LUCAS TASSONI ANDRIETTA

Uso de *Machine Learning* e dados genômicos para melhoria de características econômicas em bovinos de leite

Pirassununga
2022

LUCAS TASSONI ANDRIETTA

Uso de *Machine Learning* e dados genômicos para melhoria de características econômicas em bovinos de leite

VERSÃO CORRIGIDA

Dissertação apresentada ao Programa de Pós-Graduação em Nutrição e Produção Animal da Faculdade de Medicina Veterinária e Zootecnia da Universidade de São Paulo para a obtenção do título de Mestre em Ciências.

Departamento:

Nutrição e Produção Animal

Área de concentração:

Nutrição e Produção Animal

Orientador:

Prof. Dr. Ricardo Viera Ventura

Pirassununga

2022

DADOS INTERNACIONAIS DE CATALOGAÇÃO NA PUBLICAÇÃO

(Biblioteca Virginie Buff D'Ápice da Faculdade de Medicina Veterinária e Zootecnia da Universidade de São Paulo)

T. 4211
FMVZ

Andrietta, Lucas Tassoni
Uso de *Machine Learning* e dados genômicos para melhoria de características econômicas em bovinos de leite / Lucas Tassoni Andrietta. – 2022.
69 f. : il.

Dissertação (Mestrado) – Universidade de São Paulo. Faculdade de Medicina Veterinária e Zootecnia. Departamento de Nutrição e Produção Animal, Pirassununga, 2022.

Programa de Pós-Graduação: Nutrição e Produção Animal.

Área de concentração: Nutrição e Produção Animal.

Orientador: Prof. Dr. Ricardo Vieira Ventura.

1. Aprendizado de máquina. 2. Acasalamento dirigido. 3. Melhoramento animal. I. Título.

Ficha catalográfica elaborada pela bibliotecária Maria Aparecida Laet, CRB 5673-8, da FMVZ/USP.

Certificado da Comissão de Ética

Comissão de Ética no Uso de Animais
Faculdade de Medicina Veterinária e Zootecnia
Universidade de São Paulo

São Paulo, 26 de maio de 2022
CEUAx N 4123270220
(ID 001636)

IImo(a). Sr(a).
Responsável: Ricardo Vieira Ventura
Área: Nutrição E Produção Animal

Título da proposta: "Uso de Machine Learning e dados genômicos para melhoria de características econômicas em bovinos de leite".

CERTIFICADO (Relatório Parcial versão de 19/maio/2021)

A Comissão de Ética no Uso de Animais da Faculdade de Medicina Veterinária e Zootecnia da Universidade de São Paulo, no cumprimento das suas atribuições, analisou e **APROVOU** o Relatório Parcial (versão de 19/maio/2021) da proposta acima referenciada.

Resumo apresentado pelo pesquisador: "Estudante: Lucas CEUAx Nº 4123270220 Qual o estágio do estudo no momento? O referido estudo encontra-se em desenvolvimento, ainda sem resultados concretos publicados ou apresentados, uma vez que a estrutura de dados simulados e as respectivas formas de análise estão em desenvolvimento. Tais avanços serão apresentados no exame de qualificação no dia 14 de junho de 2021. Por quanto tempo mais o estudo se estenderá? O estudo se estenderá até a data de vigência do projeto financiado pela FAPESP, restando o período de aproximadamente 10 meses para conclusão. Resultados parciais ou totais apresentados em congresso? Um estudo adicional ao projeto principal apresentado à essa CEUA foi parcialmente apresentado entre os dias 07 e 08 de dezembro de 2020, durante o I Simpósio Internacional da Pós-graduação (XIV Simpósio de Pós-Graduação e Pesquisa em Nutrição Animal-VNP-2020) intitulado EMPREGO DE ALGORITMOS DA TEORIAS DOS GRAFOS NA IDENTIFICAÇÃO DE ANIMAIS CANDIDATOS À GENOTIPAGEM EM PROGRAMAS DE MELHORAMENTO ANIMAL. Conforme mencionado anteriormente, não existe manipulação ou uso direto de animais no presente projeto, todos os dados foram gerados via emprego de dados simulados. Resultados parciais ou totais já publicados? Referente ao estudo adicional apresentado à essa CEUA previamente mencionado, um artigo com o título Evaluation of different graph metrics to identify genotyping candidates in a small cattle population and its impact on Genome-Wide Association Studies foi submetido à revista Livestock Science e atualmente aguarda revisão dos avaliadores. Tal estudo foi gerado com dados simulados, não utilizando qualquer tipo de dados oriundos de procedimentos/manipulações que demandaram contato com os animais. ".

Comentário da CEUA: O projeto está sendo executado como proposto

Prof. Dr. Marcelo Bahia Labruna
Coordenador da Comissão de Ética no Uso de Animais
Faculdade de Medicina Veterinária e Zootecnia da Universidade de São Paulo

Camilla Mota Mendes
Vice-Coordenadora da Comissão de Ética no Uso de Animais
Faculdade de Medicina Veterinária e Zootecnia da Universidade de São Paulo

FOLHA DE AVALIAÇÃO

Autor: ANDRIETTA, Lucas Tassoni

Título: **Uso de *Machine Learning* e dados genômicos para melhoria de características econômicas em bovinos de leite**

Dissertação apresentada ao Programa de Pós-Graduação em Nutrição e Produção Animal da Faculdade de Medicina Veterinária e Zootecnia da Universidade de São Paulo para obtenção do título de Mestre em Ciências.

Data: 22/07/2022

Banca Examinadora

Prof. Dr. Ricardo Vieira Ventura
Instituição: VNP – FMVZ – USP Julgamento: APROVADO

Prof. Dr. Rafael Espigolan
Instituição: FZEA - USP Julgamento: APROVADO

Prof. Dr. Anderson Antônio Carvalho Alves
Instituição: University of Wisconsin-Madison Julgamento: APROVADO

DEDICATÓRIA

Aos meus pais, Zé e Míria.

AGRADECIMENTOS

Agradeço a Deus e a intercessão da S.S Virgem Maria pelo dom da vida, pela saúde, pelo auxílio nos momentos de maior necessidade e pelos milagres diários. A meus pais, José e Míria, por todo o aprendizado e apoio, por sempre fazerem o máximo por mim em todas as situações, nunca medindo esforços para me incentivarem em meus sonhos. A meu irmão Ilson, pelos cuidados e preocupação. À minha tia Marilda, por ser simplesmente minha segunda mãe.

A todos aqueles que tornaram este estudo possível, tendo sido fundamentais em suas considerações e sugestões: Prof. Dr. Anderson Alves, Prof. Dr. Júlio Balieiro, Prof. Dr. Mehdi Sargolzaei e Prof. Dr. Roberto Carvalheiro.

Aos Professores e todas as pessoas que colaboraram em meu crescimento técnico e pessoal ao longo de toda minha carreira acadêmica.

Aos meus amigos e professores que tive a oportunidade de conhecer pela Pós-Graduação, em especial do Laboratório BioMa, nas pessoas do Prof. Dr. Júlio Cesar de Carvalho Balieiro, Diógenes Lodi Pinto, Wecksley Leonardo de Souza, Alana Selli, Dr. Fernando de Oliveira Bussiman, Dr. Bruno Perez e a Dra. Lígia Garcia Mesquita.

À Fundação de Amparo à Pesquisa do Estado de São Paulo (FAPESP) pela bolsa (Processo Nº2020/04461-6) que viabilizou o desenvolvimento deste estudo integralmente, assim como o fomento de Projeto Jovem Pesquisador cedido ao orientador deste estudo (Processo Nº 16/19514-2).

À Universidade de São Paulo, em especial a Faculdade de Medicina Veterinária e Zootecnia (FMVZ) e o Departamento de Nutrição e Produção Animal (VNP) por ter me proporcionado todas as condições estruturais para desempenho deste.

Por fim, a meu professor, orientador, afiliado de crisma, e a pessoa que tenho o maior prazer do mundo em chamar de amigo, o Prof. Dr. Ricardo Vieira Ventura, que acreditou em mim desde o início, foi compreensivo nos momentos mais difíceis de minha jornada e proporcionou as condições possíveis e impossíveis para meu crescimento científico e pessoal. Serei eternamente grato, tendo para mim como um exemplo e fonte de inspiração.

RESUMO

ANDRIETTA, L. Tassoni **Uso de *Machine Learning* e dados genômicos para melhoria de características econômicas em bovinos de leite**. 2022. 69 f. Dissertação (Mestrado em Ciências) – Faculdade de Medicina Veterinária e Zootecnia, Universidade de São Paulo, São Paulo, 2022.

Estratégias de acasalamento dirigido são consideradas ferramentas essenciais em programas de melhoramento animal. Com o advento da Seleção Genômica na última década, em associação aos avanços nas técnicas reprodutivas, nota-se diminuição no intervalo de gerações, aumentos na acurácia de predição e na intensidade de seleção, proporcionando expressivo ganho genético para os animais das cadeias produtivas. A fim de se compreender atributos das informações genotípicas e otimizar acasalamentos, objetivou-se neste estudo, por meio da simulação de uma população de bovinos leiteiros, a exploração de diferentes abordagens de extração de atributos de informações genotípicas de indivíduos do rebanho, tendo como objetivo a avaliação do desempenho preditivo ao se empregar tais dados por meio de dois algoritmos de *Machine Learning* (*Random Forests* e *K-Nearest Neighbours)* em 11 cenários propostos referentes ao coeficiente de endogamia (Froh), valor genético, além da proposta de um método de acasalamento. O uso das abordagens propostas de extração de atributos contribuiu para a diminuição dos dados a serem empregados nos modelos em até 98%, implicando na maioria dos cenários, em resultados mais representativos quando consideradas as informações reduzidas em dimensão quando comparadas a utilização de dados íntegros. Destacou-se o uso do *algoritmo Random Forests* para os cenários de regressão propostos, em especial na predição dos valores de Froh utilizando os genótipos dos pais em comparação a informação do próprio indivíduo, sendo o resultado de r^2 do primeiro superior em 29%, quando utilizado o método de distância euclidiana proposto. Destaca-se também a abordagem visual proposta, favorecendo o desenvolvimento de estudos em busca de indivíduos a serem acasalados de acordo com os interesses relacionados a uniformidade de rebanho e a potenciais expoentes no quesito reprodução.

Palavras-chave: aprendizado de máquina. acasalamento dirigido. melhoramento animal.

ABSTRACT

TASSONI ANDRIETTA, L. **Applications of machine learning and genomic data to improve economic traits in dairy cattle**. 2022. 69 f. Dissertação (Mestrado em Ciências) – Faculdade de Medicina Veterinária e Zootecnia, Universidade de São Paulo, São Paulo, 2022.

Mating strategies are considered essential tools on animal breeding programs, playing an important role to achieve genetic progress. The advent of Genomic Selection in the last decade, in addition to the improvements on reproduction techniques, shortened the generation interval, enhanced breeding values prediction reliabilities and selection intensity, which provided an expressive genetic gain across several industries. In order to understand attributes of genotypic information and optimize matings, the objective of this study, through the simulation of a dairy cattle population, was to explore different approaches to extract attributes of genotypic information from individuals in the herd, with the objective of the evaluation of the predictive performance when using such data through two Machine Learning algorithms (Random Forests and K-Nearest Neighbors) in 11 proposed scenarios referring to the inbreeding coefficient (Froh), genetic value, in addition to the proposal of a mating strategy. The use of the proposed feature extraction methods contributed to the reduction of data by up to 98%, implying in most scenarios, in lower costs and better results when compared to the use of raw data. The Random Forests algorithm for the proposed regression scenarios showed the better results, especially in the prediction of Froh values using the genotypes of the sire and dam in comparison with the of the individual and its own information, with the result of r^2 of the former being superior by 29%, when using the proposed Euclidean distance method. Also noteworthy is the proposed visual approach, favoring the development of studies in search of individuals to be mated according to the interests related to herd uniformity and potential exponents in reproduction.

Keywords: Machine learning. Mating scheme. Animal breeding.

LISTA DE FIGURAS

Figura 1 - Esquema adotado na Seleção Genômica.....21
Figura 2 – Diferentes distribuições e pontos de truncamento.....27
Figura 3 – Esquema representativo do processo de Seleção e Extração de atributos. 31
Figura 4 - Esquema de seleção de hiperplano e as respectivas projeções.....32
Figura 5 – Esquema do processo de predição pelo algoritmo K-NN.....33
Figura 6 – Estrutura geral de árvore e os respectivos elementos presentes no algoritmo Decision Tree e Random Forests.....34
Figura 7 – Esquema das principais etapas desenvolvidas neste estudo. 36
Figura 8 - Esquema de Simulação. 37
Figura 9 - Representação do funcionamento da divisão do conjunto de dados em teste e treinamento para os modelos de Machine Learnining.....39
Figura 10 – Esquema de entrada de dados para o primeiro método de predição de F_{ROH} proposto.....41
Figura 11 - Esquema de entrada de dados para o segundo método de predição de F_{ROH} proposto.....42
Figura 12 - Esquema de entrada de dados para o terceiro método de predição de F_{ROH} proposto. 42
Figura 13 - Esquema de entrada de dados para o terceiro método de predição de F_{ROH} proposto. 43
Figura 14 - Esquema de entrada de dados para o quarto método de predição de F_{ROH} proposto. 44
Figura 15 - Esquema de entrada de dados para o quinto método de predição de F_{ROH} proposto. 44
Figura 16 - Esquema de entrada de dados para o método de predição de valor fenotípico proposto.....45
Figura 17 - Esquema de entrada de dados para o método de predição de valor genômico predito.....46
Figura 18 - Esquema de entrada de dados para o método de predição de valor genômico verdadeiro.....46
Figura 19 - Esquema de efeito de SNP para predição de valor genômico verdadeiro. 46
Figura 20 - Esquema de entrada de dados para o método de predição de rótulos...48

Figura 21 – Processo de divisão das fitas dos genótipos após faseamento, ou oriundo da saída nativa do QMSim. 48
Figura 22 – Representação da saída de dados do *software* pybioma.....49
Figura 23 - Trecho dos formatos de arquivo de entrada utilizados pelo software Pybioma. 49
Figura 24 – Decaimento médio do LD da população simulada. 50
Figura 25 – Informações dos parâmetros populacionais ao longo das gerações.....51
Figura 26 – Variabilidade do mérito genético predito entre as “Progênies Fake”. 59
Figura 27 - Variabilidade do coeficiente de endogamia predito entre as “Progênies Fake”. 59

LISTA DE TABELAS

Tabela 1 – Resultados do algoritmo Random Forests para predição de F_{ROH} usando os próprios genótipos. 53
Tabela 2 - Resultados do algoritmo KNN para predição de F_{ROH} usando os próprios genótipos. 53
Tabela 3 - Resultados do algoritmo Random Forests para predição de F_{ROH} usando genótipos dos pais. 54
Tabela 4 - Resultados do algoritmo KNN para predição de F_{ROH} usando genótipos dos pais. 55
Tabela 5 - Resultados do algoritmo *Random Forests* para predição de Fenótipos, GEBV (BLUPF90) e TBV. 56
Tabela 6 – Acurácias dos algoritmos de classificação 57
Tabela 7 – Correlação entre valores preditos (Progênie Fake) e valores reais simulados 58

SUMÁRIO

1. Introdução 14

2. Objetivos 17

3. Revisão bibliográfica 17

2.1 Seleção 17

2.1.1 Seleção Tradicional 17

2.1.2 Seleção Genômica 19

2.2 Endogamia 21

2.2.1 Coeficiente de endogamia via *pedigree* 22

2.2.2 Coeficiente de endogamia genômico 23

2.3 Acasalamento dirigido 25

2.3.1 Tradicional 26

2.3.2 Nível Genômico 28

2.4 Machine Learning 28

2.4.1 Extração e seleção de atributos 29

2.4.2 Algoritmos 31

2.4.3 Aplicação na genômica 35

3. Materiais e métodos 35

3.1 Simulação 36

3.2 Método geral para treinamento e teste dos modelos 39

3.2 Cenários considerados 40

3.2.1 Coeficiente de endogamia genômico 40

3.2.2 Valores fenotípico e genético estimados 44

3.2.3 Acasalamento genômico 48

4. Resultados e discussão 49

4.1. Validação da simulação 49

4.2. Coeficiente de endogamia 51

4.2.1 Coeficiente de endogamia predito pelo uso do próprio genótipo 52

4.2.1 Coeficiente de endogamia predito pelo genótipo dos pais 54

4.3. Predição dos valores fenotípicos, dos valores genômicos estimados e do valor genético real 55
4.4. Predição dos indivíduos otimizados 56
4.5 Resultados do acasalamento genômico 57

4. Conclusão 59

5. Referências 59

1. INTRODUÇÃO

A Seleção Genômica (*Genomic Selection* - GS) (MEUWISSEN; HAYES; GODDARD, 2001), método estabelecido em larga escala a partir do ano de 2009, foi inicialmente aplicada em rebanhos de bovinos leiteiros e, posteriormente, em rebanhos de corte (BOUQUET; JUGA, 2013). Ao se comparar com os métodos de seleção quantitativos, tradicionalmente empregados anterior ao advento da GS, destacam-se: aumento na acurácia e intensidade de seleção, e a respectiva diminuição no intervalo entre gerações (MILLER, 2010). Além disso, viabilizam-se resultados mais eficientes e menos onerosos em relação aos métodos exclusivamente quantitativos. Principalmente ao se tratar do teste de progênies em rebanhos leiteiros, sendo este, responsável pelo maior percentual dos custos operacionais, rotineiramente empregado em programas de melhoramento com a abordagem tradicional (SCHAEFFER, 2006; VANRADEN, 2020).

Os Polimorfismos de Base Única - *Single Nucleotide Polymorphisms* - (SNPs) são representados por alterações pontuais, ocorrendo aproximadamente a cada 500 - 1000 pares de bases ao longo de toda a extensão do genoma, e com frequência superior a 1% dentro da referida população (CHEN et al., 2002). Justifica-se o uso destes marcadores pela possível associação às regiões cromossômicas atribuídas a aspectos de interesse econômico (SYVÄNEN, 2001). Adiciona-se às características de interesse na eleição dos SNPs: baixa taxa de mutação, natureza bi alélica e abundante ocorrência no genoma. Soma-se a tais fatores, a diminuição no preço do processo de genotipagem via aprimoramentos em painéis de uso comercial (DAWSON, 1999; GOMPERT et al., 2010; M VERLOUW et al., 2021).

Para efetivo emprego das técnicas de seleção genômica, assume-se a existência de desequilíbrio de ligação - *Linkage Disequilibrium* - (LD), condição associada aos *Quantitative Trait Loci* (QTLs) (MEUWISSEN; HAYES; GODDARD, 2001). Os QTLs são compreendidos por regiões específicas, distribuídas continuamente ao longo do genoma e responsáveis por alterações fenotípicas. Logo, por alterarem características quantitativas, a busca desses segmentos genômicos apresenta grande relevância econômica, diretamente relacionada ao aumento do ganho genético (TOLEDO et al., 2008). Duas principais abordagens têm sido consideradas na evidenciação dos QTLs: a descoberta de genes candidatos (*Candidate genes)* e o mapeamento de QTLs (*QTL mapping*). A primeira (*Candidate*

genes) assume que mutações causais em genes específicos e descobertos via estudos meticulosos de associação genômica ampla - *Genome Wide Association Studies* - (GWAS) potencialmente explicam alterações fisiológicas e, consequentemente, fenotípicas (ZHANG et al., 2012). Em relação à segunda técnica (*QTL mapping*), as regiões cromossômicas de interesse são conhecidas, porém os genes responsáveis pelas alterações nas características quantitativas são considerados desconhecidos. Nesta abordagem, contabiliza-se a contribuição de cada locus e o percentual que cada um explica da variância das características contínuas sob investigação. Sendo, portanto, o *QTL mapping* a forma adotada nos programas de Seleção Genômica (RABIER et al., 2016; SEATON et al., 2002). Quanto ao LD, entende-se este como a associação não randômica de alelos entre dois *loci*. Condição que possibilita a seleção genômica, uma vez que a não aleatorização de determinados segmentos favorece a seleção e manutenção de regiões cromossômicas de interesse (MEUWISSEN; HAYES; GODDARD, 2001).

Ao longo dos anos, os principais esforços em relação ao aprimoramento dos métodos de GS se voltaram principalmente na otimização de aspectos preditivos (JIA; JANNINK, 2012; MEUWISSEN; GODDARD, 2010; MONTESINOS-LÓPEZ et al., 2021). Tais esforços, impulsionados pelo uso intensivo de técnicas de inseminação artificial em rebanhos leiteiros, implicaram no incremento dos níveis de intensidade de seleção, diminuição no intervalo de gerações, e consequentemente, no aumento do ganho genético. Entretanto, como aspecto negativo, tais avanços favorecem o aumento dos níveis de endogamia entre indivíduos da população, uma vez que há tendência na diminuição da variabilidade genética dos rebanhos sob regime de seleção, impulsionado pela diminuição do intervalo entre gerações (WIGGANS et al., 2017). Na tentativa de se possibilitar a produção de indivíduos mais adequados aos sistemas de produção, busca-se maximizar a ocorrência de combinações gênicas favoráveis e o controle dos níveis de endogamia da população. Neste contexto, métodos de acasalamento direcionado são cruciais para o adequado funcionamento e manutenção da cadeia produtiva (COLE; VANRADEN, 2010).

O acasalamento dirigido é definido como o processo de escolha de pares de indivíduos destinados à reprodução em um mesmo rebanho, raça ou espécie. Essa estratégia pode apresentar diferentes motivações, consistindo em forma geral, no direcionamento do fluxo gênico de acordo com o interesse da cadeia produtiva

(ALLAIRE, 1980). Em programas de melhoramento genético, o acasalamento dirigido tem como princípio favorecer combinações gênicas que atendam às principais demandas de cada sistema produtivo em questão, afetando diretamente em decisões relacionadas à escolha de gametas de determinados indivíduos, compra de animais e/ou material genético externos, critério de descarte em rebanhos, entre outros aspectos (BOURDON, 2000).

Tradicionalmente, metodologias aplicadas em acasalamentos dirigidos se baseiam na técnica conhecida por seleção de truncamento (*Truncation Selection*), que consiste em conduzir o acasalamento entre os melhores indivíduos de uma população (baseado nos valores genéticos estimados - *Estimated Breeding Value* - EBV*)*. Buscando, dessa maneira, a maximização do ganho genético (FALCONER et al., 1996). A referida escolha de indivíduos, em busca de formar pares, pode ser embasada no uso de informações variáveis via métodos quantitativos ou genômicos, associados a Índices de Seleção de interesse (KINGHORN, 2011a). Tendo o direcionamento de acasalamentos como objetivo o aumento do valor genético dos indivíduos e o controle da endogamia (HAYES; SHEPHERD; NEWMAN, 2002). Mantendo, dessa forma, a diversidade genética e favorecendo o ganho genético a longo prazo.

O uso de dados genômicos em sistemas de acasalamento direcionado pode, muitas vezes, ser dificultado pelo “excesso” de variáveis (X) em relação às observações (Y) disponíveis em determinadas situações, fenômeno nomeado de “maldição da dimensionalidade” (CHEN, 2009). Logo, relações complexas podem ser prejudicadas ao serem estimadas via emprego de modelos lineares, portanto, modelos não-lineares de *Machine Learning* se apresentam como possíveis alternativas (NAYERI; SARGOLZAEI; TULPAN, 2019).

Como reportado em estudos prévios, algoritmos de *Machine Learning* podem proporcionar vantagens quando comparados aos métodos tradicionais relacionados ao emprego e exploração de dados genômicos (ALVES et al., 2020; PÉREZ-ENCISO; ZINGARETTI, 2019; XU; JACKSON, 2019a). Portanto, tal abordagem pode ser apresentada como potencial alternativa aos estudos de acasalamento dirigido, uma vez que apresenta um maior nível de detalhamento dos dados no processo de escolha dos melhores acasalamentos. Neste contexto, é de suma importância a condução de

estudos que utilizem a detecção de padrões não tão evidenciados em busca novas abordagens que proporcionem aumento do ganho genético futuro.

2. OBJETIVOS

Avaliar o impacto de algoritmos supervisionados associados a diferentes tipos de dados como ferramenta de direcionamento de acasalamentos, buscando proporcionar estratégias que favoreçam a maior compreensão dos níveis de endogamia e do progresso genético a partir das informações de touro e vaca.

3. REVISÃO BIBLIOGRÁFICA

2.1 Seleção

2.1.1 Seleção Tradicional

Embora a escolha de indivíduos, baseada em atributos fenotípicos ou da respectiva ancestralidade, ocorresse muito antes que qualquer abordagem técnico-científica, evidenciou-se no decorrer do processo os efeitos da seleção baseado em aspectos de interesse, e a clara diferenciação entre as características qualitativas e quantitativas. Sendo, a primeira classe, governada por poucos genes, relacionada a fenótipos facilmente identificáveis, e suscetíveis a seleção baseada em atributos meramente visuais. No caso das características do segundo grupo, que normalmente estão associadas aos atributos de maior interesse econômico, são relacionadas a efeitos poligênicos. Estes, apresentam estrutura de elevada complexidade, uma vez que muitos genes (centenas ou até mesmo milhares) contribuem com pequenos efeitos em atributos fenotípicos, seguindo uma distribuição normal. Que no caso dos animais de produção, são responsáveis pela maioria das características de interesse de seleção como produção de leite, percentuais de gordura e proteína, características reprodutivas, entre outras (DEKKERS, 2012).

A maior parte das informações consideradas na seleção de indivíduos são oriundas de características sujeitas a mensuração, condição que viabilizou, ao longo do tempo, a coleta de inúmeros registros fenotípicos e do respectivo *pedigree.* Impulsionando, consequentemente, a construção de bancos de dados robustos, favorecendo desta maneira a aplicação dos referidos registros na estimativa dos valores genéticos dos indivíduos. Sendo, o valor genético, definido pela soma dos

efeitos aditivos de todos os loci que contribuem na característica de interesse (QTL) em relação à média da população.

Com os avanços no campo estatístico, especialmente em relação a análise de variâncias, possibilitou-se a separação da variância genotípica, em especial a parte dos efeitos aditivos, dos demais fatores (FISHER, 1919). Somando essas informações ao acesso dos coeficientes de parentesco (WRIGHT, 1921), e emprego e estimação da herdabilidade das características de interesse, viabilizou-se a estimação dos valores genéticos dos animais. Permitindo, dessa maneira, a comparação e classificação entre indivíduos de um mesmo rebanho, e o respectivo progresso na expressão de fenótipos de interesse, a partir das informações oriundas dos indivíduos.

Posteriormente, com o desenvolvimento da metodologia dos modelos lineares mistos, via *Best Linear Unbiased Prediction* (BLUP) (HENDERSON, 1975) associada aos avanços computacionais, houve uma revolução na genética quantitativa, uma vez que a técnica viabilizou a maximização da extração de informação dos fenótipos registrados, tanto das características de interesse, quanto de características correlacionadas no processo de estimação do valor genético (EBV) dos indivíduos candidatos à seleção. Destaca-se também o fato de o método considerar não apenas os registros do próprio indivíduo, mas também de seus pares, de acordo com o parentesco, a fim de se maximizar a acurácia dos valores genéticos preditos. Além disso, somou-se a possibilidade em se avaliar um maior número de indivíduos, tendo em vista a possibilidade de comparação entre rebanhos, que associados a estratégias de seleção, favoreceram grandes impactos nos índices de produção.

Embora a técnica tenha revolucionado a forma de seleção, apresentando resultados expressivos, também algumas dificuldades e limitações foram encontradas. Dentre elas, o maior entrave se referiu a necessidade em se coletar fenótipos dos indivíduos candidatos a seleção, ou de seus pares em momentos específicos, e por diversas vezes ao longo da vida para obtenção de valores significativos de acurácia. Somado ao fato que algumas características de interesse só poderiam ser mensuradas tardiamente podendo, muitas vezes, estarem associadas a dificuldades de mensuração (DEKKERS, 2012). Tais dificuldades foram bastante expressivas em especial nos rebanhos leiteiros, uma vez que a forma de seleção de touros seria baseada no teste de progênies. Ou seja, o valor genético dos touros só poderia ser estimado baseado na expressão fenotípica das filhas, caracterizando este método de avaliação, embora eficaz, como dispendioso,

demorado e de grandes complicações logísticas (SCHAEFFER, 2006). As referidas restrições na fenotipagem acabaram por limitar o progresso genético ao ritmo da obtenção dos mesmos, levando ao desenvolvimento de novas abordagens a fim de se acelerar o processo.

2.1.2 Seleção Genômica

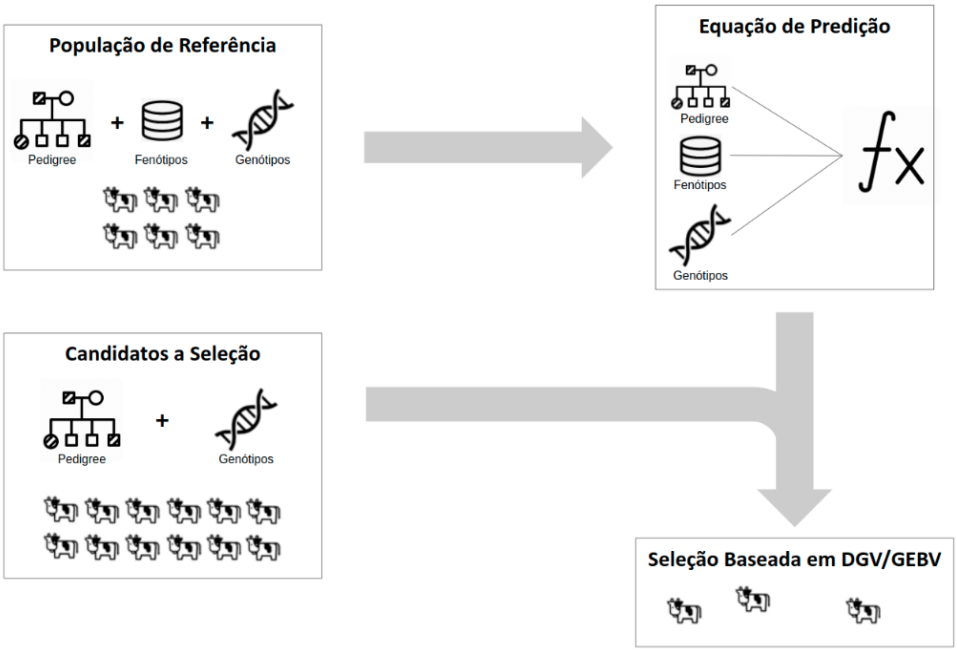
Com os avanços nas técnicas laboratoriais, possibilitando o acesso à informação ao nível do material genético dos indivíduos, novas abordagens de seleção baseadas na identificação de regiões associadas a QTLs foram desenvolvidas. A primeira forma de seleção associada a marcadores ficou conhecida como *Marker Assisted Selection* - Seleção Assistida por Marcadores - (MAS) (PEDERSEN; SØRENSEN; BERG, 2009), técnica que considerava o uso de apenas um, ou um pequeno grupo de marcadores moleculares do tipo *Single Nucleotide Polymorphism* (SNP) associados a QTLs. Porém, apresentou restrições de uso, uma vez que ao se avançar na investigação, ficaram claras as limitações da técnica, devido a maioria das características de interesse econômico serem afetadas por vários genes e, respectivamente, vários QTLs distribuídos ao longo do genoma. Comprovando o aspecto poligênico da maior parte das características de produção (CHAMBERLAIN; MCPARTLAN; GODDARD, 2007).

Tendo em vista a ineficiência do uso de poucos marcadores ao longo do genoma associados a QTLs, uma nova abordagem foi desenvolvida, sendo conhecida como *Genomic Selection* (GS) - Seleção Genômica (MEUWISSEN; HAYES; GODDARD, 2001). Diferentemente da abordagem anterior, esta, ao invés de apenas um, ou um pequeno conjunto de marcadores, passou a considerar um grande número de SNPs espalhados ao longo do genoma, de forma com que seja possível explicar uma maior fração da variância da característica de interesse. Diferentemente da MAS, a GS não leva em consideração apenas SNPs que estejam relacionados exclusivamente a efeitos causais, mas ao efeito aparente, uma vez que esses se encontram em desequilíbrio de ligação com os QTLs, viabilizando o processo de seleção. Com o advento dos programas de sequenciamento do genoma bovino, e consequentemente, o mapeamento de novas regiões de QTL, novos SNPs foram identificados, viabilizando a criação dos painéis comerciais de genotipagem com ampla cobertura do genoma (GIBBS et al., 2009).

De maneira resumida, a GS se baseia no emprego de equações de predição utilizando marcadores genômicos a fim de se estimar os valores genéticos dos indivíduos a partir dos efeitos estimados para cada marcador em relação ao fenótipo de interesse. Dessa forma, tais valores podem ser obtidos de duas principais maneiras, sendo a primeira baseada no próprio genótipo dos indivíduos a partir da somatória de efeitos estimados dos SNPs (valor genômico direto – *Direct Genomic Value* – DGV) (MEUWISSEN; HAYES; GODDARD, 2001), e a segunda a partir da associação tanto dos efeitos estimados de SNPs quanto da influência do pedigree e dos dados fenotípicos dos mesmos indivíduos (Valor Genômico Estimado - *Genomic Estimated Breeding Value* - GEBV) (VANRADEN et al., 2009). Como pode ser observado na Figura 1, ao se empregar a seleção genômica em rebanhos comerciais, faz-se necessário a existência de uma população de referência contendo informações de pedigree, de fenótipos e de genótipos para a criação da equação de predição. Esta utiliza as informações referenciadas na estimação dos efeitos de cada SNP individualmente e de acordo com as frequências alélicas ao longo da população de referência, de forma que ao se somar os efeitos estimados de acordo com os genótipos dos indivíduos, torna-se possível obter o valor genômico estimado, mesmo sem o registro dos fenótipos de interesse (KOIVULA et al., 2012).

A partir do processo descrito, viabiliza-se o uso da seleção genômica, possibilitando, dessa maneira, maior intensidade de seleção, diminuição do intervalo entre gerações e consequentemente o aumento do ganho genético ao se comparar com o método de seleção tradicional. Além da diminuição dos custos associados aos métodos de seleção, em especial no que se refere ao teste de progênies, previamente descrito em rebanhos leiteiros (HARRIS; JOHNSON, 2010; HAYES et al., 2009)

Figura 1 - Esquema adotado na Seleção Genômica



Fonte: Autoria Própria.

Embora as primeiras impressões e expectativas a respeito da seleção genômica indicassem um maior controle, e consequentemente diminuição na taxa de endogamia por geração (LUND et al., 2011), estudos envolvendo rebanhos leiteiros sob intenso regime de seleção genômica indicaram o contrário, uma vez que as taxas de endogamia se mostraram crescentes ao longo dos anos sob o referido regime de seleção (MAKANJUOLA et al., 2020).

2.2 Endogamia

Endogamia pode ser descrita, de forma resumida, como o acasalamento de animais mais intimamente relacionados do que o relacionamento médio dentro da raça ou população em questão. Tal condição pode levar a efeitos deletérios no âmbito de produção animal, uma vez que a diminuição da variabilidade genética em rebanhos pode prejudicar a taxa de ganho genético, levando a efeitos de depressão endogâmica no longo prazo (DOEKES et al., 2019). A depressão endogâmica é caracterizada pelo aumento de segmentos idênticos ao ponto de se observar menores índices reprodutivos, produtivos e de sobrevivência, quando comparados a indivíduos de menor grau de parentesco em comum. Os efeitos deletérios normalmente estão associados a genes recessivos (letais), que acabam sendo expressos a partir do aumento dos seguimentos idênticos no processo do aumento da endogamia (HOWARD et al., 2017).

Compreende-se que indivíduos aparentados apresentam maior propensão de compartilhar também características produtivas de interesse quando comparados a indivíduos menos relacionados. Cabe mencionar que o acasalamento de indivíduos não aparentados em uma população finita e sob seleção é impossível, uma vez que a longo prazo todos os descendentes tendem a ter ancestrais em comum. Desta forma, há tendencias em se fixar alelos ao longo do tempo, reduzindo dessa maneira a variância genética aditiva, assim como a resposta a seleção de características altamente selecionadas, e demais características correlacionadas (WANG; SANTIAGO; CABALLERO, 2016). Devido aos possíveis efeitos negativos da endogamia, associadas à importância do pedigree nos modelos preditivos dos valores genéticos, ao longo do processo de evolução do melhoramento animal, diferentes metodologias envolvendo a estimação do coeficiente de endogamia foram desenvolvidas. Estas são divididas em dois grupos, sendo a abordagem tradicional alimentada pelas informações de pedigree, e a molecular partir da informação de marcadores genômicos.

2.2.1 Coeficiente de endogamia via *pedigree*

O coeficiente de endogamia (F) é um parâmetro fundamental a ser explorado em populações sob regime de seleção. Originalmente, a partir das informações de pedigree, Wright, em 1922, definiu tal coeficiente como a correlação entre dois genes homólogos em um mesmo *locus* de um indivíduo diploide. Ou seja, a probabilidade em se encontrar um mesmo gene replicado em um mesmo indivíduo a partir de acasalamentos que apresentem um ou mais ancestrais em comum (WRIGHT, 1922). Posteriormente, seguindo a mesma proposta de Wright, Malécot, em 1948, introduziu uma definição alternativa a partir da estimativa da probabilidade de dois genes homólogos em um mesmo *locus* de um mesmo indivíduo, ou entre indivíduos aparentados, que compartilhassem cópias do mesmo gene a partir da condição de serem idênticos por descendência (*Identity By Descent* - IBD) de um mesmo, ou dos mesmos, ancestrais em comum (MALECOT,1948). Em 1965, Wright demostrou que as duas definições são equivalentes na maioria dos casos simples, porém a proposta do coeficiente de correlação, mostrou-se mais generalista quando comparada à proposta de Malécot (WRIGHT, 1965).

Tradicionalmente, o método de correlação proposto por Wright (1922) foi o mais adotado em rebanhos sob seleção, especificamente na composição da matriz de parentesco (Matriz A). Embora bastante eficiente, proporcionando grandes avanços no controle da endogamia em rebanhos sob seleção, algumas limitações puderam ser observadas em relação ao referido método. O fato de o cálculo do coeficiente de endogamia depender diretamente das informações de pedigree suscetibilizou a ocorrência de erros na estimação, uma vez que em lacunas referentes a genealogia dos indivíduos são bastante comuns, em especial em rebanhos pequenos, inviabilizando o cálculo de indivíduos não conectados à estrutura do pedigree (CORTES-HERNÁNDEZ et al., 2021).

Em rebanhos da raça holandesa, reportou-se na literatura percentuais de erro de paternidade associados a falhas em registros de pedigree consideráveis, evidenciando, dessa maneira, potenciais resultados espúrios tanto no que diz respeito a estimação de valores genéticos, quanto na otimização de acasalamentos. (BRADFORD et al., 2019; GARCÍA-RUIZ; WIGGANS; RUIZ-LÓPEZ, 2019; VISSCHER et al., 2002). Com os avanços no campo da ciência molecular, novas formas de se avaliar o grau de relacionamento entre indivíduos foram propostas, viabilizando o cálculo do coeficiente de endogamia via informação genômica, favorecendo resultados mais acurados.

2.2.2 Coeficiente de endogamia genômico

Com o advento dos painéis de genotipagem de marcadores do tipo SNP, e o emprego da seleção genômica, um número cada vez maior de indivíduos de ambos os sexos são genotipados, especialmente em rebanhos leiteiros. No que se refere à raça holandesa, sendo a mais expressiva quanto a produção de leite, mais de 5,8 milhões de indivíduos foram genotipados globalmente de acordo com os dados fornecidos pela USCDCB (acesso 20 de abril de 2022). Destaca-se que o emprego da seleção genômica favoreceu no aumento de animais a serem avaliados, uma vez que os custos associados são inferiores ao tradicional teste de progênies. Porém, mesmo com o aumento de candidatos sob avaliação, não necessariamente houve mudanças significativas na diversidade genética do rebanho, uma vez que indivíduos aparentados com animais de elite tendem a se beneficiar das informações coletadas

previamente de seus ancestrais na estimação dos respectivos valores genômicos (LOURENCO et al., 2014).

Uma das abordagens mais simples e comumente utilizadas empregando dados genômicos a fim de se estimar o coeficiente de endogamia dos indivíduos genotipados são as corridas de homozigosidade. Embora via genotipagem não seja possível constatar diretamente os segmentos cromossômicos idênticos por descendência (IBD), é suscetível a constatação do segmento pela condição de Idêntico por Estado (I*dentical By State* - IBS), baseado na ocorrência de homozigosidade em múltiplos marcadores em sequência. Desta maneira, quando dois segmentos cromossômicos ocorrem em um mesmo indivíduo, a condição é conhecida como corrida de homozigosidade ou *Runs of Homozigosity* (ROH) (GIBSON; MORTON; COLLINS, 2006; MACLEOD et al., 2009). Ou seja, são regiões contínuas de ocorrências de genótipos homozigotos presentes nas duas fitas do indivíduo, oriundas de haplótipos idênticos transmitidos pelos respectivos pais.

Uma vez que os ROHs estão presentes ao longo do genoma, torna-se possível estimar o número de gerações associadas ao ancestral comum com que os pais dos indivíduos genotipados foram originados. Tal constatação é possível a partir do tamanho dos ROHs encontrados, pois devido ao processo de segregação mendeliana, com o passar das gerações, os segmentos tendem a ficar cada vez menores. Desta maneira, associa-se segmentos maiores com ancestralidade endogâmica recente, enquanto os segmentos menores estão associados com endogamia em acasalamentos passados, ou até mesmo ao processo de formação da população (FORUTAN et al., 2018). Alguns estudos estimaram o tamanho das corridas de homozigosidade associadas aos eventos de meiose, e correlacionaram ROHs de comprimentos de 25, 10 e 2,5 Mb como oriundos de 2, 5 e 20 gerações anteriores, respectivamente (HOWARD et al., 2015; PURFIELD et al., 2012). No entanto, faz-se necessário mencionar que longos ROHs podem persistir por fatores como mutações não usuais, desequilíbrio de ligação e diferentes taxas de recombinação em determinadas regiões cromossômicas (GIBSON; MORTON; COLLINS, 2006). Diferentes abordagens podem ser empregadas a fim de se estimar o nível de relacionamento entre indivíduos a partir das informações genotípicas. Porém, ao se tratar do cálculo do nível endogâmico dos indivíduos via uso dos ROHs, a forma mais considerada é o modelo proposto por (PURFIELD et al., 2012), onde:

$$F_{ROH} = \Sigma \frac{L_{ROH}}{L_{TOTAL}}$$

Sendo L_{ROH} referente ao somatório dos ROHs identificadas ao longo do genoma a partir do critério estabelecido, e L_{TOTAL} condizente com o comprimento total dos cromossomos autossômicos de acordo com a cobertura do painel de genotipagem empregado. A partir da fórmula demonstrada, torna-se possível acessar os valores referentes ao grau de endogamia dos indivíduos a partir das corridas de homozigosidade (F_{ROH}).

Estudos demonstraram a importância em se considerar os níveis de F_{ROH} nas populações de interesse, uma vez que se evidenciaram a ocorrência de desordens associadas a genes recessivos e respectivos efeitos deletérios, que comumente estão alocados em regiões de ROH (BISCARINI et al., 2016; MÉSZÁROS et al., 2015). Desta maneira, é de extrema importância a investigação do coeficiente de endogamia genômico, uma vez que as informações podem auxiliar diretamente no direcionamento de acasalamentos, evitando possíveis aspectos deletérios, além de colaborar com a manutenção da variabilidade genética do rebanho (TORO; VARONA, 2010).

2.3 Acasalamento dirigido

Uma vez aplicadas as metodologias de seleção, sejam estas baseadas no método tradicional quantitativo ou na seleção genômica, a segunda etapa é relacionada ao direcionamento de acasalamentos. A escolha de indivíduos e a combinação dos pares a serem acasalados não é uma tarefa fácil, tendo em vista que muitas informações devem ser levadas em consideração simultaneamente. Dentre estas, destacam-se o ganho genético, o coeficiente de endogamia, e complicações logísticas e de custo operacional. Geralmente, alguns componentes são invariáveis em qualquer uma das abordagens de acasalamento empregadas, como o ranqueamento de indivíduos a partir dos valores genéticos ou genômicos estimados, e a respectiva preferência por aqueles mais bem ranqueados (KINGHORN, 2000).

Uma vez escolhido o percentual de indivíduos a ser considerado como reprodutores, baseando-se no ranqueamento de valores genéticos estimados, duas principais abordagens de acasalamento podem ser consideradas, sendo essas o acasalamento aleatório entre indivíduos selecionados e o acasalamento direcionado.

Em vista de aumentos mais expressivos do progresso genético, controle nos níveis de endogamia e consequente manutenção da variabilidade genética, prioriza-se o direcionamento de acasalamentos como estratégia de interesse (HAMMACK, 2011).

Em rebanhos bovinos, destacam-se duas formas acasalamento direcionado, sendo a primeira classificada como acasalamento preferencial positivo, e a segunda como acasalamento preferencial negativo. No acasalamento preferencial positivo, indivíduos de maior mérito genético são direcionados a serem acasalados entre si. Esta abordagem é priorizada em rebanhos de elite, uma vez que as progênies terão maior probabilidade de apresentarem valores genéticos superior ao dos pais, sendo esses, os pais e mães de touros das próximas gerações. Quanto ao acasalamento preferencial negativo, compreende-se como uma estratégia empregada quanto ao uso de touros de alto valor genético em rebanhos comerciais. Ou seja, touros provados e acasalados com vacas de produção em busca de se aumentar a média produtiva e a uniformidade do rebanho (DE REZENDE NEVES et al., 2009).

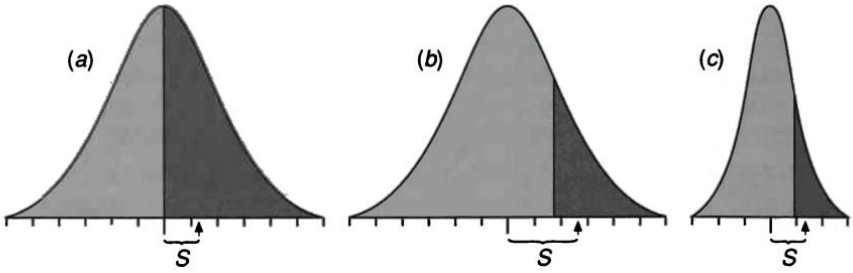
De forma geral, seja no método tradicional ou genômico, o acasalamento dirigido pode apresentar diferentes estratégias. Estas possivelmente associadas à diferentes índices de seleção e a algoritmos de otimização, sendo tais abordagens classificadas como Sistemas de Acasalamento. Quanto aos dados empregados, podem ser consideradas informações de pedigree, valor genético estimado, e também, em alguns casos, informações em nível genômico, como a presença de genes deletérios, regiões de homozigose em comum, entre outras abordagens (AKDEMIR; SÁNCHEZ, 2016a; BENGTSSON et al., 2022; KINGHORN, 2011b). Para melhor elucidar o direcionamento de acasalamentos, segmentamos um resumo das abordagens possíveis em dois principais grupos: forma tradicional e em nível genômico.

2.3.1 Tradicional

Partindo do pressuposto de que os melhores indivíduos possibilitarão progênies superiores, desenvolveu-se uma das primeiras técnicas de direcionamento de acasalamentos, denominada truncamento por fenótipos, e posteriormente por EBVs (BOURDON, 1995). Resumidamente, esta abordagem consiste no ranqueamento de indivíduos a partir das estimativas dos valores genéticos, e tem como critério de seleção um ponto de truncamento na distribuição dos valores dos indivíduos, como representado na

Figura 2. Sendo possível o referido ponto de truncamento variar de acordo com a intensidade de seleção, e os parâmetros populacionais de interesse (FALCONER; MACKAY, 1996).

Figura 2 – Diferentes distribuições e pontos de truncamento (S – Intensidade de Seleção).



Fonte: Introduction to Quantitative Genetics (FALCONER; MACKAY,1996).

Embora possa apresentar resultados superiores aos métodos de acasalamentos aleatorizados, a técnica de truncamento exclusivamente associada a um parâmetro de interesse (fenótipo ou EBV) não se caracteriza como a melhor estratégia de maximização do ganho genético a longo prazo. Pois, apresenta como aspecto negativo o aumento dos coeficientes de endogamia e, consequentemente, a diminuição da diversidade genética (KINGHORN, 2011b)

A fim de se otimizar o referido processo, diferentes abordagens foram propostas na literatura a respeito de métodos que valorizassem indivíduos tanto pelo alto mérito genético, quanto pelo menor grau de relacionamento (via pedigree) com os demais, otimizando as combinações entre touros e matrizes (ALLAIRE, 1980; FERNÁNDEZ; CABALLERO, 2001; JANSEN; WILTON, 1985; MEUWISSEN, 1996).

Outra forma proposta de bastante relevância a fim de se maximizar a predição do mérito genético das progênies foi denominada por *Mating Selection*, ou “seleção de acasalamentos”. A referida abordagem é composta por dois principais componentes, sendo (i) um índice de seleção de acasalamentos – *mate selection index* (MSI) e (ii) um algoritmo de seleção de acasalamentos, utilizado com a finalidade de se maximizar o índice (i). Sendo possível a partir do índice predizer o mérito genético da progênie a partir da média dos pais, e em alguns casos associar aos ganhos financeiros associados aos referidos acasalamentos, levando sempre em consideração a proporção do mérito genético em relação à endogamia (AKDEMIR; SÁNCHEZ, 2016b).

2.3.2 Nível Genômico

A partir do uso da seleção genômica em larga escala, viabilizou-se o emprego das referidas informações nos modelos de acasalamento tradicionalmente utilizados. Dessa maneira, com o aumento da acurácia do mérito genético dos indivíduos e o acesso mais preciso dos níveis de endogamia (F_{ROH}), graças ao uso de marcadores, observou-se ganhos genéticos ainda mais expressivos. Porém, explicitou-se no decorrer das gerações, que métodos que consideram apenas ganhos genéticos e controle da endogamia são incompletos, uma vez que se ignora a variância dos valores genéticos, não capturando todo o potencial dos possíveis acasalamentos a longo prazo (AKDEMIR; SÁNCHEZ, 2016a).

Embora os níveis de endogamia oriundos da seleção genômica normalmente sejam inferiores quando comparados ao tradicional teste de progênies, a diminuição do intervalo de gerações pode levar a aumentos expressivos nos referidos coeficientes (DAETWYLER et al., 2007; LILLEHAMMER; MEUWISSEN; SONESSON, 2011). Além disso, assim como na seleção tradicional, a seleção genômica tende a fixar regiões próximas a QTLs na população sob seleção, levando ao aparecimento de grandes segmentos em homozigosidade, e consequentemente a diminuição da variabilidade genética e a problemas associados a genes deletérios. Impulsionando, dessa maneira, abordagens que otimizem o processo de seleção e favoreçam a manutenção do progresso genético a longo prazo (COLE, 2015; SONESSON; WOOLLIAMS; MEUWISSEN, 2012).

2.4 Machine Learning

Machine Learning (ML) ou Aprendizado de Máquina é compreendido como um ramo da inteligência artificial que busca o emprego de algoritmos para reconhecimento, classificação e regressão, possibilitando abordagens inovadoras associadas ao processamento de dados (BZDOK; ALTMAN; KRZYWINSKI, 2018). Dentre os algoritmos, destacam-se dois principais grupos, os classificados como supervisionados e os não supervisionados.

Os algoritmos supervisionados visam a obtenção de resultados dado um conjunto de informações de entrada que descrevam as características dos resultados de interesse. Para este fim, o aprendizado supervisionado otimiza o modelo preditivo a partir parâmetros ajustáveis, visando o treinamento prévio com um conjunto de dados de treinamento rotulados (*Labels*) para ajuste. Os referidos dados rotulados

consistem em informações de entrada e de saída (resultados a serem preditos) correspondentes, de forma com que o algoritmo se comporte de acordo com os referidos padrões pré-estabelecidos (VAN DIJK et al., 2021). No caso dos algoritmos não supervisionados, dispensa-se o uso de dados rotulados, tendo em vista que os algoritmos dessa categoria são encarregados na detecção de padrões de forma independente, a partir de associações próprias. Porém, relacionam-se a esses algoritmos certa imprevisibilidade de resultados (XU; JACKSON, 2019b).

Tradicionalmente, os métodos preditivos adotados no melhoramento animal são, em sua maioria, embasados em modelos lineares (HENDERSON, 1975; MEUWISSEN; HAYES; GODDARD, 2001). Considerando as possibilidades associadas aos modelos de *Machine Learning*, estes oferecem novas abordagens para a exploração do conjunto de dados biológicos e respectivos estudos ômicos, que podem favorecer a detecção de padrões não-lineares, principalmente em características com estrutura complexa, baixa herdabilidade e que envolva efeitos não aditivos (NAYERI; SARGOLZAEI; TULPAN, 2019).

Diversos algoritmos de ML foram reportados na literatura quanto à respectiva aplicação em estudos ômicos, sendo amplamente utilizados principalmente como métodos preditivos e de Associação Genômica Ampla (PÉREZ-ENCISO; ZINGARETTI, 2019). Porém, raras menções foram feitas na literatura a respeito da exploração desses algoritmos aplicados ao direcionamento de acasalamentos. Para o presente estudo, buscaremos explorar diferentes classes de algoritmos, uma vez que a proposta se baseia no maior entendimento das *features* (atributos), consideradas até o presente momento. Como ponto de partida, um algoritmo não supervisionado (Análise de componentes principais) e dois supervisionados (K-Nearest Neighbours e Random Forests) foram considerados nesta revisão a fim de se compreender melhor o respectivo funcionamento e a potencial aplicação desses no presente estudo.

2.4.1 Extração e seleção de atributos

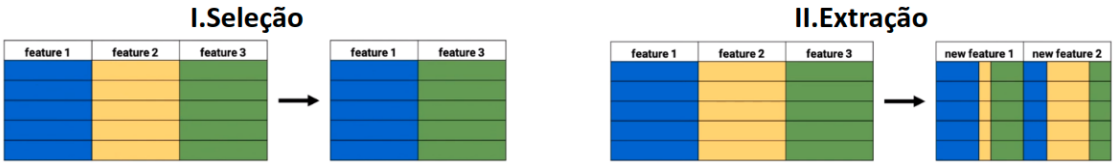
Atributos (*features)* são características intrínsecas ao conjunto de dados sob análise, que podem ser mensuradas de forma que representem um conjunto de informações independentes e passíveis de serem utilizadas na diferenciação entre as variáveis (BLUMA; LANGLEY, 1997). Ao considerarmos a tendência de aumento no

volume de dados referentes à produção animal, especialmente no melhoramento de bovinos leiteiros, além do volume crescente de dados genômicos que exclusivamente apresentam dimensões cada vez mais desafiadoras à eficiência computacional, acrescenta-se à tendência em se considerar e coletar o máximo de informações, incluindo novos fenótipos, muitas vezes associados a diferentes formatos de arquivo de áudio, imagem e vídeo (VENTURA et al., 2020). A vastidão de informações exclusivas para cada indivíduo explicita o entrave chamado de “maldição da dimensionalidade” situação essa caracterizada por ser computacionalmente onerosa (CABESTANY et al., 2005). Como alternativa, dentro do ML, diferentes técnicas, processos e algoritmos foram criados para facilitar a manipulação e extração de informações dos conjuntos de dados, a fim de se reduzir a dimensionalidade das informações em uma etapa de pré-processamento, previamente à implementação do conjunto de dados de *input* nos modelos a serem testados. Dentro de tais processos, destacam-se a extração e a seleção de atributos (*Feature Extraction* e *Selection*) que consistem, de maneira simplificada, na remoção de informações redundantes e irrelevantes. Aumentando dessa maneira a acurácia e a compreensão dos atributos mais importantes a serem consideradas do conjunto de dados sob análise, impactando diretamente a performance do modelo (KHALID, 2014).

Feature Selection, juntamente com a preparação do banco de dados (limpeza de resultados espúrios), se enquadram normalmente como uma das primeiras e mais importantes etapas prévias ao delineamento dos modelos de ML a serem empregados. A referida seleção de atributos é um processo que pode ser realizado tanto manualmente, de acordo com critérios individuais, quanto automaticamente via algoritmos pré-estabelecidos, como por exemplo: *Laplacian Score*, *Fisher Score*, *Gini Index*, entre outros (GINI; C., 1912; HE; CAI; NIYOGI, 2005; ZHAO et al., 2010). Algoritmos estes, que consistem na escolha de um conjunto de dados dos atributos disponíveis em busca apenas das informações mais representativas, como representado no item I.Seleção da Figura 3. O emprego de seleção de atributos normalmente se associa a benefícios relacionados a redução de *overfitting* (termo usual para mencionar que o modelo de ML considerado se ajusta muito ao conjunto de dados de treinamento, apresentando ineficiência ao se predizer novos valores), aumento da acurácia e diminuição no tempo de treinamento do modelo (ZHAO et al., 2010).

No caso da *Feature Extraction*, técnica mais generalista, esta consiste na transformação do conjunto de dados, sendo caracterizado de forma simplificada como um processo de redução da dimensionalidade, como pode ser observado no esquema representado no item II.Extração da Figura 3. Mantém-se no processo, o conjunto de informações mais relevantes do conjunto de dados, porém associado à perda de informações quanto à contribuição original de cada atributo (MOTODA; LIU, 2002).

Figura 3 – Esquema representativo do processo de Seleção e Extração de atributos.



Fonte: Autoria própria

2.4.2 Algoritmos

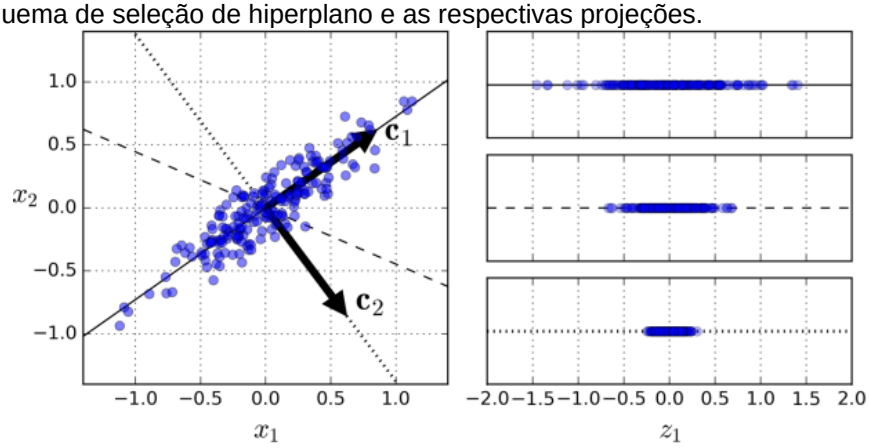
2.4.2.1 Análise de componentes principais

A análise de componentes principais (*Principal Component Analysis* (PCA)), classificada como um algoritmo não supervisionado, tem como principal objetivo a redução da dimensionalidade de um conjunto de dados multivariados. Buscando, dessa maneira, condensar a informação original em componentes estatísticos (componentes principais), a partir da transformação ortogonal dos dados via maximização da variância, associada à perda mínima de informação (JOLLIFE; CADIMA, 2016). Dessa forma, podemos obter maior compreensão de dados estruturalmente complexos, com base nas variâncias e covariâncias existentes entre as variáveis, principalmente em situações em que haja o desequilíbrio das variáveis independentes (X) em relação às observações (Y). Tal metodologia tem sido amplamente usada como parte inicial das análises exploratórias dos dados, especialmente em áreas como a bioinformática (MA; DAI, 2011).

De forma simplificada, os componentes principais (PC) consistem na combinação das variáveis originais após a transformação linear. São determinados ao se solucionar a equação de variâncias e covariâncias em busca dos autovalores (*eigenvalues*) e autovetores (*eigenvectors*) (STEIN et al., 2006). Por meio destes, obtém-se a contribuição de cada componente principal, expresso em porcentagem, em busca da proporção de variância total explicada por cada componente principal. Uma representação simplificada do funcionamento do referido algoritmo pode ser

observada na Figura 4, onde o hiperplano mais próximo do conjunto de dados é identificado, e a projeção que mais se adequa é aquela que preserva o máximo da variância (dados mais dispersos no eixo ortogonal) (BOEHMKE; GREENWELL, 2019).

Figura 4 - Esquema de seleção de hiperplano e as respectivas projeções.



Fonte: Hands-On Machine Learning (BOEHMKE; GREENWELL, 2019).

2.4.2.2 K-Nearest Neighbours

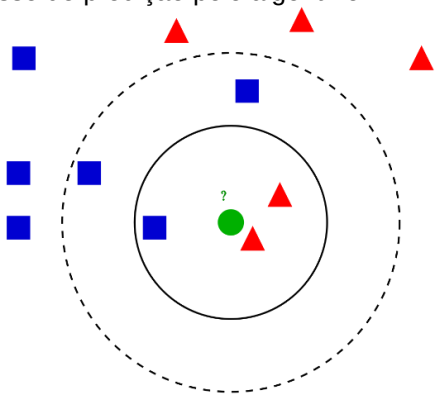
K-Nearest Neighbours (KNN) é um algoritmo não-paramétrico, considerado da classe de algoritmos supervisionados e de estrutura pouco complexa, apresentando resultados bastante satisfatórios (tanto na classificação como para fins de regressão) quando comparado a algoritmos caracterizados por estruturas mais robustas tanto em relação a valores de acurácia, quanto performance computacional (WU et al., 2008). O KNN apresenta como principal diferencial a não necessidade de treinamento de um conjunto de informações prévias, pois as predições de quaisquer novas instâncias são feitas a partir do conjunto de dados completo (*lazy learning algorithms*). Além disso, nenhuma pressuposição prévia a respeito da distribuição dos dados se faz necessária ao se empregar este algoritmo. O que pode apresentar tanto aspectos positivos, quanto negativos, uma vez que em todos os processos preditivos, exige-se o processamento do conjunto de informações por completo (WETTSCHERECK; AHA; MOHRI, 1997).

A lógica de funcionamento do KNN é embasada no pressuposto de que "ocorrências similares ocorrem em proximidade", justificando, portanto, o nome atribuído a este algoritmo ("K - vizinho mais próximo"). Este método é embasado em dois principais parâmetros, sendo o primeiro associado à função de medidas de distância a ser considerada para predição, e o segundo relacionado ao número de elementos vizinhos (K - "*Neighbours*") da instância a ser predita. No referido processo,

a função de distâncias é adotada de acordo com a métrica determinada previamente para “encontrar” as instâncias previamente rotuladas no conjunto de dados de teste, de maior proximidade, ou seja, os K-Nearest Neighbours. As funções de distâncias mais comumente utilizadas são: as distâncias Euclidiana, Manhattan e Minkowski (PRASATH et al., 2017). O número K pode variar de acordo com o critério adotado, uma vez que podem ser diferentes as considerações quanto ao “número ideal”.

Portanto, a partir da maioria das instâncias mais próximas e com o mesmo rótulo no conjunto teste, de acordo com K, o algoritmo assume o novo rótulo à instância a ser predita. Assumindo, a fim de se ilustrar o funcionamento do algoritmo, o exemplo da Figura 5, e se baseando no funcionamento do algoritmo anteriormente citado, inferimos que a classificação a ser adotada pela instância em questão seria a classe “vermelho”. Processo que exemplifica as etapas lógicas do algoritmo, situação que se repete ao longo de todos os pontos a serem classificados em quaisquer conjuntos de dados.

Figura 5 – Esquema do processo de predição pelo algoritmo K-NN.



Fonte: https://cdn.analyticsvidhya.com/wp-content/uploads/2018/03/knn3-300x271.png.

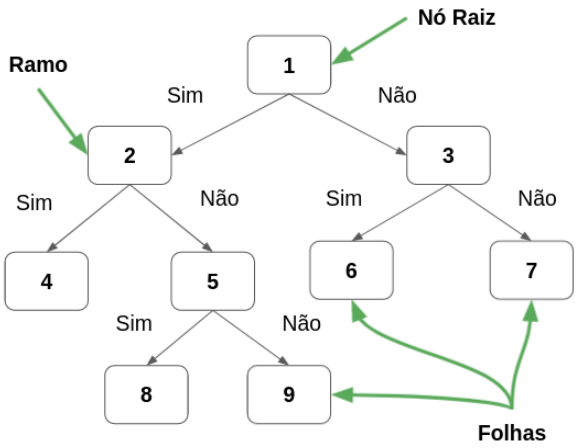
2.4.2.3 Random Forests

Random Forests (RF) é um algoritmo normalmente supervisionado e não-paramétrico, sendo considerado um dos algoritmos mais populares e versáteis, tanto como método de classificação, quanto regressão (ENGLUND; VERIKAS, 2012). A versatilidade e relevância deste algoritmo são justificados pela estrutura de *ensemble*, caracterizada pela possibilidade de o algoritmo estruturar e associar outros diversos algoritmos simultaneamente, a fim de se aumentar a performance e capacidade preditiva do mesmo. Dessa forma, podemos ressaltar que o RF é um algoritmo composto, pois implementa em sua base a funcionalidade de outros algoritmos simultaneamente.

Como exemplo, podemos citar as árvores de decisão (*Decision Tree),* implementadas massivamente dentro do escopo de atuação do RF.

Decision Tree é um algoritmo que atua como classificador e regressor, sendo baseado em modelos sequenciais que combinam uma sequência de árvores de decisões (estrutura semelhante a um fluxograma). Cada árvore de decisão é constituída por um nó raiz, ramos e folhas (SOMVANSHI et al., 2017), como pode ser observado na representação da Figura 6. Outro parâmetro que se destaca é referente ao estabelecimento da métrica responsável pela hierarquia da estrutura das árvores de decisão, ou seja, a forma com que o fluxograma será ordenado. Dentro de tais métricas, destacam-se a Entropia, Ganho de informação, Erro de Classificação, e principalmente o Índice Gini, métrica bastante utilizada como *Feature Selection* e caracterizada por ser computacionalmente eficiente (SONG; LU, 2015).

Figura 6 – Estrutura geral de árvore e os respectivos elementos presentes no algoritmo Decision Tree e Random Forests.



Fonte: autoria própria

Portanto, a menor estrutura possível que se estabelece no algoritmo *Random Forests* é composta de ao menos uma *Decision Tree*, podendo, via método *ensemble,* ser composta por inúmeros modelos destes simultaneamente (BREIMAN, 2001). Vale ressaltar que os resultados obtidos via *Random Forests* são oriundos da média de valores obtidos em cada *Decision Tree* (para estudos de regressão) ou da moda das classes (para problemas de classificação).

No que diz respeito aos principais hiper parâmetros a serem considerados no algoritmo RF, destacam-se como os mais importantes: (i) número de estimadores a serem considerados (número de árvores na floresta), diretamente associados a

eficiência computacional do algoritmo, (ii) a profundidade das árvores (*depth*) associado a captura de informação, sendo o ponto mais profundo as folhas, e (iii) o número de amostras aleatoriamente selecionadas em cada etapa. Outra característica importante a ser mencionada a respeito do algoritmo é o emprego do *bootstrap aggregating* (*Bagging),* caracterizado pela seleção aleatória de observações do conjunto de dados de treinamento com reposição para cada modelo considerado (BREIMAN, 1996). Tal medida, aumenta a performance do modelo, uma vez que via reposição, a variância é diminuída sem que haja aumento do viés, proporcionando menor *overfitting* e aumento da acurácia.

2.4.3 Aplicação na genômica

O uso de dados genômico pode, muitas vezes, ser dificultado pelo “excesso” de variáveis (X) em relação às observações (Y) disponíveis em determinadas situações, fenômeno nomeado de “maldição da dimensionalidade” (CHEN, 2009). Considerando o influxo de informações oriundas de material genético (genótipos, informações metabólicas, expressão gênica), e de novos fenótipos (associados a dados oriundos de sensores), a capacidade de capturar o efeito de relações não lineares ocultas pode ser comprometida pelo uso de modelos lineares. Portanto, modelos não-lineares de *Machine Learning* se apresentam como possíveis alternativas (NAYERI; SARGOLZAEI; TULPAN, 2019).

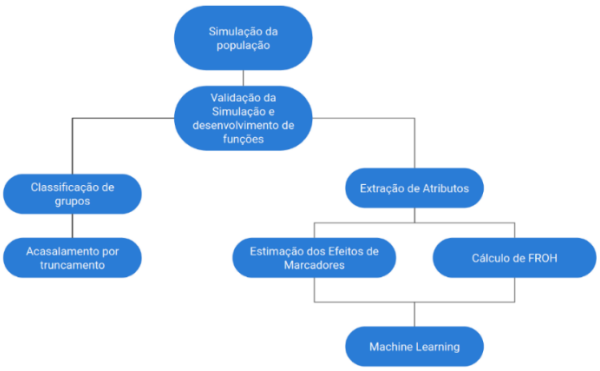
Como reportado em estudos prévios, algoritmos de *Machine Learning* podem proporcionar vantagens quando comparados aos métodos tradicionais relacionados ao emprego e exploração de dados genômicos em associação com fenótipos não convencionais (PÉREZ-ENCISO; ZINGARETTI, 2019; XU; JACKSON, 2019a). Dentre às abordagens empregando técnicas de ML, destacam-se principalmente os estudos acerca de características de baixa herdabilidade, como problemas relacionados a sanidade de rebanhos (HIDALGO et al., 2018), fertilidade (SHAHINFAR et al., 2014), nutrição, produção (GONZÁLEZ-RECIO; JIMÉNEZ-MONTERO; ALENDA, 2013) e seleção (LI et al., 2018).

3. MATERIAIS E MÉTODOS

O desenvolvimento do projeto envolveu diversas etapas, que para maior compreensão, estão dispostas na Figura 7 de acordo com a ordem cronológica de

execução. As estratégias adotadas e as funções desenvolvidas durante este estudo se encontram disponibilizadas com maior nível de detalhamento nos tópicos abaixo.

Figura 7 – Esquema das principais etapas desenvolvidas neste estudo.



Fonte: Autoria própria.

De forma resumida, os principais passos contaram com a simulação de uma população com parâmetros populacionais associados a um rebanho leiteiro sob seleção. Posteriormente, estratégias para validação dos resultados da simulação foram desenvolvidas, a fim de se atestar a empregabilidade dos dados. De maneira geral, duas principais etapas foram desenvolvidas posteriormente, sendo essas associadas a classificação dos dados a partir de valores tradicionalmente obtidos, como coeficiente de endogamia via pedigree, e valores fenotípicos e de EBVs. E a outra relacionada a extração de atributos oriundos das informações genômicas, levando em consideração valores da estimativa dos efeitos de marcadores e do cálculo de F_{ROH}, estes usados como rótulos nas abordagens supervisionadas.

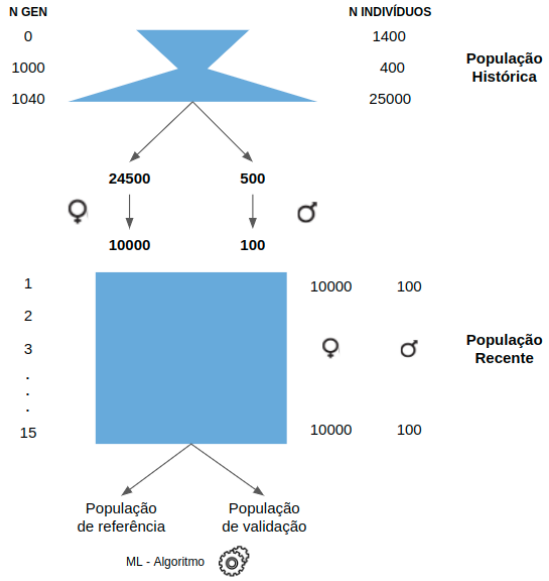
A fim de se validar os dados, construir modelos e extrair atributos, dois softwares em linguagem Python foram desenvolvidos. O primeiro (GASP3) foi empregado na extração de atributos e aplicação de modelos de *machine learning*, enquanto o segundo (PyBioma) teve como finalidade a simulação de acasalamentos entre indivíduos de interesse.

3.1 Simulação

Em busca de se mimetizar as características de uma população de bovinos de leite, optou-se como modelo, uma estrutura de simulação já reconhecida em periódicos internacionais, com alterações pontuais, conforme reportado (DE OLIVEIRA et al., 2019). Para maior compreensão, ilustramos na Figura 8, o esquema da estrutura de simulação adotada, e os respectivos parâmetros na Tabela 1.

De forma sucinta, o processo seguiu os passos corriqueiramente encontrados em rotinas de simulação, sendo o QMSim (SARGOLZAEI; SCHENKEL, 2009) o software escolhido, caracterizado por duas etapas distintas: a formação da população histórica e a formação da população recente. Na primeira etapa, foi considerada uma população inicial de 1400 indivíduos, sendo proporcional o número de machos e fêmeas. Estes, acasalados entre si por mil gerações, de forma com que o número total de indivíduos decrescesse à 400 ao todo. Posteriormente, expandiu-se a população para 25000 indivíduos, em apenas 40 gerações, buscando dessa maneira, mimetizar o processo de deriva genética no processo evolutivo da espécie. Ressalta-se que na última geração da população histórica o número total de machos foi fixado em 500, enquanto o de fêmeas foi de 24500. Na segunda etapa, para criação da população recente, 10000 fêmeas e 100 machos foram selecionados da última geração da população histórica, assumindo como critério de escolha indivíduos de maior valor genético.

Figura 8 - Esquema de Simulação.



Fonte: Autoria própria.

Quanto à população recente, 15 gerações foram consideradas a taxas constantes de reposição de 20% e 60% para fêmeas e machos, respectivamente. Sendo considerados fixos os números de progênie por acasalamento (n=1) e o sexo das progênies seguindo o percentual de 50% para machos e fêmeas. Logo, o número de filhos por geração foi fixo de 10000, sendo usados os valores de EBVs, embasados no método de Henderson (1975) (já implementado no software QMSim) como critério de seleção. Onde os indivíduos selecionados foram escolhidos pelo maior EBV, e os

descartados de com o menor EBV. No que se diz respeito ao design adotado de acasalamento, este foi considerado como aleatório entre os indivíduos selecionados, buscando mimetizar a situação real, onde os melhores acasalamentos não são contemplados em rebanhos comerciais leiteiros. Pois, muitas vezes, a decisão final fica com o produtor e seus critérios pessoais.

Quadro 1 - Parâmetros gerais da simulação.

População Histórica	
N.º Total de gerações	1040
N.º Machos na última geração	500
N.º Fêmeas na última geração	24500
População Recente	
N.º de gerações	15
N.º Machos	100
N.º Fêmeas	10000
N.º de progênies por matriz	1
Proporção de sexo nas progênies	50%
Design de acasalamento	Aleatório
Taxa de Reposição de Machos	60%
Taxa de Reposição de Fêmeas	20%
Critério de Seleção e Descarte	EBV (BLUP)
Herdabilidade	0,30
Variância Fenotípica	100
Genoma	
N.º de Cromossomos	29
N.º de Marcadores	57024
Distribuição de Marcadores	Equidistantes
Número de QTLs	1979
Distribuição de QTLs	Equidistantes

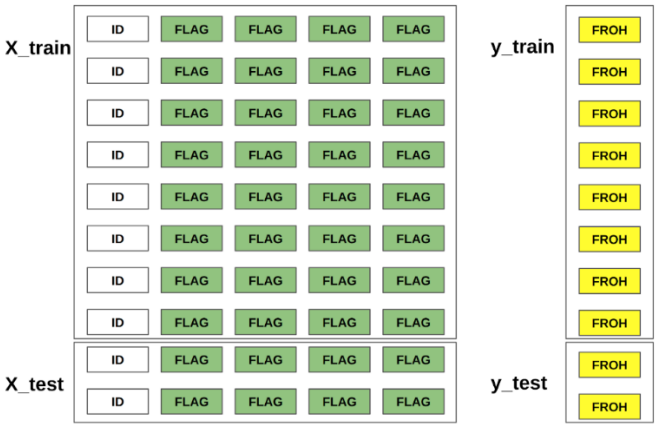
Em relação aos genótipos, todos os indivíduos a partir da sexta geração foram considerados, onde cada um dos 29 cromossomos autossômicos foi ajustado de acordo com o tamanho reportado pelo *National Center for Biotechnology Information website* (NCBI, https://www.ncbi.nlm.nih.gov/genome/?term=cow). Contendo número variável de SNPs por cromossomo, sendo considerados ao todo 57024 SNPs, representando desta forma um chip de genotipagem comercial de média densidade. O número de QTLs reportados para a produção de leite também foi ajustado por

cromossomo de acordo com o *Animal QTLDatabase* (https://www.animalgenome.org/cgi-bin/QTLdb/BT/traitmap?trait_ID=1044). Sendo considerados ao todo 1,979 QTLs dispersos ao longo dos cromossomos autossômicos. No que diz respeito à taxa de mutação, tanto para SNPs, quanto QTLs, considerou-se a taxa fixa e recorrente de 1 × 10-4, e apenas um cenário de herdabilidade (h2=0.3). Sendo descartada a ocorrência de erros de genotipagem ao longo da simulação.

3.2 Método geral para treinamento e teste dos modelos

Como qualquer outro tipo de análise de regressão ou classificação, faz-se também necessário, ao se conduzir estudos que envolvam o uso de *Machine Learning*, a divisão dos dados em conjunto de treinamento e de teste, a fim de se avaliar o desempenho dos modelos empregados. Ao longo deste estudo, foi usada a biblioteca Python Scikit-Learn (PEDREGOSA et al., 2011), tanto para o preparo dos dados, quanto para a aplicação dos algoritmos. No que diz respeito a divisão do conjunto de dados para treinamento e teste, foi utilizada a função *split_train_test_dataset*, que randomiza os dados a serem treinados e testados, diminuindo o viés na criação dos grupos, como representado na Figura 9. Conforme reportado na literatura, uma taxa de 20% do conjunto de dados foi considerada para teste do modelo (SALAZAR et al., 2022), assim como para a obtenção das métricas de avaliação de desempenho de erro quadrado médio e r-quadrado, respectivamente representados por MSE (mean squared error) e r^2.

Figura 9 - Representação do funcionamento da divisão do conjunto de dados em teste e treinamento para os modelos de Machine Learning.



Fonte: Autoria própria.

A fim de se uniformizar os modelos e os cenários considerados, um conjunto total de 10.000 genótipos aleatoriamente escolhidos entre as gerações seis e doze da simulação foi gerado para treinamento e teste dos modelos. Os indivíduos das gerações de 12 a 15 foram reservados para posteriores testes e validações dos acasalamentos propostos.

Dentre os métodos considerados, podemos enumerar as abordagens em quatro principais grupos, sendo estes referentes ao (I) coeficiente de endogamia genômico (F_{ROH}), (II) a estimação do valor fenotípico, (III) o valor genômico estimado (GEBV), (IV) a estimação do valor genômico real. Além disso, em um quinto tópico, considerou-se o desenvolvimento de um método para validação e avaliação de potenciais acasalamentos entre os indivíduos. Tais cenários foram conjecturados a fim de se avaliar a empregabilidade de diferentes atributos extraídos de dados genômicos brutos em comparação a métodos já consolidados, como forma de se avaliar o desempenho dos algoritmos de *Machine Learning*.

3.2 Cenários considerados

3.2.1 Coeficiente de endogamia genômico

No que diz respeito aos coeficientes de endogamia, diferentes abordagens foram consideradas a fim de se avaliar os impactos preditivos de acordo com os atributos extraídos. Uma vez que existem poucas referências bibliográficas acerca da estimação do coeficiente de endogamia em dados genômicos empregando algoritmos de Machine Learning, a fim de se avaliar o desempenho dessa tecnologia, adotou-se como alvo para os referidos modelos, valores de endogamia estimados por métodos já consolidados.

Os valores de ROH dos genótipos foram calculados via software PLINK (PURCELL et al., 2007), sendo considerados os seguintes parâmetros: (i) janela de detecção de 50 SNPs, (ii) tamanho mínimo de ROH igual a 50 SNPs e 1.000 kb, (iii) máxima lacuna entre SNPs dentro da ROH igual a 1.000 kb, (iv) densidade mínima de 1 SNP a cada 100 kb, (v) máximo de 1 SNP heterozigoto na ROH e janela. Posteriormente, utilizou-se o *software* DetectRuns (BISCARINI, F. et al., 2018) para o cálculo do Coeficiente de Endogamia Genômico (F_{ROH}) em todos os genótipos obtidos no processo de simulação, como reportado na fórmula citada previamente.

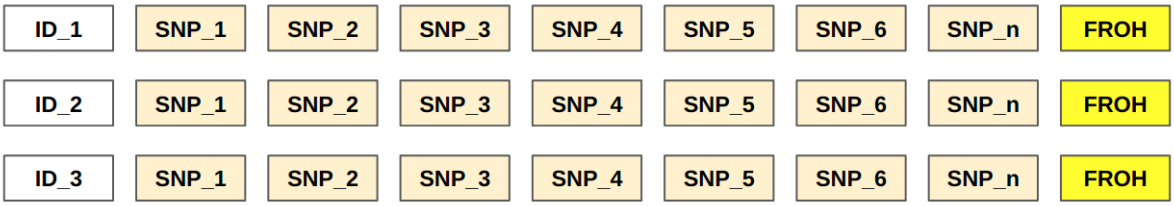
Foi desenvolvido o software GASP3, em *Python,* com a finalidade de se explorar diferentes estratégias de extração de atributos e manipulação de genótipos.

Seis abordagens foram propostas, sendo duas baseadas nas informações dos genótipos dos próprios indivíduos, e três a partir dos genótipos dos pais, como forma de se avaliar possíveis acasalamentos. Seguem os esquemas e maiores informações a respeito de cada abordagem.

3.2.1.1 Método contendo os genótipos íntegros dos indivíduos

O primeiro cenário de extração de atributos se refere a estimação dos coeficientes de endogamia a partir da totalidade de marcadores contidos nos genótipos. Como representado no esquema da Figura 10, os marcadores (SNP) dos indivíduos de interesse (ID) foram importados individualmente e em sua totalidade, tendo como alvo o valor de F_{ROH}, previamente calculado via uso dos *softwares PLINK* e *DetectRuns,* a fim de se avaliar o desempenho preditivo do referido modelo empregado.

Figura 10 – Esquema de entrada de dados para o primeiro método de predição de F_{ROH} proposto.

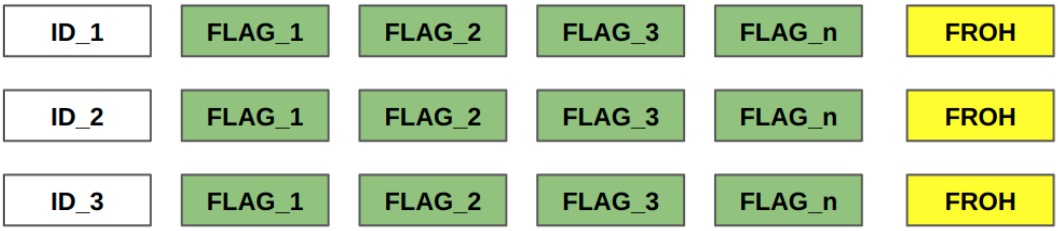


Fonte: Autoria própria.

3.2.1.2 Método contemplando a extração de atributos dos genótipos

O segundo cenário, representado na Figura 11, trata-se de uma entrada de dados a partir da extração de atributos das informações genotípicas referentes a cada indivíduo. Neste modelo, os genótipos de interesse foram submetidos a uma segmentação em blocos a fim de se diminuir o número de variáveis inseridas nos algoritmos, reduzindo a dimensionalidade dos dados. Para tanto, foi considerado o número fixo de 50 marcadores por segmento, avaliando-se o percentual de homozigotos, e posteriormente atribuindo uma informação numérica binária (0 ou 1), a fim de se reduzir a dimensionalidade das informações por bloco.

Figura 11 - Esquema de entrada de dados para o segundo método de predição de F_{ROH} proposto.



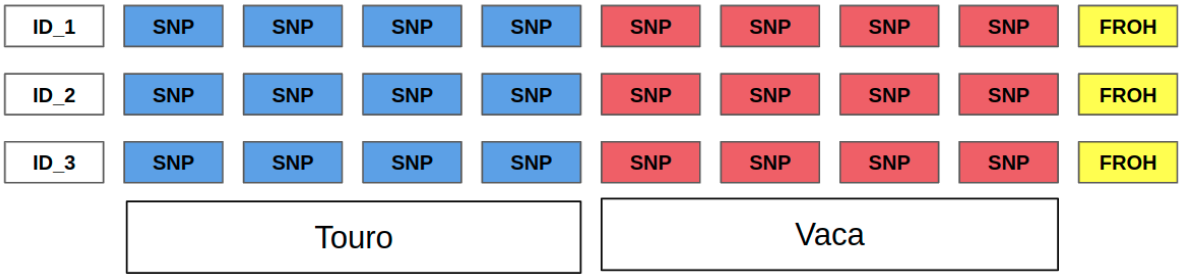
Fonte: Autoria própria.

Nos segmentos contendo mais de 96% dos marcadores homozigotos, atribuiu-se o número um, sendo a janela identificada como “Flag de ROH”, nos demais segmentos empregou-se o número zero, seguindo o referido padrão ao longo de todos os genótipos. Reduzindo, dessa maneira, o número de variáveis em até 98%, quando comparado aos genótipos integralmente considerados.

3.2.1.3 Método considerando os genótipos integralmente dos pais dos indivíduos

No terceiro cenário proposto, sendo este o primeiro com a finalidade de se predizer o coeficiente de endogamia dos filhos a partir da informação genotípica dos pais, considerou-se os genótipos completos tanto do touro, quanto da vaca agrupados sequencialmente, como representado na Figura 12**.** Assim como no método contendo os genótipos íntegros dos indivíduos apresentado anteriormente, a medida adotada empregou cada SNP com uma variável independente, tendo como alvo principal de treinamento e predição o F_{ROH} calculado previamente a partir do genótipo dos indivíduos de interesse.

Figura 12 - Esquema de entrada de dados para o terceiro método de predição de F_{ROH} proposto.

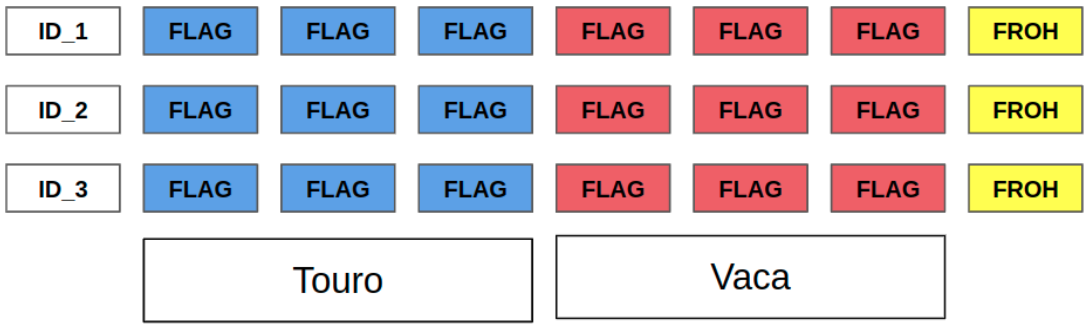


Fonte: Autoria própria.

3.2.1.4 Método contemplando os atributos extraídos dos genótipos dos pais dos indivíduos de interesse

No quarto cenário, levou-se em consideração a mesma abordagem de extração de atributos como apresentado anteriormente no item 3.2.1.2. Mas, desta vez, considerando as informações oriundas dos genótipos dos pais dos referidos indivíduos a serem avaliados. O mesmo método de divisão de genótipos em blocos de 50 marcadores foi considerado, tanto para o touro, quanto para a vaca, empregando os indicadores binários (0 e 1), baseados nos mesmos parâmetros descritos anteriormente (>96% homozigotos, 1, <96%, 0). Diminuindo dessa maneira a dimensão dos dados avaliados em 98% em relação ao cenário do item 3.2.1.3. Cabe ressaltar que as informações foram dispostas de maneira completa para o touro, e posteriormente para a vaca, como representado na Figura 13.

Figura 13 - Esquema de entrada de dados para o terceiro método de predição de F_{ROH} proposto.



Fonte: Autoria própria.

3.2.1.5 Método contemplando os atributos extraídos dos genótipos dos pais dos indivíduos de interesse com alternância de posição

No quinto cenário, o procedimento de extração dos atributos foi exatamente igual ao método apresentado anteriormente no item 3.2.1.4, mas com alterações na disposição dos atributos. Neste caso, os atributos (FLAGs) extraídos dos pais do indivíduo sob análise foram dispostos de forma alternada, ou seja, informações paternas intercaladas por informações maternas, como representado na Figura 14.

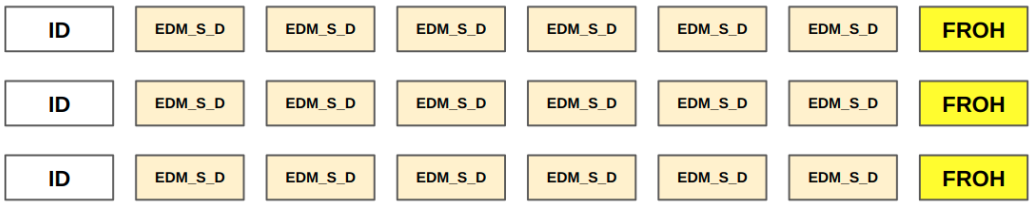
Figura 14 - Esquema de entrada de dados para o quarto método de predição de F_{ROH} proposto.

Fonte: Autoria própria.

3.2.1.6 Método empregando a distância Euclidiana entre genótipos dos pais dos indivíduos de interesse

No sexto e último cenário de atributos para a estimação de F_{ROH}, considerou-se mais uma vez os genótipos dos pais dos indivíduos, tendo como alvo os coeficientes de endogamia previamente calculados via métodos tradicionais para cada genótipo. Da mesma forma abordada anteriormente, os genótipos dos touros e das vacas foram segmentados em blocos contendo 50 marcadores cada, sendo posteriormente calculada a distância euclidiana entre os segmentos a partir da codificação genômica presente com a respectiva numeração (012), gerando um conjunto de atributos como representado na Figura 15**.**

Figura 15 - Esquema de entrada de dados para o quinto método de predição de F_{ROH} proposto.



Fonte: Autoria própria.

3.2.2 Valores fenotípico e genético estimados

3.2.2.1 Método para estimação do valor fenotípico

Como forma de se avaliar o potencial preditivo dos fenótipos simulados no QMSim a partir do conteúdo genotípico, considerou-se as informações fenotípicas exclusivas das fêmeas, uma vez que o esquema de simulação contemplou um rebanho leiteiro. Dessa maneira, apenas os genótipos das fêmeas foram filtrados contemplando uma escolha aleatória de 10 mil indivíduos entre as gerações 6 e 12, a fim de se treinar e testar os modelos de interesse, e posteriormente avaliar as gerações posteriores e respectivos acasalamentos. Por se tratar de dados oriundos

de uma simulação, não foram necessários ajustes nos dados fenotípicos, e estes foram empregados diretamente como alvos a serem treinados e posteriormente preditos. De maneira geral, a entrada de dados nos modelos foi baseada no esquema apresentado na Figura 16, sendo os marcadores importados individualmente com os respectivos fenótipos.

Figura 16 - Esquema de entrada de dados para o método de predição de valor fenotípico proposto.

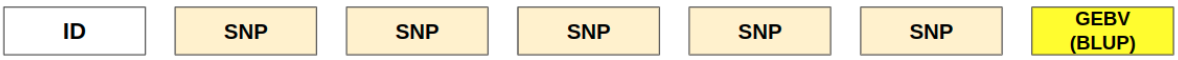
Fonte: Autoria própria.

3.2.2.2 Método para estimação do valor genômico estimado

Em relação ao cenário de estimativa de valores genômicos, buscou-se apenas aferir a viabilidade de uso das abordagens propostas e os respectivos atributos em algoritmos de aprendizado de máquina, não tendo como objetivo propor novos métodos de predição. Considerou-se uma abordagem semelhante ao que foi desenvolvido e apresentado anteriormente, mas com a intenção de se obter o valor genômico dos indivíduos. Como passo prévio a estimação dos valores genômicos (GEBVs) e os respectivos efeitos de SNPs, empregou-se a família de softwares BLUPF90 (MISZTAL; TSURUTA, 2015) para o referido cálculo. De maneira resumida, o modelo assumiu como efeitos fixos a geração dos indivíduos genotipados e os marcadores como efeitos aleatórios, considerando 25000 indivíduos aleatórios, da sexta à décima segunda geração para estimação dos efeitos de marcadores e dos respectivos valores genéticos. Destaca-se que o parentesco via pedigree foi desconsiderado, conservando apenas o parentesco genômico, a fim de se evitar possíveis interferências oriundas de um processo de *Single Step* (AGUILAR et al., 2010).

Posterior a estimação dos efeitos, multiplicou-se os marcadores dos indivíduos de interesse pelos respectivos efeitos estimados para obtenção do somatório dos efeitos, considerado aqui como o valor genômico estimado, sendo esta informação tratada como o alvo dos modelos de *Machine Learning* a serem empregados. Quanto as demais informações, considerou-se os SNPs individualmente como dado de entrada, assim como representado na Figura 17**.**

Figura 17 - Esquema de entrada de dados para o método de predição de valor genômico predito.

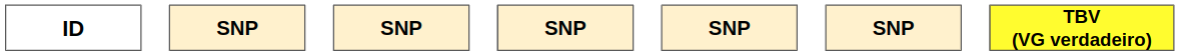


Fonte: Autoria própria.

3.2.2.3 Método para estimação do valor genômico verdadeiro (TBV)

O referido cenário seguiu a mesma estrutura das demais propostas apresentadas nos itens 3.2.2.1 e 3.2.2.2, porém com a variação relacionada ao alvo de treinamento e predição, sendo neste caso, o valor genômico verdadeiro (*True Breeding Value* – TBV), como representado no esquema da Figura 18. Cabe ressaltar que a referida informação só é possível em estudos de simulação, uma vez que não se tem acesso ao valor genético real dos indivíduos corriqueiramente, mesmo em situações que envolvam altas acurácias. Portanto, a abordagem foi considerada e adotada como referência em comparação com as demais cenários

Figura 18 - Esquema de entrada de dados para o método de predição de valor genômico verdadeiro.

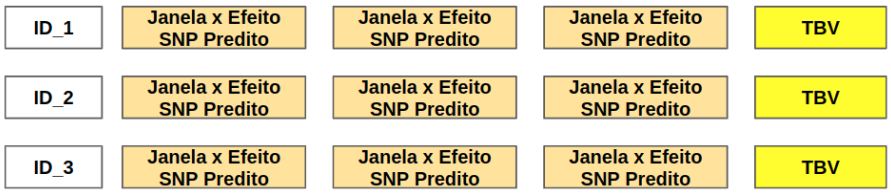


Fonte: Autoria própria.

3.2.2.4 Método de estimação do valor genômico verdadeiro (TBV) via efeitos de SNP estimados

Como forma de se avaliar a empregabilidade dos efeitos estimados dos SNPs via GBLUP, por meio da família de softwares BLUPF90 com o mesmo cenário abordado no item 3.2.2.2 (MISZTAL; TSURUTA, 2015), para estimação do TBV, considerou-se o cenário representado pela Figura 19. Posterior a estimação dos efeitos, multiplicou-se os marcadores dos indivíduos de interesse pelos respectivos valores estimados, agrupando os valores obtidos em janelas de 50 marcadores. Posteriormente, considerou-se como alvo do modelo a ser treinado os valores de TBV obtidos via simulação.

Figura 19 - Esquema de efeito de SNP para predição de valor genômico verdadeiro.



Fonte: Autoria própria.

3.2.2.5 Predição otimizando-se F_{ROH} e valor genético

A fim de se propor um cenário de otimização, avaliando a capacidade preditiva dos algoritmos quanto a classificação dos melhores indivíduos nos aspectos de baixa endogamia e alto valor genético, considerou-se o cenário descrito na Figura 20. Para tanto, foi criada uma função (*create_class)* no *software* GASP3 a fim de se classificar os indivíduos da população simulada a partir de limites (*thresholds*) populacionais de interesse. Empregou-se, de maneira geral, um rótulo binário (0/1) para se referir aos indivíduos, de forma que fossem contemplados tanto os aspectos de baixa endogamia, quanto o de alto valor genético. A rotulação foi realizada pela função mencionada anteriormente de acordo com os seguintes passos:

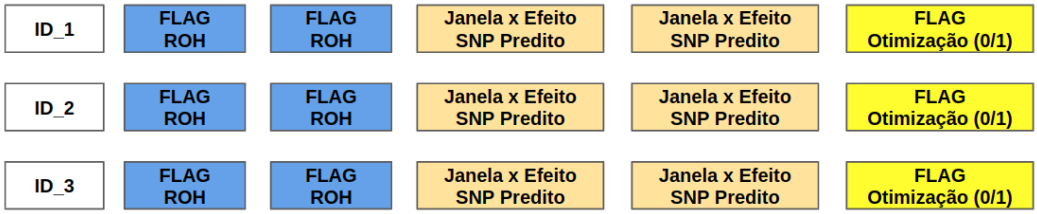
1. Indicação do *threshold* de TBV, assumindo que todos os indivíduos de valor genético acima de 75% do grupo de interesse recebessem o rótulo “1”, e os demais, “0”.

2. Indicação do *threshold* de F_{ROH}, assumindo que todos os indivíduos com o coeficiente de endogamia acima de 50% da população de interesse recebessem o rótulo “1”, e os demais, “0”.

3. Criação da rotulação binária final de acordo com os resultados obtidos nos passos anteriores, de forma que os indivíduos considerados de alvo estivessem dentro das condições “alto valor genético” (“1”) e baixo coeficiente de endogamia (“0”). Proporcionando dessa maneira a criação da rotulação final, sendo “1” para os indivíduos alvo e “0” para os indivíduos indesejáveis.

Dessa maneira, o modelo considerado foi treinado com as informações relacionadas aos atributos considerados no que diz respeito às informações de endogamia (FLAG ROH) e de Efeito de marcadores (“Janela x Efeito SNP Predito”).

Figura 20 - Esquema de entrada de dados para o método de predição de rótulos.

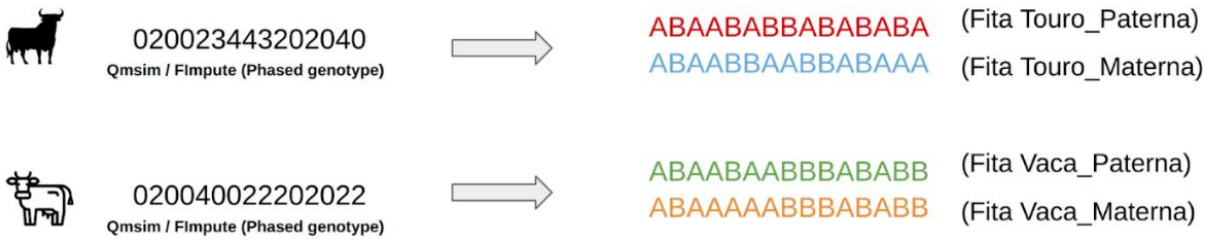


Fonte: Autoria própria.

3.2.3 Acasalamento genômico

O *software* Pybioma foi desenvolvido em linguagem Python com a finalidade de ser um simulador de acasalamentos entre indivíduos de interesse. De forma sucinta, o programa não contemplou o mérito da segregação mendeliana, objetivando apenas estimar as combinações possíveis entre as fitas maternas e paternas integralmente, a partir do conhecimento da origem de cada segmento (genótipos faseados) como representado na Figura 21. Sendo, as referidas fitas, posteriormente associadas de acordo com todas as combinações possíveis entre as informações oriundas do touro quanto da vaca, conforme o esquema de combinação demonstrado no Figura 22.

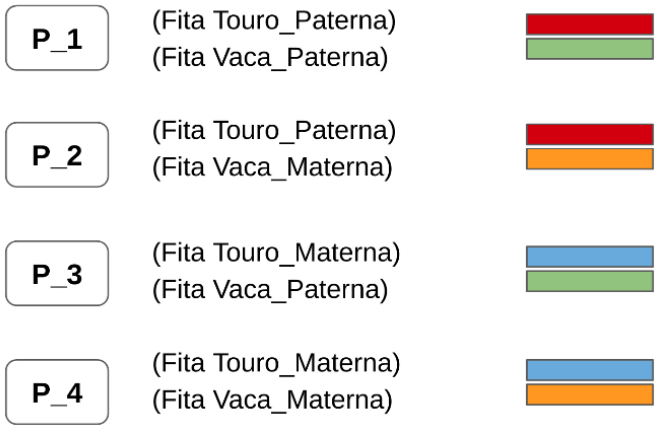
Figura 21 – Processo de divisão das fitas dos genótipos após faseamento, ou oriundo da saída nativa do QMSim.



Fonte: Autoria própria.

O esquema empregado na simulação dos acasalamentos, considerou todas as associações possíveis entre as fitas, tanto do touro, quanto da vaca, gerando como arquivo de saída quatro possíveis combinações (*outputs*), nomeadas neste estudo como “Progênies Fake”.

Figura 22 – Representação da saída de dados do *software* pybioma.



Fonte: Autoria própria.

Os arquivos de entrada do programa Pybioma consistiu em três principais arquivos, sendo esses o pedigree, os genótipos faseados dos touros e matrizes e o arquivo de mapa dos genótipos. Cabe mencionar o formato dos referidos arquivos idênticos originados no *software* de simulação QMSim, conforme apresentados na Figura 23.

Figura 23 - Trecho dos formatos de arquivo de entrada utilizados pelo software Pybioma.

Pedigree

```
Progeny Sire Dam
1 0 0
2 0 0
3 0 0
4 0 0
5 1 2
6 3 4
7 5 6
8 5 6
9 5 6
10 5 6
```

Genótipo

```
ID Call
1 0000000000000000000
2 0000000000000022222
3 1112222222222222222
4 2222222222222222222
5 0000000000000033333
6 2222222222222222222
7 0004444444444422222
8 4444444444444442222
9 4444444444444444444
10 0004444444444444222
```

Mapa

```
ID Chr Pos
M1 1 42252
M2 1 48050
M3 1 34347
M4 1 60503
M5 1 75230
M6 1 75546
M7 1 97842
M8 1 106857
M9 1 121280
M10 1 128404
M11 2 130738
M12 2 133684
```

Fonte: Autoria própria.

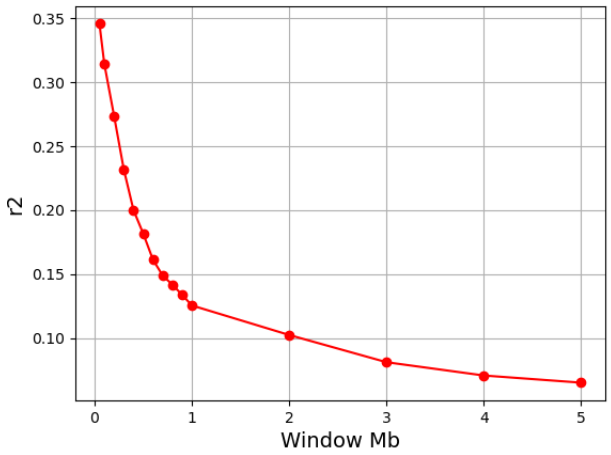
4. RESULTADOS E DISCUSSÃO

4.1. Validação da simulação

Embora o esquema de simulação adotado no corrente estudo tenha sido empregado em publicações prévias, desenvolveu-se aqui uma série de funções de validação no *software* GASP3, em linguagem Python. A partir dos arquivos de saída do simulador QMSim, empregou-se abordagens visuais para compreensão dos dados. Sendo a primeira informação validada representada na Figura 24, referente ao

decaimento médio do desequilíbrio de ligação (R2) em relação a distância em Megabases (Mb) da população simulada recente. O decaimento de LD é a principal metodologia empregada em estudos que envolvam a simulação de populações com dados genômicos, e de acordo com o padrão de decaimento aqui apresentado, foi possível validar os resultados de acordo com informações reais de populações de bovinos leiteiros Holstein descritas na literatura (MARQUES et al., 2008; SARGOLZAEI et al., 2008).

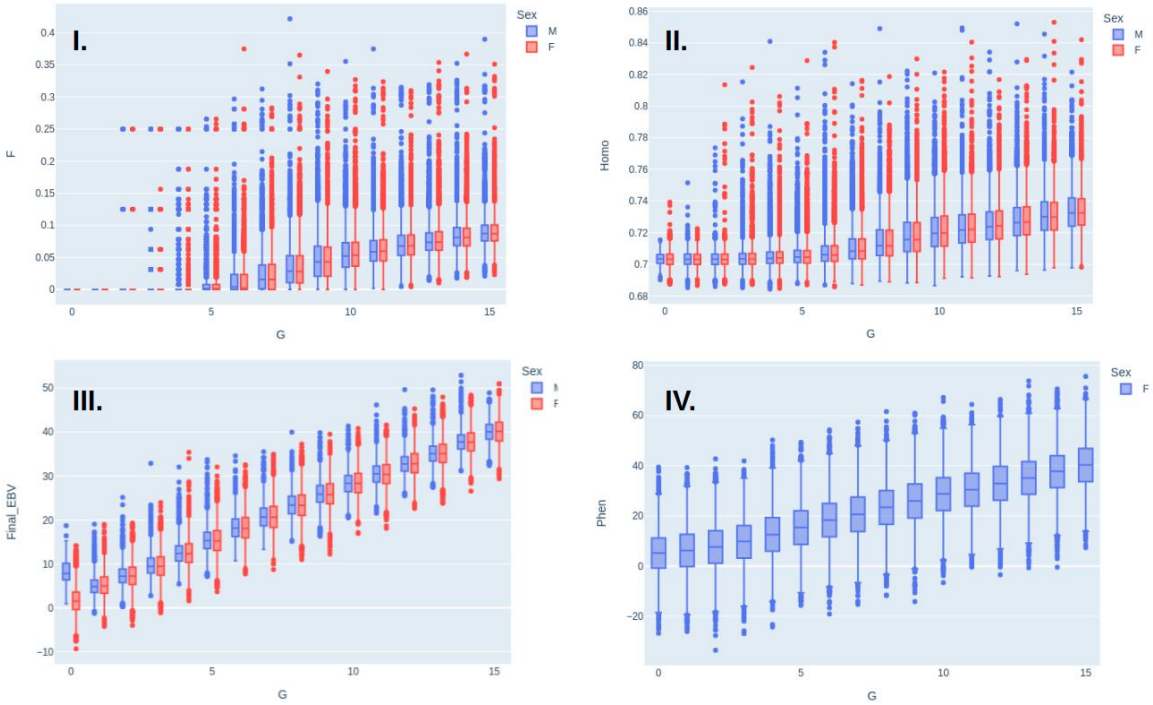
Figura 24 – Decaimento médio do LD da população simulada.



Fonte: Autoria própria.

Após a validação das informações referentes ao LD, buscou-se compreender o comportamento dos demais dados simulados relacionados aos demais aspectos genotípicos e fenotípicos. Dessa maneira, segmentou-se as informações a partir de dois principais grupos: (i) sexo (tendo vista as diferentes intensidades de seleção aplicadas) e (ii) o número da geração de origem das referidas informações. Representamos na Figura 25 alguns dos gráficos construídos a fim de ter uma maior compreensão da estrutura populacional simulada sob o processo de seleção, sendo esses referentes ao I. Coeficiente de endogamia, II. Percentual de ocorrência de homozigotos nos genótipos, III. Valores genéticos estimados via BLUP e IV. Fenótipo (exclusivo ao sexo), em relação ao número de gerações (G).

Figura 25 – Informações dos parâmetros populacionais ao longo das gerações.



Fonte: Autoria própria.

Observou-se, como esperado, o aumento gradual dos referidos parâmetros investigados ao longo das gerações, indicando dessa maneira, níveis de coeficientes de endogamia via parentesco condizentes aos reportados na literatura para a raça Holstein (VANRADEN et al., 2011). Inferiu-se também nas demais informações avaliadas características adequadas à uma população sob seleção, atendendo aos requisitos esperados e observados em situações reais, como aumentos no percentual endogâmico e, consequentemente, na proporção de homozigotos, nos valores de EBV e na expressão dos fenótipos.

4.2. Coeficiente de endogamia

Uma vez consolidados os cenários, a fim de se predizer o coeficiente de endogamia, dois algoritmos (*Random Forests* e *K-Nearest Neighbours*) foram avaliados quanto a viabilidade e eficiência preditiva. Como forma de se facilitar a compreensão e comparação dos resultados, estes foram dispostos em dois principais grupos, sendo o primeiro relacionado a predição do F_{ROH} dos indivíduos a partir da própria informação genotípica (Tabela 1 e Tabela 2), e o segundo a partir da informação genotípica dos pais, empregando o uso dos genótipos tanto do touro, quanto da vaca, a fim de se predizer os valores dos filhos (Tabela 3 e Tabela 4). Duas

tabelas foram empregadas para cada grupo a fim de se reportar tanto os resultados oriundos do algoritmo *Random Forests,* quanto do algoritmo *K-Nearest Neighbours*.

No que diz respeito ao uso do algoritmo RF, diferentes cenários quanto ao número de árvores de decisão foram considerados, sendo esses de 50, 100, 200 e 300 árvores. No caso do algoritmo KNN, diferentes números de vizinhos foram considerados para a predição dos resultados a fim de se avaliar os melhores parâmetros. Sendo quatro cenários considerados, e o respectivo número de vizinhos de 3, 5, 11 e 31. Cabe mencionar que embora o procedimento adotado para esse algoritmo foi o de regressão, o número de vizinhos considerado para a predição baseou-se em números ímpares, a fim de se evitar potenciais “empates preditivos”, situação comum em desafios de classificação (ISLAM et al., 2008). Cabe mencionar também que o algoritmo KNN é caracterizado como do tipo *lazy learner*, exigindo maior custo computacional, especialmente de memória *ram*, inviabilizando o emprego deste em conjunto de dados de maior dimensionalidade (ZHU; ZHANG; HUANG, 2014).

4.2.1 Coeficiente de endogamia predito pelo uso do próprio genótipo

Conforme apresentado no item 3.2.1.1 (Método contendo os genótipos íntegros dos indivíduos) e no item 3.2.1.2 (Método contemplando a extração de atributos dos genótipos), 10 mil genótipos entre as gerações 6 e 12, foram selecionados aleatoriamente e estimados quanto ao nível de F_{ROH} pelo *software* PLINK, empregando os referidos resultados como alvo de treinamento (80% dos dados) e predição (20% dos dados). Dois algoritmos foram eleitos para o processo preditivo, sendo eles o KNN e o RF. Cabe ressaltar que os referidos algoritmos foram empregados sem que houvesse ajuste fino dos hiper parâmetros.

Os resultados quanto ao desempenho dos algoritmos estão descritos nas Tabelas 1 e 2, sendo respectivamente associados aos algoritmos RF e KNN. Os resultados apresentados nas tabelas estão divididos em duas principais colunas, sendo elas nomeadas “Genótipo e F_{ROH}” e “Flag ROH e F_{ROH}”, referentes aos cenários 3.2.1.1 e 3.2.1.2, respectivamente. E os resultados reportados quanto ao erro quadrado médio (MSE) e ao r-quadrado(r^2).

Tabela 1 – Resultados do algoritmo Random Forests para predição de F_{ROH} usando os próprios genótipos.

Random Forests	**Genótipo e F_{ROH}**		**Flag ROH e F_{ROH}**	
Nº árvores	MSE	r^2	MSE	r^2
50	0.001551	0.310340	0.000621	0.690011
100	0.001517	0.325520	0.000599	0.700773
200	0.001511	0.328465	0.000593	0.703975
300	0.001506	0.330451	0.000594	0.703196

Legenda: Erro Quadrado Médio (Mean Squared Error - MSE)
Fonte: Autoria própria.

Para os resultados apresentados na Tabela 1, notou-se em geral um melhor desempenho em ambos os cenários quando considerado um maior número de árvores de decisão, assim como esperado, tendo em vista resultados prévios reportados na literatura quanto ao desempenho deste algoritmo (CHEN; ISHWARAN, 2012). Quanto ao tipo de entrada de dados, evidenciou-se a superioridade dos resultados preditos quando considerado o cenário de extração de atributos "Flag ROH e F_{ROH}" em comparação ao cenário "Genótipo e F_{ROH}", com superioridade do primeiro em relação ao segundo de 0.372745 de diferença quanto ao r^2 para o cenário de maior número de árvores de decisão.

Tabela 2 - Resultados do algoritmo KNN para predição de F_{ROH} usando os próprios genótipos.

KNN	**Genótipo e F_{ROH}**		**Flag F_{ROH} e F_{ROH}**	
Nº Neighbours	MSE	r^2	MSE	r^2
3	0.002269	-0.008939	0.002660	-0.327937
5	0.002053	0.087077	0.002727	-0.361250
11	0.002093	0.069287	0.002856	-0.426023
31	0.001881	0.163449	0.003144	-0.569545

Legenda: MSE (Mean Squared Error - Erro Quadrado Médio). Coeficiente de correlação (r^2).
Fonte: Autoria própria.

Quanto aos resultados do algoritmo KNN (Tabela 2), evidenciou-se superioridade na predição dos valores de F_{ROH} quando considerado o cenário "Genótipo e F_{ROH}" em comparação ao método de extração de atributos ("Genótipo e F_{ROH}"). Sendo este inclusive associado a r^2 de valores negativos, evidenciando a ineficiência algoritmo para o cenário envolvendo a extração de atributos na maneira proposta. Comparados os resultados dos algoritmos RF e KNN, evidenciou-se o melhor desempenho do primeiro associado aos resultados para ambos os cenários propostos.

4.2.1 Coeficiente de endogamia predito pelo genótipo dos pais

Assim como reportado anteriormente (item 4.2.1) os mesmos 10 mil indivíduos entre as gerações 6 e 12, foram considerados, assim como o nível de F_{ROH} calculados previamente. Mas, desta vez, considerados os genótipos dos pais dos indivíduos (touro e vaca) para a predição do F_{ROH} dos mesmos, de acordo com os quatro métodos distintos. Novamente, os dois algoritmos KNN e o RF foram empregados para avaliação dos cenários preditivos propostos.

O desempenho dos algoritmos está descrito nas Tabelas 3 e 4, sendo associadas aos algoritmos RF e KNN, respectivamente. Os resultados apresentados nas tabelas estão divididos em quatro principais colunas, sendo nomeadas “Genótipo Touro e Vaca”, “FlagROH Touro Vaca”, “FlagROH Touro Vaca Alternado”, “EDM Touro e Vaca”, referentes aos cenários descritos nos itens 3.2.1.3, 3.2.1.4, 3.2.1.5 e 3.2.1.6, respectivamente.

Tabela 3 - Resultados do algoritmo Random Forests para predição de F_{ROH} usando genótipos dos pais.

Random Forests	**Genótipo Touro e Vaca**		**FlagROH Touro Vaca**		**FlagROH Touro Vaca Alternado**		**EDM Touro e Vaca**	
Nº árvores	MSE	r^2	MSE	r^2	MSE	r^2	MSE	r^2
50	0.001397	0.277996	0.001885	0.132346	0.001871	0.138995	0.001298	0.422952
100	0.001397	0.278393	0.001840	0.153277	0.001875	0.136898	0.001290	0.426274
200	0.001376	0.289064	0.001850	0.148503	0.001831	0.157575	0.001277	0.432211
300	0.001367	0.293550	0.001860	0.144212	0.001836	0.155084	0.001280	0.430866

Legenda: MSE - Mean Squared Error (Erro Quadrado Médio). Coeficiente de correlação (r^2).
Fonte: Autoria própria.

Para os resultados apresentados na Tabela 3, referentes ao desempenho do algoritmo RF, notou-se em geral, melhor performance referente ao cenário “EDM Touro e Vaca”, sendo o r^2 deste superior em 0.137316 ao segundo melhor cenário (“Genótipo Touro e Vaca”), e 0.286654 em relação ao cenário menos eficiente (“FlagROH Touro Vaca), porém a um maior custo computacional, tendo em vista o cálculo da distância euclidiana entre as janelas dos genótipos do touro e da vaca exigindo maior tempo de processamento. Cabe também ressaltar a superioridade na predição dos valores de F_{ROH} ao se empregar os genótipos brutos dos pais (“Genótipo Touro e Vaca”), em comparação ao cenário de Flags (“FlagROH Touro Vaca”), reportado o r^2 superior em 0.149338. Quanto aos resultados relacionados aos modelos descritos nos itens 3.2.1.4 e 3.2.1.5 (“FlagROH Touro Vaca” e “FlagROH Touro Vaca

Alternado", respectivamente), notou-se superioridade preditiva do modelo *Random Forests* quanto aos resultados de r^2 em 0.010872 ao se alternar as janelas de "FlagROH" oriundas do touro e da vaca, em comparação ao cenário adotado sem intercalar as referidas informações ("FlagROH Touro Vaca").

Ao compararmos os resultados dos cenários considerados na Tabela 4 (referentes ao desempenho do algoritmo KNN para os cenários considerados), observou-se a manutenção do ranqueamento de acordo com os resultados do algoritmo RF, demonstrados na Tabela 3, porém com menor eficiência preditiva do algoritmo KNN para todos os cenários considerados.

Tabela 4 - Resultados do algoritmo KNN para predição de F_{ROH} usando genótipos dos pais.

KNN	**Genótipo Touro e Vaca**		**FlagROH Touro Vaca**		**FlagROH Touro Vaca Alternado**		**EDM entre Touro e Vaca**	
Nº Neighbours	MSE	r^2	MSE	r^2	MSE	r^2	MSE	r^2
3	0.001709	0.117029	0.002326	-0.070576	0.002326	-0.070576	0.001341	0.403512
5	0.001525	0.212040	0.002243	-0.032416	0.002243	-0.032416	0.001239	0.448845
11	0.001472	0.239231	0.002057	0.053487	0.002057	0.053487	0.001253	0.442884
31	0.001449	0.251215	0.001998	0.080320	0.001998	0.080320	0.001374	0.389008

Legenda: MSE - Mean Squared Error (Erro Quadrado Médio). Coeficiente de correlação (r^2).
Fonte: Autoria própria.

4.3. Predição dos valores fenotípicos, dos valores genômicos estimados e do valor genético real

No que diz respeito aos cenários apresentados nos itens 3.2.2.1 (valor fenotípico), 3.2.2.2 (valor genômico estimado), 3.2.2.3 (valor genômico verdadeiro) e 3.2.2.4 (valor genômico verdadeiro via efeitos de SNP estimados), considerou-se os mesmos indivíduos empregados nos cenários de predição do coeficiente de endogamia, descritos anteriormente (10 mil indivíduos entre as gerações 6 e 12).

Considerou-se, para esse cenário, apenas o algoritmo *Random Forests*. Os resultados referentes aos cenários descritos se encontram na Tabela 5, juntamente dos valores de MSE e r^2 para os diferentes números de árvores considerados (50, 100, 200 e 300).

Tabela 5 - Resultados do algoritmo *Random Forests* para predição de Fenótipos, GEBV (BLUPF90) e TBV.

Random Forests	**Genótipo Pred.Fenótipo**		**Genótipo pred. GEBV(BLUPF90)**		**Genótipo pred. TBV**		**Efeito SNP pred. TBV**	
Nº árvores	**MSE**	**r^2**	**MSE**	**r^2**	**MSE**	**r^2**	**MSE**	**r^2**
50	114.811049	0.107895	49.730008	0.065280	101.625642	0.150552	21.965684	0.14372398
100	114.80410	0.107949	49.064169	0.077796	101.030600	0.155526	21.508958	0.16152826
200	114.143887	0.113079	48.184759	0.094325	101.406866	0.152381	21.567772	0.15923753
300	113.423267	0.118679	47.581255	0.098325	101.52209	0.151418	21.486724	0.16239504

Legenda: MSE - *Mean Squared Error* (Erro Quadrado Médio). Coeficiente de correlação (r^2).
Fonte: Autoria própria.

O primeiro cenário (3.2.2.1), reportado na Tabela 5 como “Genótipo Pred.Fenótipo”, foi obtido utilizando informações genotípicas e fenotípicas de fêmeas das gerações mencionadas anteriormente, desconsiderando qualquer outra informação. Dessa maneira, pode-se mencionar que os resultados de r^2 foram relevantes, porém justificados pelo fato de os dados serem ajustados automaticamente pelo referido processo de simulação descrito, podendo sofrer variações ao se considerar um cenário preditivo de uma população real. Quanto ao segundo resultado (3.2.2.2), reportado na tabela como “Genótipo pred. GEBV(BLUPF90)”, os valores de r^2 foram os de menor relevância dentro desse conjunto de resultados apresentado, possivelmente justificado pelo fato dos valores de GEBV terem sido calculados externamente, via BLUPF90. Quanto aos cenários (3.2.2.3 e 3.2.2.4) de estimação do valor genético verdadeiro (TBV), ambos apresentaram valores de r^2 próximos, com leve superioridade (r^2 = 0.01097704) ao se empregar os efeitos de SNP estimados via BLUPF90.

4.4. Predição dos indivíduos otimizados

O cenário considerado no item 3.2.2.5, a fim de se classificar os indivíduos mais otimizados de acordo com os parâmetros previamente descritos, foi empregado nos algoritmos de classificação KNN, considerando quatro cenários referentes ao número de vizinhos (3, 5, 11 e 31), e no algoritmo *Random Forests*, considerando também quatro cenários quanto ao número de árvores (50,100,200 e 300).

Para este cenário em específico (3.2.2.5), foram consideradas apenas informações oriundas da geração 14, totalizando 10 mil indivíduos ao todo. A justificativa em se empregar apenas indivíduos de uma única geração se deve ao fato de a população estar sob o regime de seleção. Dessa maneira, há diferenças

expressivas tanto no que se diz respeito ao coeficiente de endogamia, quanto ao valor genético entre as gerações, como reportado na Figura 25, situação que afeta diretamente os parâmetros classificatórios empregados neste cenário, como descrito no item 3.2.2.5. Dessa maneira, empregou-se a classificação da geração anterior a mais recente, a fim de se predizer os resultados desta.

Ao se empregar o método de classificação, utilizou-se como alvo os indivíduos que atendessem níveis de coeficiente de endogamia e de valor genético inferiores a 50% e superiores a 75% aos valores da geração, respectivamente. Dessa maneira, dos 10 mil indivíduos, apenas 10,54% apresentaram os parâmetros descritos, sendo esses classificados com o rótulo de interesse (1), conforme reportado no item 3.2.2.5.

Previamente ao processo de classificação de ambos os algoritmos, padronizou-se os dados de entrada com a função *StandardScaler()* da biblioteca *scikit-learn* e, posteriormente, foi realizado o processo de predição. Os resultados referentes à acurácia estão descritos na Tabela 6, demonstrado que não houve diferença expressiva entre os algoritmos, neste quesito, porém com resultados bastante distintos ao se avaliar a detecção das ocorrências de interesse.

Tabela 6 – Acurácias dos algoritmos de classificação

KNN Classificador		**Random Forests Classificador**	
Nº Neighbours	**Acurácia**	**Nº Árvores**	**Acurácia**
3	0.84	50	0.89
5	0.86	100	0.89
11	0.89	200	0.89
31	0.89	300	0.89

Fonte: Autoria própria.

Observou-se superioridade do algoritmo KNN em relação ao *Random Forests*, quanto a precisão, sendo o primeiro capaz de detectar 21% das ocorrências atribuídas ao rótulo de interesse (‘1’), enquanto o segundo não apresentou nenhum resultado de interesse. Dessa maneira, para a entrada de dados e para os parâmetros considerados, compreende-se que o algoritmo KNN empregado para classificação apresentou resultados mais satisfatórios em comparação ao RF.

4.5 Resultados do acasalamento genômico

Após realizadas as etapas descritas no método de acasalamento genômico proposto (item 3.2.3), foram considerados 200 acasalamentos da última geração (15)

utilizando as informações exclusivas das gerações anteriores (touro e vaca). Empregou-se essa abordagem a fim de se comparar os resultados preditos com os resultados obtidos das simulações. Empregou-se cenários descritos nos itens 3.2.1.4 (Flag FROH), 3.2.2.1 (Fenótipo) e 3.2.2.4 (TBV), e os respectivos modelos treinados, a fim de se avaliar a correlação entre a média dos resultados preditos e os resultados obtidos na simulação. Os resultados podem ser observados na Tabela 7, contemplando a correlação entre o ranqueamento das “progênies fake” ordenadas pelos cenários propostos e o ranqueamento dos mesmos indivíduos, de acordo com as informações da simulação.

Tabela 7 – Correlação entre valores preditos (Progênie Fake) e valores reais simulados

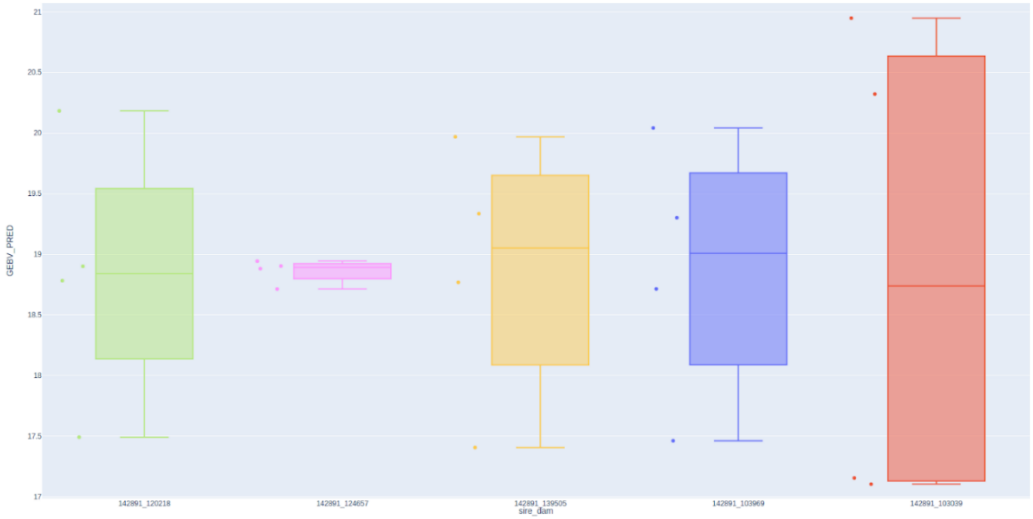
		Valor Real (QMSim) - r^2
Predição	**Flag FROH**	0,358394
	Fenótipo	0,42455
	TBV	0,411632

Fonte: Autoria Própria.

Levando em consideração que a segregação mendeliana não foi contemplada para o método de acasalamento proposto, considerou-se também a avaliação da variabilidade apresentada entre as possíveis progênies de acordo com as combinações possíveis entre as fitas do DNA (Figura 22). Dessa maneira, foi possível avaliar os acasalamentos que podem propiciar maior ou menor variabilidade de progênies, favorecendo as estratégias de acasalamento em situações em que são exigidas menor variabilidade, relacionadas a rebanhos em busca de homogeneidade, ou maior variabilidade, a fim de se obter indivíduos mais diversos e com a possibilidade de mérito genético potencialmente superior.

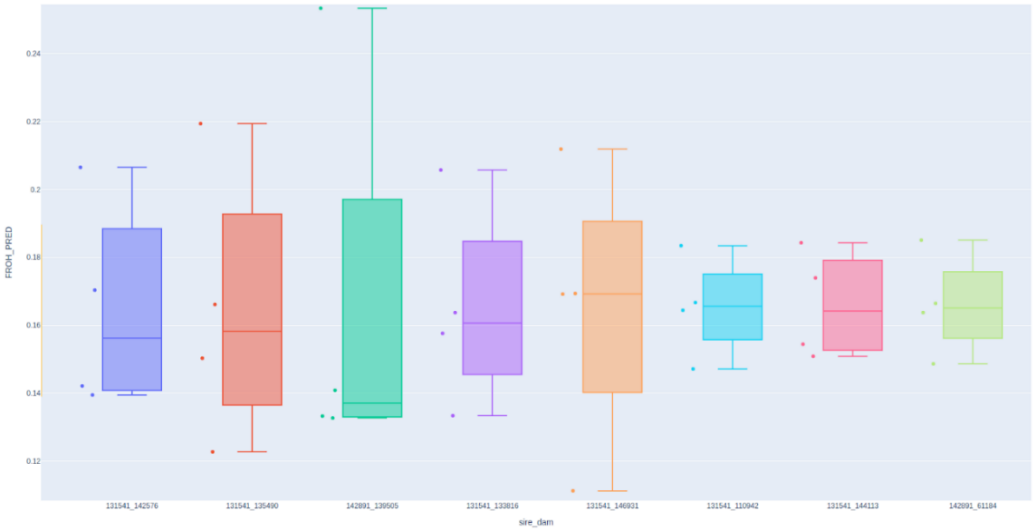
Dessa forma, representou-se na Figura 26 o esquema visual da variabilidade entre as possíveis progênies dos acasalamentos entre os indivíduos de interesse, evidenciando que determinados acasalamentos tendem a apresentar maior ou menor variação, sendo esta uma estratégia a ser considerada de acordo com os interesses produtivos. O mesmo pode ser adotado para o coeficiente de endogamia, como representado na Figura 27, contemplando a variação possível entre as progênies oriundas do acasalamento proposto.

Figura 26 – Variabilidade do mérito genético predito entre as “Progênies Fake”.



Fonte: Autoria Própria.

Figura 27 - Variabilidade do coeficiente de endogamia predito entre as “Progênies Fake”.



Fonte: Autoria Própria.

O aprimoramento e uso das tais abordagens, pode trazer benefícios quanto a tomada de decisão nos acasalamentos, além da possibilidade em se utilizar o método referente ao item 3.2.2.5 (Método de classificação de indivíduos otimizados), viabilizando a otimização de acasalamentos, controlando a endogamia e favorecendo o aumento do mérito genético.

4. CONCLUSÃO

Em conclusão, compreende-se que os resultados apresentados neste trabalho propõem uma nova abordagem quanto ao uso de algoritmos de aprendizado de máquina a fim de se estimar parâmetros de grande importância em programas de

melhoramento animal. Resultados referentes a predição do coeficiente de endogamia e valor genético associados se apresentaram como uma potencial estratégia de direcionamento de acasalamentos a fim de se otimizar o ganho genético e controlar a endogamia. Porém, cabe ressaltar que estudos adicionais serão necessários, com o intuito de se aprimorar o uso dos referidos cenários propostos, explorando com maior nível de detalhamento os hiper parâmetros dos algoritmos associados ao método de predição otimizando-se Froh e valor genético proposto, e posteriormente validar e o potencial uso das técnicas aqui descritas em dados reais.

5. REFERÊNCIAS

AGUILAR, I. et al. Hot topic: A unified approach to utilize phenotypic, full pedigree, and genomic information for genetic evaluation of Holstein final score. **Journal of Dairy Science**, v. 93, n. 2, p. 743–752, fev. 2010.

AKDEMIR, D.; SÁNCHEZ, J. I. Efficient breeding by genomic mating. **Frontiers in Genetics**, v. 7, n. NOV, 29 nov. 2016a.

AKDEMIR, D.; SÁNCHEZ, J. I. Efficient breeding by genomic mating. **Frontiers in Genetics**, v. 7, n. NOV, 29 nov. 2016b.

ALLAIRE, F. R. Mate selection by selection index theory. **Theoretical and Applied Genetics**, v. 57, n. 6, p. 267–272, 1980.

ALVES, A. A. C. et al. Genome-wide prediction for complex traits under the presence of dominance effects in simulated populations using GBLUP and machine learning methods. **Journal of Animal Science**, v. 98, n. 6, p. 1–11, 1 jun. 2020.

BENGTSSON, C. et al. Mating allocations in Nordic Red Dairy Cattle using genomic information. **Journal of Dairy Science**, v. 105, n. 2, p. 1281–1297, 1 fev. 2022.

BISCARINI, F. et al. Using runs of homozygosity to detect genomic regions associated with susceptibility to infectious and metabolic diseases in dairy cows under intensive farming conditions. 26 jan. 2016.

BLUMA, A. L.; LANGLEY, P. **Artificial Intelligence Selection of relevant features and examples in machineAmficial Intelligence**. [s.l: s.n.].

BOEHMKE, B.; GREENWELL, B. **Hands-On Machine Learning with R**. [s.l: s.n.].

BOUQUET, A.; JUGA, J. Integrating genomic selection into dairy cattle breeding programmes: A review. **Animal**, v. 7, n. 5, p. 705–713, maio 2013.

BOURDON, R. M. **Understanding animal breeding**. [s.l: s.n.].

BRADFORD, H. L. et al. Modeling pedigree accuracy and uncertain parentage in single-step genomic evaluations of simulated and US Holstein datasets. **Journal of Dairy Science**, v. 102, n. 3, p. 2308–2318, 1 mar. 2019.

BREIMAN, L. Bagging predictors. **Risks**, v. 8, n. 3, p. 1–26, 1996.

BREIMAN, L. Random forests. **Random Forests**, p. 1–122, 2001.

BZDOK, D.; ALTMAN, N.; KRZYWINSKI, M. Points of Significance: Statistics versus machine learning. **Nature Methods**, v. 15, n. 4, p. 233–234, 2018.

CABESTANY, J. et al. **The Curse of Dimensionality in Data Mining and Time Series PredictionLNCS**. [s.l: s.n.]. Disponível em: <www.ucl.ac.be/mlg>.

CHAMBERLAIN, A. J.; MCPARTLAN, H. C.; GODDARD, M. E. The number of loci that affect milk production traits in dairy cattle. **Genetics**, v. 177, n. 2, p. 1117–1123, out. 2007.

CHEN, L. Curse of Dimensionality. In: LIU, L.; ÖZSU, M. T. (Eds.). . **Encyclopedia of Database Systems**. Boston, MA: Springer US, 2009. p. 545–546.

CHEN, L. Y. Y. et al. Single nucleotide polymorphism mapping using genome-wide unique sequences. **Genome research**, v. 12, n. 7, p. 1106–11, 2002.

CHEN, X.; ISHWARAN, H. **Random forests for genomic data analysisGenomics**, jun. 2012.

COLE, J. B. A simple strategy for managing many recessive disorders in a dairy cattle breeding program. **Genetics Selection Evolution**, v. 47, n. 1, 30 nov. 2015.

COLE, J. B.; VANRADEN, P. M. Visualization of results from genomic evaluations. **Journal of Dairy Science**, v. 93, n. 6, p. 2727–2740, jun. 2010.

CORTES-HERNÁNDEZ, J. et al. Correlation of genomic and pedigree inbreeding coefficients in small cattle populations. **Animals**, v. 11, n. 11, 1 nov. 2021.

DAETWYLER, H. D. et al. Inbreeding in genome-wide selection. **Journal of Animal Breeding and Genetics**, v. 124, n. 6, p. 369–376, 7 dez. 2007.

DAWSON. SNP maps : more markers needed ? Making sense of the human proteome. **Molecular Medicine**, v. 5, n. October, p. 419–420, 1999.

DE OLIVEIRA, H. R. et al. Impact of including information from bulls and their daughters in the training population of multiple-step genomic evaluations in dairy cattle: A simulation study. **Journal of Animal Breeding and Genetics**, v. 136, n. 6, p. 441–452, 1 nov. 2019.

DE REZENDE NEVES, H. H. et al. Acasalamento dirigido para aumentar a produção de animais geneticamente superiores e reduzir a variabilidade da progênie em bovinos. **Revista Brasileira de Zootecnia**, v. 38, n. 7, p. 1201–1204, jul. 2009.

DEKKERS, J. C. M. Application of Genomics Tools to Animal Breeding. **Current Genomics**, v. 13, n. 3, p. 207–212, 2012.

DOEKES, H. P. et al. Inbreeding depression due to recent and ancient inbreeding in Dutch Holstein-Friesian dairy cattle. **Genetics Selection Evolution**, v. 51, n. 1, p. 1–16, 2019.

ENGLUND, C.; VERIKAS, A. A novel approach to estimate proximity in a random forest: An exploratory study. **Expert Systems with Applications**, v. 39, n. 17, p. 13046–13050, 2012.

FALCONER, D. S.; MACKAY, T. F. C. **Introduction to Quantitative Genetics**. 4. ed. ed. [s.l.] Benjamin-Cummings Pub Co; Subsequent edition , 1996.

FERNÁNDEZ, J.; CABALLERO, A. **A comparison of management strategies for conservation with regard to population fitnessConservation Genetics**. [s.l: s.n.].

FISHER, R. A. XV.—The Correlation between Relatives on the Supposition of Mendelian Inheritance. **Transactions of the Royal Society of Edinburgh**, v. 52, n. 2, p. 399–433, 1919.

FORUTAN, M. et al. Inbreeding and runs of homozygosity before and after genomic selection in North American Holstein cattle. **BMC Genomics**, v. 19, n. 1, 27 jan. 2018.

GARCÍA-RUIZ, A.; WIGGANS, G. R.; RUIZ-LÓPEZ, F. J. Pedigree verification and parentage assignment using genomic information in the Mexican Holstein population. **Journal of Dairy Science**, v. 102, n. 2, p. 1806–1810, 1 fev. 2019.

GIBBS, R. A. et al. Genome-wide survey of SNP variation uncovers the genetic structure of cattle breeds. **Science**, v. 324, n. 5926, p. 528–532, 24 abr. 2009.

GIBSON, J.; MORTON, N. E.; COLLINS, A. Extended tracts of homozygosity in outbred human populations. **Human Molecular Genetics**, v. 15, n. 5, p. 789–795, 1 mar. 2006.

GINI, C.; C. Variabilità e mutabilità. **vamu**, 1912.

GOMPERT, Z. et al. Bayesian analysis of molecular variance in pyrosequences quantifies population genetic structure across the genome of Lycaeides butterflies. **Molecular Ecology**, v. 19, n. 12, p. 2455–2473, 2010.

GONZÁLEZ-RECIO, O.; JIMÉNEZ-MONTERO, J. A.; ALENDA, R. The gradient boosting algorithm and random boosting for genome-assisted evaluation in large data sets. **Journal of Dairy Science**, v. 96, n. 1, p. 614–624, 2013.

HAMMACK, S. P. Breeding Systems for Beef Production. 2011.

HARRIS, B. L.; JOHNSON, D. L. Genomic predictions for New Zealand dairy bulls and integration with national genetic evaluation. **Journal of Dairy Science**, v. 93, n. 3, p. 1243–1252, mar. 2010.

HAYES, B. J. et al. **Invited review: Genomic selection in dairy cattle: Progress and challengesJournal of Dairy Science**American Dairy Science Association, , 2009.

HAYES, B.; SHEPHERD, R. K.; NEWMAN, S. Look ahead mate selection schemes for multi-breed beef populations. **Animal Science**, v. 74, n. 1, p. 13–23, 2002.

HE, X.; CAI, D.; NIYOGI, P. Laplacian Score for feature selection. **Advances in Neural Information Processing Systems**, p. 507–514, 2005.

HENDERSON, C. R. Best linear unbiased estimation and prediction under a selection model published by : international biometric society stable. **Biometrics**, v. 31, n. 2, p. 423–447, 1975.

HIDALGO, A. et al. Prediction of postpartum diseases of dairy cattle using machine learning. **Proceedings of the World Congress on Genetics Applied to Livestock Production**, v. 11, n. February, p. 104, 2018.

HOWARD, J. T. et al. Investigation of regions impacting inbreeding depression and their association with the additive genetic effect for United States and Australia Jersey dairy cattle. **BMC Genomics**, v. 16, n. 1, 19 out. 2015.

HOWARD, J. T. et al. Invited review: Inbreeding in the genomics era: Inbreeding, inbreeding depression, and management of genomic variability. **Journal of Dairy Science**, v. 100, n. 8, p. 6009–6024, 1 ago. 2017.

ISLAM, M. J. et al. **Investigating the Performance of Naive- Bayes Classifiers and K- Nearest Neighbor Classifiers**. Institute of Electrical and Electronics Engineers (IEEE), 28 abr. 2008.

JANSEN, G. B.; WILTON, J. W. Selecting Mating Pairs with Linear Programming Techniques. **Journal of Dairy Science**, v. 68, n. 5, p. 1302–1305, 1985.

JIA, Y.; JANNINK, J. L. Multiple-trait genomic selection methods increase genetic value prediction accuracy. **Genetics**, v. 192, n. 4, p. 1513–1522, 2012.

JOLLIFE, I. T.; CADIMA, J. Principal component analysis: A review and recent developments. **Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences**, v. 374, n. 2065, 2016.

KHALID, S. **A Survey of Feature Selection and Feature Extraction Techniques in Machine Learning**. [s.l: s.n.]. Disponível em: <www.conference.thesai.org>.

KINGHORN, B. The tactical approach to implementing breeding programs. **Animal breeding - use of new technologies**, 2000.

KINGHORN, B. P. An algorithm for efficient constrained mate selection. **Genetics Selection Evolution**, v. 43, n. 1, p. 1–9, 2011a.

KINGHORN, B. P. An algorithm for efficient constrained mate selection. **Genetics Selection Evolution**, v. 43, n. 1, 2011b.

KOIVULA, M. et al. Different methods to calculate genomic predictions-Comparisons of BLUP at the single nucleotide polymorphism level (SNP-BLUP), BLUP at the individual level (G-BLUP), and the one-step approach (H-BLUP). **Journal of Dairy Science**, v. 95, n. 7, p. 4065–4073, jul. 2012.

LI, B. et al. Genomic prediction of breeding values using a subset of SNPs identified by three machine learning methods. **Frontiers in Genetics**, v. 9, n. JUL, p. 1–20, 2018.

LILLEHAMMER, M.; MEUWISSEN, T. H. E.; SONESSON, A. K. A comparison of dairy cattle breeding designs that use genomic selection. **Journal of Dairy Science**, v. 94, n. 1, p. 493–500, jan. 2011.

LOURENCO, D. A. L. et al. Are evaluations on young genotyped animals benefiting from the past generations? **Journal of Dairy Science**, v. 97, n. 6, p. 3930–3942, 2014.

LUND, M. S. et al. A common reference population from four European Holstein populations increases reliability of genomic predictions. **Genetics Selection Evolution**, v. 43, n. 1, p. 1–8, 2011.

M VERLOUW, J. A. et al. A comparison of genotyping arrays. **European Journal of Human Genetics**, v. 29, p. 1611–1624, 2021.

MA, S.; DAI, Y. Principal component analysis based Methods in bioinformatics studies. **Briefings in Bioinformatics**, v. 12, n. 6, p. 714–722, 2011.

MACLEOD, I. M. et al. A novel predictor of multilocus haplotype homozygosity: Comparison with existing predictors. **Genetics Research**, v. 91, n. 6, p. 413–426, 2009.

MAKANJUOLA, B. O. et al. Effect of genomic selection on rate of inbreeding and coancestry and effective population size of Holstein and Jersey cattle populations. **Journal of Dairy Science**, v. 103, n. 6, p. 5183–5199, 1 jun. 2020.

MARQUES, E. et al. High density linkage disequilibrium maps of chromosome 14 in Holstein and Angus cattle. **BMC Genetics**, v. 9, p. 1–12, 2008.

MÉSZÁROS, G. et al. Genomic background of entropion in Fleckvieh cattle. **Poljoprivreda**, v. 21, n. 1, p. 48–51, 1 jun. 2015.

MEUWISSEN, T.; GODDARD, M. Accurate prediction of genetic values for complex traits by whole-genome resequencing. **Genetics**, v. 185, n. 2, p. 623–631, 2010.

MEUWISSEN, T. H. E. **Maximizing the Response of Selection with a Predefined Rate of Inbreeding 1**. [s.l: s.n.].

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

MILLER, S. Genetic improvement of beef cattle through opportunities in genomics. **Revista Brasileira de Zootecnia**, v. 39, n. SUPPL. 1, p. 247–255, 2010.

MISZTAL, I.; TSURUTA, S. Manual for BLUPF90 family of programs. 2015.

MONTESINOS-LÓPEZ, O. A. et al. **A review of deep learning applications for genomic selectionBMC Genomics**BioMed Central Ltd, , 1 dez. 2021.

MOTODA, H.; LIU, H. Feature selection, extraction and construction. **Communication of IICM**, v. 5, p. 67–72, 2002.

NAYERI, S.; SARGOLZAEI, M.; TULPAN, D. A review of traditional and machine learning methods applied to animal breeding. **Animal Health Research Reviews**, n. Dl, p. 31–46, 2019.

PEDERSEN, L. D.; SØRENSEN, A. C.; BERG, P. Marker-assisted selection can reduce true as well as pedigree-estimated inbreeding. **Journal of Dairy Science**, v. 92, n. 5, p. 2214–2223, 2009.

PEDREGOSA FABIANPEDREGOSA, F. et al. **Scikit-learn: Machine Learning in Python Gaël Varoquaux Bertrand Thirion Vincent Dubourg Alexandre Passos PEDREGOSA, VAROQUAUX, GRAMFORT ET AL. Matthieu PerrotJournal of Machine Learning Research**. [s.l: s.n.]. Disponível em: <http://scikit-learn.sourceforge.net.>.

PÉREZ-ENCISO, M.; ZINGARETTI, L. M. A guide for using deep learning for complex trait genomic prediction. **Genes**, v. 10, n. 7, 2019.

PRASATH, V. B. S. et al. Distance and Similarity Measures Effect on the Performance of K-Nearest Neighbor Classifier -- A Review. p. 1–39, 2017.

PURCELL, S. et al. PLINK: A tool set for whole-genome association and population-based linkage analyses. **American Journal of Human Genetics**, v. 81, n. 3, p. 559–575, 2007.

PURFIELD, D. C. et al. Runs of homozygosity and population history in cattle. **BMC Genetics**, v. 13, 2012.

RABIER, C. E. et al. On the accuracy of genomic selection. **PLoS ONE**, v. 11, n. 6, 1 jun. 2016.

SALAZAR, J. J. et al. Fair train-test split in machine learning: Mitigating spatial autocorrelation for improved prediction accuracy. **Journal of Petroleum Science and Engineering**, v. 209, p. 109885, 1 fev. 2022.

SARGOLZAEI, M. et al. Extent of linkage disequilibrium in Holstein cattle in North America. **Journal of Dairy Science**, v. 91, n. 5, p. 2106–2117, 2008.

SARGOLZAEI, M.; SCHENKEL, F. S. QMSim: A large-scale genome simulator for livestock. **Bioinformatics**, v. 25, n. 5, p. 680–681, 2009.

SCHAEFFER, L. R. Strategy for applying genome-wide selection in dairy cattle. **Journal of Animal Breeding and Genetics**, v. 123, n. 4, p. 218–223, 2006.

SEATON, G. et al. QTL Express: Mapping quantitative trait loci in simple and complex pedigrees. **Bioinformatics**, v. 18, n. 2, p. 339–340, 2002.

SHAHINFAR, S. et al. Prediction of insemination outcomes in Holstein dairy cattle using alternative machine learning algorithms. **Journal of Dairy Science**, v. 97, n. 2, p. 731–742, 2014.

SOMVANSHI, M. et al. A review of machine learning techniques using decision tree and support vector machine. **Proceedings - 2nd International Conference on Computing, Communication, Control and Automation, ICCUBEA 2016**, 2017.

SONESSON, A. K.; WOOLLIAMS, J. A.; MEUWISSEN, T. H. E. Genomic selection requires genomic control of inbreeding. **Genetics Selection Evolution**, v. 44, n. 1, 2012.

SONG, Y. Y.; LU, Y. Decision tree methods: applications for classification and prediction. **Shanghai Archives of Psychiatry**, v. 27, n. 2, p. 130–135, 2015.

STEIN, S. A. M. et al. Chapter 13 Principal Components Analysis: A Review of its Application on Molecular Dynamics Data. **Annual Reports in Computational Chemistry**, v. 2, n. C, p. 233–261, 2006.

SYVÄNEN, A. C. Accessing genetic variation: Genotyping single nucleotide polymorphisms. **Nature Reviews Genetics**, v. 2, n. 12, p. 930–942, 2001.

TOLEDO, E. R. DE et al. Mapeamento de QTLS: uma abordagem Bayesiana. **Revista Brasileira de Biometria**, v. 26, p. 107–114, 2008.

TORO, M. A.; VARONA, L. **A note on mate allocation for dominance handling in genomic selection**. [s.l: s.n.]. Disponível em: <http://www.gsejournal.org/content/42/1/33>.

VAN DIJK, A. D. J. et al. Machine learning in plant science and plant breeding. **iScience**, v. 24, n. 1, p. 101890, 2021.

VANRADEN, P. Symposium review: How to implement genomic selection. **Journal of Dairy Science**, v. 103, p. 5291–5301, 2020.

VANRADEN, P. M. et al. **Invited review: Reliability of genomic predictions for North American Holstein bullsJournal of Dairy Science**American Dairy Science Association, , 2009.

VANRADEN, P. M. et al. Genomic inbreeding and relationships among Holsteins, Jerseys, and Brown Swiss. **Journal of Dairy Science**, v. 94, n. 11, p. 5673–5682, 2011.

VENTURA, R. V. et al. Opportunities and challenges of phenomics applied to livestock and aquaculture breeding in South America. **Animal Frontiers**, v. 10, n. 2, p. 45–52, 1 abr. 2020.

VISSCHER, P. M. et al. Estimation of pedigree errors in the UK dairy population using microsatellite markers and the impact on selection. **Journal of Dairy Science**, v. 85, n. 9, p. 2368–2375, 2002.

WANG, J.; SANTIAGO, E.; CABALLERO, A. **Prediction and estimation of effective population sizeHeredity**Nature Publishing Group, , 1 out. 2016.

WETTSCHERECK, D.; AHA, D. W.; MOHRI, T. A Review and Empirical Evaluation of Feature Weighting Methods for a Class of Lazy Learning Algorithms. **Artificial Intelligence Review**, v. 11, n. 1/5, p. 273–314, 1997.

WIGGANS, G. R. et al. Genomic Selection in Dairy Cattle: The USDA Experience. **Annual Review of Animal Biosciences**, v. 5, n. 1, p. 309–327, 2017.

WRIGHT, S. SYSTEMS OF MATING. I. THE BIOMETRIC RELATIONS BETWEEN PARENT AND OFFSPRING. **Genetics**, v. 6, n. 2, p. 111–123, 1921.

WRIGHT, S. Coefficients of Inbreeding and Relationship Author (s): Sewall Wright Source : The American Naturalist , Vol . 56 , No . 645 (Jul . - Aug ., 1922), pp . 330-338 Published by : The University of Chicago Press for The American Society of Naturalists Sta. **The American Naturalist**, v. 56, n. 645, p. 330–338, 1922.

WRIGHT, S. **The Interpretation of Population Structure by F-Statistics with Special Regard to Systems of MatingSource: Evolution**. [s.l: s.n.].

WU, X. et al. Top 10 algorithms in data mining. **Knowledge and Information Systems**, v. 14, n. 1, p. 1–37, 4 dez. 2008.

XU, C.; JACKSON, S. A. Machine learning and com plex biological data. **Genome Biology**, v. 20, n. 1, p. 1–4, 2019a.

XU, C.; JACKSON, S. A. **Machine learning and complex biological dataGenome Biology**BioMed Central Ltd., , 16 abr. 2019b.

ZHANG, H. et al. **Progress of genome wide association study in domestic animalsJournal of Animal Science and Biotechnology**, 22 ago. 2012.

ZHAO, Z. et al. Advancing Feature Selection Research. **ASU Feature Selection Repository Arizona State University**, p. 1–28, 2010.

ZHU, X.; ZHANG, L.; HUANG, Z. A sparse embedding and least variance encoding approach to hashing. **IEEE Transactions on Image Processing**, v. 23, n. 9, p. 3737–3750, 2014.