrix containing attribute values associated with markers, with values of -1, 0, and 1 representing the AA, Aa, and aa forms, respectively; W^1 is the weight matrix of the first hidden layer. Each element in this matrix corresponds to the weight associated with the connection between neurons in the input layer and the first hidden layer; b_1 is a bias vector for the first hidden layer. Each element in this vector serves as an additional adjustment term that enables the model to capture nonlinear behavior; f represents a nonlinear activation function applied to the input values of

[3]The effect of the λ penalty function is more easily noticeable in the matrix notation of the RR-Blup model.

neurons in the network. In our study, we employed the hyperbolic tangent function; W^2 is the weight matrix of the second hidden layer (or output layer, depending on the architecture). Each element in this matrix signifies the weight associated with the connection between neurons from the preceding layer and the current layer; b_2 is the bias vector for the second hidden layer (or output layer). Each element in this vector functions as an additional adjustment term.

The equation of the MLP model encapsulates the weighted combination of inputs, traversing hidden layers with nonlinear activation functions until it reaches the output layer, where the final prediction or classification is generated. During model training, the weights (W) and biases (b) are adjusted to optimize network performance and achieve the best fit with the training data.

The Neural Network topology can capture complex relationships between input and output data, primarily due to the introduction of nonlinearity[4] through the activation function[5]. The number of hidden layers and neurons (n) in each network can vary depending on the dataset presented to the model. Typically, the number of neurons is adjusted empirically, and fine-tuned until an optimal solution is achieved. However, it is important to exercise caution and avoid using an excessive number of neurons, as it may lead to overfitting [10]. The MLP model in this study utilized the backpropagation training algorithm, with the input layer consisting of the matrix of molecular markers and the output layer producing the predicted phenotypic value for each individual.

The Radial Basis Function Neural Network (RBF) presents a hybrid network model with two stages [11, 12]: a) An unsupervised stage for clustering the explanatory variables, which also determines the number of neurons in the hidden layer; b) A supervised stage that closely resembles the MLP, with the exception that the activation function should exhibit symmetric spread and/or Gaussian characteristics.

The term "radial" arises from the fact that the spreading function assigns greater weight to information within the cluster and lower weight to other data[6]. The architecture of the RBF consists of a feedforward design, encompassing a single input layer, an intermediate layer, and an output layer.

Tree-based prediction methods fall under the category of Supervised Learning, and one of the most conventional approaches in Machine Learning is the Decision Tree (DT). DT employs a top-down (greedy) methodology, and for the tree to be

[4]Nonlinearity in Biometrics is associated with the term epistasis, which is a genetic term to designate the interaction between alleles of different loci that, besides the additive-dominant effects, determines the genotypic expression of a character.

[5]In this study, the activation function used was the hyperbolic tangent, defined by the following equation: $\tanh(x) = \frac{1-e^{-2x}}{1+e^{-2x}}$

[6]Radial functions are a special class of functions whose value decreases or increases concerning the distance from a central point.

predictive, it must adhere to a branching criterion. This criterion induces a form of partitioning within the selected variable, typically involving group means or variables y and RSS (Residual Sum of Squares).

Each predictor, represented by a genetic marker, facilitates the creation of regions in the response variable. These regions are determined through sorting and partitioning, to minimize the RSS. The RSS is calculated using the following equation:

$$RSS = \sum_{m=1}^{M} \sum_{i \in R_m} (y_i - \hat{y}_{Rm})^2 \tag{2.4}$$

Where: $R_1, R_2, \ldots, R_M$ are regions in the dependent variable (usually the division is binary), y_i are the phenotypic values of the response variable, and $\hat{y}_{Rm}$ is the average of the response variable of the training observations belonging to the m-th region.

Two particularly intriguing aspects of tree-based structures in the context of Biometrics are: a) the biosynthetic pathway, which elucidates the role of genes in a metabolic pathway leading to the creation of biochemical intermediates and end products. These substances govern diverse phenotypic patterns of a trait and can be represented in the form of a tree; and b) the tree's ability to rank the most significant variables, facilitating dimensionality reduction. Consequently, its various extensions, including Random Forest (RF), Boosting (BOO), and Bagging (BAG), find extensive application in Biometrics. There is a hypothesis that this paradigm represents the optimal choice for fitting prediction models.

The Bagging (BAG) method refines the decision tree model by incorporating the Bootstrap technique. The core concept is to acknowledge that a portion of the output error in a single regression tree results from the specific selection of the training dataset. To address this, multiple similar datasets are generated through resampling (bootstrapping). Each of these datasets is then used to create individual regression trees without pruning. The final result is an average of these trees, leading to the generation of multiple models [13, 14]. Therefore, a total of B models are obtained: $\hat{f}^1(x), \hat{f}^2(x), \ldots, \hat{f}^B(x)$. These generated models are used to obtain an average model, as follows:

$$\hat{f}_{mean}(x) = \frac{1}{B} \sum_{b=1}^{B} \hat{f}^b(x) \tag{2.5}$$

The generated models are employed to mitigate the variability observed in the decision trees and calculate the mean, which serves as the final model. The Bagging technique is utilized to construct robust models that are less susceptible to overfitting, accomplishing this by leveraging a series of trees to stabilize errors [15]. In this study, a fixed number of 500 trees were sampled for bagging.

The Random Forest (RF) method employs Bootstrap samples to construct multiple

trees, each of which uses a random subset of predictors and observations. RF operates on the same principle as Bagging but introduces a variation by randomly selecting the number of predictor variables $(m < p)$ used in each partition, where m represents the number of predictors and p is the total number of predictors. RF aims to obtain independent predicted values by reducing the variability observed in the decision trees. It is recommended that the number of predictor variables used in each partition be set to $(m = p/3)$ for regression trees [16].

This approach minimizes the correlation between tree predictions, ensuring that the same variable is not consistently at the top of every tree. The RF is configured with 500 trees, and they are grown to their maximum size without any pruning. The aggregation of results is achieved through the collection of these trees.

In the Boosting (BOO) method, trees are constructed sequentially, incorporating information from previous trees [15]. It is an iterative approach trained on the same sample, where at each iteration, a prediction error measure is computed for each SNP. In the subsequent iteration, SNPs that yielded higher errors are given greater weight in the model training. This method serves as a numerical optimization technique aimed at minimizing the loss function by iteratively adding new trees that best reduce the loss function [17]. The prediction is conducted by weighting the results of all regression trees, and in this procedure, 500 trees were sampled.

Multivariate Adaptive Regression Splines (MARS) is a forecasting methodology regarded as an extension of linear models. It automatically models phenomena characterized by nonlinearities and interactions between variables. In essence, MARS incorporates a set of basis functions (BFs), resulting in a flexible model capable of handling both linear and nonlinear behavior.

The linear MARS model (MARS1) is described by the following equation:

$$y = \beta_0 + \sum_{m=1}^{M} B_m h_m(X) \tag{2.6}$$

Where: y is the value of the response variable, $_0$ is the regression constant, $_m$ with $m = 1, 2, \ldots, M$, are the regression coefficients, and $h_m(X)$ is a function or product of functions contained in C. The estimation of parameters β_0 and β_m $(m = 1, 2, \ldots, M)$ is based on minimizing the sum of squares of the residuals.

When considering the set of observed points for the jth predictor variable (molecular marker) $X_j, \Omega = \{x_{1_j}, x_{2_j}, \ldots, x_{r_j}\}$, collections of basis functions are derived from the set of points $C = \{(x - t)_+, (t - x)_+\}_{t \in \Omega, j=1,2,\ldots,p}$, where p is the total number of predictors (or molecular markers). In the multiplicative model (MARS2), a predictor element is introduced, established as the product of the effects of two predictor variables.

The RR-Blup, DT, BAG, RF, BOO, MARS1, and MARS2 methods were imple-

mented using the GENES software integrated with R. On the other hand, the MLP and RBF methods based on neural networks were executed using the GENES software integrated with Matlab [18].

2.2.2 Datasets and empirical procedures

The dataset was created through simulation using the Genes software [18] and comprised a sample size of 1000 individuals representative of an F2 population. The data-generating equation incorporated epistasis by considering the multiplicative effects of pairs of loci $(jandj+1)$, which is characteristic of biological systems where genes act sequentially in biosynthetic pathways. The equation employed was as follows:

$$Y_i = \mu + \sum_{j=1}^{n_q} p_j \alpha_j + \sum_{j=1}^{n_q} p_j \alpha_j \alpha_{j+1} + e_i \tag{2.7}$$

Where: Y_i represents the phenotypic value for observation i; μ is the general average; p_j is the contribution of locus j to the expression of the trait, with each locus having an equal contribution established by a uniform distribution; α_j, α'_j, and α_{j+1} take the values $\mu + a_j$, $\mu + d_j$ and $\mu - a_j$, respectively, for the genotypes associated with classes AA, Aa, and aa, respectively. Here, μ denotes the average of the homozygotes, α_j is half the difference in genotypic value between both homozygotes, and d_j represents the difference between the genotypic value of the heterozygote and the average of the homozygotes. Prediction processes should thus be sensitive to these effects described in the $\sum_{j=1}^{n_q} p_j \alpha_j \alpha_{j+1}$ component of the model.

Eight (08) response variables (Y_i with $i = 1, \ldots, 8$) with predetermined mean values and heritability were considered. The heritability values ranged from 20% to 80%. Additionally, 2010 predictor variables representing molecular information in the form of SNP (Single Nucleotide Polymorphism) markers were generated. Figure 2.1 illustrates the correlation matrix of these 2010 variables, revealing blocks of associations representing gene linkages established by groups of 201 markers distributed across ten linkage groups or basic chromosomes of the simulated species.

Each marker was encoded with values of 2, 1, and 0, representing a discrete multicategory variable. These markers exhibited the expected Mendelian segregation in a 1:2:1 ratio. They were distributed across a genome consisting of 10 linkage groups, mirroring the genetic structure of a diploid species with $2n = 2x = 20$ chromosomes. Each linkage group spanned a length of 100 centimorgans and contained 201 markers evenly spaced.

The continuous response variables followed normal distributions and were influenced by environmental effects of varying magnitude. They were determined by the action of alleles from either 40 or 44 QTLs (Quantitative Trait Loci) selected from

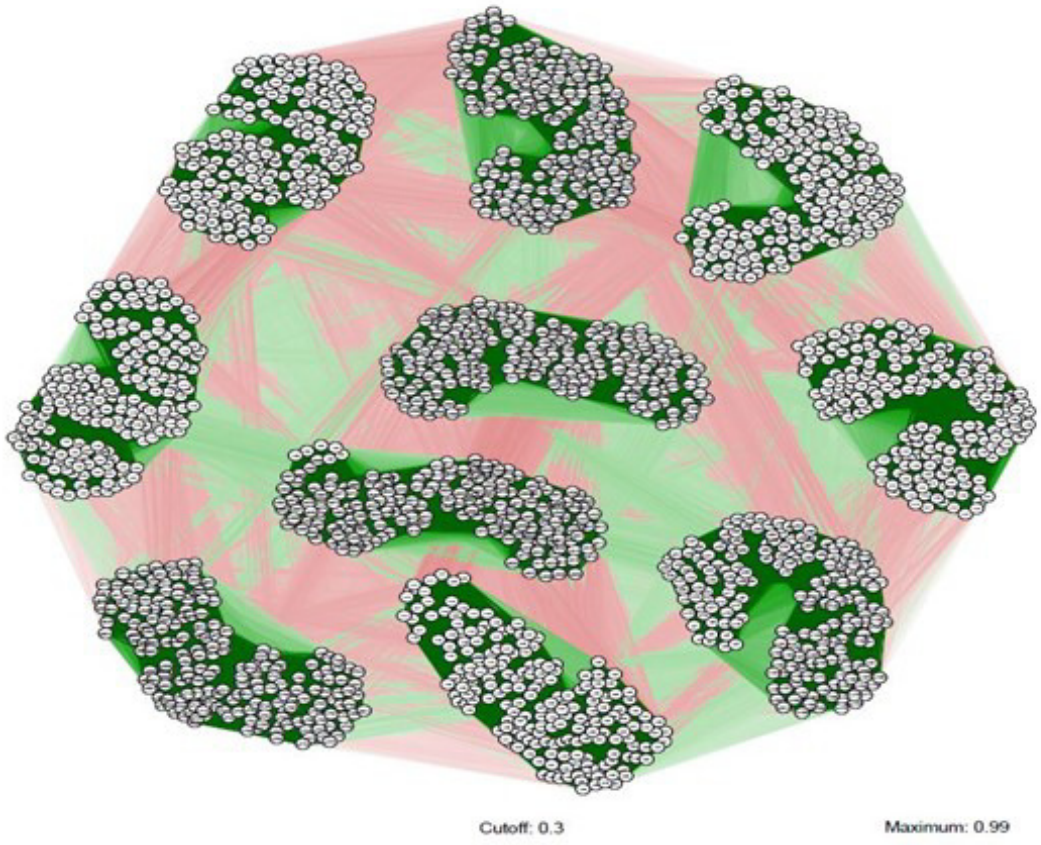

Figure 2.1: Correlation matrix 2010x2010 of predictor variables (SNPs).

among the 2010 previously genotyped markers. These QTLs had differential additive effects and weights denoting their importance in the total genotypic variability of the trait. These weights were established based on a uniform distribution.

The dataset in this study was divided into two parts, namely the training and validation sets, using the K-Fold Cross-Validation method. As suggested by [15], this division entails: a) the first part of the dataset is allocated for training purposes, where predictors are generated; b) the second part of the dataset is reserved for validation, assessing the predictive capability of the model by employing the predictor generated in the first part to make predictions for Y using the data from this second part. This approach is commonly employed when the objective is to evaluate the efficiency of prediction models or validation techniques, which is the primary focus of this study.

In the K-Fold Cross-Validation method, the sample of size n is divided into k parts or K-Folds, with the training sample obtained from the remaining $k-1$ parts. Consequently, after k iterations, all the data is used for both training and validation purposes [19]. Each of the nine (09) techniques presented in this article was implemented with Cross-Validation, employing a value of k equal to 5 ($k = 5$).

The predictive efficiency of the models was quantified using selective accuracy, or

the coefficient of determination (R^2), calculated as the simple mean of the estimates for each fold (k) in each fitted model. Selective accuracy can be defined in several ways and, in Quantitative Genetics, this measure is described as the square of the correlation between the estimated values and the true values [20]. It represents the degree to which the obtained estimate is related to the true value of the parameter, as shown in the equation:

$$R^2 = (cor(\hat{y}, y))^2 \tag{2.8}$$

Where: $\hat{y}$ represents the predicted values generated by the model in both the training and validation sets and y represents the observed values.

2.3 Results and Discussion

The results of the prediction models were obtained, considering various factors. Each model used eight response variables, subject to different noise levels (0.20, 0.40, 0.60, and 0.80), and the influence of two groups of markers (40 and 44). Additionally, a 5-fold Cross-Validation approach ($k = 5$) was employed, dividing the data into five sets (80% for training and 20% for validation). The selection metric used was the mean validation R^2 value across these five datasets.

The results of the fitted models are presented in Table 2.1. The top three prediction models were RBF, BAG, and BOO. When examining Figure 2.2, it becomes clear that predictive accuracy increases as noise decreases, which corresponds to higher heritability values. In simpler terms, response variables with higher noise levels tend to yield lower R^2 values for all nine models considered.

Further investigation of noise levels in response variables (Figure 2.2) reveals that, for high noise levels in the data (0 to 0.50 range), RBF (Y_1, Y_3, and Y_4) and BAG (Y_2) exhibited the highest predictive efficiency[7]. Conversely, for low noise levels in the data (0.50 to 0.80 range), RBF (Y_5) and BOO (Y_6, Y_7, and Y_8) demonstrated the highest predictive efficiency.

Additionally, a few general considerations about the model estimates (09) should be noted. The inclusion of 'major gene effect markers' (40 versus 44 markers) enhanced predictive efficiency. The choice of the Cross-Validation parameter ($k = 5$) with a relatively small sample size (1000) and a fixed k model warrants further reflection in future studies. Numerous studies advocate for a value of $k = 10$ as it strikes a balance between computationally intensive fitting procedures and the bias-variance trade-off [21, 22, 15]. However, as k increases, the fraction of data allocated

[7] In the context of this article, predictive efficiency is indeed synonymous with selective accuracy. It is quantified by calculating the mean R^2 value across five validation sets.

Table 2.1: Prediction Models (mean validation R^2 values and $k = 5$).

Label/Model	DT	BAG	RF	BOO	MLP	RBF	MARS1	MARS2	RRBLUP
$Y_1(h^2 = 0.20, ng = 40)$	0.0201	0.1376	0.1310	0.1023	0.0545	0.1471	0.0564	0.0300	0.0811
$Y_2(h^2 = 0.20, ng = 44)$	0.0341	0.1948	0.1908	0.1519	0.0666	0.1851	0.0774	0.0650	0.1041
$Y_3(h^2 = 0.40, ng = 40)$	0.0730	0.2463	0.2463	0.2174	0.1066	0.2781	0.1408	0.1072	0.1833
$Y_4(h^2 = 0.40, ng = 44)$	0.1646	0.3260	0.3256	0.2960	0.1356	0.3328	0.2101	0.1665	0.2268
$Y_5(h^2 = 0.60, ng = 40)$	0.1504	0.4545	0.4484	0.4355	0.2317	0.4568	0.2726	0.3148	0.3502
$Y_6(h^2 = 0.60, ng = 44)$	0.2424	0.4776	0.4860	0.5370	0.2971	0.5024	0.3721	0.3855	0.3981
$Y_7(h^2 = 0.80, ng = 40)$	0.2496	0.5647	0.5751	0.6523	0.3362	0.5893	0.3394	0.4560	0.4513
$Y_8(h^2 = 0.80, ng = 44)$	0.3129	0.6035	0.6104	0.6966	0.3450	0.6096	0.3836	0.5306	0.4686

Note: DT (Decision Tree), BAG (Bootstrap Aggregating), RF (Randon Forest), BOO (Boosting), MLP (Multi-Layer Perceptron), RBF (Radial Basis Function Network), MARS1 and MARS2 (Multivariate Adaptative Spline linear and interacting), and RRBlup (Ridge Regression Best Linear Unbiased Predictor).

for validation decreases, potentially compromising representativeness and predictive efficiency.

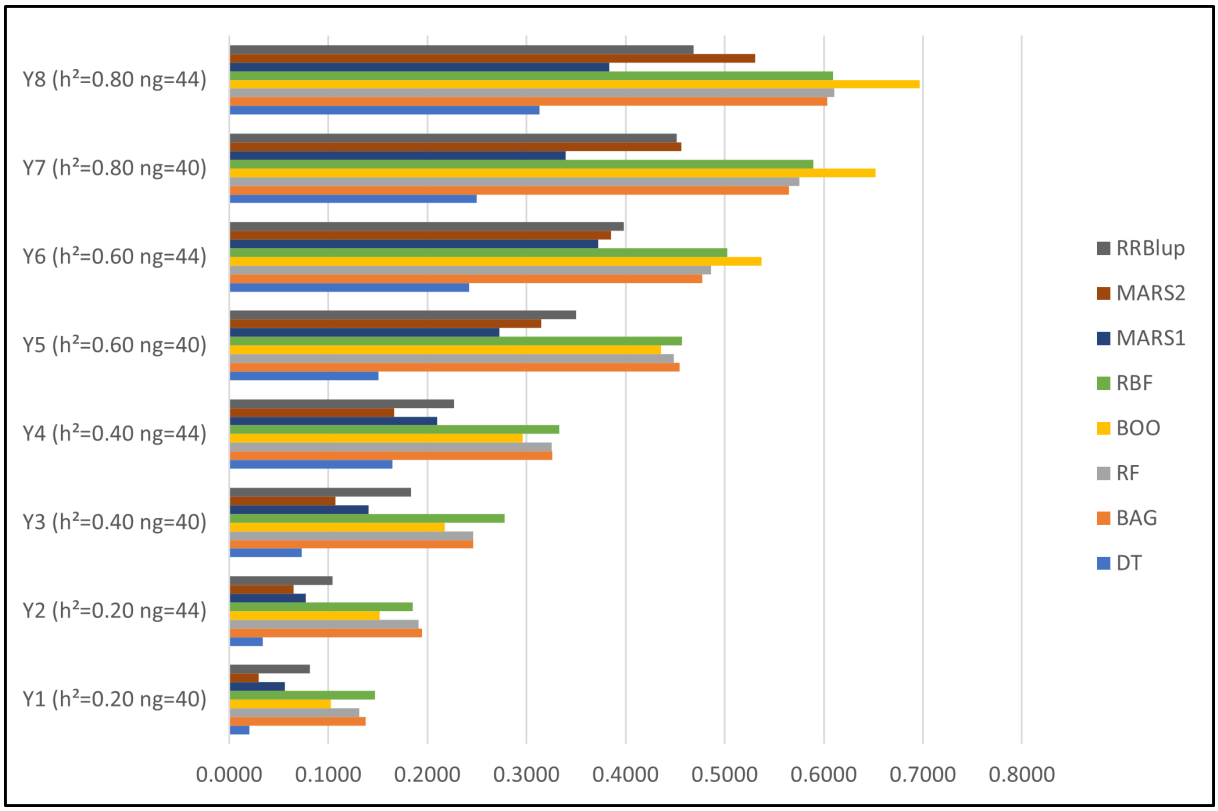

Figure 2.2: Prediction accuracy (R2) achieved by different techniques for variables with differences in gene control and environmental influence.
Note: DT (Decision Tree), BAG (Bootstrap Aggregating), RF (Randon Forest), BOO (Boosting), MLP (Multi-Layer Perceptron), RBF (Radial Basis Function Network), MARS1 and MARS2 (Multivariate Adaptative Spline linear and interacting), and RRBlup (Ridge Regression Best Linear Unbiased Predictor).

It is worth noting that techniques grouped under the same paradigm exhibited

significantly different performances (e.g., MLP versus RBF or DT versus refinements). A more in-depth understanding of the dataset's statistical issues (nonlinearity, multicollinearity, and dimensionality) on each model's performance can provide insights into these variations.

In summary, when faced with the task of selecting an approach for analyzing SNP datasets, experienced biometricians may opt for tree-based models. These models are particularly well-suited for data that mirrors gene action in genetic mechanisms involving biosynthetic pathways characterized by step-by-step formation of trait-determining products. Furthermore, the binary partitions and splits inherent in tree-building procedures can effectively represent relationships between alleles within the same gene (dominance) or between different genes (epistasis). This underscores the importance of domain-specific knowledge in model selection.

However, it is crucial to emphasize that choosing an appropriate model for predicting genetic data should not rely solely on domain-specific knowledge. While valuable, other factors such as problem complexity, dataset size and quality, algorithm characteristics, and data-handling capabilities must also be considered. Therefore, while tree-based models hold promise, it is advisable to explore various approaches and modeling techniques, including linear models, neural networks, or machine learning-based methods, to determine which model best fits the dataset and analysis objectives. Model selection should be based on experimentation and rigorous result evaluation.

2.4 Conclusion

In this study, the selective accuracy of nine prediction models was evaluated in the context of data presenting statistical challenges (nonlinearity, multicollinearity, and data dimensionality), varying noise levels (heritability values of 0.20, 0.40, 0.60, and 0.80), and specificities associated with Genome-Wide Selection (GWS).

Among these models, two stood out as the best fits for the simulated dataset, exhibiting the highest selective accuracy for the response variables: RBF and BOO. RBF, associated with the computational intelligence paradigm or neural network, demonstrated the highest selective accuracy in the presence of high data noise (with an honorable mention of the BAG model in Y_2). On the other hand, BOO, aligned with the tree paradigm, excelled in low-noise data settings (or high heritability scenarios). A significant generalization was drawn as follows: as data noise decreases, predictive efficiency increases for all considered models, indicating a positive correlation between heritability and selective accuracy.

Our study also identified a question worthy of future research: the impact of sample size versus the value of k in the K-Fold Cross-Validation method on selective accuracy. Given the relatively small sample size in this study (1000 individuals) and

the use of $k = 5$, these factors may have introduced issues related to representativeness (statistical and genetic) in dataset partitioning, which should be further explored in future works concerning SNP analysis.

References

[1] Bo Li, Nanxi Zhang, You-Gan Wang, Andrew W. George, Antonio Reverter, and Yutao Li. Genomic Prediction of Breeding Values Using a Subset of SNPs Identified by Three Machine Learning Methods. *Frontiers in Genetics*, 9:237, July 2018. ISSN 1664-8021. doi: 10.3389/fgene.2018.00237. URL https://www.frontiersin.org/article/10.3389/fgene.2018.00237/full.

[2] Weverton Gomes da Costa, Maurício de Oliveira Celeri, Ivan de Paiva Barbosa, Gabi Nunes Silva, Camila Ferreira Azevedo, Aluizio Borem, Moysés Nascimento, and Cosme Damião Cruz. Genomic prediction through machine learning and neural networks for traits with epistasis. *Computational and Structural Biotechnology Journal*, 20:5490–5499, January 2022. ISSN 2001-0370. doi: 10.1016/j.csbj.2022.09.029. URL https://www.sciencedirect.com/science/article/pii/S2001037022004342.

[3] Christina B Azodi, Emily Bolger, Andrew McCarren, Mark Roantree, Gustavo De Los Campos, and Shin-Han Shiu. Benchmarking Parametric and Machine Learning Models for Genomic Prediction of Complex Traits. *G3 Genes | Genomes | Genetics*, 9(11):3691–3702, November 2019. ISSN 2160-1836. doi: 10.1534/g3.119.400498. URL https://academic.oup.com/g3journal/article/9/11/3691/6026746.

[4] Rafael Izbicki and Tiago Mendonça dos Santos. *Aprendizado de máquina : uma abordagem estatística*. RI, São Carlos, 2020. ISBN 978-65-00-02410-4. URL http://www.rizbicki.ufscar.br/AME.pdf.

[5] Jeffrey B. Endelman. Ridge Regression and Other Kernels for Genomic Selection with R Package rrBLUP. *The Plant Genome*, 4(3):250–255, November 2011. ISSN 19403372. doi: 10.3835/plantgenome2011.08.0024. URL http://doi.wiley.com/10.3835/plantgenome2011.08.0024.

[6] Cosme Damião Cruz, Caio Césio Salgado, and Leonardo Lopes Bhering. *Genômica Aplicada*. Suprema, Visconde do Rio Branco, 2013.

[7] T. H. Meuwissen, B. J. Hayes, and M. E. Goddard. Prediction of total genetic value using genome-wide dense marker maps. *Genetics*, 157(4):1819–1829, April 2001. ISSN 0016-6731.

[8] Yann LeCun, Yoshua Bengio, and Geoffrey Hinton. Deep learning. *Nature*, 521 (7553):436–444, May 2015. ISSN 0028-0836, 1476-4687. doi: 10.1038/nature14539. URL https://www.nature.com/articles/nature14539.

[9] Osval Antonio Montesinos López, Abelardo Montesinos López, and José Crossa. *Multivariate Statistical Machine Learning Methods for Genomic Prediction*. Springer International Publishing, Cham, 2022. ISBN 978-3-030-89009-4 978-3-030-89010-0. doi: 10.1007/978-3-030-89010-0. URL https://link.springer.com/10.1007/978-3-030-89010-0.

[10] Cosme Damião Cruz and Moysés Nascimento. *Inteligência computacional aplicada ao melhoramento genético*. UFV, Viçosa, 2018.

[11] J. Park and I. W. Sandberg. Universal Approximation Using Radial-Basis-Function Networks. *Neural Computation*, 3(2):246–257, June 1991. ISSN 0899-7667, 1530-888X. doi: 10.1162/neco.1991.3.2.246. URL https://direct.mit.edu/neco/article/3/2/246-257/5580.

[12] Simon S. Haykin. *Neural networks and learning machines*. Prentice Hall, New York, 3rd ed edition, 2009. ISBN 978-0-13-147139-9. OCLC: ocn237325326.

[13] Leo Breiman. Bagging predictors. *Machine Learning*, 24(2):123–140, August 1996. ISSN 0885-6125, 1573-0565. doi: 10.1007/BF00058655. URL http://link.springer.com/10.1007/BF00058655.

[14] Anantha M. Prasad, Louis R. Iverson, and Andy Liaw. Newer Classification and Regression Tree Techniques: Bagging and Random Forests for Ecological Prediction. *Ecosystems*, 9(2):181–199, March 2006. ISSN 1432-9840, 1435-0629. doi: 10.1007/s10021-005-0054-1. URL http://link.springer.com/10.1007/s10021-005-0054-1.

[15] Gareth James, Daniela Witten, Trevor Hastie, and Robert Tibshirani. *An introduction to statistical learning: with applications in R*. Springer texts in statistics. Springer, New York NY, second edition edition, 2021. ISBN 978-1-07-161417-4 978-1-07-161420-4. doi: 10.1007/978-1-0716-1418-1.

[16] Trevor Hastie, Robert Tibshirani, and J. H. Friedman. *The elements of statistical learning: data mining, inference, and prediction*. Springer series in statistics. Springer, New York, NY, 2nd ed edition, 2009. ISBN 978-0-387-84857-0 978-0-387-84858-7.

[17] Farhad Ghafouri-Kesbi, Ghodratollah Rahimi-Mianji, Mahmood Honarvar, and Ardeshir Nejati-Javaremi. Predictive ability of Random Forests, Boosting, Support Vector Machines and Genomic Best Linear Unbiased Prediction in different

scenarios of genomic evaluation. *Animal Production Science*, 57(2):229, 2017. ISSN 1836-0939. doi: 10.1071/AN15538. URL http://www.publish.csiro.au/?paper=AN15538.

[18] Cosme Damião Cruz. Genes Software – extended and integrated with the R, Matlab and Selegen. *Acta Scientiarum*, 38(4):547–552, 2016. ISSN 1807-8621. doi: 10.4025/actasciagron.v38i4.32629. URL https://www.scielo.br/j/asagr/a/sLvDYF5MYv9kWR5MKgxb6sL/.

[19] Prabir Burman. A Comparative Study of Ordinary Cross-Validation, v-Fold Cross-Validation and the Repeated Learning-Testing Methods. *Biometrika*, 76 (3):503, September 1989. ISSN 00063444. doi: 10.2307/2336116. URL https://www.jstor.org/stable/2336116?origin=crossref.

[20] Cosme D. Cruz. *Princípios de genética quantitativa*. UFV, Viçosa, 2005. ISBN 978-85-7269-207-6. OCLC: 77542943.

[21] Ron Kohavi. A Study of Cross-Validation and Bootstrap for Accuracy Estimation and Model Selection. In *Proceedings of the 14th International Joint Conference on Artificial Intelligence - IJCAI'95*, volume 2, pages 1137–1143, Montreal, Quebec, Canada, 1995. Morgan Kaufmann Publishers Inc. ISBN 1-55860-363-8. URL https://dl.acm.org/doi/10.5555/1643031.1643047.

[22] Ji-Hyun Kim. Estimating classification error rate: Repeated cross-validation, repeated hold-out and bootstrap. *Computational Statistics & Data Analysis*, 53(11): 3735–3745, September 2009. ISSN 01679473. doi: 10.1016/j.csda.2009.04.009. URL https://linkinghub.elsevier.com/retrieve/pii/S0167947309001601.

Capítulo 3

Feature Selection em Biometria

Redução de Dimensionalidade (RD) tem se tornado uma ferramenta fundamental para melhorar a eficiência de modelos de Predição e mais especificamente ainda a classe de modelos do tipo *Feature Selection*. Visando contribuir com esse processo, este artigo faz uso de três (03) técnicas de Feature Selection (*Bagging*, Sonda e *Stepwiswe*) aplicadas a um modelo do tipo *(Random Forest - RF)* e fazendo uso de dados de simulação do tipo marcadores moleculares (*Single Nucleotide Polymorphisms - SNPs*) no contexto da Biometria. Os resultados obtidos sinalizam melhorias na acurácia seletiva (R^2) entre 10% e 28% e uma melhor adequação de dois (02) modelos (*Bagging* e *Stepwiswe*) em detrimento a um modelo (Sonda), em captar de maneira mais adequada as situações que envolvem o controle gênico de uma característica. Por fim, a medida que o ruído diminui, a acurácia seletiva aumenta (isso vale para modelos com e sem RD) e as taxas de crescimento da acurácia seletiva se tornam decrescentes.

3.1 Introdução

Redução de Dimensionalidade (RD) ou Redução de Dimensão ou Projeção ou até muitas vezes Mudança de Representação são conceitos usados como sinônimos para descrever uma de série de métodos utilizados em dados de alta dimensão[1]. A dimensionalidade de um conjunto de dados é definida como a cardinalidade (#) dos números de colunas/features de uma matriz [1]. O assunto tem sido um problema recorrente em grupos de pesquisa de Computação (e mais especificamente ainda em problemas de Bioinformática devido ao tipo dos dados utilizados nesses estudos) e muitas vezes um procedimento fundamental para viabilizar e melhorar a eficiência de certos tipos de modelos.

Conforme a dimensionalidade dos dados de entrada aumenta, o número de amostras a serem consideradas durante o treinamento de um modelo aumenta exponencialmente e este fenômeno é chamado de "Maldição da Dimensionalidade" [2]. A

[1]Em Bioinformática, por exemplo, os dados são considerados de alta dimensão menos em termo de número de observação (amostra) e mais em termos de suas características (número de variáveis ou atributos ou *features* ou dimensões).

maioria dos conjuntos de dados com o chamado problema "grande p, pequeno n" têm mais chances de overfitting [3, 4][2].

Problemas como esparsidade dos dados, escalabilidade, redundância nas covariáveis, razão de dimensionalidade intrínseca, dificuldade em processamento e/ou estimação geram a necessidade de buscar métodos de Redução da Dimensão. Em linhas gerais, as técnicas de Redução de Dimensionalidade (RD) são classificadas como problema de Aprendizado não supervisionado porque envolvem aprender mais sobre a estrutura de dados utilizada na ausência de rótulos[3]. Em dados de alta dimensão aprender mais sobre a estrutura de dados implica em investigar redundância nas covariáveis e esparsidade e enfrentar o problema da maldição da dimensionalidade [6].

Os métodos de Redução de Dimensionalidade (RD) muitas vezes são divididos seguindo diversas estratificações ou taxonomias. Por exemplo, existem os métodos de seleção versus extração de recursos e dentro desses últimos, existem os métodos lineares versus os não lineares.

A utilização de métodos de RD é justificada por sete (07) motivos, conforme listado por [1]: a) para remover dados faltantes, redundantes e ruídos; b) para reduzir a necessidade de espaço de armazenamento; c) para melhorar a velocidade de aprendizagem em modelos; d) para construir modelos com mais precisão; e) para diminuir a complexidade de modelos; f) para reduzir o sobre ajuste; e por fim; g) para permitir a visualização dos dados e observar seus padrões com mais clareza. A escolha do método de RD mais adequado para as necessidades de contexto de uso é um desafio, mesmo para usuários "experientes", devido ao aumento no número de técnicas de RD, como enfatizado por [7].

A primeira consideração a ser feita em relação a RD é que existem muitas técnicas que fazem esse papel, mas não são consideradas como tais. Por exemplo, Redes Neurais ou os modelos do tipo Árvore. Então, o pesquisador pode usar um modelo tradicional que faça RD e a partir do mesmo estabelecer uma nova topologia para estimação. Uma outra abordagem é diretamente do conjunto de dados fazer RD e a partir daí partir para uso de diferentes topologias. Essa reflexão leva a questão de quão independente são o conjunto de dados utilizados na pesquisa e a escolha do modelo, por exemplo, a ser utilizado para Predição.

Uma segunda consideração é que técnicas de RD implicam que dada uma matriz (mxn), ao se aplicar uma técnica de projeção e produzir uma nova matriz (por exemplo, (mxd)), por definição $d <<< n$ (d tem que ser muito menor que n). É importante ressaltar, conforme mencionado por [6], que a simples diminuição das features/colunas de uma matriz (em sua coluna n, por exemplo na matriz mxn) é meramente uma

[2]Onde p é o número de características e n é o número de amostras.

[3]Os métodos de RD supervisionados são em menor números e podem ser facilmente adaptados para operar dessa maneira [5].

Mudança de Representação e não uma RD (e que métodos de extração podem levar inclusive ao aumento de dimensão dos dados).

Uma pergunta inicial para quem busca técnicas de RD seria quantas existem? Ao revisar a literatura de Machine Learning (ML) e Visualização de Dados (Infovis), [7] listam cerca de noventa (90) técnicas de RD. Uma possível taxionomia para classificar métodos de RD leva em conta os seguintes atributos [5]: tipos de dados, linearidade, flexibilidade para supervisão, capacidade para lidar com vários níveis estruturas, localidade, dirigibilidade, estabilidade, capacidade de manuseio dados fora do núcleo (OOC) e complexidade.

As técnicas de RD podem ser agrupadas de uma maneira mais simples e nesse contexto são divididas em dois (02) tipos [4]: *Feature Selection* (FS - Seleção de dados) e Extração de dados. *Feature Selection (FS)* é um processo de redução do espaço dimensional selecionando um subconjunto de dados ideal (relevantes e não redundantes) com base em algum critério fundamental para a análise [8], enquanto a extração de dados é um processo de geração de novos dados por combinação ou transformação dos existentes. Muitas vezes a técnica de *Feature Selection* é baseada no conceito de navalha de Occam [1][4].

Seja $X = \{x_1, x_2, \ldots, x_n\}$ as dimensões dos dados iniciais. A tarefa de um modelo de *Feature Selection* é encontrar o melhor subconjunto $Y = \{y_1, y_2, \ldots, y_m\}$ onde $m < n$, que otimiza o modelo. *Feature Selection* deve avaliar 2^n subconjuntos para encontrar o conjunto ideal de dados. Para grandes valores de n, é computacionalmente caro e é considerado um problema NP-difícil (ou NP-Complexo).

Uma das áreas de pesquisa que tem feito extenso uso de *Feature Selection (FS)* é a Biometria. A Biometria é uma área de pesquisa do Melhoramento Genético que se concentra na análise, processamento e interpretação de dados biológicos. A Biometria faz uso de conceitos da Genética Quantitativa para desenvolver modelos e arquiteturas computacionais que permitem a compreensão e previsão de fenômenos biológicos. Os modelos de Predição ajustados por biometristas, em linhas gerais, fazem uso de marcadores moleculares do tipo SNP (*Single Nucleotide Polymorphisms*) e tais marcadores cobrem amplamente o genoma permitindo incorporar informações moleculares na Predição do mérito genético dos indivíduos [9].

O uso de marcadores moleculares na Seleção Genômica Ampla (*Genome Wide Selection – GWS*) via modelos de Predição exige superar uma série de desafios estatísticos (dimensionalidade dos dados, multicolinearidade e não-linearidade ou interação gênica/epistasia) e a utilização de métodos de Redução de Dimensionalidade do tipo *Feature Selection* tem contribuído com esse processo.

Com essas considerações iniciais e visando contribuir com essa discussão, este ar-

[4] A navalha de Occam é uma filosofia que sugere que se selecione a mais simples entre várias hipóteses concorrentes que fazem as mesmas previsões e com menos suposições.

tigo pretende confrontar a eficiência de modelos de Predição utilizando tanto conjuntos de dados originais quanto aqueles submetidos à Redução de Dimensionalidade (RD). Para tanto são utilizadas três (03) técnicas de Redução de Dimensionalidade (RD) do tipo *Feature Selection* (*Bagging*, Sonda e *Stepwiswe*) aplicadas a um modelo do tipo *(Random Forest - RF)* e faz-se uso de dados de simulação do tipo marcadores moleculares (*Single Nucleotide Polymorphisms - SNPs*) no contexto da Biometria.

3.2 Material e Métodos

Esse item apresenta a metodologia de Predição (*Random Forest*) que será utilizada nesse estudo para mensurar se houve alteração da eficiência (acurácia seletiva) sem e com Redução de Dimensionalidade (RD). Posteriormente, serão apresentados os métodos de Redução de Dimensionalidade (RD) ou mais especificamente de *Feature Selection*. Nessa categoria serão utlizados nesse estudo três (03) métodos de *Feature Selection*: Bagging *(Bootstrap Aggregation)*, Sonda e *Stepwise*.

3.2.1 Método Random Forest

O método *Random Forest (RF)* é um algoritmo ensemble de Aprendizado de Máquina utilizado para realizar Predições. O método RF tenta encontrar alternativas para o problema da correlação existente entre as predições de modelos de árvores que usam todas as variáveis e faz isso modificando o mecanismo de criação das árvores para que essas se tornem diferentes umas das outras. [6, 10]. Esta abordagem minimiza a correlação entre as previsões das árvores, garantindo que a mesma variável não esteja consistentemente no topo de cada árvore.

O RF utiliza amostras *bootstrap* para construir múltiplas árvores e cada uma delas utiliza um subconjunto aleatório de covariáveis. Ao invés de escolher qual das p covariáveis será utilizada em cada um dos nós da árvore, o método permite apenas que em cada passo seja escolhida $m < p$ covariáveis. Estas m covariáveis são escolhidas aleatoriamente dentre as covariáveis originais e, a cada nó criado, um novo subconjunto de covariáveis é sorteado.

O objetivo do RF é obter valores preditos independentes, reduzindo a variabilidade observada nas árvores de decisão. Recomenda-se que o número de variáveis preditoras usadas em cada partição seja definido como (m = p/3) para árvores de regressão [11]. Nesse estudo, o RF foi configurado com 500 árvores e estas são aumentadas até ao seu tamanho máximo sem qualquer poda. O resultado final do método é obtido através da combinação destas árvores que compõem o modelo.

3.2.2 Métodos de Feature Selection

Bagging

Árvores de decisão, quando treinadas com diferentes partes de um mesmo conjunto de dados, podem produzir resultados variáveis. O método *Bagging (Bootstrap Aggregation)* aperfeiçoa o modelo de árvore de decisão através da incorporação da técnica *bootstrap*. O ponto principal consiste em reconhecer que uma parte do erro de uma árvore de regressão resulta da seleção específica do conjunto de dados de treino. Para resolver este problema, são gerados vários conjuntos de dados semelhantes através de reamostragem (*bootstrapping*). Cada um destes conjuntos de dados é depois utilizado para criar árvores de regressão individuais sem poda.

Isso resulta em uma redução da variabilidade das previsões. Por isso, o *Bagging* é frequentemente referido como um método de aprendizado por agrupamento, usado para diminuir a variância em conjuntos de dados com ruídos. O resultado final é uma média destas árvores, o que leva à geração de modelos múltiplos [12, 13]. Por conseguinte, obtém-se um total de B modelos: $\hat{f}^1(x), \hat{f}^2(x), \ldots, \hat{f}^B(x)$. Estes modelos gerados são utilizados para obter um modelo médio conforme a Equação 3.1 proposta por [14]:

$$\hat{f}_{mean}(x) = \frac{1}{B} \sum_{b=1}^{B} \hat{f}^b(x) \tag{3.1}$$

O *Bagging* é uma técnica que reduz a variância das previsões ao combinar vários classificadores treinados em diferentes sub-amostras do mesmo conjunto de dados. A técnica de *Bagging* é utilizada para construir modelos robustos que são menos susceptíveis ao sobreajuste (*overfitting*), conseguindo-o através da utilização de uma série de árvores para estabilizar os erros [10]. A quantidade de árvores usadas no *Bagging* é determinada de forma a estabilizar o erro do modelo e neste estudo um número de 500 árvores foi utilizado[5].

O método *Bagging* foi também utilizado nesse estudo para determinar a importância das variáveis explicativas e posterior eliminação em proporção fixa[6]. O procedimento seguiu as diretrizes propostas em [16, 15] e o ponto de partida é estabelecer uma medida de precisão.

A medida é calculada a partir de permutação de dados OOB (*Out-of-Bag*), ou seja, para cada árvore é registrado o erro de previsão na parte dos dados fora do saco (taxa de erro para classificação, erro quadrático médio (*Mean squared error - MSE)* para

[5]Conjuntos de dados com muito ruído ou vários preditores podem precisar de um número de árvores maior [14].

[6]O procedimento foi implementado através do pacote RandomForest [15] atualizado em 24/01/2022 e na modalidade diminuição média da precisão (*mean decrease in accuracy*) que é a opção um (1) do pacote.

regressão). Adicionalmente, o mesmo é feito após a permutação de cada variável de previsão. A diferença entre os dois (02) é então calculada como média de todas as árvores e normalizada pelo desvio padrão das diferenças. Se o desvio padrão das diferenças for igual a 0 para uma variável, a divisão não é efetuada (mas a média é quase sempre igual a zero nesse caso).

O método da importância da caraterística de permutação é utilizado para determinar os efeitos das variáveis no modelo *Bagging*. Este método calcula o aumento do erro de previsão (MSE) após a permutação dos valores das características. Se a permutação não alterar o erro do modelo, a caraterística relacionada é considerada sem importância. Com esse procedimento é possível listar na matriz de *Features* as variáveis mais importantes e a partir dessa listagem fazer a eliminação de variáveis em proporção fixa utilizada nesse estudo.

Sonda

O Sonda é um método linear (regressão) de seleção de variáveis (subconjuntos) com inspiração em Algoritmo Genético (busca exaustiva) e foi proposto por [17]. A implementação do método é composta de duas etapas: a) o estabelecimento de um conjunto finitos de soluções, e posteriormente, b) a recombinação das melhores soluções para uma solução única.

Na primeira etapa, um número n de subamostras (ou Sondas) representando uma amostra de variáveis explicativas do tipo SNPs (ou *Features*) é retirada de maneira aleatória do conjunto original de variáveis[7]. No final dessa etapa é estabelecido um índice que indica a importância relativa de cada variável conforme a Equação 3.2:

$$IR_i = \frac{1}{n_i} \sum_{j=1}^{k} \beta_{ij} R_j^2 \text{ para } i = 1, 2, ..., m \tag{3.2}$$

Onde: k é o número de vezes que uma variável ou marcador x_i participou das n Sondas efetuadas; β_{ij} é o coeficiente de regressão associado a variável x_i incluído no modelo de regressão obtido em uma determinada Sonda S_j que tenha sido incluída; R_j^2 é o coeficiente de determinação obtido pela regressão ajustada na *j-ésima* Sonda; m é o número total de marcas estudadas.

A segunda etapa do método Sonda consiste na recombinação dos resultados para uma montagem de uma nova regressão e levando em conta as regressões que tiveram o melhor desempenho. Isso é feito a partir do rankeamento fornecido pelo índice IR_i e a sua utilização permite selecionar as n melhores variáveis (marcadores) para o cálculo da nova regressão (múltipla) e do novo coeficiente de determinação R_{IR}^2.

[7] Os parâmetros iniciais do método Sonda são o tamanho da Sonda (200, 400 ou 500 nesse estudo) e o número de buscas (1000).

Stepwise

O método *Stepwise* é uma metodologia linear de seleção de variáveis (regressão) e seu foco é selecionar as variáveis preditivas que mais influenciam o conjunto de saída, de modo a reduzir o modelo de regressão [18, 10]. Preditores são adicionados ou excluídos do modelo com base no *teste-F* e isso é feito com a combinação de dois (02) procedimentos: a) *Forward* - com a inclusão de variáveis; b) *Backward* - com a exclusão de variáveis do modelo.

No passo *Forward* avalia-se se uma nova variável deve ser adicionada ao modelo. Isso é feito considerando a correlação dessa variável com a resposta, ajustada pelas outras variáveis já no modelo. Se essa correlação for significativa (com *p-valor* menor que 10%), a variável é adicionada. No passo *Backward*, que ocorre após a adição de uma variável, verifica-se se todas as variáveis no modelo ainda são necessárias. Isso é feito usando um critério semelhante ao do passo *Forward*, mas agora considerando a remoção de variáveis. Se a exclusão de uma variável não reduz significativamente a qualidade do modelo, a mesma é removida.

3.2.3 Conjunto de dados e procedimentos empíricos

O conjunto de dados utilizado nesse artigo foi gerado por meio de simulação através do programa Genes [19] e representa uma população F_2 com tamanho efetivo de 900 indivíduos. A equação geradora dos dados incorporou a epistasia ao considerar os efeitos multiplicativos de pares de *loci* (j e $j+1$), o que é característico de sistemas biológicos onde os genes atuam sequencialmente em vias biossintéticas. A Equação (3.3) foi utilizada nesse processo de geração dos dados e apresenta a seguinte formulação:

$$Y_i = \mu + \sum_{j=1}^{n_q} p_j \alpha_j + \sum_{j=1}^{n_q} p_j \alpha_j \alpha_{j+1} + e_i \tag{3.3}$$

Onde: Y_i representa o valor fenotípico para a observação i; μ é uma média geral dos homozigotos; p_j é a contribuição do *locus* j para a expressão da característica, tendo cada *locus* uma contribuição igual estabelecida por uma distribuição uniforme; α_j, α'_j, e α_{j+1} assumem os valores $\mu + a_j$, $\mu + d_j$ and $\mu - a_j$ para os genótipos associados às classes AA, Aa, e aa, respectivamente. Adicionalmente, α_j é a metade da diferença do valor genotípico entre ambos os homozigotos; d_j representa a diferença entre o valor genotípico do heterozigoto e a média dos homozigotos; e, e_i representa o efeito ambiental (ou resíduo) e sua estrutura de variância é dada por $e_i \sim N(0, V_e)$, onde $V_e = ((1-h^2) \times V_g)/h^2$, sendo V_e a variância residual, V_g a variância genotípica e h^2 a herdabilidade.

Na Equação (3.3), o primeiro somatório representa á contribuição individual por meio do efeito aditivo e de dominância. O segundo somatório ($\sum_{j=1}^{n_q} p_j \alpha_j \alpha_{j+1}$) re-

presenta os efeitos multiplicativos correspondentes às interações epistáticas entre os pares de *loci* e portanto, os processos de predição devem ser sensíveis a estes efeitos descritos na componente do modelo.

Foram consideradas três (03) variáveis de resposta (Y_i com $i = 1, \ldots, 3$) com valores médios e herdabilidade pré-determinados. Os valores de herdabilidade utilizados nesse estudos foram de 25% (Y_1), 50%(Y_2) e 75%(Y_3), ou seja, a matriz original dos Y_{is} teve dimensão de $900x3$ e em cada coluna um valor de herdabilidade diferente. Adicionalmente, foram geradas 2000 variáveis preditoras que representam informações moleculares na forma de marcadores SNP (*Single Nucleotide Polymorphism*), ou seja, a matriz original de *Features* nesse estudo teve dimensão de $900x2000$.

Cada marcador foi codificado com valores de 2, 1 e 0, representando uma variável discreta de multicategoria. Estes marcadores apresentaram a segregação mendeliana esperada numa razão de 1:2:1. Foram distribuídos por um genoma constituído por 10 grupos de ligação, reproduzindo a estrutura genética de uma espécie diplóide com $2n = 2x = 20$ cromossomas. Cada grupo de ligação abrangeu um comprimento de 100 centimorgans e continha 200 marcadores uniformemente espaçados.

As variáveis de resposta contínua seguiram distribuições normais e foram influenciadas por efeitos ambientais de magnitude variável (25%, 50% e 75%). Foram determinadas pela ação de alelos de 40 QTLs (*Quantitative Trait Loci*) selecionados entre os 2000 marcadores previamente genotipados conforme a Figura **??**. Estes QTLs tiveram efeitos aditivos diferenciais e pesos que denotam a sua importância na variabilidade genotípica total da caraterística. Estes pesos foram estabelecidos com base numa distribuição uniforme.

p

O conjunto de dados deste estudo foi dividido em duas partes, especificamente os conjuntos de treinamento e validação, usando o método *K-Fold Cross-Validation*. Conforme sugerido por [10], essa divisão implica: a) a primeira parte do conjunto de dados é alocada para fins de treinamento, onde são gerados preditores; b) a segunda parte do conjunto de dados é reservada para validação, avaliando a capacidade preditiva do modelo ao empregar o preditor gerado na primeira parte para fazer previsões para Y usando os dados dessa segunda parte. Esta abordagem é normalmente utilizada quando o objetivo é avaliar a eficiência de modelos de previsão ou técnicas de validação, que é o foco principal deste estudo.

No método de validação cruzada *K-Fold*, a amostra de tamanho n é dividida em k partes ou *K-Folds*, sendo a amostra de treino obtida a partir das $k - 1$ partes restantes. Por conseguinte, após k iterações, todos os dados são utilizados para efeitos de formação e validação [20]. Cada um dos modelos de *Feature Selection* acoplados ao *Random Forest* apresentados nesse artigo foram implementados com *Cross-Validation*,

empregando um valor de k igual a 3 ($k = 3$)[8].

A eficiência preditiva dos modelos foi quantificada utilizando a acurácia seletiva e a métrica escolhida para representar essa medida foi o coeficiente de determinação (R^2), calculado como a média simples das estimativas para cada *Fold* (k) em cada modelo ajustado. A acurácia seletiva pode ser definida de várias formas [21] e especificamente em Genética Quantitativa, esta medida é descrita como o quadrado da correlação entre os valores estimados e os valores verdadeiros. Essa medida representa o grau em que a estimativa obtida está relacionada com o valor verdadeiro do parâmetro[9], conforme a Equação (3.4):

$$R^2 = (cor(\hat{y}, y))^2 \tag{3.4}$$

Onde: $\hat{y}$ representa os valores estimados pelo modelo nos conjuntos de treino e validação e y representa os valores observados (ou verdadeiros).

3.3 Resultado e Discussão

Os estudos sobre Redução de Dimensionalidade (RD) tem sido de grande interesse em Predições Genômicas em razão de que o conjunto relativamente grande de variáveis preditoras, que neste estudo são representadas por marcadores do tipo SNP, pode provocar consideráveis problemas relativos à dimensionalidade e a multicolinearidade. Assim, a presença destes marcadores não-significativos para a expressão da variável em estudo pode provocar redução da eficiência da Predição e exigir do processamento maior tempo de análise.

O primeiro ponto a se ressaltar nos resultados é que a estratégia foi investigar essas questões fazendo uso de métodos de Redução de Dimensionalidade (RD) do tipo *Feature Selection (FS)*, ou seja, busca-se trabalhar com um número reduzido de variáveis após selecionar variáveis de maior relevância ou descartar aquelas de menor importância para explicar as variações da variável resposta.

Quando se faz descartes ou seleção de variáveis, uma série de questões devem ser apresentadas e debatidas. Uma primeira questão está relacionada à averiguação do impacto da Redução da Dimensionalidade sobre o resultado da análise. É necessário avaliar se a redução conduz apenas ao efeito benéfico da redução do tempo ou do custo computacional ou se há algum benefício em termos de aumento da acurácia seletiva ou preditiva de variáveis respostas controladas por diferentes arquiteturas genéticas e submetidas a diferentes efeitos ambientais.

[8] No presente artigo que contém 900 observações (indivíduos), cerca de 600 ($n \times (k-1)/k$) são utilizadas para treinamento e 300 (n/k) para validação dos modelos em cada rodada das estimativas.

[9] Em Gnética Quantitativa a acurácia seletiva medida pelo quadrado da correlação ou mais especificamente pelo coeficiente de determinação, expressa a herdabilidade da característica.

Com isso em vista, os resultados foram obtidos levando em conta as seguintes considerações. Cada um dos modelos foi ajustado para três (03) tipos de variáveis respostas com diferentes níveis de ruído/herdabilidade (Y_1 - 0.25, Y_2 - 0.50 e Y_3 - 0.75)[10] e o efeito de apenas um grupo de marcadores moleculares (40).

Esse último ponto implica que as variáveis respostas são controladas por um número menor de marcadores moleculares (40) e, portanto, há presença de variáveis que não tem relevância direta sobre a característica e pouca influência indireta, por meio das correlações genéticas estabelecidas pelos blocos de ligação gênica, tendo em vista o distanciamento entre os marcadores genéticos, considerados no mapa de ligação que serviu de base para geração dos dados que devem manter as características e propriedades de características obtidas em estudos experimentais.

Adicionalmente, foi utilizado o método de Validação Cruzada com $k = 3$ (*3-Fold Cross-Validation*), ou seja, 03 conjuntos de dados (com 80% para treinamento e 20% para validação). A métrica usada para comparação dos modelos foi o R^2 médio desses três (03) conjuntos de dados listados anteriormente. Com essas premissas, os resultados dos modelos ajustados de *Feature Selection (FS)* são apresentados nas Tabelas 3.1 e 3.2.

Tabela 3.1: Matriz de *Features*: sem RD versus com RD.

Modelo/Variáveis	Y_1/X	Y_2/X	Y_3/X
Random Forest	900x2000	900x2000	900x2000
Bag10, Sonda10	900x200	900x200	900x200
Bag20, Sonda20	900x400	900x400	900x400
Bag25, Sonda25	900x500	900x500	900x500
Stepwise	900x42	900x57	900x51

Nota: RF (Randon Forest - método sem RD), Bag10 $\vee$ Bag20 $\vee$ Bag25 (RD Bagging fixo de 10% ou 20% ou 25% da amostra de X), Sonda10 $\vee$ Sonda20 $\vee$ Sonda25 (RD Sonda fixo de 10% ou 20% ou 25% da amostra de X) e Stepwise (RD Stepwise flexível de acordo com a amostra X).

A Tabela 3.1 sintetiza o processo de Redução de Dimensionalidade realizado na matriz de *Features* - X desse estudo. A primeira linha indica a matriz original X de dimensão $900x2000$. O segundo bloco (das linhas 2-4), apresenta os métodos de Redução de Dimensionalidade (RD) de proporção fixa (ou seja, as variáveis foram rankeadas pelos referidos métodos e mantidas numa proporção fixa de 10%, 20% e 25% da matriz de *Features* original). Por fim, a última linha da Tabela 3.1 indica o método de proporção variável (ou ajustamento ótimo - *Stepwise*).

Na Tabela 3.1 há também uma notação (Y_i/X) que ressalta a importância de Y no processo de rankeamento e posterior eliminação das variáveis da matriz X. Mais

[10]Se a herdabilidade de uma característica é 25%, isso significa que 25% da variação observada nessa característica é atribuída à variação genética e o restante (75%) é explicado por outro fatores como ambiente, interações gene-ambiente, ou mesmo erro de medição. Quanto menor a herdabilidade, maior o ruído nos dados.

especificamente ainda, é importante ressaltar que embora o número de variáveis seja comum nas linhas dos métodos de proporção fixa (Bag e Sonda) da matriz X a medida que se varia o Y, o tipo de variável pode ser comum ou não e sua permanência depende do rankeamento feito a cada índice i ($i = 1, \ldots, 3$) da variável Y.

A partir dessa Redução de Dimensionalidade (RD) ou *Features Selection (FS)* foi gerada a Tabela 3.2. O primeiro ponto a se considerar na Tabela 3.2 é que a medida que o ruído diminui (ou a herdabilidade aumenta), a acurácia seletiva aumenta (maior valor do coeficiente de determinação R^2) e isso vale para modelos sem e com Redução de Dimensionalidade (RD). Pode ser observado nas Figuras 3.1 e 3.2 que os maiores valores de acurácia seletiva (R^2) foram obtidos para as variáveis de maior herdabilidade.

Tabela 3.2: Modelos de *Feature Selection (FS)* aplicados em Random Forest (RF).

Variável / Modelo	Y_1				Y_2				Y_3			
	R^2	DP	REQM	DP	R^2	DP	REQM	DP	R^2	DP	REQM	DP
Random Forest	0.1152	0.0282	182.5824	10.8328	0.2361	0.0621	123.7531	3.1368	0.3598	0.0830	92.4970	5.8755
RF_Bag10	0.1437	0.0205	179.4950	9.5762	0.2720	0.0347	119.7891	2.6108	0.3795	0.0641	89.5900	6.1534
RF_Bag20	0.1390	0.0382	180.1225	9.9491	0.2768	0.0457	120.4465	2.8556	0.3960	0.0756	90.0617	5.9971
RF_Bag25	0.1467	0.0320	179.9037	10.2170	0.2671	0.0432	121.0362	2.4885	0.3876	0.0664	90.4034	6.1122
RF_Sonda10	0.0709	0.0133	186.3486	9.5870	0.1672	0.0143	127.2121	1.9574	0.2052	0.0561	97.6459	5.3493
RF_Sonda20	0.0990	0.0226	184.2660	9.8127	0.1993	0.0423	125.0971	2.2462	0.3187	0.0646	93.8638	5.5556
RF_Sonda25	0.1135	0.0257	182.7646	9.8702	0.2130	0.0470	124.7178	2.190	0.3281	0.0611	93.8118	5.9228
RF_Stepwise	0.1229	0.0051	181.3665	9.1822	0.2479	0.0041	121.8150	1.8258	0.3688	0.0511	90.2985	5.8234

Nota: RF (Randon Forest), RF_Bag10 $\vee$ RF_Bag20 $\vee$ RF_Bag25 (Random Forest com método de RD Bagging fixo de 10% ou 20% ou 25% da amostra de X), RF_Sonda10 $\vee$ RF_Sonda20 $\vee$ RF_Sonda25 (Random Forest com método de RD Sonda fixo de 10% ou 20% ou 25% da amostra de X), RF_Stepwise (Random Forest com método RD Stepwise flexível de acordo com a amostra X).

O segundo ponto a se considerar na Tabela 3.2 é que houve aumento da acurácia seletiva (R^2) por meio do procedimento de Redução de Dimensionalidade (RD) em todos os níveis de ruído considerado na variável Y, ou seja, é possível aumentar a acurácia seletiva por meio de procedimentos de descarte de variáveis. O ponto de partida para se mensurar essa ganho é considerar o modelo (*Random Forest - RF*) que usa a totalidade das variáveis da matriz de *Features* ($900x2000$) e comparar o valor da acurácia seletiva (R^2) com esse mesmo modelo sendo posteriormente ajustado a partir de diferentes matrizes geradas no processo de Redução de Dimensionalidade.

Esse processo dá origem a sete (07) modelos de Redução de Dimensionalidade (RD) conforme mostram as Figuras 3.1 e 3.2.

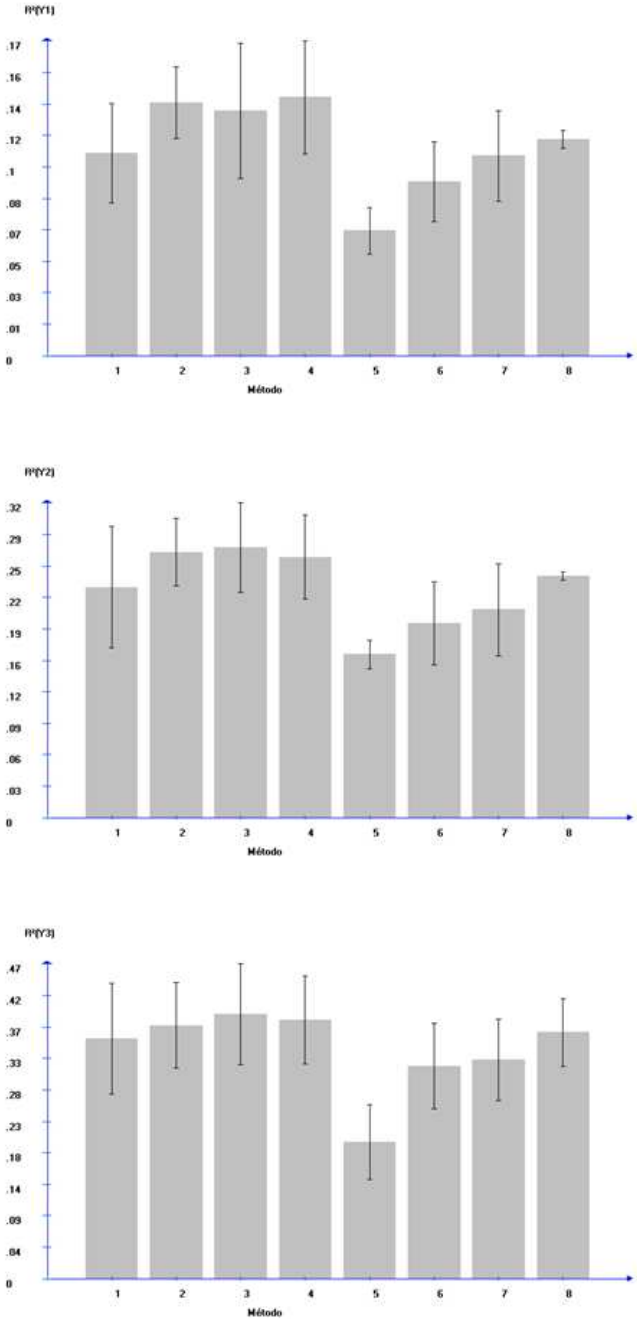

Figura 3.1: Valores da acurácia seletiva (R^2) em estudos de Predição considerando análises sem e com Redução da Dimensionalidade (RD) estabelecida por diferentes estratégias de seleção de variáveis.

Nota: 1.Random Forest (RF), 2.RF_Bag10, 3.RF_Bag20, 4.RF_Bag25, 5.RF_Sonda10, 6.RF_Sonda20, 7.RF_Sonda25 e 8.RF_Stepwise.

As Figuras 3.1 e 3.2 sinalizam também que apenas dois (02) métodos (*Bagging e Stepwise*) foram efetivos em melhorar a acurácia seletiva fazendo uso de Redução de Dimensionalidade (RD) e o modelo *Bagging* foi a melhor escolha em todos os cenários. Em Y_1 esse aumento foi de 27.35% ($RF\_Bag25 - 0.1467 > RF\_Bag10 - 0.1437 > RF -$

0.1152), em Y_2 foi de 17.25% ($RF\_Bag20 - 0.2768 > RF\_Bag25 - 0.2671 > RF - 0.2361$) e em Y_3 foi de 10.07% ($RF\_Bag20 - 0.3960 > RF\_Bag25 - 0.3876 > RF - 0.3598$). Os ganhos em termos de taxa de crescimento da acurácia seletiva (R^2) são decrescentes a medida que a herdabilidade aumenta.

Abordagens estatísticas que são capazes de identificar a relevância de preditores dentro do complexo de ação e interação das variáveis (situações de controle gênico) parecem ser as mais apropriadas na análise de acurácia seletiva em Biometria e nesse contexto que o modelo *Bagging* se destaca. Independente do nível de herdabilidade, o modelo parece captar melhor as situações que o controle gênico da característica envolve tais como: ação aditiva, ação de interação intra-alélica (dominância) e ação de interação inter-alélica (epistasia ou interação epistática não-linear). Por outro lado, a menor eficiência das técnicas de RD, *Stepwise* e Sonda, parece estar associada ao fato que os modelos baseiam-se em estruturas lineares tanto nos efeitos diretos quanto nas interrelações entre preditores, e isso compromete a capacidade desses modelos em expressar as *Features* mais relevantes para a variável resposta.

Apesar do método *Stepwise* não ser o mais eficiente em RD, houve uma melhoria na acurácia seletiva (R^2) em todos os cenários em relação ao modelo sem RD. O método também melhorou a acurácia seletiva (R^2) recorrendo a um número relativamente menor de variáveis descritoras do que os outros dois (02) métodos.

Sonda e *Stepwise* baseiam-se em regressão linear múltipla, mas o critério de estabelecer relevância de variáveis por meio de probabilidade que expressam relevância e as estatísticas do F parcial e correlação parcial foram eficazes em selecionar variáveis contemplando critérios de relevância por importância e redundância de forma mais efetiva nesse último método. Por isso, pelo fato do *Stepwise* ser um método de fácil implementação, pode ser uma boa escolha em processos de descarte menos acentuados ou mesmo para uma primeira exploração dos dados em outros estudos.

Por outro lado, o método Sonda não promoveu melhoria da acurácia seletiva em nenhum cenário de RD, mas pode ser uma ferramenta útil se for factível um *trade-off* de acurácia seletiva versus simplificação do modelo (por exemplo, o descarte de variáveis pode otimizar o tempo de processamento do modelo ou reduzir a capacidade computacional necessária).

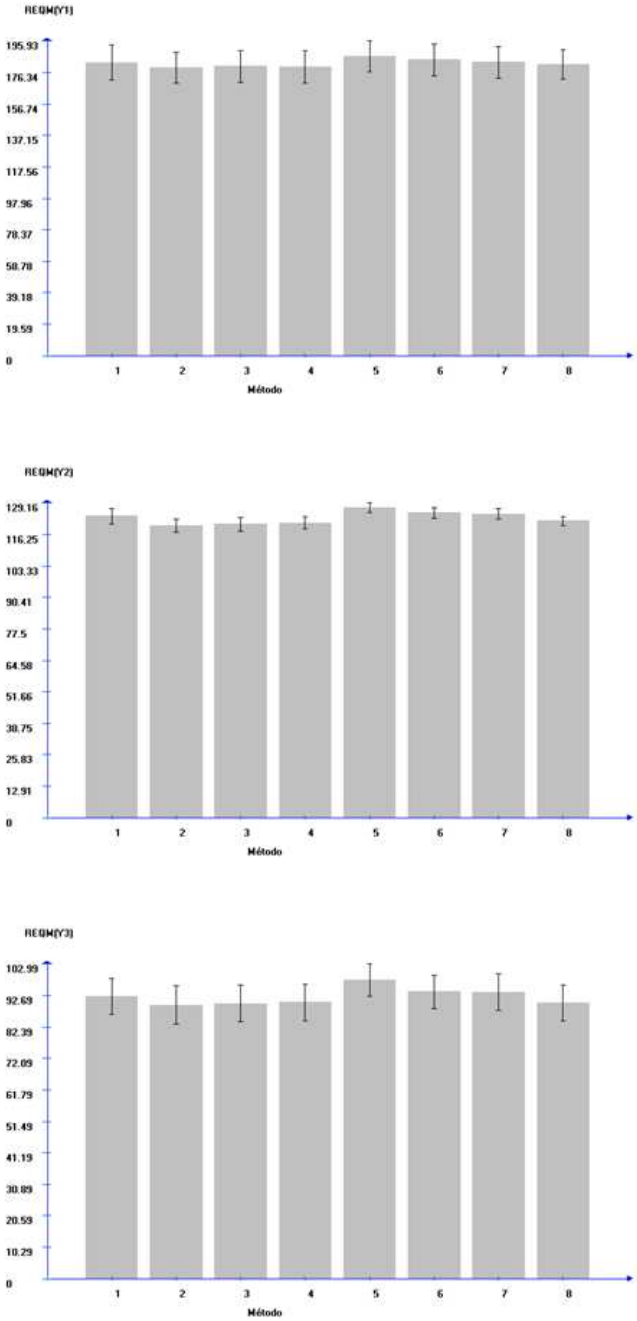

Figura 3.2: Valores da acurácia seletiva (*REQM* - Raiz do erro quadrático médio) em estudos de Predição considerando análises sem e com Redução da Dimensionalidade (RD) estabelecida por diferentes estratégias de seleção de variáveis.

Nota: 1.Random Forest (RF), 2.RF_Bag10, 3.RF_Bag20, 4.RF_Bag25, 5.RF_Sonda10, 6.RF_Sonda20, 7.RF_Sonda25 e 8.RF_Stepwise.

O ruído ambiental impacta consideravelmente os resultados de Predição e, assim, o pesquisador deve estar sempre atento em adotar procedimento ou conduta para que os dados obtidos estejam sob influência mínima dos ruídos indesejáveis proporcionados por erros aleatórios. Entretanto, o efeito ambiental sobre as técnicas de descarte

ou *Feature Selection* foram homogêneos, ou seja, as inferências feitas sobre as comparações de técnicas são praticamente as mesmas para as variáveis Y_1, Y_2 e Y_3, em termos de comportamento, de forma que nas três (03) situações pode-se generalizar que o procedimento *Bagging* mostrou desempenho superior, as técnicas da Sonda desempenho inferior e o método *Stepwise* de comportamento intermediário.

Uma outra discussão que merece destaque diz respeito à quantidade e a qualidade de informação que devem ser mantida no modelo de Predição. Fazendo uma análise nos resultados de maior acurácia seletiva promovido pelo procedimento *Bagging* (métodos 2, 3 e 4) é possível observar que deve existir uma quantidade ótima de marcadores para cada situação de forma que a redução conduzindo à 10%, 20% ou 25% proporcionaram resultados similares, e todos eficiente, mas sem que pudesse ser generalizado um número ótimo ou uma tendência.

Assim, tanto o controle genético da característica quanto o ruído ambiental influenciam o grau de Redução da Dimensionalidade (RD) e deve, portanto, ser avaliado previamente para cada variável em estudo. Deve ser ressaltado que uma RD como a apresentada neste estudo, de níveis entre 10% e 25%, podem ser útil no contexto prático de redução de recursos e tempo computacional e até resolver problemas de dimensionalidade permitindo o uso de procedimentos estatísticos bem mais simples como regressões polinomiais ou procedimento mais elaborados, como redes neurais com maiores números de camadas ou de neurônios por camadas.

3.4 Conclusão

Esse estudo demonstrou que a Redução de Dimensionalidade (RD) do tipo *Feature Selection* pode conduzir ao aumento na acurácia seletiva (R^2), além de promover um conjunto simplificado de dados demandando menor tempo computacional de análise e processamento.

A medida que o ruído diminui (ou a herdabilidade aumenta), a acurácia seletiva aumenta e isso vale para modelos sem e com Redução de Dimensionalidade (RD). Esse parece um efeito da via Biossintética que também se manifesta em estudos que envolvem RD versus Biometria.

O aumento da acurácia seletiva (R^2) por meio do procedimento de Redução de Dimensionalidade (RD) ocorreu em todos os níveis de ruído considerado na variável Y, ou seja, é possível aumentar a acurácia seletiva por meio de procedimentos de descarte de variáveis. Nesse artigo, o método de RD *Bagging* foi o mais eficaz em promover melhorias na acurácia seletiva (entre 10% e 28%), e portanto, a melhor escolha em todos os cenários considerados no artigo.

Levando em conta todos os três (03) métodos de RD utilizados nesse estudo pode-se generalizar que o procedimento *Bagging* mostrou desempenho superior, as técnicas

da Sonda desempenho inferior e o método *Stepwise* de comportamento intermediário.

Referências

[1] V. Lakshmi Chetana, Soma Sekhar Kolisetty, and K. Amogh. A Short Survey of Dimensionality Reduction Techniques. In Anupama Namburu and Soubhagya Sankar Barpanda, editors, *Recent Advances in Computer Based Systems, Processes and Applications*, pages 3–14. CRC Press, 1 edition, June 2020. ISBN 978-1-00-304398-0. doi: 10.1201/9781003043980-2. URL https://www.taylorfrancis.com/books/9781000221664/chapters/10.1201/9781003043980-2.

[2] Richard Bellman. *Adaptive Control Processes: A Guided Tour*. Princeton University Press, 1961.

[3] Zena M. Hira and Duncan F. Gillies. A Review of Feature Selection and Feature Extraction Methods Applied on Microarray Data. *Advances in Bioinformatics*, 2015: 1–13, June 2015. ISSN 1687-8027, 1687-8035. doi: 10.1155/2015/198363. URL https://www.hindawi.com/journals/abi/2015/198363/.

[4] Mahdieh Labani, Parham Moradi, Fardin Ahmadizar, and Mahdi Jalili. A novel multivariate filter method for feature selection in text classification problems. *Engineering Applications of Artificial Intelligence*, 70:25–37, April 2018. ISSN 09521976. doi: 10.1016/j.engappai.2017.12.014. URL https://linkinghub.elsevier.com/retrieve/pii/S0952197617303172.

[5] Luis Gustavo Nonato and Michael Aupetit. Multidimensional Projection for Visual Analytics: Linking Techniques with Distortions, Tasks, and Layout Enrichment. *IEEE Transactions on Visualization and Computer Graphics*, 25(8):2650–2673, August 2019. ISSN 1077-2626, 1941-0506, 2160-9306. doi: 10.1109/TVCG.2018.2846735. URL https://ieeexplore.ieee.org/document/8383983/.

[6] Rafael Izbicki and Tiago Mendonça dos Santos. *Aprendizado de máquina : uma abordagem estatística*. RI, São Carlos, 2020. ISBN 978-65-00-02410-4. URL http://www.rizbicki.ufscar.br/AME.pdf.

[7] Mateus Espadoto, Rafael M. Martins, Andreas Kerren, Nina S. T. Hirata, and Alexandru C. Telea. Toward a Quantitative Survey of Dimension Reduction Techniques. *IEEE Transactions on Visualization and Computer Graphics*, 27(3):2153–2173, March 2021. ISSN 1077-2626, 1941-0506, 2160-9306. doi: 10.1109/TVCG.2019.2944182. URL https://ieeexplore.ieee.org/document/8851280/.

[8] Yun Li, Tao Li, and Huan Liu. Recent advances in feature selection and its applications. *Knowledge and Information Systems*, 53(3):551–577, December 2017. ISSN 0219-1377, 0219-3116. doi: 10.1007/s10115-017-1059-8. URL http://link.springer.com/10.1007/s10115-017-1059-8.

[9] Weverton Gomes da Costa, Maurício de Oliveira Celeri, Ivan de Paiva Barbosa, Gabi Nunes Silva, Camila Ferreira Azevedo, Aluizio Borem, Moysés Nascimento, and Cosme Damião Cruz. Genomic prediction through machine learning and neural networks for traits with epistasis. *Computational and Structural Biotechnology Journal*, 20:5490–5499, January 2022. ISSN 2001-0370. doi: 10.1016/j.csbj.2022.09.029. URL https://www.sciencedirect.com/science/article/pii/S2001037022004342.

[10] Gareth James, Daniela Witten, Trevor Hastie, and Robert Tibshirani. *An introduction to statistical learning: with applications in R*. Springer texts in statistics. Springer, New York NY, second edition edition, 2021. ISBN 978-1-07-161417-4 978-1-07-161420-4. doi: 10.1007/978-1-0716-1418-1.

[11] Trevor Hastie, Robert Tibshirani, and J. H. Friedman. *The elements of statistical learning: data mining, inference, and prediction*. Springer series in statistics. Springer, New York, NY, 2nd ed edition, 2009. ISBN 978-0-387-84857-0 978-0-387-84858-7.

[12] Leo Breiman. Bagging predictors. *Machine Learning*, 24(2):123–140, August 1996. ISSN 0885-6125, 1573-0565. doi: 10.1007/BF00058655. URL http://link.springer.com/10.1007/BF00058655.

[13] Anantha M. Prasad, Louis R. Iverson, and Andy Liaw. Newer Classification and Regression Tree Techniques: Bagging and Random Forests for Ecological Prediction. *Ecosystems*, 9(2):181–199, March 2006. ISSN 1432-9840, 1435-0629. doi: 10.1007/s10021-005-0054-1. URL http://link.springer.com/10.1007/s10021-005-0054-1.

[14] Brad Boehmke and Brandon M. Greenwell. *Hands-on machine learning with R*. Chapman & Hall/CRC the R series. CRC Press, Boca Raton, 2019. ISBN 978-1-138-49568-5 978-0-367-41829-8.

[15] Andy Liaw and Matthew Wiener. Classification and regression by randomforest. *R News*, 2(3):18–22, 2002. URL https://CRAN.R-project.org/doc/Rnews/.

[16] Leo Breiman. Random Forests. *Machine Learning*, 45(1):5–32, 2001. ISSN 08856125. doi: 10.1023/A:1010933404324. URL http://link.springer.com/10.1023/A:1010933404324.

[17] G.N. Silva, I.C. Sant'Anna, C.D. Cruz, M. Nascimento, C.F. Azevedo, and L.S. Glória. Research Article Neural networks and dimensionality reduction to increase predictive efficiency for complex traits. *Genetics and Molecular Research*, 21(1), 2022. ISSN 16765680. doi: 10.4238/gmr18982. URL http://www.funpecrp.com.br/gmr/articles/year2022/vol21-1/pdf/gmr18982_-_neural-networks-and-dimensionality-reduction-increase-predictive-efficiency-co pdf.

[18] Cosme Damião Cruz. *Estatística Genômica Aplicada a Populações Derivadas De Cruzamentos Controlados*. Editora UFV, March 2008. ISBN 978-85-7269-339-4.

[19] Cosme Damião Cruz. Genes Software – extended and integrated with the R, Matlab and Selegen. *Acta Scientiarum*, 38(4):547–552, 2016. ISSN 1807-8621. doi: 10.4025/actasciagron.v38i4.32629. URL https://www.scielo.br/j/asagr/a/sLvDYF5MYv9kWR5MKgxb6sL/.

[20] Prabir Burman. A Comparative Study of Ordinary Cross-Validation, v-Fold Cross-Validation and the Repeated Learning-Testing Methods. *Biometrika*, 76 (3):503, September 1989. ISSN 00063444. doi: 10.2307/2336116. URL https://www.jstor.org/stable/2336116?origin=crossref.

[21] Cosme D. Cruz. *Princípios de genética quantitativa*. UFV, Viçosa, 2005. ISBN 978-85-7269-207-6. OCLC: 77542943.

Capítulo 4

Conclusão

Neste trabalho foram desenvolvidos modelos de Redução de Dimensionalidade (ou mais especificamente *Feature Selection*) no contexto da Biometria e isso foi feito através dos dois (02) artigos apresentados. Uma questão importante presente em ambos os artigos é que a natureza do problema tem profundas implicações nos resultados obtidos.

No primeiro artigo (Capítulo 2) o foco foi desenvolver diferentes modelos de Predição (09 modelos seguindo diferentes paradigmas) e analisar como cada um era aderente a dados do tipo marcadores moleculares (SNPs). O principal resultado do estudo foi associar eficiência preditiva versus ruído e observar que a seleção do melhor modelo também está associada ao problema do ruído. O artigo também apontou que em trabalhos futuros seria oportuno se aprofundar no método de escolha de k (Método de Validação Cruzada) por causa do dilema Representatividade Estatística versus Representatividade Genética.

No segundo artigo (Capítulo 3) o foco foi desenvolver modelos de Redução de Dimensionalidade (RD) ou mais especificamente *Feature Selection* para enfrentar problemas recorrentes (alta dimensionalidade, não-linearidade e multicolinearidade) no conjunto de dados (SNPs) e com isso melhorar a acurácia seletiva (e/ou outros benefícios complementares: simplificação dos modelos, menor tempo de processamento, interpretabilidade mais simples, redução do sobreajuste, etc....). Novamente, os principais resultados obtidos (relação entre diminuição do ruído versus aumento acurácia seletiva ou a taxa de crescimento da acurácia seletiva em modelos sem e com RD) também são associados ao problema do ruído.