UNIVERSIDADE FEDERAL DE JUIZ DE FORA
INSTITUTO DE CIÊNCIAS EXATAS
PROGRAMA DE PÓS-GRADUAÇÃO EM CIÊNCIA DA COMPUTAÇÃO

Luiz Afonso Glatzl Junior

Predição de linhagens em gado de leite para a redução da endogamia visando o aconselhamento de cruzamentos em programas de melhoramento genético animal

Juiz de Fora
2023

Luiz Afonso Glatzl Junior

Predição de linhagens em gado de leite para a redução da endogamia visando o aconselhamento de cruzamentos em programas de melhoramento genético animal

Dissertação apresentada ao Programa de Pós-Graduação em Ciência da Computação da Universidade Federal de Juiz de Fora como requisito parcial à obtenção do título de Mestre em Ciência da Computação.

Orientador: Prof. Dr. Saulo Moraes Villela

Coorientador: Dr. Marcos Vinicius Gualberto Barbosa da Silva

Juiz de Fora

2023

Ficha catalográfica elaborada através do programa de geração
automática da Biblioteca Universitária da UFJF,
com os dados fornecidos pelo(a) autor(a)

Glatzl Junior, Luiz Afonso.
Predição de linhagens em gado de leite para a redução da endogamia visando o aconselhamento de cruzamentos em programas de melhoramento genético animal / Luiz Afonso Glatzl Junior. -- 2023.
59 p.

Orientador: Saulo Moraes Villela
Coorientador: Marcos Vinicius Gualberto Barbosa da Silva
Dissertação (mestrado acadêmico) - Universidade Federal de Juiz de Fora, Instituto de Ciências Exatas. Programa de Pós-Graduação em Ciência da Computação, 2023.

1. Predição de Linhagens. 2. Clusterização. 3. Genômica Animal. 4. Aprendizado de Máquina. 5. Bioinformática. I. Villela, Saulo Moraes, orient. II. da Silva, Marcos Vinicius Gualberto Barbosa, coorient. III. Título.

Luiz Afonso Glatzl Junior

Predição de linhagens em gado de leite para a redução da endogamia visando o aconselhamento de cruzamentos em programas de melhoramento genético animal

Dissertação apresentada ao Programa de Pós-graduação em Ciência da Computação da Universidade Federal de Juiz de Fora como requisito parcial à obtenção do título de Mestre em Ciência da Computação. Área de concentração: Ciência da Computação.

Aprovada em 16 de novembro de 2023.

BANCA EXAMINADORA

Prof. Dr. Saulo Moraes Villela - Orientador
Universidade Federal de Juiz de Fora

Prof. Dr. Marcos Vinicius Gualberto Barbosa da Silva - Coorientador
Empresa Brasileira de Pesquisa Agropecuária

Prof. Dr. Carlos Cristiano Hasenclever Borges
Universidade Federal de Juiz de Fora

Prof. Dr. Ronney Moreira de Castro
Universidade Federal de Juiz de Fora

Prof. Dr. João Cláudio do Carmo Panetto
Empresa Brasileira de Pesquisa Agropecuária

Juiz de Fora, 20/03/2025.

Documento assinado eletronicamente por **Saulo Moraes Villela**, **Professor(a)**, em 20/03/2025, às 17:16, conforme horário oficial de Brasília, com fundamento no § 3º do art. 4º do Decreto nº 10.543, de 13 de novembro de 2020.

Documento assinado eletronicamente por **Marcos Vinicius Gualberto Barbosa da Silva**, **Usuário Externo**, em 21/03/2025, às 10:29, conforme horário oficial de Brasília, com fundamento no § 3º do art. 4º do Decreto nº 10.543, de 13 de novembro de 2020.

Documento assinado eletronicamente por **Joao Claudio do Carmo Panetto**, **Usuário Externo**, em 27/08/2026, às 08:58, conforme horário oficial de Brasília, com fundamento no § 3º do art. 4º do Decreto nº 10.543, de 13 de novembro de 2020.

Documento assinado eletronicamente por **Ronney Moreira de Castro**, **Professor(a)**, em 08/09/2026, às 14:27, conforme horário oficial de Brasília, com fundamento no § 3º do art. 4º do Decreto nº 10.543, de 13 de novembro de 2020.

Documento assinado eletronicamente por **Carlos Cristiano Hasenclever Borges**, **Professor(a)**, em 08/09/2026, às 16:40, conforme horário oficial de Brasília, com fundamento no § 3º do art. 4º do Decreto nº 10.543, de 13 de

novembro de 2020.

A autenticidade deste documento pode ser conferida no Portal do SEI-Ufjf (www2.ufjf.br/SEI) através do ícone Conferência de Documentos, informando o código verificador **2304207** e o código CRC **AE988EB4**.

Dedico a oportunidade de ter realizado este trabalho, bem como as consequentes oportunidades, a memória de minha mãe, Elizabeth, que já se foi, mas que se faz presente em todos os dias da minha vida e que gostaria muito de estar presente neste momento. No entanto, sei que, de algum lugar, ela olha por mim e se orgulha com minha trajetória.

AGRADECIMENTOS

A Deus, por me sondar, orientar e sustentar cada vez mais ao longo de minha vida acadêmica, profissional e pessoal.

Aos meu pais, Elizabeth e Luiz Afonso, meus irmãos, Elizangela e Fabiano, e minhas sobrinhas, Amanda e Thaís, pelo amor, incentivo e apoio incondicional de sempre.

Aos meus amigos do CTU, quase irmãos, que estiveram presentes desde o Ensino Médio tornando a vida mais leve, até mesmo no momento mais difícil pelo qual já passei.

A minha companheira de vida, Leandra, pelo carinho, apoio e luz sempre que as esperanças se esvaíam, e aos seus pais, Selma e Carlos, por aparecem desde sempre como anjos da guarda em minha caminhada.

A esta universidade, seus docentes e servidores por propiciarem um ambiente de transformação pessoal e intelectual.

Ao meu orientador Saulo, pelo suporte, incentivo, paciência e compreensão no pouco tempo que tivemos, bem como por suas sugestões e correções na realização deste trabalho.

Ao meu supervisor de estágio e coorientador Marcos Vinicius, por instigar minha veia acadêmica e fazer enxergar a Ciência da Computação como meio de transformação em qualquer área do conhecimento. Além de todo o suporte e compreensão de sempre, em todos os aspectos.

Aos colegas da UFJF, por percorrerem juntamente esse caminho e auxiliar a superar os diversos obstáculos.

E a todos que direta ou indiretamente fizeram parte da minha formação, o meu muito obrigado.

RESUMO

Os avanços na genética molecular possibilitaram um processo de genotipagem eficiente em programas de melhoramento genético, utilizando milhares de marcadores polimórficos de nucleotídeo único (SNPs). Essas informações aceleram a seleção genética, estimando com maior precisão o valor genético de animais sem registro fenotípico próprio. No Brasil, programas de melhoramento genético animal visam identificar animais superiores para multiplicação genética, garantindo a sustentabilidade econômica e ambiental da pecuária leiteira. As avaliações genéticas empregam modelos e estatísticas para calcular valores genéticos. A necessidade de lidar com a endogamia surge no contexto do melhoramento genético, especialmente ao selecionar animais de alto desempenho para realização de cruzamentos e construção de rebanhos. Este estudo focaliza a predição de linhagens em gado leiteiro, utilizando aprendizado de máquina. Análises supervisionadas, baseadas em pedigree, e não supervisionadas, explorando apenas os marcadores moleculares, foram empregadas. Os modelos demonstraram alta acurácia na classificação de animais provenientes dos cinco touros de maior progênie em linhagens, validados nos conjuntos de treinamento e teste. A abordagem não supervisionada ofereceu visões adicionais sobre a composição genética da população, além de indicar caminhos promissores para estudos futuros. Destaca-se a relevância da anotação genealógica na identificação de linhagens. Modelos supervisionados mostraram promissora melhora na identificação de novos animais com conhecimento prévio da população. Adicionalmente, verificou-se a possibilidade de futura implementação dos modelos como apoio para lidar com a diversidade genética e aprimorar a seleção de animais em programas de melhoramento genético animal.

Palavras-chave: Aprendizado de Máquina; Bioinformática; Clusterização; Genômica Animal; Predição de Linhagens.

ABSTRACT

Advancements in molecular genetics have enabled an efficient genotyping process in animal breeding programs, utilizing thousands of single nucleotide polymorphism (SNP) markers. These data expedite genetic selection, providing more accurate estimates of the genetic value of animals without their own phenotypic records. In Brazil, animal breeding programs aim to identify superior animals for genetic multiplication, ensuring economic and environmental sustainability in dairy farming. Genetic evaluations employ models and statistics to calculate genetic values. The need to address inbreeding arises in the context of genetic improvement, especially when selecting high-performance animals for crossbreeding and herd development. This study focuses on predicting lineages in dairy cattle using machine learning. Supervised analyses, based on pedigree, and unsupervised analyses, solely exploring molecular markers, were employed. The models exhibited high accuracy in classifying animals from the top five progeny bulls into lineages, validated in training and testing datasets. The unsupervised approach provided additional insights into the genetic composition of the population, indicating promising paths for future studies. The importance of genealogical annotation in lineage identification is emphasized. Supervised models showed promising improvements in identifying new animals with prior knowledge of the population. Additionally, the possibility of future implementation of these models as support for handling genetic diversity and enhancing animal selection in animal breeding programs was verified.

Keywords: Animal Genomics; Bioinformatics; Clustering; Lineage Prediction; Machine Learning.

LISTA DE FIGURAS

LISTA DE TABELAS

LISTA DE ABREVIATURAS E SIGLAS

ABCGIL	Associação Brasileira dos Criadores de Gir Leiteiro
ABCZ	Associação Brasileira dos Criadores de Zebu
CH	Cluster Hierárquico
CNPGL	Centro Nacional de Pesquisa de Gado de Leite
DNA	Ácido Desoxirribonucleico
EMBRAPA	Empresa Brasileira de Pesquisa Agropecuária
GEO	*Gene Expression Omnibus*
GWAS	Estudos de Associação Genômica Ampla
MAF	*Minor Alele Frequency*
MLP	*Multilayer Perceptron*
PCA	Análise de Componentes Principais
PNMGL	Programa Nacional de Melhoramento do Gir Leiteiro
RF	*Random Forest*
RNAs	Redes Neurais Artificiais
SNP	*Single Nucleotide Polymorphism*
SVM	*Support Vector Machine*
TAE	Teoria do Aprendizado Estatístic

SUMÁRIO

1 INTRODUÇÃO

1.1 MOTIVAÇÃO

Nas últimas décadas, a intensificação do desenvolvimento tecnológico, atuando direta ou indiretamente, tem proporcionado avanços significativos em diversas áreas do conhecimento, ocasionando a criação e estabelecimento de campos de estudo e pesquisa fundamentalmente interdisciplinares (SANTOS; COELHO; FERNANDES, 2020).

A união entre as ciências exatas, biológicas, humanas, sociais, da saúde e engenharias tem gerado soluções para problemas cada vez mais complexos, atuando desde aplicações aeroespaciais e industriais, até diagnósticos de doenças, aconselhamento genético e melhoramento genético de plantas e animais de interesse produtivo (SHINDE; SHAH, 2018).

Os avanços nas técnicas de genética molecular tornaram possível, aos programas de melhoramento de plantas e animais, genotipar indivíduos a baixo custo em relação a muitos milhares de marcadores polimórficos de nucleotídeo único (em inglês *single nucleotide polymorphism*, SNP) espalhados por todo o genoma (GARRICK; TAYLOR; FERNANDO, 2014). O uso de informações genotípicas possibilita reduzir o tempo de geração de seleção e estimar, com maior precisão, o valor genético de indivíduos que não possuem registro fenotípico próprio e não tem descendência (GODDARD; HAYES, 2009).

No Brasil, programas delineados de melhoramento genético animal foram implantados a partir de 1976, com a criação do Centro Nacional de Pesquisa de Gado de Leite (CNPGL), atual Embrapa Gado de Leite (AUAD; AL., 2010). Possuindo como objetivos a identificação de indivíduos superiores, a multiplicação genética de forma orientada, a avaliação de várias características econômicas e a promoção da sustentabilidade da atividade leiteira, são avaliadas diversas características, incluindo medidas produtivas, de conformação e de manejo, além de medidas genotípicas (MEUWISSEN; HAYES; GODDARD, 2001). As avaliações genéticas são realizadas utilizando procedimentos do modelo animal, aliado a metodologias de estimação estatística para se calcular os valores genéticos para as características de interesse (GODDARD, 2009).

Os Programas envolvem estudos em genômica, visando obter modelos de predição do valor genômico para os animais, gerando grande impacto sobre ganhos genéticos, redução do intervalo de geração e do custo do teste de progênie. A evolução e os avanços recentes em biotecnologia e bioinformática possibilitaram a incorporação de informações de marcadores moleculares através dos chips de SNP (OTTO; AL., 2023).

Como resultado dos programas de melhoramento genético e das avaliações, é gerado um *ranking* dos animais (top 10%) para as características de interesse. Objetivando uma melhora no rebanho, os melhores colocados nos ranqueamentos são amplamente utilizados

para geração de novos indivíduos através de inseminação artificial, elevando o índice de endogamia ou consanguinidade, o que implica diretamente na perda da variabilidade genética, se tornando um sério problema no caso de doenças e anomalias desenvolvidas a partir de alelos recessivos, além de diminuir a velocidade dos ganhos no melhoramento da raça (GROSZ; KAKI, 2007).

A identificação de linhagens normalmente é realizada com base apenas no *pedigree*, ou em árvores genealógicas. No entanto, os registros genealógicos de animais mantido pelas associações de criadores podem possuir falhas de cadastro, dados faltantes, inconsistências em informações de antepassados e limitações para animais nascidos antes do início do registro. Desta forma, a correta identificação de linhagens na população possibilita a busca de animais mais distantes geneticamente e que satisfaçam o *trade off* entre critérios de características de interesse produtivo e a diversidade genética para desenvolvimento da população, evitando doenças e anomalias como citado anteriomente (OTTO; AL., 2023).

1.2 DEFINIÇÃO DO PROBLEMA

O problema abordado no trabalho busca a partir do genoma bovino sequenciado através de um chip de SNPs em cada animal do rebanho disponibilizado pela Embrapa, realizar a separação destes animais em linhagens (na raça Gir Leiteiro) para aconselhar os cruzamentos, auxiliando na identificação de animais pertencentes a linhagens mais distantes geneticamente, com objetivo de se obter maior diversidade genética na população, contribuindo para o desenvolvimento dos programas de melhoramento. Para isso, serão realizadas comparações de desempenho de técnicas de Aprendizado de Máquina em duas etapas: a primeira baseada em aprendizado supervisionado, utilizando as linhagens paternas e maternas definidas através do pedigree, ou genealogia dos animais, como rótulos para classificação e a segunda através do aprendizado não-supervisionado, buscando compreender melhor a composição genética da população através somente da verificação de seus marcadores moleculares.

1.3 OBJETIVOS

O presente trabalho desenvolve modelos baseados em aprendizado de máquina utilizando os genótipos como base para realizar a predição de linhagens (normalmente identificadas apenas com base no *pedigree*, ou árvores genealógicas) em gado leiteiro objetivando auxiliar no planejamento de cruzamentos, em programas de melhoramento genético animal, que minimizem os níveis de endogamia através da identificação de indivíduos pertencentes a linhagens mais distantes geneticamente.

1.4 TRABALHOS RELACIONADOS

Em Lopez-Cortes et al. (2020) foram utilizados métodos de agrupamento de aprendizado de máquina, especificamente K-means e agrupamento hierárquico, em conjunto com técnicas de redução de dimensionalidade linear e não linear, como autoencoder profundo (DeepAE) e análise de componentes principais (PCA). A análise concentrou-se na estrutura genética e na alocação individual de linhagens de milho, incluindo milho dentado de campo (n = 97) e pipoca (n = 86). Os resultados indicaram que a abordagem de agrupamento hierárquico, combinada com o pré-processamento de dados utilizando DeepAE (DeepAE-HC), apresentou uma eficácia superior na atribuição individual aos clusters, atingindo uma taxa de 96% de acurácia, superando outras metodologias. Este estudo busca focar na aplicabilidade da redução dimensional baseada em aprendizado profundo, aliada a métodos de agrupamento na identificação de grupos geneticamente distintos.

Algoritmos de RNAs têm sido amplamente empregados na análise de dados genômicos, com foco em SNPs, variações genéticas prevalentes no genoma humano associadas a diversas condições patológicas. O desenvolvimento de RNAs para manipulação desse tipo de dados é considerado um marco significativo na área médica. Entretanto, a alta dimensionalidade dos dados genômicos e a limitação amostral podem complicar a tarefa de aprendizado. Em Soumare et al. (2021), propõe-se um novo método de classificação de RNAs fundamentado em perturbação de entrada. A abordagem incorpora a Decomposição em Valores Singulares (SVD) para redução da dimensionalidade dos dados de entrada, seguida pelo treinamento de uma rede de classificação. Os erros de previsão são mitigados por meio da perturbação da matriz de projeção SVD. Avaliado em dados de indivíduos com diversas origens ancestrais, os resultados experimentais destacam a eficácia da metodologia, alcançando uma precisão de classificação de até 96,23%.

Em Ausmees et al. (2022) é proposto um modelo de aprendizado profundo baseado em uma arquitetura de autoencoder convolucional para a redução de dimensionalidade de dados genotípicos. Avaliado em uma base diversificada de amostras humanas, o modelo identifica agrupamentos populacionais, fornecendo informações visuais aprimoradas em comparação com a análise de componentes principais, ao mesmo tempo que preserva a geometria global mais eficientemente que o t-SNE e o UMAP. Além disso, o método é apresentado como uma abordagem de agrupamento genético, com resultados comparados ao software ADMIXTURE frequentemente utilizado em estudos genéticos populacionais.

Identificar loci de susceptibilidade associados a doenças complexas é um desafio crucial na modelagem biomédica. Abordagens existentes para a descoberta de biomarcadores enfrentam limitações como detecção subdimensionada, falta de consideração para interações entre variantes e dependência restritiva de conhecimento biológico prévio. Como abordagem para superar esses desafios, Silva et al. (2022) propõe descobrir loci de sus-

ceptibilidade associados a doenças por meio da integração de random forest e análise de agrupamento em estudos de associação genômica ampla (GWAS) anteriores. O framework integrado proposto é aplicado a dados de GWAS para a soroclearância do antígeno de superfície do vírus da hepatite B (HBsAg). Análises de agrupamento foram realizadas em SNPs significativos no GWAS e SNPs com os maiores escores de importância de características obtidos através de uma random forest. Os conjuntos resultantes de SNPs das análises de agrupamento foram testados quanto à associação com características. Três loci de susceptibilidade possivelmente associados à soroclearância do HBsAg foram identificados. Os resultados destacam o potencial do framework integrado proposto na identificação de loci de susceptibilidade associados a doenças.

Em Pirmoradi et al. (2020) são abordados os desafios na detecção de SNPs e na classificação de amostras saudáveis e de pacientes usando algoritmos de Aprendizado de Máquina. A maldição da dimensionalidade, tamanhos de amostra limitados em comparação com SNPs e possíveis desequilíbrios nas amostras saudáveis e de pacientes representam obstáculos significativos. O estudo propõe um método que envolve o uso de Codificação Média para converter dados nominais de SNPs em numéricos, um filtro em dois passos para seleção de características e um autoencoder profundo para classificação automática. A avaliação em cinco conjuntos diversos de dados de SNPs, incluindo câncer de tireoide, retardamento mental, câncer de mama, câncer colorretal e autismo do conjunto de dados Gene Expression Omnibus (GEO), demonstra o sucesso do método na seleção de características e classificação. Altas taxas de precisão de 100%, 94.4%, 100%, 96% e 99.1% são alcançadas para câncer de tireoide, retardamento mental, câncer de mama, câncer colorretal e autismo, respectivamente.

1.5 ORGANIZAÇÃO DO TEXTO

O trabalho aqui apresentado está organizado em sete capítulos, o primeiro é referente a introdução, composta pela motivação, definição do problema abordado, objetivos de estudo, trabalhos relacionados e organização do texto. O segundo e terceiro capítulos contém uma fundamentação teórica necessária para a construção deste trabalho sobre Melhoramento Genético Animal e Aprendizado de Máquina, respectivamente. No quarto, a metodologia proposta é exposta, explicando os tratamentos de dados necessários e fornecendo uma visão geral do fluxo de processos na abordagem supervisionada aplicada aos touros de maior progênie e na abordagem não supervisionada para caracterização das linhagens na população como um todo. O quinto, exibe os resultados obtidos na etapa de desenvolvimento para cada uma das abordagens. E no sexto capítulo são descritas as conclusões sobre o trabalho e as possibilidades de futuros aperfeiçoamentos e modificações.

2 MELHORAMENTO GENÉTICO ANIMAL

2.1 Fundamentos Biológicos

As mutações representam mudanças na sequência de nucleotídeos do DNA, sendo a fonte primária de diversidade genética. Podem ocorrer devido a várias causas, como erros na replicação do DNA, exposição a radiações ou produtos químicos mutagênicos, ou até mesmo mutações espontâneas (WATSON; AL., 2008; ALBERTS; AL., 2014). Possuem diferentes formas de atuação, afetando um único nucleotídeo (mutações pontuais) ou envolvendo inserções, deleções e rearranjos de segmentos maiores do DNA.

Um exemplo clássico de mutação é a substituição de um único nucleotídeo, denominada Polimorfismos de Nucleotídeo Único (SNP) (CHAKRAVARTI, 2001). Uma mutação pontual pode ter efeitos variados, desde ser neutra, ou seja, não ter impacto na função da proteína codificada, até causar doenças genéticas hereditárias ou contribuir para a evolução de características complexas. As mutações são consideradas a base da variabilidade genética no processo de seleção natural.

SNPs são marcadores moleculares amplamente utilizados em Genômica Animal (ANDERSSON, 2001), representam variações na sequência de DNA que envolvem a substituição de um único nucleotídeo por outro em uma determinada posição do genoma, caracterizados como a forma mais comum de variação genética entre indivíduos de uma mesma espécie. Podem ocorrer em várias regiões do genoma, incluindo regiões codificadoras (exons) e regiões não codificadoras (introns) (MORIN; AL., 2004). A principal característica dos SNPs é que eles são bialélicos, ou seja, existem apenas duas variantes alélicas possíveis em uma determinada posição do genoma (BROOKES, 1999). Ao longo do tempo têm sido fundamentais em estudos de associação genômica (HIRSCHHORN; DALY, 2002), mapeamento de loci de características de interesse produtivo (GODDARD; HAYES, 2009) e genética de populações (LUIKART; AL., 2003).

O sequenciamento de DNA é uma ferramenta fundamental na genômica, permitindo a análise da sequência de nucleotídeos no genoma de um organismo (SHENDURE; AL., 2008). Uma abordagem eficaz e de alto rendimento para a genotipagem de indivíduos é o uso do SNP Chip.

Os SNP Chips, ou microarranjos de SNPs (Polimorfismos de Nucleotídeo Único), são uma tecnologia que permite a genotipagem de milhares a milhões de SNPs a partir de uma única amostra de DNA (OLIPHANT; AL., 2002). Os chips contêm sondas de DNA complementares para as variantes alélicas dos SNPs de interesse. O processo de sequenciamento envolve a extração do DNA da amostra, a preparação do DNA e a hibridização com o chip. Durante a hibridização, o DNA da amostra se liga às sondas no chip, revelando as variantes alélicas presentes no genoma do indivíduo (FAN; AL., 2006).

2.2 **Seleção**

A Seleção é um dos componentes principais do Melhoramento Genético Animal. É uma estratégia pela qual os animais são escolhidos como reprodutores a partir de critérios genéticos específicos. Visa aprimorar características desejadas no rebanho, como produção de leite, ganho de peso, resistência a doenças, conformação, eficiência reprodutiva, dentre outras (SMITH, 2005).

Os critérios de seleção são fundamentais para o sucesso do processo de melhoramento. Eles são baseados nas características de interesse, muitas das quais são quantitativas e complexas. Essas características podem ser influenciadas por características baseadas em múltiplos genes e fatores ambientais, bem como suas complexas interações. Os índices de seleção combinam várias características de importância relativa e variável em apenas um único valor, permitindo uma abordagem mais ampla na seleção dos animais (JACKSON; RESENDE; RESENDE, 2017; VANRADEN, 2016).

Vários métodos de seleção têm sido desenvolvidos ao longo do tempo, cada um deles com seus próprios objetivos, vantagens e desafios. A partir do contexto de seleção em genética animal é possível citar quatro métodos: seleção por antepassados, seleção por parentes, seleção por progênie e seleção individual (GODDARD, 2009; MEUWISSEN; HAYES; GODDARD, 2001).

A seleção por antepassados, ou seleção histórica, é baseada na análise das características dos animais anteriores de uma linhagem. Essa abordagem é interessante quando há informações completas sobre o desempenho de animais anteriores, como pais, avós e bisavós. A ideia intrínseca é que animais com antepassados bem-sucedidos têm uma maior probabilidade de herdar características desejáveis. No entanto, a seleção por antepassados possui certas limitações, uma vez que supõe que o ambiente de criação permanece sem alterações ao longo das gerações (GARRICK; TAYLOR; FERNANDO, 2014).

A seleção por parentes envolve a avaliação do desempenho de parentes próximos de um animal, como irmãos, meio-irmãos ou parentes de primeira geração. Essa abordagem é útil quando não há informações muito detalhadas sobre antepassados ou quando a variabilidade genética na população é limitada. A seleção por parentes explora a relação genética entre os animais e pode ser eficaz na identificação de características desejáveis para serem transmitidas geneticamente (MACKAY; LYMAN, 2001).

A seleção por progênie é um método que se concentra na avaliação das características dos descendentes de um determinado animal, relevante em rebanhos com registros detalhados de progênie, permitindo uma análise precisa do desempenho dos descendentes. A seleção por progênie é importante na identificação dos animais que transmitem as melhores características para suas crias, no entanto pode ser demorada devido à necessidade de acompanhar o desempenho das gerações futuras (HAZEL, 1943; VANRADEN; AL.,

2007).

A seleção individual é o método mais direto, foca na avaliação das características de cada animal individualmente, não dependendo das informações de parentesco ou antepassados. Cada animal é avaliado com base em critérios específicos, e os melhores são selecionados para reprodução. A seleção individual permite um controle mais direto sobre as características desejadas, mas requer a disponibilidade de dados detalhados e a capacidade de mensurar com precisão as características de interesse (GIANOLA; FOULLEY, 1986; VANRADEN; O'CONNELL; WIGGANS, 2008).

2.3 **Linhagens**

Linhagens, no contexto do melhoramento genético animal, referem-se a grupos de animais que compartilham ancestrais comuns, próximos ou distantes, e características genéticas específicas. Essas características podem incluir características produtivas, morfológicas e reprodutivas. Geralmente identificadas com base em pedigrees (árvores genealógicas) que rastreiam a ascendência dos animais ao longo das gerações. Cada linhagem é representada por animais que demonstram características genéticas semelhantes e são descendentes diretos de ancestrais comuns.

O estudo das linhagens desempenham diversos papeis na produção animal:

- Conservação da genética: linhagens bem definidas ajudam a conservar e preservar características genéticas específicas. Isso é fundamental para a manutenção de raças de interesse produtivo e também a conservação da diversidade genética (GROSZ; KAKI, 2007).
- Aprimoramento de características: linhagens com características desejáveis podem ser selecionadas para reprodução seletiva, aprimorando o desenvolvimento dessas características no rebanho ao longo das gerações (GJEDREM, 2009).
- Avaliação de desempenho: através da avaliação do desempenho dos animais de uma linhagem específica, é possível determinar quais indivíduos são os mais adequados para a reprodução, impulsionando características-chave (VANRADEN et al., 2004).
- Transparência na seleção: a identificação de linhagens permite uma seleção mais transparente e direcionada. Os criadores podem escolher animais com base nas linhagens e características desejadas (DAVIS, 2008).
- Genética de populações: linhagens são cruciais para estudos de genética de populações, permitindo a análise da estrutura genética e relacionamentos dentro de uma população de animais (FALCONER; MACKAY, 1996).

A correta identificação de linhagens possibilita a seleção de animais reprodutores que possuam características produtivas de interesse econômico e bem estar animal, possibilitando a manutenção da diversidade genética necessária para manutenção dos ganhos de produção desejados, controlando a endogamia e evitando tanto doenças quanto anomalias (VANRADEN et al., 2004).

3 APRENDIZADO DE MÁQUINA

Assim como toda área do conhecimento, o aprendizado de máquina possui seus principais fundamentos, notações e nomenclaturas. Esta seção procura introduzir o assunto através de uma abordagem pela visão da modelagem matemática e computacional, entretanto o tema já vem sendo estudado amplamente e discutido a vários anos pela estatística inferencial (VAPNIK, 1998). A grande retomada dos conceitos e aplicações é motivada principalmente pelo aumento da capacidade de processamento, armazenamento e geração de dados para análise, onde termos como aprendizado de máquina, *Big Data*, *Deep Learning*, entre outros, vêm se tornando cada vez mais parte do dia a dia de um profissional da área tecnológica (LECUN; BENGIO; HINTON, 2015).

O aprendizado de máquina tem como objetivo obter conclusões genéricas através de um conjunto particular de dados, ou seja, a obtenção de informações sobre os parâmetros da distribuição de probabilidades da população a partir das estatísticas das amostras disponíveis. Desta forma, é empregado um princípio de inferência denominado indução, e este aprendizado indutivo pode ser dividido essencialmente em: supervisionado e não-supervisionado (MITCHELL, 1997).

No aprendizado supervisionado, a figura de um professor externo apresenta o conhecimento inerente ao ambiente através de conjuntos de exemplos. Desta forma, o modelo desenvolvido deve ser capaz de produzir saídas corretas para entradas não conhecidas anteriormente.

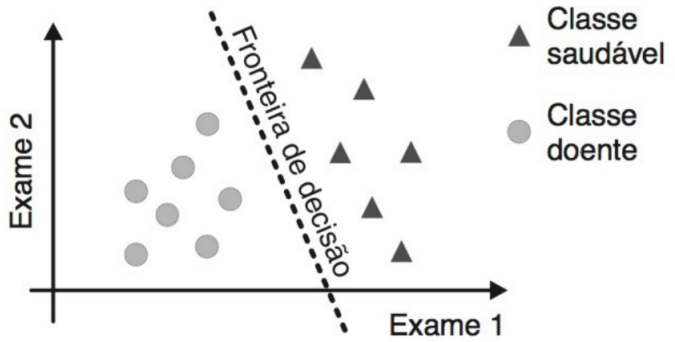
Entretanto, no aprendizado não-supervisionado, não há treinamento por exemplos através da figura de um professor, desta forma o algoritmo busca agrupar as entradas utilizando uma medida de qualidade pré-determinada, como por exemplo, distância, similaridade, dissimilaridade, densidade, hierarquia, entre outras. Técnicas desde tipo são utilizadas principalmente quando é desejada a identificação de tendências ou padrões que expliquem o comportamento analisado.

No presente trabalho, são explicadas e utilizadas técnicas de aprendizado supervisionado, onde buscamos encontrar os parâmetros ótimos de ajuste para os modelos, de forma que este seja capaz de prever rótulos desconhecidos em amostras não vistas anterioremente em seu processo de treinamento.

Essencialmente, se o rótulo for determinado através de um conjunto finito, normalmente dividido em classes, a tarefa realizada é denominada classificação, entretanto, se os rótulos forem representados por números reais tem-se um problema de regressão.

Na Figura 1 é possível perceber duas regiões predominantes entre as duas classes de dados, representadas pelas figuras de círculos e triângulos, sendo facilmente separadas através de uma reta (um hiperplano em mais dimensões) traçada entre elas. Este é um exemplo típico de um problema de classificação.

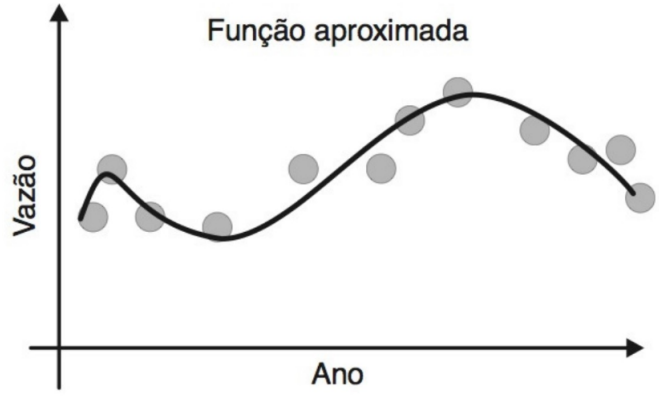
Figura 1 – Exemplo de um problema de classificação.



Fonte: adaptado de (GAMA et al., 2011)

Quando os rótulos são números reais e representam o comportamento de uma curva, como os pontos ilustrados na Figura 2, temos um caso clássico de regressão.

Figura 2 – Exemplo de um problema de regressão.



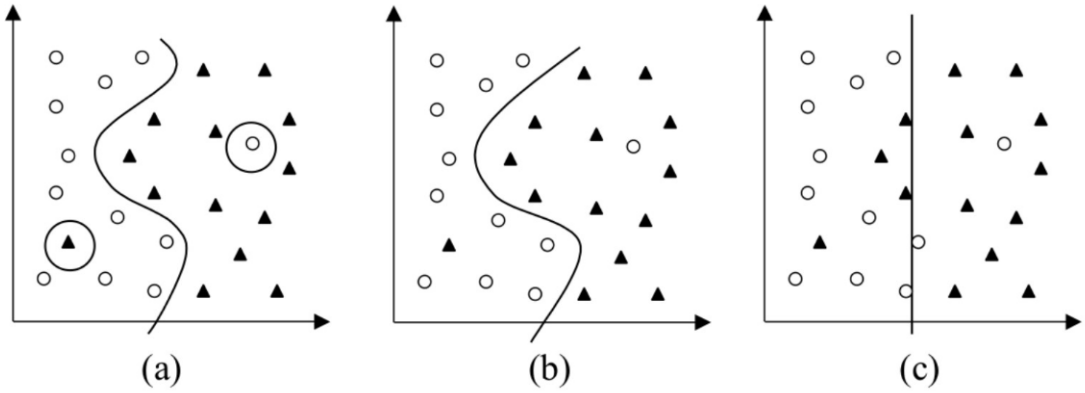
Fonte: adaptado de (GAMA et al., 2011)

Os modelos de predição que utilizam aprendizado supervisionado tem como principal objetivo determinar corretamente os rótulos para dados nunca vistos anteriormente tomando como base exemplos em que os rótulos foram previamente determinados, como detalhado nas seções anteriores (MITCHELL, 1997).

No entanto, alguns problemas podem surgir no decorrer deste processo. Como por exemplo, a partir do conceito de generalização, a capacidade de prever corretamente os rótulos dos novos dados, o modelo pode se especializar nos dados utilizados em seu treinamento, apresentando uma baixa taxa de acertos quando confrontado com novas entradas, o que é denominado super-ajustamento ou *overfitting*. Quando o conjunto de treinamento não é muito significativo ou a técnica utilizada é muito simples, podemos ter um baixa taxa de acertos nos dados já conhecidos, ou seja, um sub-ajustamento ou *underfitting*. Considerando o exemplo de conjunto de treinamento ilustrado na Figura 3 podemos verificar os diferentes formatos das fronteiras de decisão nos casos ideal (b),

overfitting (a) e *underfitting* (c) (MITCHELL, 1997).

Figura 3 – Diferentes formas de fronteiras de decisão.



Fonte: adaptado de (GAMA et al., 2011)

Com objetivo de minimizar esses problemas e estabelecer condições matemáticas que auxiliem na escolha de um classificador para um problema em particular, foi desenvolvida a Teoria do Aprendizado Estatístico (TAE), onde são levados em conta o desempenho no conjunto de treinamento e sua complexidade associada (GAMA et al., 2011).

Nas próximas seções deste capítulo, alguns conceitos da TAE são inseridos no decorrer do texto, buscando ilustrar o desenvolvimento do campo do aprendizado de máquina ao longo do tempo, bem como justificar os argumentos por trás da criação de modelos como a máquina de vetor suporte a partir do problema de maximização de margem encontrado no perceptron.

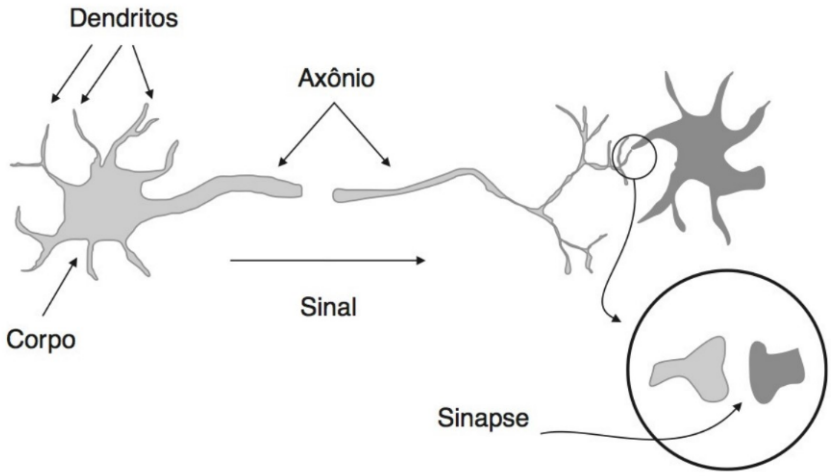
3.1 **Aprendizado Supervisionado**

3.1.1 Redes Neurais Artificiais

Segundo Lent (LENT, 2010), uma rede neural artificial é uma estrutura bioinspirada, por isso inicialmente o funcionamento do neurônio biológico como unidade fundamental de processamento de informação no sistema nervoso central dos eucariotos deve ser compreendido. O neurônio biológico é descrito como uma estrutura celular que responde a estímulos elétricos e pode ser divido em três partes principais: dendritos, corpo e axônio, como ilustrado na Figura 4.

Cada componente do neurônio possui uma função distinta para seu funcionamento final. Os dentritos, pequenas ramificações, captam os sinais de entrada, que são recebidos, combinados e processados no corpo, gerando uma resposta de saída propagada através de uma longa estrutura para outras células, o axônio. A comunicação interneural é realizada através de sinapses químicas, onde a liberação de substâncias neurotransmissoras excitam

Figura 4 – Topologia simplificada do neurônio biológico.



Fonte: adaptado de (LENT, 2010)

ou inibem a propagação dos sinais elétricos entre o axônio de um nerônio e os dentritos de outro neurônio.

O neurônio pode ser compreendido como o componente básico para o funcionamento do cérebro. Estes se combinam em redes complexas, atuando em atividades sensoriais e motoras.

Em 1943, McCulloch e Pitts propuseram um modelo matemático de neurônio artificial, no qual cada neurônio era uma unidade lógica simples, podendo executar funções diferentes. Eles mostraram que a combinação de vários neurônios artificiais possui um grande poder computacional, atuando como um aproximador para funções que possam ser representadas como uma combinação linear de várias funções (HAYKIN, 2007). No entanto, essas redes iniciais não possuíam capacidade de aprendizado.

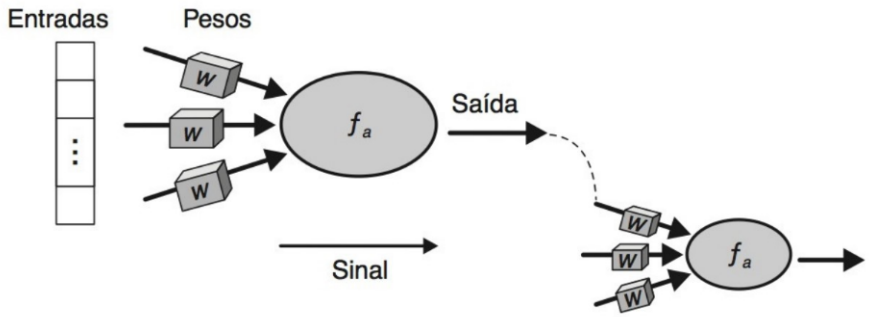
Rosenblatt, nas décadas de 1950 e 1960, desenvolveu uma modelagem matemática sobre o funcionamento do neurônio baseado nos trabalhos citados anteriormente. Atualmente não é tão utilizada, mas permite uma clara compreensão do componente de unidade de uma rede neural em termos matemáticos (HAYKIN, 2007).

De acordo com Gama et al., (2011) na Figura 5 é apresentado um modelo simples para o neurônio artificial. Cada terminal de entrada, mimetizando os dendritos, recebe um valor, que são combinados e ponderados por função matemática, gerando a resposta de saído do neurônio.

Supondo uma entrada X com n atributos, tem-se o vetor X = $[x_1, x_2, \ldots, x_n]^T$, e um neurônio também com n entradas com os pesos representados vetorialmente como W=$[w_1, w_2, \ldots, w_n]$. Definindo a equação,

$$u = \sum_{i=1}^{n} x_i w_i \tag{3.1}$$

Figura 5 – Topologia do neurônio artificial.

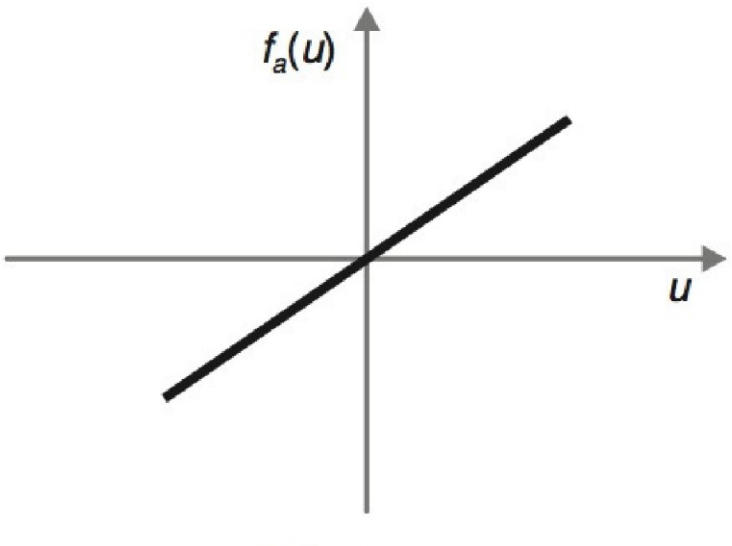


Fonte: adaptado de (GAMA et al., 2011)

A saída final do neurônio é obtida pela aplicação de uma função de ativação após o processo de combinação definido pela Equação 3.1. Existem diversas funções propostas na literatura, dentre elas se destacam a linear, limiar e sigmoidal (GAMA et al., 2011).

A função linear identidade possui a característica de retornar como saída os mesmos valores de entrada, como na Figura 6.

Figura 6 – Função de ativação linear.

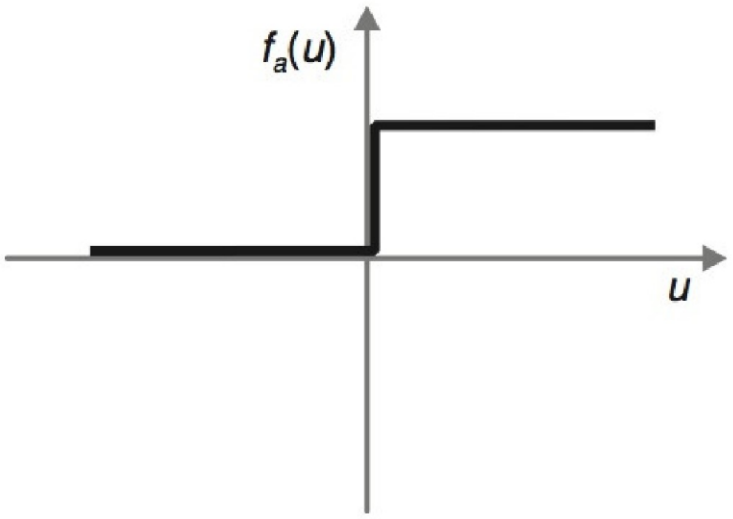


Fonte: adaptado de (GAMA et al., 2011)

Na função limiar representada na Figura 7 e empregada no modelo de neurônio artificial de McCulloch e Pitts, o valor escolhido como limiar define quando a saída será ativa ou não. Quando a soma das entradas recebidas ultrapassa o limiar estabelecido, a saída do neurônio possui o valor 1, no caso contrário, o valor 0, possuindo também alternativamente a versão com saída 1 e -1.

Na função sigmoidal, podem ser empregadas diferentes inclinações, representando essencialmente uma aproximação diferenciável e contínua da função limiar, como represen-

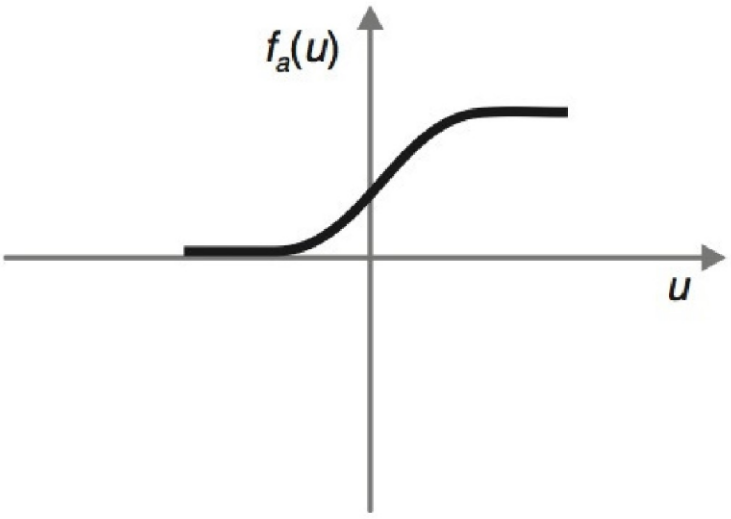
Figura 7 – Função de ativação limiar.



Fonte: adaptado de (GAMA et al., 2011)

tado na Figura 8.

Figura 8 – Função de ativação sigmoidal.



Fonte: adaptado de (GAMA et al., 2011)

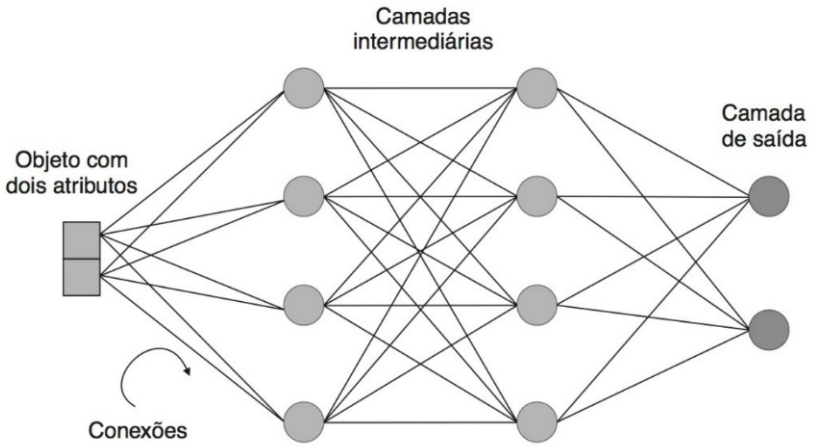
No treinamento supervisionado, um conjunto de entradas rotuladas estão disponíveis para a realização da modelagem, regendo o direcionamento na definição dos pesos associados a cada entrada.

Segundo Gama et al. (2011), em uma rede neural artificial, os neurônios podem estar dispostos em uma ou mais camadas. No caso em que duas ou mais camadas são utilizadas, é comum denominar a estrutura obtida como Perceptron Multicamadas, do inglês *Multilayer Perceptron* (MLP). Nesta configuração, os valores de entrada de um

neurônio das camadas interiores são as saídas obtidas na camada anterior, gerando uma rede de unidades simples de processamento.

Na Figura 9 está presente uma rede de três camadas, composta por duas camadas intermediárias, também denominadas ocultas, escondidas ou interiores, e uma camada de saída, com o mesmo número de atributos presentes na entrada. Com o aumento da quantidade de camadas, a complexidade do modelo cresce e métodos mais sofisticados de determinação dos pesos associados as redes neurais se tornam necessários. Várias estratégias foram propostas na literatura, sendo a principal e mais difundida o algoritmo de *backpropagation*, onde os erros em cada camada são utilizados com objetivo de recalibrar os pesos associados, partindo inicialmente de valores aleatórios (LECUN; BENGIO; HINTON, 2015).

Figura 9 – Rede neural com três camadas.



Fonte: adaptado de (GAMA et al., 2011)

3.1.2 Máquinas de Vetor Suporte

Na classificação utilizando o perceptron simples, a operação de treinamento ajusta um hiperplano separador entre as classes. No entanto, esta operação é realizada através de uma lei de atualização iterativa com critério de parada baseado na primeira iteração que separa corretamente os dados de acordo com os rótulos previamente definidos.

Este hiperplano separador pode não ser o mais eficiente na tarefa de prever corretamente as classes para dados não conhecidos, desta forma torna-se necessário o estudo de características que auxiliem na escolha do melhor classificador para uma determinada base de dados.

A Teoria do Aprendizado Estatístico é responsável por definir as condições matemáticas para a obtenção do melhor classificador a partir dos dados de treinamento.

Suposições sobre a geração de forma independente e a distribuição de probabilidades dos dados identicamente distribuídas necessitam ser realizadas para iniciar o

desenvolvimento da TAE.

O erro, ou risco esperado é definido pela Equação 3.2 com objetivo de medir a capacidade de generalização de um classificador em particular para os dados de teste. A integral a ser calculada depende de uma função de custo, normalmente retornando 1 quando corretamente classificado e 0 caso contrário (JAMES et al., 2013).

$$R(f) = \int c(f(x), y) dP(x, y) \tag{3.2}$$

Uma vez que intrinsecamente a função dos classificadores é estimar a distribuição de probabilidade P(x,y) que associa os dados aos seus rótulos, não é possível minimizar diretamente o risco esperado através da Equação 3.2. Contudo, a partir dos dados de treinamento, também amostrados de P(x,y) é possível utilizar o princípio de indução para inferir uma função que minimize o erro entre esses dados, esperando que obtenha-se um menor erro sobre os dados de teste. Torna-se necessária então a medição do desempenho do classificador nos dados de treinamento, denominada risco empírico e quantificada pela Equação 3.3 (JAMES et al., 2013).

$$R_{\text{emp}}(f) = \frac{1}{n} \sum_{i=1}^{n} c(f(x_{\text{i}}), y_{\text{i}}) \tag{3.3}$$

Com base na capacidade do conjunto de classificadores, ou seja, o quão mais complexas podem ser as funções de classificação induzidas para um dado problema, é possível estabelecer condições para que o algoritmo de aprendizado garanta a obtenção de classificadores em que o risco empírico convirja assintoticamente para o risco esperado. Tal parâmetro é estimado utilizando a dimensão de Vapnik-Chervonenkis descrita por Vapnik (VAPNIK, 1998). Para a escolha do melhor classificador, existem resultados que associam o risco esperado ao conceito de margem (VAPNIK, 1995). A margem de um classificador é definida com a relação a distância de um exemplo à fronteira de decisão induzida. Para um problema binário, por exemplo, no qual o rótulo é definido como 1 ou -1, os exemplos são denominados x_{i}, e a função f, tem-se que a margem com que o dado é classificado por f pode ser calculada pela Equação 3.4, onde um valor positivo de margem denota uma classificação correta e um valor negativo uma classificação incorreta.

$$\delta(f(x_{\text{i}}), y_{\text{i}}) = y_{\text{i}} f(x_{\text{i}}) \tag{3.4}$$

Novamente é possível estabelecer utilizando o conceito de margem como uma relação entre o risco esperado, o erro marginal, definido como a proporção de exemplos de treinamento cuja margem de confiança é inferior a uma determinada constante positiva, e um termo de capacidade. Uma maior margem implica diretamente em um menor termo de capacidade. No entanto, uma margem muito alta pode levar a erros de classificação, pois torna-se difícil atender o critério para a distância de todos os exemplos em relação

ao hiperplano separador induzido. Deve-se buscar então na geração de um classificador linear, um hiperplano que possua margem elevada e cometa poucos erros de treinamento, minimizando simultaneamente o erro entre os dados de teste e treinamento, obtendo assim o hiperplano denominado ótimo (VAPNIK, 1998).

Como uma aplicação direta dos resultados obtidos pela TAE surgiram as SVMs (CORTES; VAPNIK, 1995). Lidando inicialmente com problemas onde a fronteira de decisão é linearmente separável, são desenvolvidas duas abordagens que caracterizam a SVM linear: margem rígida e margem flexível.

Nas SVMs lineares com margens rígidas, as fronteiras de decisão são construídas a partir de dados linearmente separáveis. Como resultado, são obtidos hiperplanos separadores, com a seguinte fórmula geral:

$$f(\mathbf{x}) = \mathbf{w} \cdot \mathbf{x} + b = 0 \tag{3.5}$$

Desta forma, o espaço fica subdividido em duas regiões, onde a utilização de uma função sinal nos dá as classificações de maneira resumida:

$$+1, \mathbf{w} \cdot \mathbf{x_i} + b > 0 \tag{3.6}$$

$$-1, \mathbf{w} \cdot \mathbf{x_i} + b < 0 \tag{3.7}$$

Resultando em,

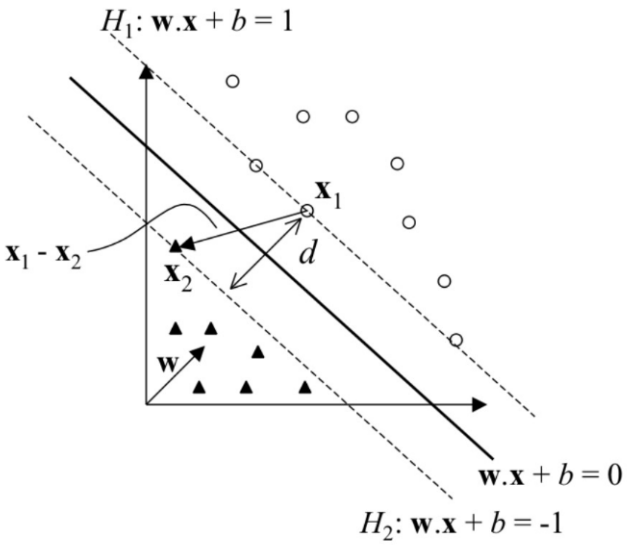
$$y_\mathrm{i}(\mathbf{w} \cdot \mathbf{x_i} + b) - 1 \geq 0 \tag{3.8}$$

Para se obter o hiperplano ótimo, ou seja, de maior margem, devemos maximizar a distância entre o hiperplano e os pontos mais próximos por classe, como ilustrado na Figura 10. O que resulta em um problema de otimização quadrático, resolvido por Boser (BOSER; GUYON; VAPNIK, 1992) através do uso da função Lagrangeana em sua forma dual e primal, obtendo os parâmetros do hiperplano.

Em aplicações reais, dificilmente são encontrados dados linearmente separáveis, assim torna-se necessária a obtenção de fronteiras de decisão que possam lidar melhor com os dados em análise (GAMA et al., 2011).

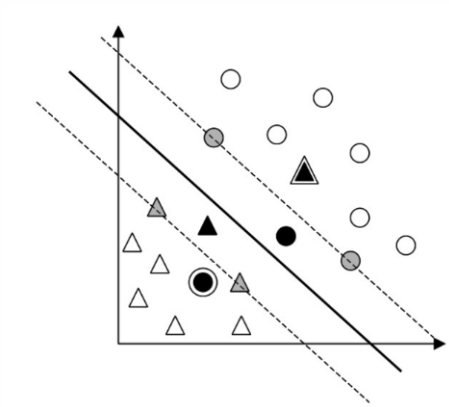
Para resolver esta situação, como descrito por Smola et al. (1999), é utilizada uma variável de folga na Equação 3.8, resultando em outro problema de otimização, que é resolvido de maneira análoga para a margem rígida. Esta estratégia caracteriza a SVM de margem suave, ilustrada pela Figura 11, onde é possível o tratamento de dados com alguns ruídos e/ou *outliers*.

Figura 10 – Hiperplano ótimo



Fonte: adaptado de (GAMA et al., 2011)

Figura 11 – SVM margem suave

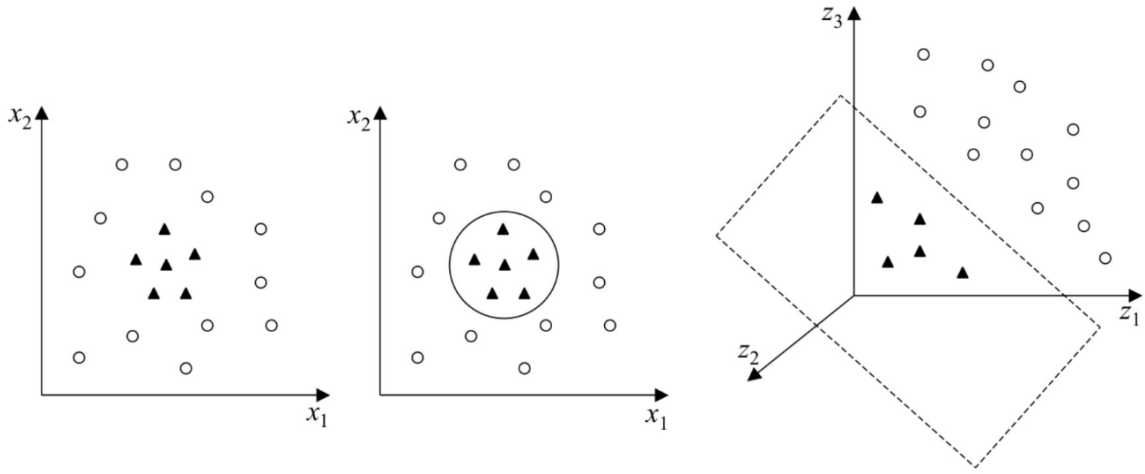


Fonte: adaptado de (GAMA et al., 2011)

Há casos onde a flexibilização de margem não é o tratamento mais adequado, pois as fronteiras de decisão possuem comportamento altamente não linear, o que motivou o desenvolvimento de mapeamentos do conjunto de treinamento original para um novo espaço de maior dimensão, denominado espaço de características (SMOLA et al., 1999).

A ilustração do funcionamento do mapeamento para o caso de uma fronteira circular pode ser observado na Figura 12. As funções que realizam esse mapeamento são denominadas na literatura funções *kernel* (SMOLA; SCHOLKOPF, 2002), sendo as mais utilizadas na prática sumarizadas na Tabela 1.

Figura 12 – Mapeamento no espaço de características



Fonte: adaptado de (GAMA et al., 2011)

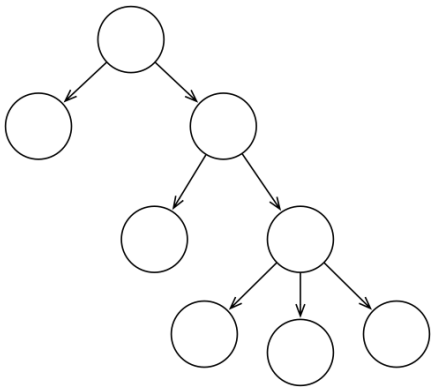
Tabela 1 – Funções *kernel* mais comuns.

Tipo	**Função K(x_i,y_i)**	**Parâmetros**
Polinomial	$(\delta(x_i \cdot y_i)) + \kappa)^d$	δ, κ e d
Gaussiano	$exp(-\sigma\|\|x_i - x_j\|\|^2)$	σ
Sigmoidal	$tanh(\delta(x_i \cdot y_i)) + \kappa)$	δ e κ

3.1.3 Random Forest

Uma árvore pode ser compreendida como uma coleção de nós sujeitos a uma relação hierárquica de "paternidade", onde o nó inicial é denominado raiz, e os seguintes podem ser definidos recursivamente como nós pais e nós filhos como ilustrado pela Figura 13 (GAMA et al., 2011).

Figura 13 – Exemplo de uma árvore



Fonte: elaborada pelo autor

No modelo de Árvore de Decisão cada um dos nós representa um ponto de decisão que é construído de acordo com os valores correspondentes a cada uma das variáveis

presentes na base de dados, sendo interpretado como divisões no espaço de características (JAMES et al., 2013).

O conceito de Entropia pode ser compreendido como um cálculo de ganho de informação baseado em uma medida, muito utilizado na Teoria da Informação. Em um conjunto de dados é uma medida da falta de homogeneidade dos dados de entrada em relação a classificação das amostras, sendo máxima quando os dados são heterogêneos (SHANNON, 1948).

Na construção de uma Árvore de Decisão são perseguidos três principais objetivos: minimizar a entropia, possuir consistência com o conjunto de dados na modelagem e possuir o menor número possível de nós (QUINLAN, 1986).

Desenvolvido em 1912, o Índice Gini atua também na medição do grau de heterogeneidade citado anteriormente. Gerando assim, valores de impureza para um determinado nó, quanto mais próximo de zero, mais puro é este nó. Ao se aproximar do valor um, é denominado impuro, aumentando o número de classes uniformemente distribuídas deste nó (GINI, 1912).

O cálculo do índice de Gini é definido como:

$$Gini(S) = 1 - \sum_{i=1}^{c}(p_i)^2 \tag{3.9}$$

onde $Gini(S)$ é o índice de Gini do conjunto de dados S, c é o número de classes, p_i é a proporção de exemplos da classe i.

Ao utilizar a Entropia em árvores de decisão com partições binárias, resulta em um maior balanceamento no número de registros em cada ramo, já seguindo com o índice de GINI, a tendência dos registros das classes mais frequentes se isolarem em ramos é maior.

No treinamento da árvore através da recursão de partições, a profundidade da árvore é expandida até que modele completamente o conjunto de dados de treinamento. Quando há ruído no conjunto, ou não é bem representativo, o processo pode se especializar excessivamente nos dados, gerando *overfitting*. O número ótimo de nós pode ser obtido graficamente através do erro percentual no treinamento e o número de nós terminais, buscando o momento em que o erro começa a crescer. Para prevenção do problema é indicado como solução realizar a poda da árvore.

A técnica Random Forest (RF) pode ser compreendida como um conjunto de Árvores de Decisão que são treinadas com amostras aleatórias do conjunto de dados. As partições de cada árvore da floresta são feitas com base em subconjuntos aleatórios, contando também amostras repetidas, objetivando criar árvores com pequenas diferenças (JAMES et al., 2013).

É importante notar que pela característica das árvores da floresta serem distintas

entre si, o caminho de decisão percorrido, bem como os atributos considerados na divisão de cada nó não são os mesmos, aumentando a probabilidade de atingir nós folhas de diferentes classes. As classificações são realizadas pelo voto da maioria das árvores, minimizando o *overfitting* comparado a utilização de uma única Árvore de Decisão.

Toda árvore recebe um subconjunto de dados aleatório e único para o processo de indução. Por exemplo, dado um conjunto de treino com elementos enumerados de 0 à 9, onde 3/4 dos dados são utilizados e o restante é excluído. Um possível subconjunto poderia ser: 0, 1, 3, 5, 7, 9, note que alguns elementos não foram utilizados. O subconjunto final deve ter tamanho igual ao do conjunto de treino original, desta forma, são repetidos aleatoriamente elementos até obter tamanhos iguais. Desta forma, um possível subconjunto válido seria: 0, 0, 1, 3, 5, 5, 7, 7, 9, 9. Este processo é denominado bootstraping.

3.2 **Aprendizado Não Supervisionado**

3.2.1 K-modes

K-means é um algoritmo de clusterização que objetiva agrupar dados em K clusters, onde K é um valor predefinido (JAIN; MURTY; FLYNN, 2010). Matematicamente, o algoritmo busca minimizar a função de custo, definida como a soma dos quadrados das distâncias entre os pontos de dados e os centroides dos clusters correspondentes:

$$J = \sum_{i=1}^{K} \sum_{x \in C_i} ||x - \mu_i||^2 \tag{3.10}$$

Aqui, C_i representa o conjunto de pontos no i-ésimo cluster e μ_i é o centróide do i-ésimo cluster.

A minimização da função é realizada por meio de iterações que atualizam os centroides e reavaliam cada um dos pontos de dados como pertencentes aos clusters até atingir a convergência. Resultando em separações de dados com grande dependência da inicialização aleatória ou específica dos centróides.

A atualização dos centroides é realizada da seguinte maneira:

$$\mu_i = \frac{1}{|C_i|} \sum_{x \in C_i} x \tag{3.11}$$

Essa equação calcula o centróide do i-ésimo cluster como a média dos pontos de dados dentro desse cluster. O processo de alocação de pontos de dados a clusters envolve a minimização da distância euclidiana entre os pontos e os centroides dos clusters.

O algoritmo K-modes é uma extensão do K-means desenvolvida para trabalhar com dados categóricos, ao contrário do foco em dados numéricos do K-means (JAIN; MURTY; FLYNN, 2010). Ao invés de realizar a minimização da soma dos quadrados das distâncias

euclidianas, o K-modes busca minimizar a dissimilaridade entre dados categóricos. Isso é feito através de uma métrica de dissimilaridade, como a distância de Hamming (NOROUZI; FLEET; SALAKHUTDINOV, 2012), que avalia a diferença entre duas variáveis contando o número de posições em que elas diferem. A distância de Hamming é definida como:

$$d_H(x, y) = \sum_{i=1}^{n} \delta(x_i, y_i) \tag{3.12}$$

onde x e y são os vetores de características categóricas a serem comparados, n é o número de características e $\delta(x_i, y_i)$ é uma função delta que retorna 0 quando $x_i = y_i$ (características iguais) e 1 quando $x_i \neq y_i$ (características diferentes).

3.2.2 Cluster Hierárquico

O cluster hierárquico é uma técnica que organiza dados em uma estrutura de árvore ou dendrograma, onde os clusters são formados de forma aninhada (JAMES et al., 2013). Existem dois principais focos de construção: o aglomerativo e o divisivo .

No método aglomerativo, cada ponto dos dados é inicialmente considerado como um cluster individual. Os clusters são gradualmente mesclados com base na distância entre eles. A medição dessa distância pode ser realizada de várias formas, a distância euclidiana é um exemplo. A aglomeração dos clusters é construída através da atualização da matriz de ligação (ou *linkage matrix*), que representa as distâncias entre todos os clusters em cada um dos passos. Esse processo continua até que todos os pontos de dados estejam em um único cluster, formando uma hierarquia de clusters aninhada (KARYPIS; HAN; KUMAR, 1999).

Suponha quatro pontos de dados A, B, C e D, com as seguintes distâncias iniciais:

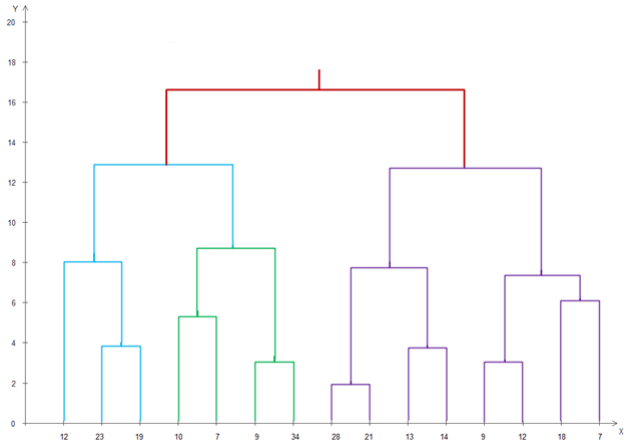
$$\begin{array}{c|cccc} & A & B & C & D \\ A & - & 4 & 5 & 8 \\ B & & - & 3 & 7 \\ C & & & - & 6 \\ D & & & & - \end{array} \tag{3.13}$$

No primeiro passo, os pontos B e C são os mais próximos, logo eles são mesclados em um novo cluster, BC, e a matriz de distâncias é atualizada:

$$\begin{array}{c|ccc} & A & BC & D \\ A & - & 4.5 & 8 \\ BC & & - & 6 \\ D & & & - \end{array} \tag{3.14}$$

Este processo é repetido até que todos os pontos de dados estejam em um único cluster, formando um dendrograma que ilustra a hierarquia de clusters aninhados, como na Figura 14.

Figura 14 – Exemplo de um dendrograma



Fonte: adaptado de (GAMA et al., 2011)

No método divisivo é realizado o processo completamente oposto, todos os dados são considerados inicialmente em um único cluster. A análise tem início pela divisão desse cluster em clusters menores. Isso é realizado através da divisão da matriz de ligação, de forma que a matriz original é subdividida em submatrizes que representam clusters cada vez menores. Esse processo de divisão continua até que cada ponto do conjunto de dados esteja em um cluster separado (KARYPIS; HAN; KUMAR, 1999).

Suponha um único cluster contendo os pontos de dados A, B, C e D. A métrica de dissimilaridade identifica que os pontos B e C são menos semelhantes aos pontos A e D. Desta forma, a divisão tem início com a criação de dois clusters: A, D e B, C. Os clusters menores criados podem posteriormente ser divididos em outros enquanto o processo continuar.

4 METODOLOGIA PROPOSTA

4.1 DADOS

As amostras de genótipos utilizadas neste trabalho foram disponibilizadas pela Empresa Brasileira de Pesquisa Agropecuária (Embrapa), unidade Gado de Leite. Os animais que originaram os dados são bovinos zebuínos da raça Gir Leiteiro e fazem parte do Programa Nacional de Melhoramento do Gir Leiteiro (PNMGL), desenvolvido pela Embrapa em parceria com a Associação Brasileira dos Criadores de Zebu (ABCZ) e Associação Brasileira dos Criadosres de Gir Leiteiro (ABCGIL).

Os SNP chips utilizados variam quanto ao conjunto de marcadores selecionados e a densidade. Foram obtidos dados de 5 diferentes fontes: *Illumina Bovine HD BeadChip* com 777K (HD), *GeneSeek Genomic Profiler Indicus BeadChip* com 35K e 54K (GGP Ind 35K e 54K), e *GeneSeek Genomic Profiler Indicus-LD BeadChip* com 27K (ZChip). Em todos os chips o mapa utilizado é referente ao mapeamento do genoma bovino ARS-UCD1.2 (GenBank: NKLS00000000.2). Na Tabela 2 é possível verificar a quantidade de animais genotipados por cada um dos conjuntos.

Tabela 2 – Animais por SNP Chip.

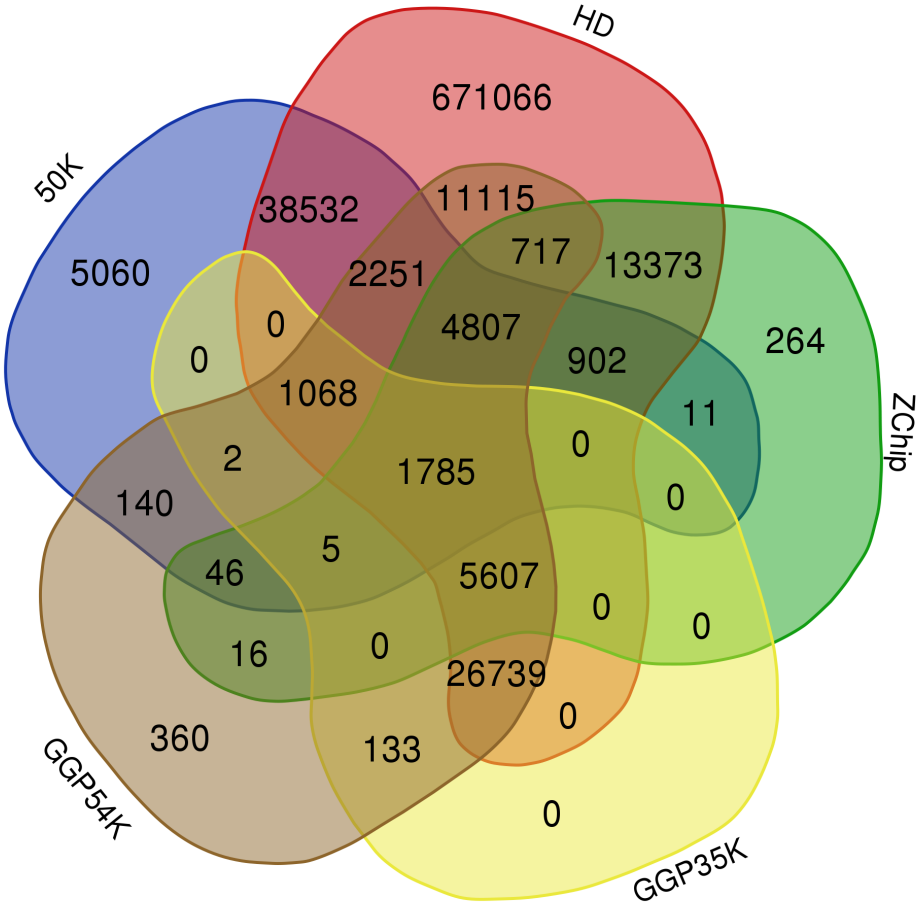
SNP Chip	Número de Animais
50K	1652
GGP Ind 35K	11238
GGP Ind 54K	1744
HD	598
ZChip	1617

O controle de qualidade dos genótipos foi realizado utilizando o pacote SNPstats (SOLE et al., 2006) e o *software* PLINK (PURCELL et al., 2007). Foram mantidas somente as amostras com uma *call rate* maior que 0.9, *GC score average* maior que 0.7, desvio do equilíbrio de Hardy-Weinberg (P > 10-6) e *minor alele frequency* (MAF) maior que 0.05.

Dado que cada SNP chip possui uma combinação diferente de marcadores, torna-se necessário uma avaliação das melhores composições de amostras na construção da base de dados para o estudo. Com objetivo de conservar apenas SNPs provenientes originalmente do processo de sequenciamento, ou seja, sem considerar imputação de valores, inicialmente foi realizada uma comparação de sobreposição entre cada um dos SNP chips, presente na Figura 15. É possível perceber que o chip HD possui um grande número de marcadores em comum com todos os outros SNP chips, no entanto, possui poucas amostras, o que traria a necessidade da utilização de imputação de valores.

Com objetivo de utilizar apenas as informações originais disponibilizadas através da combinação de marcadores de cada SNP chip, foi realizada a avaliação de marcadores

Figura 15 – Comparação das possíveis combinações de SNPs em comum entre os chips



Fonte: elaborada pelo autor

em comum ao retirar as amostras pertencentes a cada um dos tipos. Esta avaliação é de suma importância na consolidação dos dados, pois viabiliza a união entre amostras com diferentes combinações de SNPs, mantendo um subconjunto comparável e evitando a exclusão de amostras que podem ser fundamentais para caracterização da linhagem, como por exemplo, touros com mais descendentes. A partir desta análise, a combinação com o maior número de amostras foi obtida na combinação de todos os SNP chips, resultando em 16849 animais com 1785 marcadores, como é possível observar na Tabela 3.

Tabela 3 – Análise de sobreposição entre SNP chips

SNP Chip Retirado	Marcadores Comuns	Número de Animais
50K	7392	15197
GGP Ind 35K	6592	5611
GGP Ind 54K	1785	15105
HD	1790	16251
ZChip	2853	15232
Nenhum	1785	16849

Diversas técnicas baseadas em Aprendizado de Máquina não trabalham de maneira satisfatória com variáveis categóricas. Desta forma, o processo de *feature encoding* é fundamental para uma correta modelagem em problemas de Aprendizado Supervisionado e Não Supervisionado. Dados provenientes da genotipagem por chips de SNP são naturalmente categóricos, pois representam qual o alelo presente em determinado lócus.

Existem diversos modelos de codificação para dados de SNPs e cada um busca evidenciar diferentes características presentes no genótipo. Como é possível observar na Tabela 4 modelo detalhado contém a informação de cada base nitrogenada (A,C,T ou G) codificada de forma binária. Já os modelos baseados nos pares de alelos presentes para cada lócus, podem dar maior enfoque a parcela aditiva, recessiva e dominância, ou ao genótipo no geral.

Tabela 4 – Análise de sobreposição entre SNP chips

Modelo	Codificação	Descrição
Detalhado	(ACTG)	0000, 0001, ..., 1111
Aditivo	Nominal	AA, AB=BA e BB
Aditivo	Numérico	0, 1 e 2
Recessivo/Dominância	Nominal	AA, AB, BA e BB
Recessivo/Dominância	Binário	00, 01, 10 e 11
Genotípico	Binário	100, 010 e 001
Genotípico	Numérico	-1, 0, e 1

Dada a natureza das diferentes técnicas utilizadas para avaliação dos dados, modelos que criassem alguma relação de distância entre as diferentes classes poderiam influenciar no resultado das análises. Desta forma, foi selecionado o modelo genotípico binário, denominado na literatura como *one-hot encoding*.

Complementando as informações dos genótipos, foram disponibilizadas as genealogias dos animais, construídas a partir dos registros mantidos pelas associações de criadores. A base de dados denominada comumente como *pedrigree*, contém os códigos identificadores da amostra, e dos seus respectivos antepassados, bem como o sexo de registro.

A genealogia compreende uma série de elementos essenciais para a compreensão das relações de parentesco e características hereditárias dos animais. Cada entrada na base de dados é composta pelos seguintes atributos:

- **Código identificador único:** cada animal na amostra possui um código identificador único associado, o qual permite rastrear sua linhagem e obter informações de produção, desempenhos passados em características produtivas, localização e criador.
- **Antepassados diretos:** Além dos códigos identificadores individuais, a base de dados inclui também os códigos de identificação dos antepassados diretos de cada ani-

mal. Essa informação possibilita traçar a ascendência genética de forma abrangente, identificando progenitores, avós e bisavós.

- **Sexo de Registro:** Cada entrada na base de dados é acompanhada pelo sexo do animal registrado. Essa informação é crucial para análises subsequentes, permitindo a consideração do sexo como variável relevante nas avaliações genéticas e hereditárias.

4.2 TOUROS DE MAIOR PROGÊNIE

Como descrito anteriormente, a depender da disponibilidade de rótulos para os dados, o problema de aprendizado por ser supervisionado ou não supervisionado. Com objetivo de cobrir o caso em que existem registros genealógicos bem estruturados foi desenvolvida uma abordagem supervisionada para o problema.

Quanto maior o conjunto de descendentes de um indivíduo, maior é a possibilidade de consolidação de suas características na população, atuando diretamente na conservação de uma genética de interesse (GROSZ; KAKI, 2007). Desta forma, foram avaliadas na base de dados as maiores progênies através da contagem de indivíduos. Os touros correspondentes foram utilizados como índice em cada uma das cinco classes rotuladas, representando linhagens baseadas nestes animais. O número de indivíduos em cada uma das classes pode ser notado na Tabela 5.

Tabela 5 – Número de indivíduos por linhagem

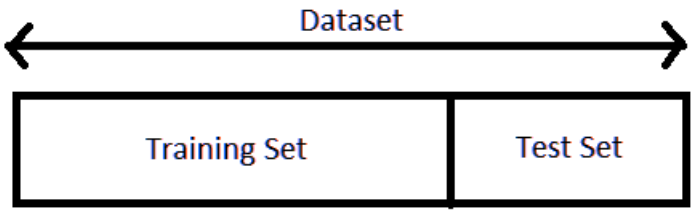
Linhagem	Número de Animais
L1437	303
L805	144
L703	113
L702	26
L981	18

4.2.1 Processamento de Dados

Para o treinamento dos modelos é necessário realizar o pré-processamento dos dados. Como representado pela Figura 16, inicialmente divide-se a base em dois conjuntos principais: treinamento e teste. Essa divisão pode ser realizada considerando diversos percentuais, 50%/50%, 75%/25% e 70%/30% são valores comuns em problemas de aprendizado. Para a aplicação proposta, baseado no número de animais da menor classe, o conjunto de treinamento foi composto por 2/3 das amostras, e teste, contando com o 1/3 restante. Além disso, a partir do conjunto de treinamento foram subdivididos dados de validação para aplicação do procedimento de K-fold.

O processo descrito pode ser compreendido como uma amostragem estatística, onde os conjuntos de treinamento e teste são amostras obtidas da base completa, a

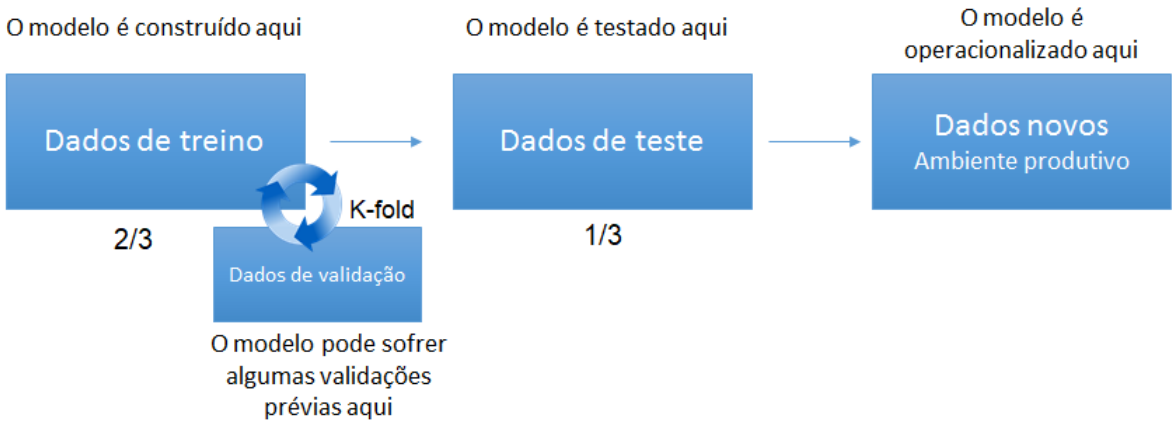
Figura 16 – Divisão da base de dados em treinamento e teste.



Fonte: do próprio autor.

população. No entanto, para geração de um modelo satisfatório e realização de uma validação realista, as amostras devem ser selecionadas de maneira estratificada, ou seja, mantendo a composição representativa de cada classe das instâncias.

Figura 17 – Processo correto de divisão de dados.



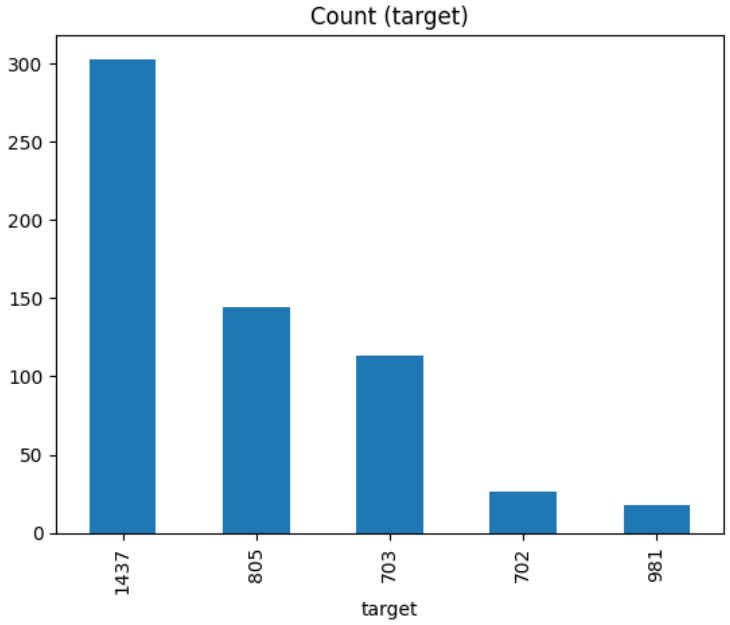
Fonte: do próprio autor.

Com os conjuntos corretamente divididos, como exemplificado pela Figura 17, ainda há o problema referente a quantidade desigual de ocorrências entre as classes. A Figura 18 ilustra o histograma de frequências para a base desbalanceada, enquanto a Figura 19 traz a distribuição desejada após o processo de balanceamento.

A distribuição heterogênea das classes pode ocasionar problemas na aplicação das técnicas de aprendizado, pois em determinadas situações o alto nível de acurácia obtido no treinamento pode significar um super ajuste do modelo aos dados, reduzindo a capacidade de predição para novas instâncias, seu principal objetivo.

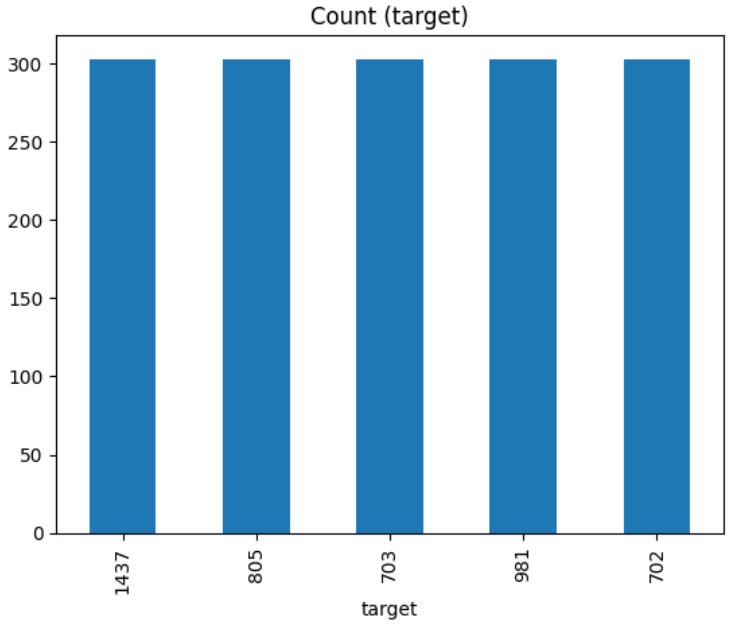
No presente trabalho optou-se pela realização do *oversampling* para correção do balanceamento. Técnica na qual as amostras das classes presentes em menor quantidade são replicadas de forma aleatória até equipararem-se em número com as maiores.

Figura 18 – Exemplo de histograma para base de dados desbalanceada.



Fonte: do próprio autor.

Figura 19 – Exemplo de histograma para base de dados balanceada.



Fonte: do próprio autor.

4.2.2 Modelagem

O conjunto de dados pré-processado foi utilizado para o treinamento da MLP completamente conectada através do algoritmo de *Backpropagation* para ajuste dos pesos sinápticos, sendo adotada uma tolerância de erro quadrático médio igual a 10^{-5}. Foram utilizados números diferentes de neurônios e camadas para avaliação do modelo, além de funções de ativação.

Para o treinamento da máquina de vetor suporte, foram utilizados dois tipos de *kernel* nas avaliações, o gaussiano ou RBF e o linear, além de diferentes valores para o parâmetro C, ligado diretamente a margem ajustada através dos vetores suporte de forma inversamente proporcional. Já no treinamento da *random forest*, foram avaliados dois principais parâmetros: o número de estimadores e a profundidade máxima das árvores.

Ao lidar com métodos baseados em distâncias, uma grande disparidade na amplitude estatística das instâncias pode impactar diretamente na obtenção de um melhor classificador, tornando-se necessária a normalização dos dados para uma faixa entre zero e um. Esta normalização é baseada nos valores mínimos e máximos, através da seguinte transformação:

$$x_{\text{norm}} = \frac{x - x_{\text{min}}}{x_{\text{max}} - x_{\text{min}}} \tag{4.1}$$

A implementação da MLP, SVM e *Random Forest*, bem como o pré-processamento dos dados, foram realizados utilizando a linguagem Python em sua versão 3.9 através das bibliotecas Numpy (OLIPHANT, 2006), Pandas (MCKINNEY, 2010) e Scikit Learn (PEDREGOSA et al., 2011).

4.3 CARACTERIZAÇÃO DE LINHAGENS NA POPULAÇÃO

Para a caracterização de linhagens através apenas dos genótipos dos animais uma abordagem não supervisionada é proposta. Foram realizadas duas análises, a primeira avaliando a base completa, como não existem rótulos nesta versão da base de dados, métricas específicas foram utilizadas. Já segunda segue um *pipeline* de processamento idêntico a anterior, mas considerando somente os animais das cinco linhagens definidas através dos touros de maior progênie, buscando a identificação e confirmação de uma estrutura particular dentro da população completa analisada no primeiro momento. A avaliação dos *clusters* é construída através de métricas internas e externas.

4.3.1 Processamento de Dados

A base de dados completa utilizada na primeira avaliação é composta por 16845 amostras e 1785 características, codificadas nas categorias, “A_A”, “A_B” e “B_B”. Já a segunda avaliação utiliza a mesma base da abordagem supervisionada, contendo 604 amostras e 1785 características.

A presença de dados faltantes demandou uma etapa de limpeza, eliminando 4 amostras incompletas para assegurar a robustez dos resultados e possibilitar os cálculos de distâncias.

Inicialmente é calculada uma matriz de distâncias entre as amostras para utilização nos algoritmos de clusterização. Para preservar a natureza categórica das variáveis, foi

adotada a distância de Gower na construção da matriz de distâncias entre os pontos. Essa escolha é interessante, pois a distância de Gower é especialmente eficaz em lidar com dados heterogêneos, atuando tanto com variáveis categóricas quanto com numéricas de forma coesa.

A distância de Gower (G_{ij}) entre dois pontos i e j é definida por:

$$G_{ij} = \frac{\sum_{l=1}^{L} g_{ijl}}{\sum_{l=1}^{L} w_{ijl}} \tag{4.2}$$

Onde:

- g_{ijl} é a contribuição da característica l para a distância entre i e j,
- w_{ijl} é o peso associado à característica l,
- L é o número total de características.

4.3.2 Modelagem

No modelo K-modes, o parâmetro K foi variado entre 2 e 20, representando o número de clusters. A avaliação foi realizada através das métricas de índice de Silhouette, Calinski Harabasz e Davies Bouldin.

O Índice de Silhouette (ROUSSEEUW, 1987) oferece uma medida da coesão e separação dos clusters, atribuindo a cada amostra um valor entre -1 e 1. Um valor próximo a 1 indica que a amostra está bem ajustada ao seu próprio cluster e distante dos clusters vizinhos, enquanto valores próximos a -1 sugerem uma possível atribuição incorreta.

A equação do índice de Silhouette é dada por:

$$S(i) = \frac{b(i) - a(i)}{\max\{a(i), b(i)\}} \tag{4.3}$$

Onde:

- $S(i)$ é o índice de Silhouette para a amostra i,
- $a(i)$ é a média da distância intra-cluster,
- $b(i)$ é a média da distância inter-cluster.

O Calinski Harabasz (DAVIES; BOULDIN, 1979), conhecido como critério de Razão de Variância, foca na relação entre a dispersão entre clusters e a dispersão intra-cluster. Uma pontuação mais alta indica clusters mais compactos e bem separados, sugerindo uma estrutura de clusterização eficaz.

A métrica Calinski Harabasz é definida como:

$$CH = \frac{B}{W} \times \frac{N-K}{K-1} \tag{4.4}$$

Onde:

- B é a dispersão entre clusters,
- W é a dispersão intra-cluster,
- N é o número total de amostras,
- K é o número de clusters.

O Índice Davies Bouldin (CALIńSKI; HARABASZ, 1974) avalia a compacidade e separação dos clusters, comparando a distância média dentro de um cluster com a distância média para o cluster mais próximo. Um valor mais baixo sugere clusters mais compactos e bem definidos.

A métrica Davies Bouldin é dada por:

$$DB = \frac{1}{K} \sum_{i=1}^{K} \max_{j \neq i} \left(\frac{a(i) + a(j)}{d(i,j)} \right) \tag{4.5}$$

Onde:

- $d(i,j)$ é a distância média entre os clusters i e j.

Ao considerar o modelo de Cluster Hierárquico, variamos o parâmetro K entre 2 e 20, e o parâmetro de ligação (*linkage*) entre single, average e complete. As métricas de índice de Silhouette, Calinski Harabasz e Davies Bouldin foram novamente utilizadas para avaliação.

Na análise para o conjunto de dados particular compostos pelas linhagens definidas pelos touros de maior progênie foi verificada a confirmação dos cinco *clusters* propostos inicialmente, utilizando o mesmo pipeline de processamento de dados anterior, como ilustrado em FIGURA FLUXOGRAMA 2. No entanto, serão avaliadas métricas internas, como índice de Silhouette, Calinski Harabasz e Davies Bouldin, e também externas, através da construção da matriz de confusão para comparação das classes identificadas.

Em ambas avaliações a implementação do K-modes e Cluster Hierárquico, bem como o pré-processamento dos dados, foram realizados utilizando a linguagem Python em sua versão 3.9 através das bibliotecas Numpy (OLIPHANT, 2006), Pandas (MCKINNEY, 2010) e Scikit Learn (PEDREGOSA et al., 2011).

5 ANÁLISE EXPERIMENTAL E RESULTADOS

5.1 TOUROS DE MAIOR PROGÊNIE

Na etapa de treinamento da MLP para a classificação das 5 maiores linhagens, após o pré-processamento da base, foram avaliadas métricas de desempenho como acurácia de treinamento, porcentagem de acertos no conjunto de teste, bem como a curva ROC e a área abaixo desta, com objetivo de avaliar a presença de falsos positivos e *overfitting* nas soluções.

Variando-se o número de camadas e neurônios em cada camada, na Tabela 6 é possível observar os resultados para os conjuntos de teste e treinamento. Na Tabela 7 foram realizados testes modificando as funções de ativação.

Tabela 6 – Resultado para variações de números de neurônios utilizando MLP.

Camadas	Neurônios	Acurácia Treino	Desvio Padrão	Acurácia Teste	ROC
3	3	0.9901	0.0043	92.8040%	0.9550
3	5	0.9987	0.0009	99.0074%	0.9938
3	10	0.9980	0.0028	98.7593%	0.9922
4	3	0.9934	0.0009	95.2854%	0.9705
4	5	0.9954	0.0025	95.2854%	0.9705
4	10	0.9987	0.0019	99.7519%	0.9984
5	3	0.9987	0.0009	99.2556%	0.9953
5	5	0.9987	0.0009	100.0000%	1.0000
5	10	0.9987	0.0009	99.2556%	0.9953

Tabela 7 – Resultado para variações na escolha da função de ativação utilizando MLP.

Função	Acurácia Treino	Desvio Padrão	Acurácia Teste	ROC
Identidade	0.9993	0.0009	100.0000%	1.0000
Logística	0.6125	0.0126	73.2010%	0.8325
TanH	1.0000	0.0000	99.7519%	0.9984
relu	0.9980	0.0016	100.0000%	1.0000

Desta forma, é possível verificar que os melhores resultados para a MLP foram obtidos com 5 camadas ocultas, 5 neurônios por camada e função identidade como função de ativação.

Na máquina de vetor suporte, as mesmas métricas de desempenho foram avaliadas, no entanto, os parâmetros do modelo são outros, como a constante C, que tem influência direta na largura da margem da fronteira de decisão, além da presença de diferentes tipos de *kernel*, como linear e RBF. Os resultados obtidos nos conjuntos de treinamento e teste podem ser verificados na Tabela 9. Os melhores resultados para a SVM possuem parâmetro C = 1 e kernel linear.

Tabela 8 – Resultado para variações no valor de C e de escolha de kernel para SVM.

Kernel	C	Acurácia Treino	Desvio Padrão	Acurácia Teste	ROC
Linear	1	1.0000	0.0000	100.0000%	1.0000
Linear	10	1.0000	0.0000	100.0000%	1.0000
Linear	100	1.0000	0.0000	100.0000%	1.0000
Linear	1000	1.0000	0.0000	100.0000%	1.0000
RBF	1	1.0000	0.0000	100.0000%	1.0000
RBF	10	1.0000	0.0000	100.0000%	1.0000
RBF	100	1.0000	0.0000	100.0000%	1.0000
RBF	1000	1.0000	0.0000	100.0000%	1.0000

Na *random forest*, foram avaliadas também as mesmas métricas de desempenho. Os parâmetros do modelo foram número de estimadores e a profundidade máxima das árvores. Os resultados obtidos nos conjuntos de treinamento e teste podem ser verificados na Tabela 9.

Tabela 9 – Resultado para variações no valor de C e de escolha de kernel para SVM.

Estimadores	Profundidade	Acurácia Treino	Desvio Padrão	Acurácia Teste	ROC
100	3	0.9993	0.0009	100.0000%	1.0000
100	5	1.0000	0.0000	100.0000%	1.0000
100	10	1.0000	0.0000	100.0000%	1.0000
500	3	0.9967	0.0009	100.0000%	1.0000
500	5	1.0000	0.0000	100.0000%	1.0000
500	10	1.0000	0.0000	100.0000%	1.0000
1000	3	1.0000	0.0000	100.0000%	1.0000
1000	5	1.0000	0.0000	100.0000%	1.0000
1000	10	1.0000	0.0000	100.0000%	1.0000

O modelo SVM com os parâmetros C=1 e kernel linear obteve os melhores resultados gerais dentre as três técnicas analisadas.

5.2 CARACTERIZAÇÃO DE LINHAGENS NA POPULAÇÃO

Para caracterização de linhagens na população como um todo foram realizadas duas análises, uma inicial, com objetivo de avaliar o funcionamento dos algoritmos e comparação de métricas internas de avaliação: índice Silhouette, Calinski-Harabasz e Davies-Bouldin. Posteriormente, é realizada a mesma avaliação focando nas amostras que compõe as linhagens construídas a partir dos cinco touros de maior progênie, verificando as mesmas métricas anteriores e também métricas externas através da matriz de confusão.

Os resultados das métricas de clusterização para o modelo K-modes, variando o número de clusters, são apresentados na Tabela 10. Ao analisar o índice Silhouette, é possível observar que, à medida que o número de clusters aumenta de 2 para 20, a separação

entre clusters permanece relativamente baixa, indicada pelos valores oscilantes em torno de 0.0259 a 0.0369. Quanto ao índice Calinski-Harabasz, notamos uma diminuição geral à medida que o número de clusters aumenta, sugerindo uma redução na dispersão entre os clusters. Especificamente, o índice Calinski-Harabasz atinge o valor mais alto (3933.7854) com 2 clusters e diminui gradualmente à medida que o número de clusters aumenta, alcançando 1062.7237 com 20 clusters.

O índice Davies-Bouldin, que é desejável quanto mais próximo de zero, apresenta uma tendência de aumento à medida que o número de clusters aumenta, indicando uma piora na compacidade dos clusters. Os valores oscilam entre 1.6413 (com 3 clusters) e 3.1758 (com 15 clusters).

Tabela 10 – Comparação dos resultados utilizando K-modes.

Modelo	K-Clusters	Silhouette	Calinski-Harabasz	Davies-Bouldin
K-modes	2	0.0259	3933.7854	1.7371
-	3	0.0276	4260.1301	1.6413
-	4	0.0299	3553.8490	1.9106
-	5	0.0282	2870.1468	2.1162
-	6	0.0302	2493.4405	2.0651
-	7	0.0305	2247.6340	2.2860
-	8	0.0313	2091.5111	2.1747
-	9	0.0340	1888.9655	2.2640
-	10	0.0355	1722.9344	2.3584
-	11	0.0227	1649.4105	2.5665
-	12	0.0369	1569.0164	2.5050
-	13	0.0260	1479.8929	2.5129
-	14	0.0230	1334.5176	2.8460
-	15	0.0282	1299.4405	3.1758
-	16	0.0219	1276.6289	2.7613
-	17	0.0299	1261.3467	2.4006
-	18	0.0257	1158.5011	2.7167
-	19	0.0312	1139.0590	2.8233
-	20	0.0194	1062.7237	2.7216

Os resultados das métricas de clusterização para o modelo de Cluster Hierárquico - *Single*, variando o número de clusters, são apresentados na Tabela 11. O índice Silhouette, indicando a qualidade da separação dos clusters, apresenta valores que aumentam gradualmente de 0.2878 (com 2 clusters) para 0.0311 (com 20 clusters). O valor mais alto é alcançado com dois clusters (0.2878), indicando uma separação distinta. Entretanto, à medida que o número de clusters aumenta, o índice diminui gradualmente.

Analisando o índice Calinski-Harabasz, é possível observar uma tendência de diminuição à medida que o número de clusters aumenta. Especificamente, o índice atinge o valor mais alto (68.4748) com 2 clusters e diminui para 29.0100 com 20 clusters, sugerindo uma redução na dispersão entre os clusters.

O índice Davies-Bouldin, que é desejável quanto mais próximo de zero, apresenta uma tendência de aumento à medida que o número de clusters aumenta, indicando uma piora na compacidade dos clusters. Isso sugere que, com um número menor de clusters, a compacidade é mais eficiente. Os valores oscilam entre 0.1181 (com 2 clusters) e 0.6534 (com 19 clusters).

Tabela 11 – Comparação dos resultados utilizando Cluster Hierárquico - Single.

Modelo	K-Clusters	Silhouette	Calinski-Harabasz	Davies-Bouldin
CH-single	2	0.2878	68.4748	0.1181
-	3	0.1026	37.2013	0.3074
-	4	0.0953	26.9330	0.3273
-	5	0.0871	21.4721	0.3550
-	6	0.0839	19.2004	0.4631
-	7	0.0803	16.8229	0.4722
-	8	0.0784	15.1808	0.4822
-	9	0.0744	13.8162	0.4845
-	10	0.0740	14.4134	0.8962
-	11	0.0744	52.3423	0.9465
-	12	0.0730	47.5973	0.8382
-	13	0.0717	43.6598	0.6770
-	14	0.0717	41.0775	0.6517
-	15	0.0598	38.3227	0.6501
-	16	0.0548	36.1139	0.6367
-	17	0.0495	34.0296	0.6344
-	18	0.0378	32.1230	0.6484
-	19	0.0315	30.4368	0.6534
-	20	0.0311	29.0100	0.6480

De acordo com o apresentado na Tabela 12, no índice Silhouette, observa-se uma variação significativa. O valor mais alto é alcançado com dois clusters (0.2878), indicando uma separação distinta. Entretanto, à medida que o número de clusters aumenta, o índice diminui gradualmente, sugerindo uma possível complexidade desnecessária na subdivisão.

Para o índice Calinski-Harabasz, notamos uma queda progressiva à medida que mais clusters são introduzidos. O valor máximo (68.4748) ocorre com dois clusters, indicando uma melhor dispersão dos dados em configuração de um menor número de clusters.

No que diz respeito ao índice Davies-Bouldin, atinge seu valor mínimo (0.1181) com dois clusters, observa-se um aumento nos valores à medida que o número de clusters cresce. Indicando uma melhor compacidade a partir de um número menor de clusters.

A avaliação do modelo CH-complete está presente na tabela 13. A métrica Silhouette, indica que o valor mais elevado é alcançado quando K é igual a 2, com um score de 0.2878. À medida que o número de clusters aumenta, o valor da Silhouette diminui, sugerindo como nos casos anteriores, uma diminuição na coesão dos clusters.

Tabela 12 – Comparação dos resultados utilizando Cluster Hierárquico - Average.

Modelo	K-Clusters	Silhouette	Calinski-Harabasz	Davies-Bouldin
CH-average	2	0.2878	68.4748	0.1181
-	3	0.1282	39.0002	0.2508
-	4	0.0959	45.9494	0.4605
-	5	0.0784	35.6972	0.4591
-	6	0.0735	107.4662	0.7763
-	7	0.0629	90.1094	0.7480
-	8	0.0535	77.5951	0.7328
-	9	0.0472	71.4321	0.8080
-	10	0.0472	63.5204	1.1134
-	11	0.0412	57.3584	1.0775
-	12	0.0412	52.1642	1.0931
-	13	0.0308	48.4113	1.3010
-	14	0.0307	44.7019	1.2343
-	15	0.0267	42.0026	1.3671
-	16	0.0253	40.5771	1.5417
-	17	0.0239	38.3199	1.6417
-	18	0.0184	36.1328	1.6111
-	19	0.0174	34.1408	1.6429
-	20	0.0169	33.8679	1.7542

Já a métrica Calinski-Harabasz, atinge seu valor máximo quando K é 5, com um score de 560.8657. Isso sugere que, nesse ponto, a configuração do modelo gera clusters mais distintos entre si. No entanto, à medida que K continua a aumentar, o valor da métrica Calinski-Harabasz diminui.

A métrica Davies-Bouldin, que avalia a compacidade e separação dos clusters, alcança seu mínimo quando K é igual a 2 com 0.1181. À medida que K aumenta, a Davies-Bouldin também aumenta, indicando uma piora na separação dos clusters.

Na base de dados particular de cinco linhagens, os resultados das métricas de clusterização para os modelos avaliados são apresentados na Tabela 14.

Para o modelo K-modes, com 5 clusters, é possível observar um índice de Silhouette de 0.1124, indicando uma boa separação entre os clusters. O índice Calinski-Harabasz, que é mais eficaz quanto maior, alcançou 482.5787, sugerindo uma boa dispersão entre os clusters. O índice Davies-Bouldin, desejável quanto mais próximo de zero, foi de 1.1230, indicando uma medida razoável da compacidade dos clusters.

O método CH-single, também com 5 clusters, apresentou um índice de Silhouette de -0.0621, indicando alguma sobreposição entre os clusters. O índice Calinski-Harabasz foi de 0.7518, sugerindo uma dispersão limitada entre os clusters. O índice Davies-Bouldin foi de 1.2090, indicando uma medida razoável da compacidade dos clusters, embora ligeiramente menos desejável que o modelo K-modes.

Tabela 13 – Comparação dos resultados utilizando Cluster Hierárquico - Complete.

Modelo	K-Clusters	Silhouette	Calinski-Harabasz	Davies-Bouldin
CH-complete	2	0.2878	68.4748	0.1181
-	3	0.1109	40.9596	0.4732
-	4	0.0873	177.5028	0.7911
-	5	0.0085	560.8657	1.4905
-	6	0.0095	521.8643	2.1838
-	7	0.0097	435.6639	2.2219
-	8	0.0128	526.5382	2.3500
-	9	0.0133	490.3196	2.3242
-	10	0.0135	445.3565	2.8058
-	11	0.0201	778.8814	2.6623
-	12	0.0202	716.5840	2.9300
-	13	0.0207	684.2479	3.2087
-	14	0.0218	723.2525	3.6442
-	15	0.0219	688.4310	3.5757
-	16	0.0208	687.5939	4.0815
-	17	0.0211	676.6562	4.1611
-	18	0.0200	645.8217	4.0488
-	19	0.0205	612.8908	3.9896
-	20	0.0201	581.7267	4.0137

Para os métodos CH-average e CH-complete, ambos com 5 clusters, os índices de Silhouette foram de 0.1097 e 0.1083, respectivamente, indicando uma separação satisfatória entre os clusters. Os índices Calinski-Harabasz, mais eficazes quanto maiores, foram de 457.3671 e 450.1144, respectivamente, indicando uma boa dispersão entre os clusters. Os índices Davies-Bouldin, desejáveis quanto mais próximos de zero, foram de 1.1357 e 1.1805, respectivamente, sugerindo medidas razoáveis da compacidade dos clusters.

Tabela 14 – Comparação dos resultados para os touros de maior progênie

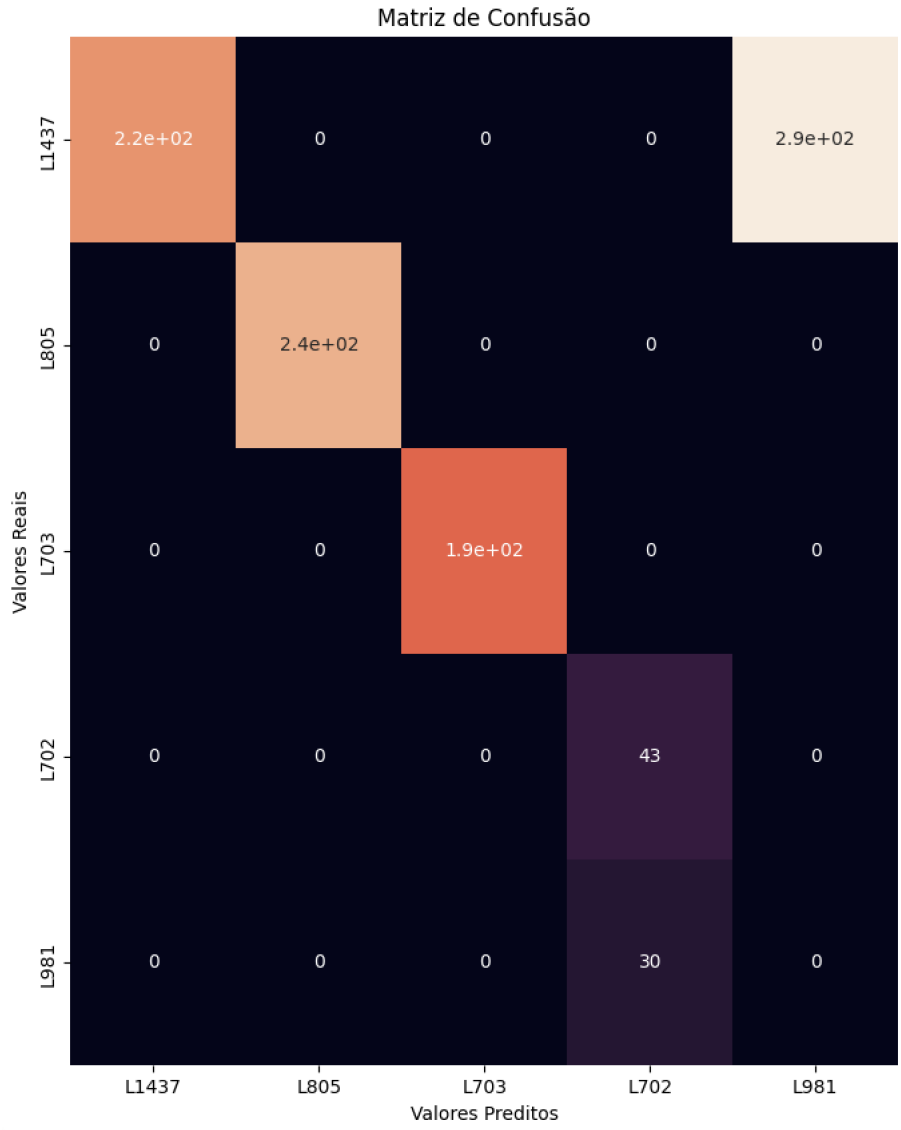
Modelo	K-Clusters	Silhouette	Calinski-Harabasz	Davies-Bouldin
K-modes	5	0.1124	482.5787	1.1230
CH-single	5	-0.0621	0.7518	1.2090
CH-average	5	0.1097	457.3671	1.1357
CH-complete	5	0.1083	450.1144	1.1805

Como forma de confirmação dos resultados obtidos, a avaliação da base de dados a partir dos rótulos das cinco linhagens construídas foi realizada, constituindo um olhar para métricas externas aos clusters neste caso particular, através dos acertos na matriz de confusão.

A matriz de confusão para o modelo K-modes, disponível na Figura 20, mostra que a identificação para as linhagens L805 e L703 foi realizada de forma correta para todas suas amostras, na linhagem L702 foram identificadas corretamente todas as amostras,

porém o modelo associou a linhagem L981 completa como L702 também. A linhagem L1437 obteve parte das amostras com identificação correta e parte das amostras como L981.

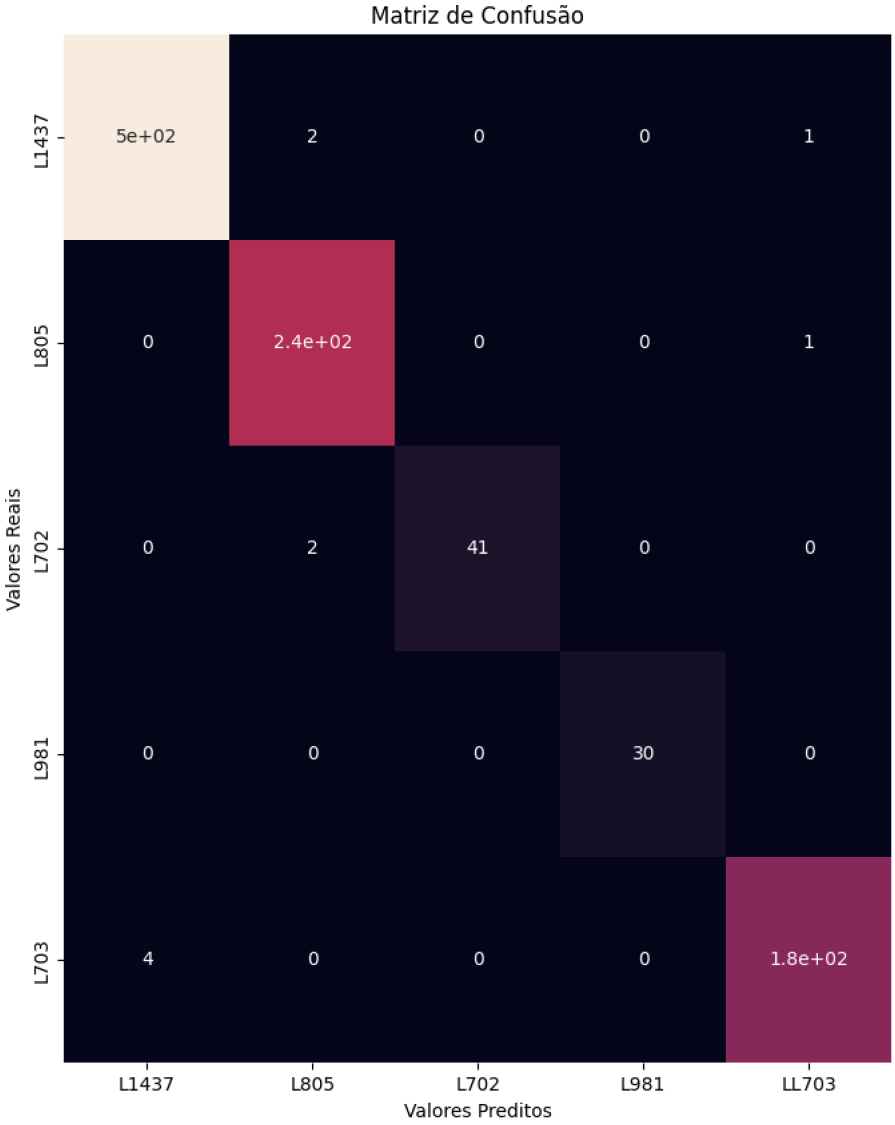
Figura 20 – Matriz de confusão para o modelo K-modes.



Fonte: do próprio autor.

Já para o modelo CH-average, como é possível avaliar na Figura 21, a identificação foi mais assertiva, associando corretamente a maior parte das amostras em todas as linhagens, havendo apenas quatro associações erradas na L703, três na L1437, duas na L702 e uma na L805. Desta forma, o modelo baseado em clusterização hierárquica se mostrou mais efetivo para a identificação das linhagens utilizando apenas dos genótipos dos animais.

Figura 21 – Matriz de confusão para o modelo CH-average.



Fonte: do próprio autor.

6 TRABALHOS FUTUROS

Como perspectiva para futuras implementações, considera-se desenvolver técnicas adicionais de clusterização, com foco em segmentos específicos da população. Além disso, motivada pela possibilidade de inclusão de dados imputados para SNP chips de maior densidade de marcadores, planeja-se avaliar a redução de dimensionalidade mediante a implementação de representação por Autoencoders e outras técnicas de seleção de características, essa abordagem surge como uma extensão das técnicas convencionais, como PCA, MCA e da avaliação da importância das variáveis através de uma *Random Forest*, as quais demonstraram configurações de características com valores aproximados à magnitude da base de dados completa.

Outro ponto de interesse é a exploração de um sistema de reconhecimento de linhagens através da produtização de modelos construídos a partir da abordagem proposta. Esse sistema seria baseado na inclusão do genótipo no conjunto comum de marcadores estudados via *clusters* previamente identificados, proporcionando informações sobre a distância entre animais, sua linhagem e identificando potenciais candidatos para seleção e cruzamento.

Buscando maior robustez nos modelos desenvolvidos, é interessante a inclusão de ruídos nos dados para estudo da influência nos resultados obtidos. Essa avaliação pode ser realizada através da inclusão de animais provenientes de raças diferentes das de objetivo.

Por fim, pretende-se aprimorar a etapa de anotação genealógica por meio de um estudo mais aprofundado sobre a população, fundamentado em seu histórico. Esse estudo visa destacar linhagens mais locais que possam enriquecer o processo de construção de novos classificadores, contribuindo para uma abordagem mais refinada e contextualizada através da identificação de prováveis linhagens.

7 CONCLUSÃO

O presente trabalho buscou desenvolver modelos de aprendizado de máquina voltados para a predição de linhagens em gado leiteiro considerando linhagens pré estabelecidas ou apenas os genótipos dos animais, possibilitando suporte ao planejamento de cruzamentos em programas de melhoramento genético animal com o objetivo de minimizar a endogamia presente na população. O trabalho se dividiu em duas análises distintas: uma fundamentada em aprendizado supervisionado e outra em aprendizado não supervisionado.

Os modelos resultantes da classificação de animais nas linhagens formadas pelos touros de maior progênie demonstraram um alto grau de acurácia. A identificação precisa realizada por esses modelos evidenciou métricas satisfatórias tanto nos conjuntos de treinamento quanto nos de teste, atestando sua eficácia na solução do problema proposto.

A abordagem não supervisionada proporcionou uma compreensão mais aprofundada sobre o funcionamento das técnicas e métricas de avaliação, corroborando com os resultados obtidos na primeira abordagem e se mostrando como uma possibilidade para colaboração em estudos sobre a composição e estrutura de uma população.

Destaca-se a importância da anotação genealógica no processo de identificação de linhagens. Os modelos baseados em aprendizado supervisionado revelaram resultados promissores, indicando que o conhecimento prévio da população desempenha um papel significativo na melhoria da identificação de novos animais.

Em síntese, os processos e modelos desenvolvidos revelam um potencial considerável para aprimoramento e aplicação em larga escala. A escolha da linguagem de programação Python e o uso de bibliotecas de código aberto facilitam a implementação em ambientes de computação em nuvem, possibilitando desempenho e acessibilidade na utilização dos modelos. Essa versatilidade confirma a viabilidade e a utilidade prática dessas ferramentas para a comunidade envolvida no programa de melhoramento.

REFERÊNCIAS

ALBERTS, B.; AL. et. **Molecular Biology of the Cell**. [S.l.]: Garland Science, 2014.

ANDERSSON, L. Genetic dissection of phenotypic diversity in farm animals. **Nature Reviews Genetics**, v. 2, n. 2, p. 130–138, 2001.

AUAD, A.M.; AL. et. **Manual de bovinocultura de leite.** [S.l.]: LK Editora, 2010.

BOSER, B. E.; GUYON, I. M.; VAPNIK, V. N. A training algorithm for optimal margin classiers. **Proceedings of the 5th Annual Workshop on Computational Learning Theory**, ACM Press, v. 1, p. 144–152, 1992.

BROOKES, A. J. The essence of snps. **Gene**, v. 234, n. 2, p. 177–186, 1999.

CALIńSKI, T.; HARABASZ, J. A dendrite method for cluster analysis. **Communications in Statistics**, Taylor Francis, v. 3, n. 1, p. 1–27, 1974.

CHAKRAVARTI, A. Single nucleotide polymorphisms—to a future of genetic medicine. **Nature**, v. 409, n. 6822, p. 822–823, 2001.

CORTES, C.; VAPNIK, V. Support-vector networks. **Machine Learning**, Kluwer Academic Publishers, v. 20, p. 273–297, 1995.

DAVIES, David L.; BOULDIN, Donald W. A cluster separation measure. **IEEE Transactions on Pattern Analysis and Machine Intelligence**, PAMI-1, n. 2, p. 224–227, 1979.

DAVIS, G. P. **Breeding for disease resistance in farm animals**. [S.l.]: CABI, 2008.

FALCONER, D. S.; MACKAY, T. F. **Introduction to quantitative genetics**. [S.l.]: Longman, 1996.

FAN, J. B.; AL. et. Highly parallel snp genotyping. **Cold Spring Harbor Symposia on Quantitative Biology**, v. 71, p. 139–143, 2006.

GAMA, J; FACELI, K; LORENA, A C; CARVALHO, A C P L F DE. **Inteligência artificial: uma abordagem de aprendizado de máquina**. [S.l.]: Grupo Gen - LTC, 2011.

GARRICK, D. J.; TAYLOR, J. F.; FERNANDO, R. L. A history of genome-wide association studies in cattle. **Animal Genetics**, v. 45, n. 5, p. 580–590, 2014.

GIANOLA, D.; FOULLEY, J. L. Theory and analysis of threshold characters. **Journal of Animal Science**, v. 63, n. 2, p. 485–498, 1986.

GINI, Corrado. Variabilità e mutabilità. **Reprinted in Memorie di metodologica statistica (Ed. CGS Rome)**, v. 17, 1912.

GJEDREM, T. The importance of selective breeding in aquaculture to meet future demands for animal protein: a review. **Aquaculture**, v. 285, n. 1-4, p. 1–7, 2009.

GODDARD, M. Genomic selection: prediction of accuracy and maximisation of long term response. **Genetica**, v. 136, n. 2, p. 245–257, 2009.

GODDARD, M. E.; HAYES, B. J. Mapping genes for complex traits in domestic animals and their use in breeding programmes. **Nature Reviews Genetics**, v. 10, n. 6, p. 381–391, 2009.

GROSZ, M. D.; KAKI, A. The role of line breeding in maintaining genetic diversity. **Journal of Animal Breeding and Genetics**, v. 124, n. 2, p. 77–83, 2007.

HAYKIN, Simon. **Redes neurais: princípios e prática**. [S.l.]: Bookman Editora, 2007.

HAZEL, L. N. The use of sibling data to estimate the heritability of a character. **Heredity**, v. 34, n. 2, p. 373–381, 1943.

HIRSCHHORN, J. N.; DALY, M. J. Genome-wide association studies for common diseases and complex traits. **Nature Reviews Genetics**, v. 6, n. 2, p. 95–108, 2002.

JACKSON, R.; RESENDE, M. F. R.; RESENDE, M. D. V. Breeding for the future: genomics for breeding in the 21st century. **Trends in Plant Science**, v. 22, n. 7, p. 624–637, 2017.

JAIN, Anil K; MURTY, M Narasimha; FLYNN, Patrick J. **Data clustering: 50 years beyond K-means**. [S.l.]: Springer Science & Business Media, 2010.

JAMES, Gareth; WITTEN, Daniela; HASTIE, Trevor; TIBSHIRANI, Robert. **Introduction to Statistical Learning**. [S.l.]: Springer, 2013.

KARYPIS, George; HAN, Eui-Hong; KUMAR, Vipin. Hierarchical clustering for large databases. **Data Mining and Knowledge Discovery**, Springer, v. 3, n. 3, p. 245–276, 1999.

LECUN, Yan; BENGIO, Yoshua; HINTON, George. Deep learning. **Nature**, Springer Nature, v. 521, n. 1, p. 436—-444, 2015.

LENT, Roberto. **Cem Bilhões de Neurônios: Conceitos Fundamentais de Neurociência**. [S.l.]: Atheneu, 2010.

LUIKART, G.; AL. et. The power and promise of population genomics: from genotyping to genome typing. **Nature Reviews Genetics**, v. 4, n. 12, p. 981–994, 2003.

MACKAY, T. F.; LYMAN, R. F. Quantitative trait loci in drosophila. **Nature Reviews Genetics**, v. 2, n. 1, p. 11–20, 2001.

MCKINNEY, Wes. Data structures for statistical computing in python. SciPy, p. 51 – 56, 2010.

MEUWISSEN, T. H.; 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.

MITCHELL, T. **Machine Learning**. [S.l.]: McGraw Hill, 1997.

MORIN, P. A.; AL. et. Snps in ecology, evolution and conservation. **Trends in Ecology & Evolution**, v. 19, n. 4, p. 208–216, 2004.

NOROUZI, M.; FLEET, D. J.; SALAKHUTDINOV, R. R. Support-vector networks. **Advances in Neural Information Processing Systems**, Cambridge, p. 1061–1069, 2012.

OLIPHANT, A.; AL. et. Beadarray™ technology: enabling an accurate, cost-effective approach to high-throughput genotyping. **BioTechniques**, v. 32, n. Supplement, p. S56–S61, 2002.

OLIPHANT, Travis E. A guide to numpy. **Trelgol Publishing USA**, 2006.

OTTO, P. I.; AL. et. Aplicações da genômica na produção animal. 2023.

PEDREGOSA, F.; VAROQUAUX, G.; GRAMFORT, A.; MICHEL, V.; THIRION, B.; GRISEL, O.; BLONDEL, M.; PRETTENHOFER, P.; WEISS, R.; DUBOURG, V.; VANDERPLAS, J.; PASSOS, A.; COURNAPEAU, D.; BRUCHER, M.; PERROT, M.; DUCHESNAY, E. Scikit-learn: Machine learning in python. **Journal of Machine Learning Research**, v. 12, p. 2825–2830, 2011.

PURCELL, Shaun; NEALE, Benjamin; TODD-BROWN, Kathe; THOMAS, Lori; FERREIRA, Manuel AR; BENDER, David; MALLER, Julian; SKLAR, Pamela; BAKKER, Paul IW de; DALY, Mark J et al. Plink: A tool set for whole-genome association and population-based linkage analyses. **American Journal of Human Genetics**, Elsevier, v. 81, n. 3, p. 559–575, 2007.

QUINLAN, J. Ross. Induction of decision trees. **Machine learning**, Springer, v. 1, n. 1, p. 81–106, 1986.

ROUSSEEUW, Peter J. Silhouettes: A graphical aid to the interpretation and validation of cluster analysis. **Journal of Computational and Applied Mathematics**, v. 20, p. 53–65, 1987. ISSN 0377-0427.

SANTOS, Genário dos; COELHO, Maria Thereza Ávila Dantas; FERNANDES, Sérgio Augusto Franco. A produÇÃo cientÍfica sobre a interdisciplinaridade: Uma revisÃo integrativa. **Educação em Revista**, Faculdade de Educação da Universidade Federal de Minas Gerais, v. 36, p. e226532, 2020. ISSN 0102-4698. Disponível em: <https://doi.org/10.1590/0102-4698226532>.

SHANNON, Claude E. A mathematical theory of communication. **The Bell System Technical Journal**, American Telephone and Telegraph Company, v. 27, n. 3, p. 379–423, 1948.

SHENDURE, J.; AL. et. Next-generation dna sequencing. **Nature Biotechnology**, v. 26, n. 10, p. 1135–1145, 2008.

SHINDE, Pramila P.; SHAH, Seema. A review of machine learning and deep learning applications. p. 1–6, 2018.

SMITH, J. M. Genetic selection for milk production. **Journal of Dairy Science**, v. 88, n. Suppl 1, p. E100–E110, 2005.

SMOLA, A. J.; BARLETT, P.; SCHOLKOPF, B.; SCHUURMANS, D. Introduction to large margin classifiers. **Advances in Large Margin Classifiers**, MIT Press, p. 1–28, 1999.

SMOLA, A. J.; SCHOLKOPF, B. **Learning with Kernels**. [S.l.]: MIT Press, 2002.

SOLE, Xavier; GUINO, Elisabet; VALLS, Joan; INIESTA, Raquel; MORENO, Víctor. Snpstats: a web tool for the analysis of association studies. **Bioinformatics**, v. 22, p. 1928—-1929, 2006.

VANRADEN, P. M. Genomic measures of relationship and inbreeding. **Interbull Bulletin**, v. 49, p. 38–45, 2016.

VANRADEN, P. M.; AL. et. Genetic evaluation of stillbirth in us holsteins. **Journal of Dairy Science**, v. 90, n. 10, p. 4423–4430, 2007.

VANRADEN, P. M.; O'CONNELL, J. R.; WIGGANS, G. R. Methods to account for multiple testing in the interpretation of dna microarray data. **Journal of Dairy Science**, v. 91, n. 2, p. 471–481, 2008.

VANRADEN, P. M.; TASSELL, C. P. Van; WIGGANS, G. R.; SONSTEGARD, T. S.; SCHNABEL, R. D.; TAYLOR, J. F.; SCHENKEL, F. S. Evaluation of direct genomic breeding values for lifetime net merit of holstein sires. **Journal of Dairy Science**, v. 87, n. 6, p. 2130–2137, 2004.

VAPNIK, V. N. **The nature of statistical learning theory**. [S.l.]: Springer-Verlag, 1995.

VAPNIK, V. N. **Statistical Learning Theory**. [S.l.]: Wiley-Interscience, 1998.

WATSON, J. D.; AL. et. **Molecular Biology of the Gene**. [S.l.]: Cold Spring Harbor Laboratory Press, 2008.