FÁBIO DANILO VIEIRA

MODELOS BASEADOS EM TÉCNICAS DE MINERAÇÃO DE DADOS PARA SUPORTE À CERTIFICAÇÃO RACIAL DE OVINOS

CAMPINAS

2014

UNIVERSIDADE ESTADUAL DE CAMPINAS
FACULDADE DE ENGENHARIA AGRÍCOLA

FÁBIO DANILO VIEIRA

MODELOS BASEADOS EM TÉCNICAS DE MINERAÇÃO DE DADOS PARA SUPORTE À CERTIFICAÇÃO RACIAL DE OVINOS

Dissertação de Mestrado submetida à banca examinadora para obtenção do título de Mestre em Engenharia Agrícola, na área de concentração de Gestão de Sistemas e Desenvolvimento Rural Sustentável.

Orientador: Prof. Dr. Stanley Robson de Medeiros Oliveira
Coorientador: Dr. Samuel Rezende Paiva

ESTE EXEMPLAR CORRESPONDE À VERSÃO FINAL DA DISSERTAÇÃO
DEFENDIDA PELO ALUNO FÁBIO DANILO VIEIRA,
E ORIENTADA PELO PROF. DR. STANLEY ROBSON DE MEDEIROS OLIVEIRA

Assinatura do Orientador

_____

CAMPINAS

2014

Ficha catalográfica
Universidade Estadual de Campinas
Biblioteca da Área de Engenharia e Arquitetura
Rose Meire da Silva - CRB 8/5974

Vieira, Fábio Danilo, 1977-
V673m Modelos baseados em técnicas de mineração de dados para suporte à certificação racial de ovinos / Fábio Danilo Vieira. – Campinas, SP : [s.n.], 2014.

Orientador: Stanley Robson de Medeiros Oliveira.
Coorientador: Samuel Rezende Paiva.
Dissertação (mestrado) – Universidade Estadual de Campinas, Faculdade de Engenharia Agrícola.

1. Mineração de dados (Computação). 2. Polimorfismo de nucleotídeo único. 3. Seleção de variáveis. 4. Ovino - Criação. I. Oliveira, Stanley Robson de Medeiros. II. Paiva, Samuel Rezende. III. Universidade Estadual de Campinas. Faculdade de Engenharia Agrícola. IV. Título.

Informações para Biblioteca Digital

Título em outro idioma: Models based on data mining techniques to support breed certification testing in brazilian sheep
Palavras-chave em inglês:
Data mining (Computer)
Single nucleotide polymorphism
Variable selection
Sheep - Creation
Área de concentração: Planejamento e Desenvolvimento Rural Sustentável
Titulação: Mestre em Engenharia Agrícola
Banca examinadora:
Stanley Robson de Medeiros Oliveira [Orientador]
Roberto Hiroshi Higa
Carlos Alberto Alves Meira
Data de defesa: 19-08-2014
Programa de Pós-Graduação: Engenharia Agrícola

Este exemplar corresponde à redação final da **Dissertação de Mestrado** defendida por **Fábio Danilo Vieira**, aprovada pela Comissão Julgadora em 19 de agosto de 2014, na Faculdade de Engenharia Agrícola da Universidade Estadual de Campinas.

Prof. Dr. Stanley Robson de Medeiros Oliveira – Presidente e Orientador
FEAGRI/Unicamp

Dr. Roberto Hiroshi Higa – Membro Titular
Embrapa/CNPTIA

Prof. Dr. Carlos Alberto Alves Meira – Membro Titular
FEAGRI/Unicamp

RESUMO

As raças de ovinos localmente adaptadas descendem de animais trazidos durante o período colonial, e durante anos foram submetidas a cruzamentos indiscriminados com raças exóticas. Estas raças de ovinos são consideradas importantes por possuírem características adaptativas às diversas condições ambientais brasileiras. Para evitar a perda deste importante material genético, a Empresa Brasileira de Pesquisa Agropecuária (Embrapa) decidiu incluí-las no seu Programa de Pesquisa em Recursos Genéticos, armazenando-as em seus bancos de germoplasma, sendo que as que possuem maior destaque nacional são as raças Crioula, Morada Nova e Santa Inês. A seleção dos ovinos para compor estes bancos é realizada por meio da avaliação de características morfológicas e produtivas. Entretanto, essa avaliação está sujeita a falhas, pois alguns animais cruzados mantêm características semelhantes àquelas dos animais locais. Desta forma, identificar se os animais depositados nos bancos são ou não pertencentes a uma raça é uma tarefa que exige muita cautela. Em busca de soluções, nos últimos anos houve um aumento significativo no uso de tecnologias que utilizam marcadores moleculares SNP (do inglês *Single Nucleotide Polimorphism*). No entanto, o grande número de marcadores gerados, que pode chegar a centenas de milhares por animal, torna-se um problema crucial. Para abordar esse problema, o objetivo deste trabalho é desenvolver modelos baseados em técnicas de mineração de dados para selecionar os principais marcadores SNP para as raças Crioula, Morada Nova e Santa Inês. Os dados utilizados neste estudo foram obtidos do Consórcio Internacional de Ovinos e são compostos por 72 animais destas três raças e 49.034 marcadores SNP para cada ovino. O resultado obtido com a conclusão deste trabalho foi um conjunto de modelos preditivos baseados em técnicas de mineração de dados que selecionaram os principais marcadores SNP para identificação das raças estudadas. A partir da intersecção desses modelos identificou-se um subconjunto de 15 marcadores com maior potencial de identificação das raças. Os modelos poderão ser utilizados para certificação das raças de ovinos já depositados nos bancos de germoplasma e de novos animais a serem inclusos, além de subsidiar associações de criadores interessadas em certificar seus animais, bem como o MAPA (Ministério da Agricultura, Pecuária e Abastecimento) no controle de animais registrados. Os modelos gerados poderão ser estendidos para outras espécies animais de produção.

Palavras-chave: marcadores moleculares, polimorfismo de nucleotídeo único, seleção de atributos, aprendizado de máquina, classificação, regressão penalizada, ovinocultura, microarranjo.

ABSTRACT

The locally adapted breeds of sheep are descended from animals brought in during the colonial period, and for years were subjected to indiscriminate crossbreeding with exotic breeds. These breeds of sheep are considered important by having adaptive characteristics to several Brazilian environmental conditions. To avoid the loss of this important genetic material, the Brazilian Agricultural Research Corporation (Embrapa) decided to include them in its Programme of Research in Genetic Resources, storing them in their genebanks, while those with greater national prominence are Creole breeds, Morada Nova and Santa Ines. The selection of sheep to compose these banks is performed through the evaluation of morphological and productive characteristics. However, this assessment is subject to failures, because some crossbred maintains similar characteristics to those of the local animals. Thus, identifying if the animals deposited in banks belong or not to a breed is a challenging task. In search for solutions in recent years there has been a significant increase in the use of technologies that use molecular markers SNP (Single Nucleotide Polimorphism). However, the large number of markers generated, which can reach hundreds of thousands per animal, becomes a crucial issue. To address this problem, the aim of this study is to develop models based on data mining techniques to select the main SNP markers for Creole, Morada Nova and Santa Ines breeds. The data used in this study were obtained from the International Consortium of Sheep and consist of 72 animals e of these three breeds and 49,034 SNP markers for each sheep. The result obtained with this study was a set of predictive models based on data mining techniques to selected major SNP markers to identify the breeds studied. The intersection of the generated models identified a subset of 15 markers, with greater potential for identification of sheep breeds. The models may be used for certification of sheep breeds already deposited in genebanks and new animals to be included, apart from subsidizing breeders associations interested in certifying their animals, as well as MAPA (Ministry of Agriculture, Livestock and Food Supply) in control registered animals. The proposed models can be extended to other species of production animals.

Palavras-chave: molecular markers, single nucleotide polymorphism, feature selection, machine learning, predictive modeling, penalized regression, sheep breeding, microarray.

SUMÁRIO

Dedico
ao meu filho JOAQUIM,
à minha esposa LETÍCIA
e aos meus pais.
Agradeço muito a Deus por
fazerem parte de
minha vida.

AGRADECIMENTOS

A DEUS, que me deu todas as forças para conseguir vencer mais este desafio.

Ao Prof. Dr. Stanley Robson de Medeiros Oliveira, pela orientação, suporte, companheirismo e pelo exemplo de como se realizar um bom trabalho e se dedicar à pesquisa.

Ao Dr. Samuel Resende Paiva, pela coorientação, e em fornecer os dados que utilizei para esta dissertação e por sempre ter se colocado à disposição para tirar minhas dúvidas.

Ao Dr. Michel Beleza Yamagishi, que foi uma das pessoas, juntamente com o Dr. Stanley, a me incentivar a realizar um mestrado considerando um problema de relevância para a agricultura brasileira, que está inserido nas prioridades de pesquisa da Embrapa..

À Embrapa - Empresa Brasileira de Pesquisa Agropecuária, pelo suporte financeiro e pela oportunidade de capacitação profissional.

À Embrapa Informática Agropecuária, em nome do Dr. Kleber Xavier Sampaio de Souza e da Dra. Sílvia Maria Fonseca Silveira Massruhá, pela oportunidade de utilizar as dependências físicas e a infraestrutura computacional durante o curso.

Às bibliotecárias da Embrapa Informática Agropecuária, em especial Carla Osawa, pela presteza na obtenção de diversas referências bibliográficas utilizadas neste trabalho.

Aos funcionários dos Recursos Humanos da Embrapa Informática Agropecuária, pela eficiência e disposição na resolução de problemas relacionados ao programa de pós-graduação.

Aos colegas da Embrapa Informática Agropecuária que me apoiaram e tiraram diversas dúvidas em relação ao trabalho. Em especial ao colega Roberto e à colega Poliana, que nunca deixaram de me ajudar quando os solicitei.

Aos meus pais, Ivone e Antônio (Nico), que sempre me deram todo o apoio e incentivo para que nunca parasse de estudar. Em especial a minha mãe, que nunca mediu esforços para me ver chegar até aqui.

À minha esposa Letícia e ao nosso filho Joaquim, que sempre estiveram ao meu lado me dando amor, apoio e compreensão durante toda essa empreitada, principalmente nos momentos em que tudo parecia estar perdido.

À minha avó Lázara, por suas orações, e a todos os familiares que me deram apoio em cada etapa deste caminho.

Ao meu amigo Edgard e ao meu amigo, e irmão de coração, Rogério, que sempre estavam dispostos a ouvir minhas lamentações e me incentivar para chegar ao final deste curso.

Ao meu padrasto Paulo e meus sogros, Eliana e Ademir, que me apoiaram e incentivaram em diversos momentos.

Aos colegas Flávio e Camila, que também muito me incentivaram. Em especial ao colega Flávio, com quem pude compartilhar os problemas de pesquisa e obter importantes soluções.

Aos demais colegas, professores e funcionários da Faculdade de Engenharia Agrícola, da Universidade Estadual de Campinas, que me apoiaram diretamente e indiretamente.

LISTA DE FIGURAS

LISTA DE TABELAS

1. INTRODUÇÃO

O Brasil possui diversas raças de ovinos que se desenvolveram a partir de raças trazidas pelos colonizadores portugueses, logo após o descobrimento. Ao longo desses quase cinco séculos, essas raças foram submetidas à seleção natural em diversos ambientes, a ponto de desenvolverem características de adaptação às diversas condições ambientais brasileiras. Essas raças aqui desenvolvidas passaram a ser conhecidas como crioulas ou localmente adaptadas. A maioria dessas raças encontra-se ameaçada de extinção, principalmente devido a cruzamentos indiscriminados com animais de raças exóticas que passaram a ser importadas a partir do final do século XIX (MARIANTE *et al.*, 2009).

As raças localmente adaptadas, apesar de não possuírem o mesmo potencial produtivo das raças exóticas melhoradas, constituem uma importante fonte de informações que pode levar à descoberta de genes envolvidos com determinadas características adaptativas, tais como resistência a diversas doenças e parasitas. Essas características permitem que os animais destas raças sejam mais adaptados que ovinos de outras raças (inclusive de raças exóticas melhoradas) a regiões de ambientes mais hostis. Essas informações fornecem um caminho muito interessante para futuras investigações, principalmente no entendimento da base genética envolvida na adaptação a estes ambientes (GOUVEIA, 2013).

Para evitar a perda deste importante e insubstituível material genético, a Embrapa decidiu incluir as raças localmente adaptadas no seu Programa de Pesquisa em Recursos Genéticos. Atualmente, a conservação dos recursos genéticos animais é realizada em bancos de germoplasma, que podem ser compostos de pequenos rebanhos de animais de uma raça que ficam submetidos à seleção natural (*in situ*), ou de material genético congelado, como sêmen, embriões e ovócitos (*ex situ*). Diversas raças localmente adaptadas estão presentes nestes bancos, sendo que as que possuem maior destaque nacional são as raças Crioula, Morada Nova e Santa Inês.

A seleção dos ovinos de uma determinada raça para compor estes bancos é realizada por meio de critérios tradicionais, tais como a avaliação de características morfológicas e produtivas. Entretanto, essa avaliação está sujeita a falhas, pois alguns animais cruzados mantêm características semelhantes àquelas dos animais locais. Com isto, identificar se os animais

depositados no banco são ou não pertencentes a uma raça é uma tarefa que exige muita cautela.

Para auxiliar na busca de soluções para este tipo de problema, o emprego de tecnologias advindas das áreas da genética e da computação é fundamental para atingir resultados mais precisos e confiáveis. Nos últimos anos houve um aumento na utilização de tecnologias que empregam análise do DNA na área animal, sendo que as que fazem uso de marcadores moleculares baseados em polimorfismos de DNA se destacam entre as mais importantes.

Dentre os tipos de marcadores moleculares existentes, os do tipo SNP (*Single Nucleotide Polimorphism*) mostraram ser mais efetivos no auxílio da certificação racial de animais domésticos (PANT *et al.*, 2012; SASAZAKI *et al.*, 2011; SUEKAWA *et al.*, 2010). Atualmente, as novas tecnologias para geração destes dados moleculares fornecem metodologias que são capazes de genotipar de dezenas até centenas de milhares de marcadores SNP em microarranjos (*microarrays*) de DNA de alta densidade em um único ensaio.

Desta forma, selecionar os marcadores mais informativos para a identificação racial de um ovino torna-se um problema desafiador. Uma das formas de se realizar esta seleção é por meio de um processo de mineração de dados, cujo objetivo é encontrar padrões e tendências em grandes volumes de dados (HAN *et al.*, 2011). Esse processo permite identificar e estudar o conjunto dos principais marcadores SNP. Para tanto, deve-se utilizar técnicas específicas que combinem seleção de atributos (ou variáveis) e geração de modelos preditivos. Estas técnicas devem ser capazes de gerar modelos que classifiquem novos exemplos a partir de experiências acumuladas em problemas anteriores e de lidar com problemas em que o número de atributos (p) é muito maior que o número de observações (n) ($p \gg n$). De acordo com James *et al.* (2013), a combinação dessas técnicas contribuem para eliminação de atributos redundantes e não-informativos, simplificam o modelo preditivo e reduzem o custo de processamento do algoritmo de aprendizado de máquina para construção do modelo.

Os modelos obtidos pelo processo de mineração de dados poderão ser utilizados na certificação racial dos animais já depositados nos bancos de germoplasma, e de novos animais a serem inclusos, assim como poderão ser utilizados por diversos segmentos ligados à ovinocultura, como por exemplo, por associações de criadores interessadas em certificar seus animais, e pelo MAPA (Ministério da Agricultura, Pecuária e Abastecimento), no controle de animais registrados que apresentam alelos de outras raças, possibilitando a reclassificação ou

mesmo a revogação desses animais registrados.

Além disso, os marcadores SNP selecionados pelos modelos poderão ser empregados na construção de ferramentas de genotipagem de marcadores SNP de baixa densidade, como os microarranjos, por exemplo (ROORKIWAL *et al.*, 2013; KIM e MISRA, 2007). Cabe ressaltar que, quanto menor o número de marcadores selecionados, menor o custo total de construção destas ferramentas de genotipagem, pois a preparação de cada SNP no arranjo custa um determinado valor (CAETANO, 2009).

O restante do trabalho está organizado como segue. O capítulo 2 apresenta a hipótese científica e os objetivos geral e específicos do trabalho.

No capítulo 3 encontra-se a revisão bibliográfica da temática desse trabalho, incluindo a apresentação do panorama da ovinocultura no Brasil e no mundo, assim como a descrição das raças Crioula, Morada Nova e Santa Inês. Além disso, neste capítulo faz-se uma explanação sobre alguns conceitos básicos sobre marcadores moleculares de polimorfismos de DNA, explica-se algumas definições sobre genética populacional e frequência alélica, apresenta-se noções gerais das tarefas de mineração de dados e alguns casos da aplicação dessas técnicas na seleção de marcadores de DNA.

O capítulo 4 apresenta o material e os métodos que foram utilizados para o desenvolvimento desse trabalho, desde a aquisição dos dados até os procedimentos utilizados nas análises dos resultados.

O capítulo 5 refere-se à exposição dos resultados obtidos com a aplicação de técnicas que combinam modelos preditivos e seleção de atributos no conjunto de dados de marcadores SNP de ovinos e à discussão com base na literatura.

Por fim, no Capítulo 6 são apresentadas as principais contribuições obtidas com a realização deste trabalho e também algumas sugestões de trabalhos futuros.

2. HIPÓTESE E OBJETIVOS

2.1. HIPÓTESE CIENTÍFICA

Para a elaboração da hipótese, foi realizado um estudo prévio dos trabalhos relacionados com seleção de SNP em dados de animais domésticos. De forma geral, observou-se que os trabalhos relacionados selecionaram um número menor que 100 marcadores SNP em seus modelos finais (MOKRY *et al.*, 2013; PANT *et al.*, 2012; SASAZAKI *et al.*, 2011; SUEKAWA *et al.*, 2010; ROHRER *et al.*, 2007; HEATON *et al.*, 2005), número pelo qual este trabalho buscou se referenciar como limite para a seleção dos marcadores mais informativos. Além disso, considerou-se o possível desenvolvimento de um microarranjo de baixa densidade, que aloca múltiplos de 48 marcadores SNP em sua superfície (ROORKIWAL *et al.*, 2013; KIM e MISRA, 2007).

Diante deste contexto, a hipótese deste trabalho é a seguinte:

- É possível selecionar um conjunto dos marcadores SNP mais importantes para a certificação das raças Crioula, Morada Nova e Santa Inês, por meio de um processo de mineração de dados, em que o número de marcadores SNP selecionados seja menor que 0,2% do conjunto de dados analisado (aproximadamente 50 mil marcadores), ou seja, abaixo de 100 marcadores.

2.2. OBJETIVO GERAL

- O objetivo geral deste trabalho é desenvolver modelos baseados em técnicas de mineração de dados para selecionar os principais marcadores SNP, cujos alelos sejam específicos para as raças Crioula, Morada Nova e Santa Inês.

2.3. OBJETIVOS ESPECÍFICOS

- Investigar técnicas que combinem modelos preditivos e seleção de atributos capazes de identificar os principais marcadores moleculares das raças ovinas estudadas, considerando um conjunto de dados em que o número de atributos é maior que o seu número de instâncias (observações).
- Identificar e selecionar os principais marcadores moleculares para as raças Crioula, Morada Nova e Santa Inês.

3. REVISÃO BIBLIOGRÁFICA

Neste capítulo será realizada uma discussão sobre a atual situação da ovinocultura e suas projeções no cenário nacional e internacional, assim como as características das raças Crioula, Morada Nova e Santa Inês. Em seguida, serão apresentados alguns conceitos sobre marcadores moleculares SNP. Posteriormente, explana-se sobre genética populacional e seu relacionamento com a frequência alélica dentro de uma população. Além disso, serão apresentados alguns conceitos em relação à mineração de dados e suas aplicações na seleção de marcadores SNP.

3.1. A OVINOCULTURA NO BRASIL E NO MUNDO

3.1.1. REBANHOS NO BRASIL E NO MUNDO

A presença da ovinocultura acontece em praticamente todos os continentes, fato que se deve fortemente a seu poder de adaptação aos diferentes climas, relevos e vegetações.

Os maiores rebanhos estão distribuídos pelos países pertencentes à Ásia, África e Oceania. A China se destaca como sendo o país com maior número de animais, seguido da Austrália, Índia, Irã, Sudão e Nova Zelândia (VIANA, 2008). Observa-se na Figura 1 os países

com maiores rebanhos de ovinos do mundo.

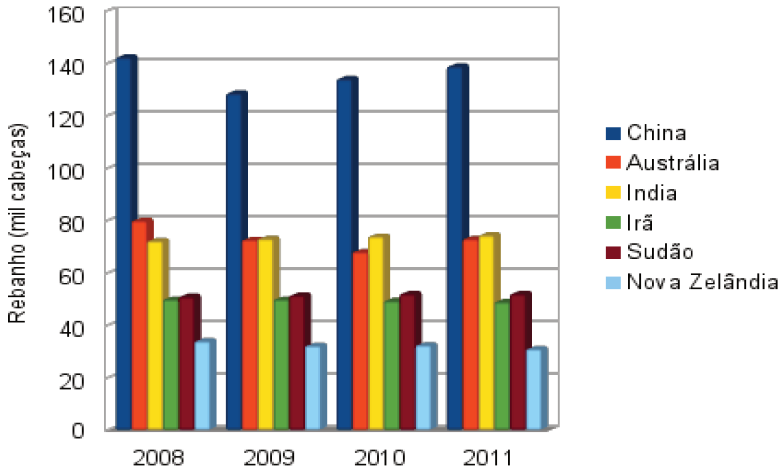

Figura 1: Principais rebanhos de ovinos do mundo, em milhões de cabeças, de 2008 a 2011. Fonte: FAO (2012).

Estima-se que a população de ovinos no mundo seja em torno de um bilhão de cabeças (MDIC, 2010). A Figura 2 ilustra a densidade populacional de ovinos (cabeças) por km^2 em todo o planeta, observando-se uma ampla difusão da espécie em todos os continentes.

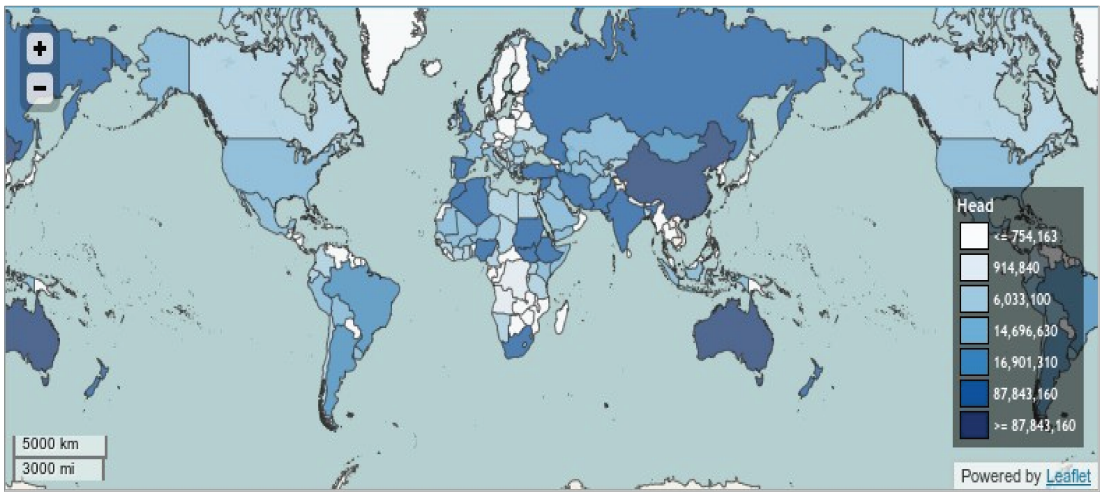

Figura 2: Densidade populacional mundial de ovinos (cabeças) por km^2. Fonte: FAO (2012).

Os seis países que possuem os maiores rebanhos ovinos concentram aproximadamente 41% do rebanho mundial. Além disso, deve-se destacar que cerca de 73% do rebanho ovino mundial está situado nos países do continente africano e asiático, o que demonstra a importância destes continentes no cenário internacional (SOUZA *et al.*, 2012).

O rebanho ovino brasileiro totalizava cerca de 17 milhões de cabeças de ovinos no ano de 2010, ocupando o décimo sétimo lugar entre os países com os maiores rebanhos do mundo. Apesar do número de cabeças ser bem menor que o dos grandes produtores mundiais, o rebanho de ovinos no Brasil, entre 2004 e 2010, cresceu aproximadamente 15,4%, enquanto que o rebanho mundial aumentou somente 1,04% neste mesmo período (REIS *et al.*, 2012). A Tabela 1 mostra o total dos rebanhos ovinos nos principais estados produtores do Brasil em 2010, em que se destacam o Rio Grande do Sul e alguns estados da região Nordeste.

Tabela 1: Estados brasileiros com os maiores rebanhos de ovinos em 2010. Fonte: IBGE (2012).

Rebanhos ovinos por estados brasileiros selecionados (Mil) - 2010	
Rio Grande do Sul	3.979
Bahia	3.125
Ceará	2.098
Pernambuco	1.622
Piauí	1.392
Brasil	**17.381**

A ovinocultura é explorada de modos diversos nas regiões geográficas do Brasil. Na região Sul, a criação ovina é composta de animais lanados, devido às temperaturas baixas na região, e da qual se obtém lã e carne. Nos estados da região Nordeste, os animais pertencem às raças deslanadas, e grande parte da produção é destinada à subsistência, produzindo carne, leite e derivados. Já na região Sudeste, os rebanhos são voltados para a produção de cortes especiais, com maior valor agregado aos mesmos (COSTA, 2007).

No início dos anos noventa, o estado do Rio Grande do Sul passou por algumas mudanças na produção de lã. Devido a um acordo entre os países produtores, no qual se definiu que a produção de lã deveria representar apenas 4% do mercado de fibras têxteis, o mercado da lã teve uma grande perda de rentabilidade, o que fez com que muitos produtores abandonassem a atividade, tendo como consequência uma redução drástica do rebanho. Diante desse fato, a atividade de produção de carnes e peles começou a ganhar maior destaque na ovinocultura brasileira (SILVA, 2002).

3.1.2. A OVINOCULTURA DE CORTE

Uma parte significativa da produção pecuária de corte no mundo é proveniente da ovinocultura (SANTOS, 2007). Ainda assim, o consumo de carne ovina é baixo em relação aos demais produtos de origem animal. Diante desta situação, o objetivo dos criadores de ovinos, no

mundo todo, está em elevar o consumo do produto, principalmente nos principais centros mundiais, o que ocasionará uma maior demanda por esta carne no mercado internacional, beneficiando os países produtores de carne de qualidade, inclusive o Brasil (VIANA, 2008).

Segundo a FAO (2012), o consumo médio mundial de carne ovina é menor que 2 kg per capita ano, entretanto, países como Mongólia, Turcomenistão, Nova Zelândia e Islândia são os maiores consumidores desta carne, com 49 kg, 26 kg, 23 kg e 19 kg per capita ano, respectivamente. Embora o consumo seja baixo no Brasil, a procura pela carne ovina é crescente e o país ainda depende de importações para supri-la. Observa-se que o mercado de carne ovina possui muitas possibilidades de crescimento como produto substituto, tendo o seu sabor diferenciado como principal característica positiva (ARAÚJO e MEDEIROS, 2003).

As tendências para o mercado de carne ovina são, portanto, promissoras. De acordo com o relatório “Perspectivas Agrícolas 2012-2021” (OECD-FAO, 2012), as importações de carnes por países em desenvolvimento deverão aumentar, impulsionadas pela urbanização e por um crescimento na renda. Dessa maneira, estima-se um crescimento anual de 1,8% na produção de carne ovina durante o período de 2012 a 2021, registrando-se essa elevação principalmente em países em desenvolvimento, os quais aumentarão sua participação global e representarão 78% da produção de carne ovina no mundo. A Figura 3 ilustra a evolução da produção mundial dos diversos tipos de carne desde 1995 e a projeção de crescimento de produção até 2021.

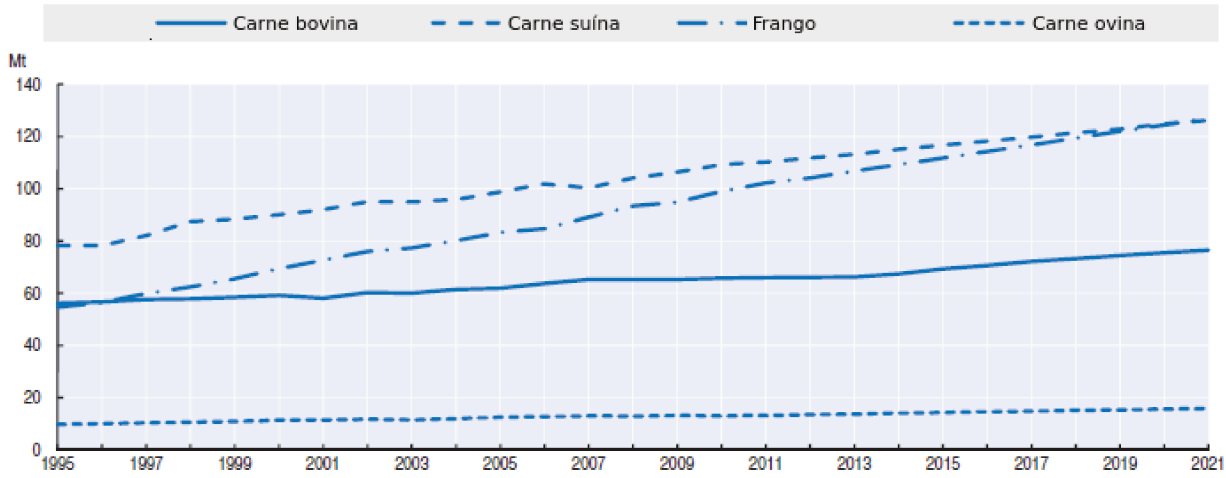

Figura 3: Crescimento da produção: por tipo de carne, 1995-2021 (equivalente peso carcaça ou pronta para preparo).

Fonte: Adaptado de (OECD-FAO, 2012).

Diante deste cenário, nosso país pode se beneficiar do crescimento da demanda de carne ovina pelos países importadores. Para tanto, algumas iniciativas devem ser tomadas para que o Brasil possa se tornar um grande exportador de carne ovina para países de maior consumo, como solucionar deficiências nos aspectos produtivos (sistema de manejo e melhoramento genético), aumentar o tamanho do rebanho nacional, combater abates clandestinos, incrementar a oferta de animais jovens para abate e fortalecer a cadeia produtiva através da organização de produtores (BRISOLA e ESPIRÍTO SANTO, 2003; VIANA, 2008).

3.1.3. AS RAÇAS CRIOULA, MORADA NOVA E SANTA INÊS

No Brasil existem variadas raças de animais domésticos que evoluíram a partir de raças que foram trazidas pelos colonizadores portugueses pouco tempo após o descobrimento. Ao longo dos séculos, essas raças passaram por um processo de seleção natural em diversos ambientes, até atingirem um ponto de apresentarem características específicas de adaptação a essas mudanças. Essas raças que aqui se desenvolveram ficaram conhecidas como “crioulas”,

“locais” ou “naturalizadas” (PAIVA, 2005).

Infelizmente, grande parte dessas raças está ameaçada de extinção, principalmente devido a cruzamentos indiscriminados com raças exóticas que começaram a ser importadas a partir do final do século XIX e inicio do século XX (MARIANTE *et al.*, 2009). Em geral, as raças naturalizadas de ovinos no Brasil são compostas de animais de pequeno porte, e, até os dias atuais, foram utilizadas em poucos processos de seleção artificial e melhoramento genético, o que faz com que ainda sejam pouco especializadas na produção intensiva de leite e/ou carne (PAIVA, 2005).

Dentre as diversas raças naturalizadas encontradas nas diferentes regiões do país, três delas serão utilizadas nessa pesquisa: Santa Inês, Morada Nova e Crioula. Essas raças foram selecionadas por se destacarem no programa de conservação da Empresa Brasileira de Pesquisa Agropecuária (Embrapa).

A raça Santa Inês, segundo Figueiredo *et al.* (1990), é originária do cruzamento da raça Bergamácia (lanada) com a Morada Nova (deslanada) e outros animais crioulos do Nordeste. Ainda de acordo com o autor, deu-se preferência pela sua criação no nordeste brasileiro devido à sua ausência de lã, a seu maior porte e também por se adaptar à vegetação arbustiva daquela região. Existe muita controvérsia acerca da origem destes animais. De acordo com Miranda (1990), a raça Bergamácia apenas chegou ao Brasil há cerca de 70 anos, o que seria um tempo insuficiente para se formar uma nova raça. Além disso, segundo Paiva (2005), a partir da década de 90, é possível notar que a morfologia externa dos animais Santa Inês apresenta algumas características da raça Somali Brasileira e de outras raças lanadas, com a raça inglesa Suffolk, principalmente.

A raça Santa Inês é destinada, principalmente, à produção de carne, sendo que seus animais são de grande porte, deslanados e com pelagem nas cores branca, vermelha, preta e chitada. Além disso, as ovelhas possuem uma excelente capacidade leiteira para criar os cordeiros e, em determinadas condições, podem ser férteis durante todo o ano (OSÓRIO e OSÓRIO, 2005).

Segundo Figueiredo *et al.* (1980), a raça Morada Nova é resultado do cruzamento de ovinos Bordaleiros, vindos de Portugal, com ovinos deslanados vindos da África época do tráfico de escravos. Os animais da raça Morada Nova constituem uma das principais raças nativas de ovinos deslanados do Nordeste brasileiro. Contudo, os rebanhos dessa raça vêm sofrendo uma

redução de tamanho nos últimos anos, pois grande parte dos criadores está preferindo a criação de outras raças, como a Dorper e, principalmente, a Santa Inês. Há, também, muitos cruzamentos indiscriminados com outras raças exóticas, comprometendo ainda mais a preservação dessa importante raça e seu genótipo (FACÓ *et al.*, 2008).

A criação de ovinos Morada Nova está voltada, essencialmente, para produção de carne e pele, a qual é muito valorizada no mercado internacional (FERNANDES, 1992). Além disso, por possuírem um porte diminuto e estarem bem adaptados às condições ambientais do semi-árido, os ovinos Morada Nova estão presentes em muitas das pequenas propriedades, constituindo uma importante fonte de alimentação da população rural (GURGEL *et al.*, 1992; FERNANDES *et al.*, 2001).

A raça Crioula lanada talvez seja a raça naturalizada brasileira que mais se assemelhe com as raças dos países ibéricos. Os ovinos da raça Crioula foram trazidos da Espanha e Portugal pelos colonizadores, e podem ser encontrados no Sul do Brasil e também em quase todos os países sul-americanos (PAIVA, 2005). De acordo com Mariante e Cavalcante (2000), é provável que os ovinos dessa raça sejam originários da raça Churra espanhola. Além disso, fornece uma lã que, apesar de possuir uma qualidade inferior à de raças especializadas, é bastante utilizada em artesanato.

Apesar dos animais dessa raça terem sobrevivido durante séculos às adversidades climáticas e nutricionais encontradas aqui no Brasil, o patrimônio genético desses ovinos encontra-se seriamente ameaçado de extinção, em consequência do cruzamento descontrolado com animais de raças exóticas e também da substituição da Crioula por outras raças mais produtivas em lã, carne e pele (VAZ, 2000).

A Figura 4 ilustra as características visuais (fenotípicas) dos animais destas três raças.

Figura 4: Algumas das raças encontradas no Brasil: Crioula, Morada Nova e Santa Inês.

Fonte: www.uniovinos.com.br.

3.2. MARCADORES MOLECULARES DE POLIMORFISMOS DE DNA

Nos últimos anos, a evolução dos conhecimentos sobre o conteúdo da informação genética, assim como as tecnologias disponíveis para o sequenciamento de genomas em larga escala, ocorreu de uma forma sem precedentes. Como resultado desta evolução, um volume muito grande de informações sobre genomas de diversos organismos foi se acumulando nos bancos de dados públicos. Com a importante ajuda da Bioinformática, principalmente pela sua multidisciplinaridade, novos produtos e tecnologias estão sendo gerados a partir destas informações, todos com boas perspectivas comerciais e com possibilidades de revolucionar a pecuária ao auxiliarem na resolução de problemas relacionados, por exemplo, à qualidade da carne.

Como resultado dessa grande evolução conjunta da biotecnologia e da bioinformática, as aplicações que utilizam análise do DNA são diversas na área animal, sendo que as que fazem uso de marcadores moleculares se incluem entre as mais importantes.

Para se definir marcadores moleculares de polimorfismos de DNA, é fundamental saber diferenciar o que são fenótipos e genótipos. Os fenótipos são as características apresentadas por um indivíduo, sejam morfológicas, comportamentais, fisiológicas, assim como características microscópicas e de natureza bioquímica, que somente podem ser identificadas após testes

específicos. Como características fenotípicas observáveis, podemos citar, no caso dos ovinos, a cor de sua pelagem, se são lanados ou deslanados, se possuem pelos longos ou curtos, etc. Os genótipos, por sua vez, são informações hereditárias de um organismo contidas em seu genoma, formando a constituição genética de um indivíduo e moldando as características fenotípicas do mesmo.

Dentre os dados genotípicos estão os marcadores moleculares de polimorfismos de DNA, que funcionam como "pontos de referência" no mapa genético. Dada a quantidade em que estão presentes, a partir da descoberta destes marcadores, o mapa do genoma, em geral, teve um incremento significativo. Isto se deve, em grande parte, à escassez de genes no genoma, sendo que muitas vezes os estudos de alguns fenótipos (principalmente os de difícil mensuração) são feitos em relação a estes marcadores em vez de serem feitos em relação a um gene (FARAH, 2007).

Dentre os diversos tipos de marcadores moleculares, os microssatélites e os SNP são os mais utilizados nos dias atuais. Antes de se detalhar um pouco mais cada um deles, deve-se saber que a diferença básica entre estes dois tipos de marcadores moleculares é que, ao se utilizar microssatélites a quantidade de marcadores é da ordem de centenas para cada animal; por outro lado, com o advento dos microarranjos de genotipagem, o número de SNP já alcança a ordem de centenas de milhares por animal, o que aumenta a área de cobertura de análise do genoma estudado.

Os microssatélites são regiões no genoma que apresentam sequências repetitivas de uma, duas, três ou quatro sequências de nucleotídeos, sendo a de duas sequências a mais comum. Devido ao seu alto grau de polimorfismo (este tipo de marcador pode possuir mais de 15 alelos), a chance de dois indivíduos possuírem o mesmo valor alélico de marcadores microssatélites é extremamente pequena. Com isso, as regiões microssatélites têm sido amplamente utilizadas em genética forense e testes de exclusão de paternidade (PAIVA, 2005).

Os marcadores SNP, por sua vez, constituem uma variação que ocorre em apenas um único nucleotídeo da cadeia de bases nitrogenadas (Adenina, Citosina, Timina e Guanina) do DNA, afetando ou não o fenótipo alvo entre os membros de uma espécie em estudo. Por exemplo, dois fragmentos de DNA sequenciados de uma mesma população, AAGC**C**TA e AAGC**T**TA, contêm uma diferença de um único nucleotídeo (Figura 5). Neste caso, pode-se dizer

que existem dois alelos: C e T.

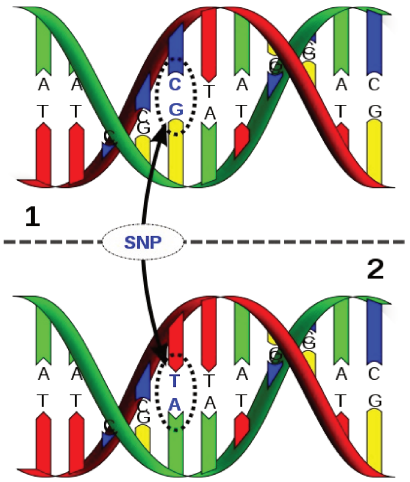

Figura 5: Diferentes valores alélicos para um determinado loco (SNP).

Fonte: https://en.wikipedia.org/wiki/Single-nucleotide_polymorphism.

Para uma variação ser considerada um SNP, a mesma deve ocorrer em no mínimo 1% de uma determinada população. Caso a frequência de uma variação seja inferior a 1%, então será considerada uma simples mutação (GIBSON e MUSE, 2009). Normalmente, os SNP estão ligados a algumas doenças hereditárias e às respostas dos organismos a fármacos, toxinas e produtos químicos (KIM e MISRA, 2007). Em pesquisas com animais domésticos, os marcadores SNP trouxeram significativos avanços em estudos direcionados para identificação de genes que controlam características de interesse econômico (CAETANO, 2009)

Os marcadores SNP aparecem, na maior parte das vezes, em espaços sem função determinada, chamados de intergênicos, entretanto, também há ocorrências de SNP em regiões codificadoras de proteínas, sendo que, segundo estudos realizados com humanos e espécies de interesse zootécnico, devem existir milhões de SNP no genoma destes indivíduos (LI *et al.*, 2009). O uso destes marcadores em pesquisas com animais domésticos está voltado para estudos de associação e mapeamento genético, assim como para testes de confirmação de paternidade e identificação individual (rastreabilidade), entre outros (CAETANO, 2009).

A nomenclatura de marcadores SNP segue um determinado padrão na maioria dos organismos. Por exemplo, o marcador OAR2_55861669.1, encontrado em ovinos, indica que o marcador está presente na espécie *Ovis Aries*, dentro do cromossomo 2 (OAR2), e sua posição dentro do cromossomo é 55.861.669. O número 1, no final do nome, indica a versão daquele marcador encontrado.

3.3. MICROARRANJO DE MARCADORES SNP

Um microarranjo de marcadores SNP consiste num arranjo pré-definido de microscópicas estruturas denominadas sondas, que ficam ligadas quimicamente a uma lâmina de vidro. As sondas são preparadas e fixadas nas lâminas por robôs altamente precisos e são comumente compostas por moléculas sintéticas de DNA (oligonucleotídeos) de 25 a 50 nucleotídeos (KIM e MISRA, 2007). Nestas sondas serão preparadas as regiões dos marcadores que se buscam encontrar nas amostras, com seus respectivos alelos (formas alternativas de um marcador que podem ocorrer em determinado loco cromossômico).

Os microarranjos de SNP são utilizados na detecção de ácidos nucleicos, como DNA genômico, originados de amostras biológicas que são colocadas para hibridizar com o DNA inserido nas sondas do arranjo. Neste processo, ocorre uma ligação por complementariedade de bases da cadeia simples de DNA da amostra com a cadeia simples da sonda, formando uma cadeia dupla por hibridização. A detecção do marcador é possível pois as amostras são marcadas com materiais corantes fluorescentes, que pode variar de uma a quatro cores numa amostra (KIM e MISRA, 2007; FARAH, 2007).

Após a hibridização e lavagem do microarranjo, os resultados são computados por um microscópio que envia um raio laser e registra a fluorescência obtida em cada ponto específico. Os dados fornecidos pelo(s) corante(s) são registrados separadamente, formando um quadro de todos os pontos que emitem fluorescência e sua intensidade relativa. Por fim, um programa de computador é responsável por combinar as imagens produzidas por cada corante (FARAH, 2007). Considerando, por exemplo, uma amostra da raça X com corante verde. Caso as células da raça X possuam um SNP homozigoto, ou seja, com um par de alelos idênticos (exemplo: AA), a

sonda que reconhece esse SNP com apenas esse alelo, terá a marcação verde mais clara, indicando que o animal analisado da raça X possui o SNP com apenas aquele alelo. Da mesma forma, um SNP que seja heterozigoto, ou seja, tenha um par de alelos diferentes um do outro (exemplo: AG), produzirá uma marcação verde mais intensa, pois mais cadeias simples (com corante) se ligam à sonda. Se o ponto no painel aparecer preto, isso significa que o SNP não foi encontrado na amostra. A Figura 6 mostra uma representação esquemática do microarranjo de SNP.

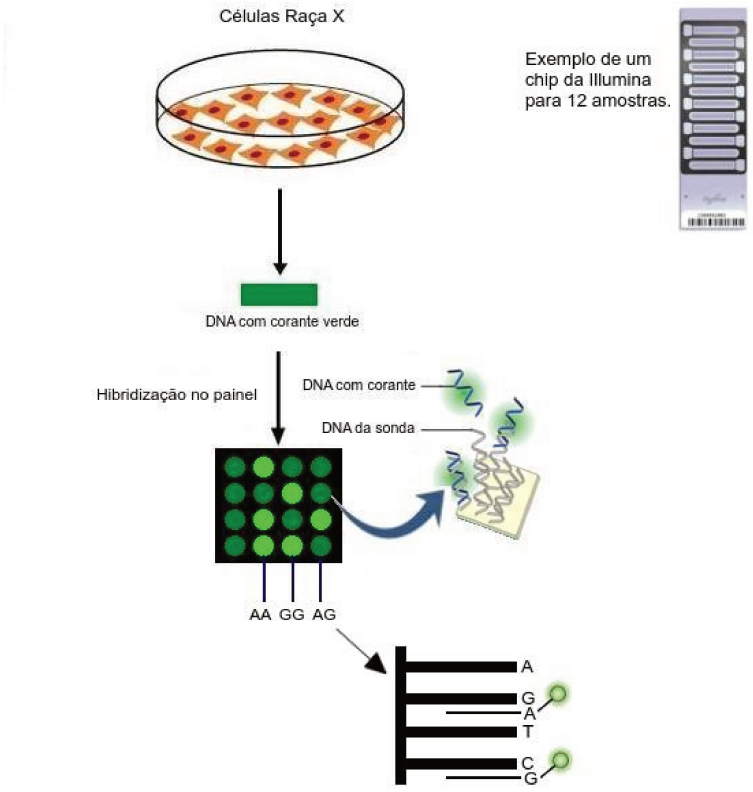

Figura 6: Representação esquemática de um microarranjo (microarray) de SNP.

Fonte: Adaptado de http://en.wikipedia.org/wiki/DNA_microarray.

3.4. GENÉTICA POPULACIONAL E FREQUÊNCIA ALÉLICA

A Genética Populacional estuda a distribuição das frequências alélicas e genotípicas (pares de alelos) nas populações e como as mesmas são mantidas ou alteradas durante as gerações. Este ramo da biologia analisa os fatores genéticos (mutação e reprodução) e ambientais (seleção e migração) que determinam a freqüência e a distribuição de determinados fenótipos, principalmente doenças, em famílias e comunidades (BEIGUELMAN, 2008).

Intuitivamente, seria possível concluir que características e alelos dominantes em uma população teriam a forte tendência de aumentar, em freqüência, em relação aos alelos recessivos, visto que a proporção de heterozigotos com alelos dominantes normalmente é muito maior que os recessivos. Entretanto, o que se observa numa população contraria tal suposição. O que ocorre, realmente, é a manutenção das proporções entre os diferentes genótipos e fenótipos por várias gerações. Este conceito é conhecido como a lei de Hardy-Weinberg (HARTL, 1997).

Desta forma, respeitadas algumas premissas básicas, tais como ausência de seleção natural e mutação no *locus* em questão, a frequência dos alelos de uma população pode ser calculada por meio da Equação 1, derivada da lei de Equilíbrio de Hardy-Weinberg.

$$Frequência\ do\ alelo = \frac{n^o\ total\ do\ alelo\ analisado\ naquele\ locus}{n^o\ total\ de\ alelos\ naquele\ locus} \quad (1)$$

Nesta equação, o número total do alelo analisado, para aquele *locus* (um marcador SNP, por exemplo), corresponde ao total de vezes que o alelo analisado aparece em uma determinada população, e número total de alelos corresponde ao total de todos alelos presentes nesta mesma população. Quando o indivíduo da população for heterozigoto, ou seja, apresentar dois alelos diferentes para o genótipo do *locus* em questão, conta-se apenas uma vez o alelo analisado. Caso o indivíduo seja homozigoto, ou seja, apresente dois alelos iguais para o genótipo do *locus* em questão, multiplica-se por dois o alelo analisado.

Como exemplo, supõe-se uma população de 30 animais da raça X e 30 animais da raça Y, distribuídos de acordo com a Tabela 2.

Tabela 2: Distribuição dos genótipos pelas raças X e Y.

Raça	Genótipo	Quantidade	Genótipo	Quantidade	Total
X	AA	25 animais	AG	5 animais	30 animais
Y	CC	20 animais	CA	10 animais	30 animais

De acordo com a Equação 1, têm-se os seguintes valores de frequência para cada alelo da raça X:

$$Frequência\ do\ alelo\ A = \frac{2\,x\,Total\ de\ AA + Total\ de\ AG}{Total\ de\ alelos\ na\ Raça\ X} = \frac{2\text{x}25+5}{60} = 0{,}92 \text{ ou } 92\%$$

$$Frequência\ do\ alelo\ G = \frac{2\,x\,Total\ de\ GG + Total\ de\ AG}{Total\ de\ alelos\ na\ Raça\ X} = \frac{5}{60} = 0{,}08 \text{ ou } 8\%$$

Para o alelo A, da raça X, têm-se 25 animais com genótipo AA e cinco animais com genótipo AG. Multiplicando-se por dois os animais homozigotos e somando este valor ao número de animais heterozigotos, obtém-se o total de 55 alelos A na raça X. Para o alelo G desta mesma raça, existem apenas cinco animais heterozigotos com genótipo AG, totalizando cinco alelos G na raça X.

Seguindo os mesmos passos para os cálculos de frequência da raça X, as frequências dos alelos da raça Y são as seguintes:

$$Frequência\ do\ alelo\ C = \frac{2\,x\,Total\ de\ CC + Total\ de\ CA}{Total\ de\ alelos\ na\ Raça\ Y} = \frac{2\text{x}20+10}{60} = 0{,}83 \text{ ou } 83\%$$

$$Frequência\ do\ alelo\ A = \frac{2\,x\,Total\ de\ AA + Total\ de\ CA}{Total\ de\ alelos\ na\ Raça\ Y} = \frac{10}{60} = 0{,}17 \text{ ou } 17\%$$

Nesse exemplo, observa-se que o alelo específico (ou alelo predominante) da Raça X é o alelo A, cuja frequência (92%) é muito maior do que na Raça Y (17%). Por outro lado, o alelo C é específico da Raça Y, na qual está presente em 83% da população e ausente na Raça X.

Entretanto, uma alta frequência alélica nem sempre significa que um alelo seja específico para aquela raça, pois pode ocorrer de outra raça também possuir uma alta frequência daquele mesmo alelo. Desta forma, verificar a frequência do alelo em todas as populações analisadas se torna necessário para confirmar se o alelo está mais presente em uma população em relação a outra.

3.5. DESCOBERTA DE CONHECIMENTO EM BANCOS DE DADOS

Nos últimos anos, observa-se que uma grande quantidade de dados cresce de forma rápida em diversos campos de conhecimento, fato que dificulta a interpretação dos mesmos, pois o ritmo de crescimento do volume destes dados é maior que o poder de interpretá-los. Desta forma, surgiu a necessidade do desenvolvimento de ferramentas e técnicas automatizadas para minimizar esta situação, as quais pudessem auxiliar o analista na transformação dos dados em conhecimento (HAN *et al.*, 2011).

Grande parte dessas técnicas e ferramentas podem ser encontradas dentro do processo de Descoberta de Conhecimento em Bases de Dados, ou somente KDD, cuja sigla em inglês significa *Knowledge Discovery in Databases*. Segundo Fayyad *et al.* (1996), a descoberta de conhecimento em bancos de dados é definida como um processo não trivial que busca identificar padrões novos, potencialmente úteis, válidos e compreensíveis, com o objetivo de melhorar o entendimento de um problema ou um procedimento de tomada de decisão.

O processo KDD se originou da intersecção de várias áreas de pesquisa, tais como aprendizado de máquina, reconhecimento de padrões, estatística, banco de dados, visualização de dados, inteligência artificial e computação de alto desempenho (FAYYAD *et al.*, 1996). Por este motivo, as técnicas existentes no KDD não devem ser consideradas substitutas de outras formas de análise (por exemplo, OLAP (*Online analytical processing*)), mas, sim, uma forma de se aperfeiçoar os resultados obtidos através das explorações realizadas pelas ferramentas atuais (REZENDE *et al.*, 2003).

As aplicações das técnicas estão presentes em praticamente todos os setores do conhecimento humano, como na área de negócios, onde existem vários casos, como por exemplo: detecção de fraudes em cartões, criação de perfis de clientes de acordo com suas compras, entre

outros; na agricultura, com sistemas de previsão de geadas, sistemas de alerta para a ferrugem do cafeeiro, sistemas de alerta para a ferrugem asiática da soja, entre outros; na medicina, onde se pode identificar terapias médicas de sucesso para diversas doenças; na bioinformática, para se buscar padrões em sequências de DNA, por exemplo; entre muitas outras possibilidades.

Segundo Fayyad *et al.* (1996), o processo de KDD é interativo e iterativo, além de envolver vários passos, exibidos na Figura 7, com muitas decisões sendo feitas pelo especialista do domínio de aplicação.

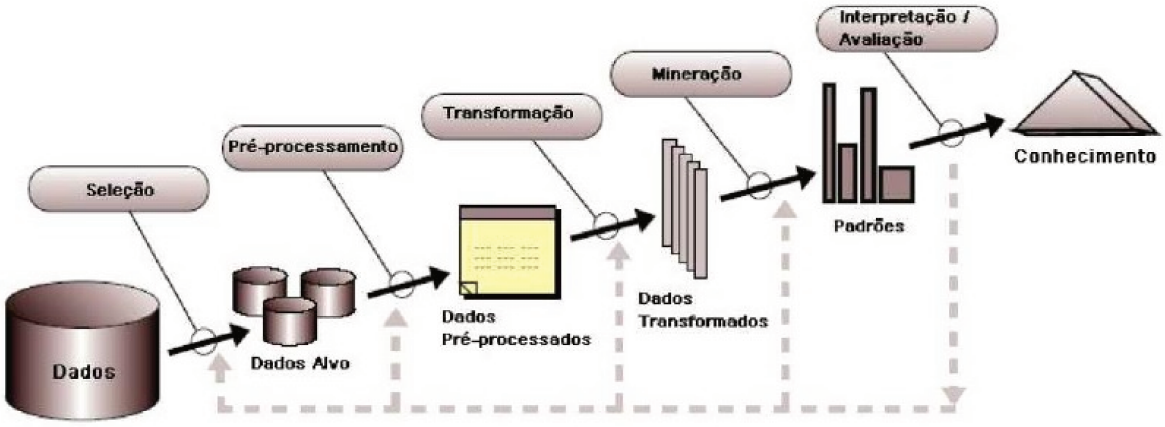

Figura 7: As fases do processo KDD.

Fonte: Adaptado de FAYYAD *et al.* (1996).

Os passos do processo KDD consistem em:

i. **Identificação do Problema**: compreensão do domínio da aplicação e do tipo de conhecimento a ser procurado, além de se identificar o objetivo do processo KDD.
ii. **Criação do conjunto de dados alvo (Seleção)**: realizar a seleção de um conjunto de dados, ou se fixar num subconjunto de registros (instâncias), onde a descoberta deve ser

feita.

iii. **Limpeza de dados e pré-processamento (Pré-processamento)**: neste passo estão operações básicas como remoção de ruídos quando necessário, coleta da informação necessária para modelar ou prever ruído, e decisão sobre quais estratégias se adotar para tratar atributos com valores faltantes.

iv. **Redução de dados e projeção (Transformação)**: busca por características úteis que possam representar os dados dependendo do objetivo da tarefa, visando à redução de dimensionalidade, ou seja, redução do número de atributos e/ou registros a serem considerados para o conjunto de dados.

v. **Mineração de dados (Mineração)**: escolha do(s) algoritmo(s) de mineração de dados e de métodos a serem aplicados para a busca por padrões de interesse numa forma particular de representação ou conjunto de representações.

vi. **Interpretação dos padrões descobertos (Interpretação/Avaliação)**: realizam-se análises dos padrões descobertos com o objetivo de descobrir se estes apresentam conhecimento novo em aplicações práticas. Algumas vezes, há a necessidade de se retornar aos passos 1-6 para avaliação posterior.

vii. **Implantação do conhecimento descoberto (Conhecimento)**: incorporação deste conhecimento à performance do sistema ou, simplesmente, documentá-lo e reportá-lo às partes interessadas.

3.5.1. TAREFAS E TÉCNICAS DE MINERAÇÃO DE DADOS

Uma tarefa de mineração de dados consiste na especificação do que se pretende buscar, ou que tipo de regularidade ou padrões interessa encontrar.

Na etapa de mineração de dados propriamente dita deve ser feita a escolha da tarefa a ser empregada, assim como a definição do algoritmo. Esta escolha deve ser baseada nos objetivos que se deseja atingir com a solução a ser encontrada. As possíveis tarefas de um algoritmo para se extrair padrões podem ser agrupadas em preditivas e descritivas (HAN *et al.*, 2011), ilustradas

na Figura 8.

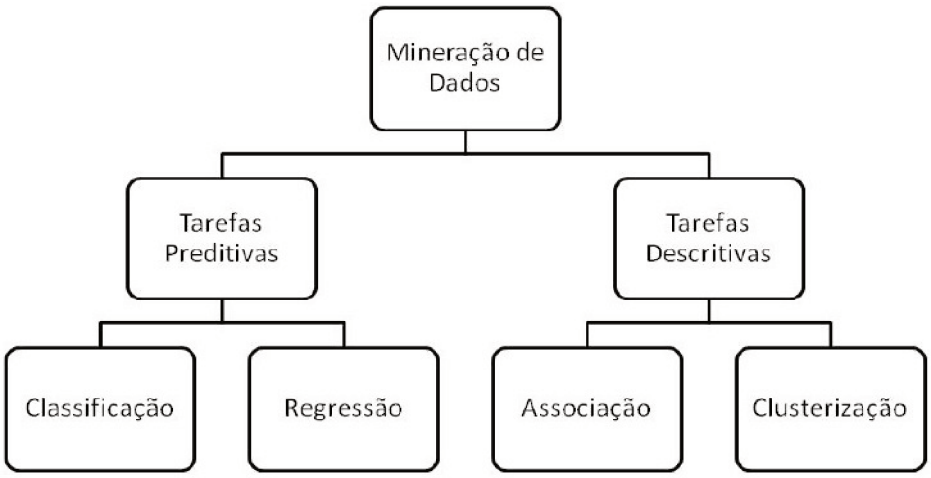

Figura 8: Tarefas de Mineração de Dados.

Fonte: Adaptado de REZENDE *et al.*, 2003.

As tarefas preditivas têm como objetivo principal a construção de modelos que possam predizer a variável resposta de um novo exemplo a partir de exemplos ou experiências passadas com respostas já conhecidas. As tarefas descritivas procuram identificar padrões intrínsecos a um conjunto de dados que não possui uma variável resposta determinada. A escolha de uma ou mais tarefas dependerá do problema a ser solucionado. As tarefas tradicionais de mineração de dados representadas na Figura 8 são brevemente descritas a seguir.

- **Classificação**: consiste na predição do valor de um atributo alvo do tipo discreto ou categórico por meio da construção de modelos e regras a partir de um conjunto de exemplos pré-classificados corretamente, para posterior classificação de exemplos novos e desconhecidos (HAN *et al.*, 2011). O grande desafio para os algoritmos de classificação é gerar modelos que possuam boa capacidade de generalização, ou seja, que estejam aptos a predizer, com alta taxa de acerto, os rótulos das classes para registros que não foram utilizados durante a construção do modelo (TAN *et al.*, 2005).
- **Regressão**: se constitui numa técnica estatística muito empregada para se realizar

predições (HILL et al, 2003). Essas predições procuram encontrar tendências de variações no conjunto de dados analisado em função dos atributos existentes. Possui um conceito semelhante à classificação, porém se aplica na predição de um valor alvo do tipo contínuo.

- **Associação**: determinam o quanto a presença de um certo conjunto de atributos nos exemplos de uma base de dados implica na presença de algum outro conjunto de atributos nos mesmos exemplos (AGRAWAL e SRIKANT, 1994). As regras de associação podem ser apresentadas no formato $L \rightarrow R$, onde L e R são, respectivamente, o lado esquerdo (*Left Hand Side*) e o lado direito (*Right Hand Side*), tal que $L \cap R = \varnothing$, de forma que representam conjuntos distintos de atributos. Basicamente, essas regras definem a relação existente entre L e R, demonstrando o quanto a presença de L implica a presença de R.
- **Agrupamento (clusterização)**: é uma tarefa descritiva que procura identificar agrupamentos (*clusters*) finitos de objetos similares entre si e dissimilares entre os grupos no conjunto de dados. De forma diferente da classificação, onde as denominações de classes são conhecidas, a clusterização analisa os dados onde as denominações de classes não estão definidas.

Cada tarefa de mineração de dados possui diferentes técnicas associadas. Dentre as mais populares estão (HAN *et al.*, 2011): árvores de decisão, redes neurais, regressão linear ou não linear, k-vizinhos mais próximos. Existem também as abordagens híbridas, que utilizam duas ou mais técnicas em conjunto.

Não existe a técnica ideal, cada uma delas possui suas vantagens e desvantagens. Assim, ao se escolher uma técnica, deve ser realizada uma análise bem apurada do problema em questão, levando em consideração o formato dos dados e como o conhecimento descoberto pode ser representado. Se necessário, pode se aplicar mais de uma técnica para solucionar o mesmo problema e no final escolher o modelo que apresente os melhores resultados.

Devido ao elevado número de atributos (marcadores SNP) e o baixo número de registros (animais), técnicas preditivas capazes de lidar com esta situação são comumente utilizadas. Entre elas estão: LASSO (*Least Absolute Shrinkage and Selection Operator*), Random Forest e

Boosting. Essas técnicas possuem um procedimento de seleção de atributos embutido na construção do próprio modelo. Diferentemente dos métodos tradicionais de seleção, como o Ganho de Informação e o Qui-quadrado, que utilizam ranqueamento dos atributos de acordo com a relevância de cada um deles para o modelo, LASSO, Random Forest e Boosting são indicadas para bases de dados genômicos, em que o número de genes e marcadores moleculares (p) é muito maior que o número de observações fenotípicas (n) ($p \gg n$), utilizando procedimentos internos para controlar a seleção de atributos redundantes (JAMES *et al.*, 2013).

3.5.2. LASSO

LASSO (TIBSHIRANI, 1997) é um método de regressão penalizada comumente utilizado em análises estatísticas, e ultimamente, também vem se constituindo uma alternativa bastante atraente para identificação de SNP relevantes em análise de dados de DNA (AYERS e CORDELL, 2010; WU *et al.*, 2009).

LASSO utiliza um algoritmo especial para reduzir os efeitos dos atributos, que não estão ligados à classe, aproximando seus coeficientes para zero ou próximo de zero. Estimar o efeito de um atributo como zero é o mesmo que excluí-lo do modelo. O método é usado normalmente para estimar os parâmetros de regressão $\beta=(\beta_1,\beta_2,\ldots,\beta_p)^T$ no modelo da Equação 2:

$$y_i=\mu+\sum_{j=1}^{p} x_{ij}\beta_j+e_i=\mu+X_i\beta+e_i \qquad (2)$$

onde, no caso deste trabalho, y_i é a raça do i-ésimo animal $(i=1,2,\ldots,n)$; μ é o coeficiente denominado intercepto, cujo valor é comum a todos os registros; x_{ij} é o valor do genótipo do marcador j $(j=1,2,\ldots,p)$ do animal i; o coeficiente β_j representa o efeito do marcador j na raça; e_i é o erro residual.

LASSO é um algoritmo inicialmente formulado para os modelos de regressão linear, nos quais o atributo alvo é contínuo. Para regressão linear, o método LASSO baseia-se na

minimização da soma de quadrados residuais (SQR) para obter os valores dos coeficientes β_j (JAMES *et al.*, 2013). LASSO também pode ser utilizado para resolver um problema de classificação, onde o atributo alvo é do tipo texto (nominal ou ordinal), como uma extensão da regressão logística, onde se desenvolvem modelos a partir de dados cujo atributo alvo contém valores do tipo categórico.

Em problemas de classificação, LASSO estima os coeficientes β_j do modelo por meio da maximização do logaritmo da função de verossimilhança, impondo a restrição de que a soma dos valores dos coeficientes absolutos seja limitada por uma constante (HASTIE *et al.*, 2011). A ideia básica por trás da maximização do logaritmo de verossimilhança é obter os coeficientes de μ e β de forma que a probabilidade $\hat{p}(x)$ de predição da classe correta (no caso, a raça) de um indivíduo, dado uma ou mais variáveis *x,* corresponda, em grande maioria das vezes, à raça do animal observado (JAMES *et al.*, 2013).

A função logística utilizada para fornecer a probabilidade de uma observação Y (animal) pertencer a uma classe l (raça), dado um ou mais atributos x, é exibida na Equação 3, a qual fornece valores de saída entre 0 e 1 (quanto mais próximo de 1, maior a probabilidade do atributo ser importante para o modelo).

$$p(x) = Pr(Y = l \mid x) = \frac{e^{\mu + \sum_{j=1}^{p} x_{ij}\beta_j}}{1 + e^{\mu + \sum_{j=1}^{p} x_{ij}\beta_j}} \tag{3}$$

Para desenvolver o modelo da Equação 3, utiliza-se a função de máxima verossimilhança (com penalização), descrita na Equação 6.

Com uma pequena manipulação na Equação 3, encontra-se a seguinte Equação 4:

$$\frac{p(x)}{1 - p(x)} = e^{\mu + \sum_{j=1}^{p} x_{ij}\beta_j} \tag{4}$$

A relação *p(x)/[1-p(x)]* é denominada *odds*, a qual indica a chance (ou probabilidade) de um evento ocorrer (neste caso, a raça) dividida pela probabilidade da não ocorrência do mesmo

evento. Aplicando-se o logaritmo nos dois lados, obtém-se a Equação 5 a seguir, denominada *logit*:

$$\log\left(\frac{p(x)}{1-p(x)}\right)=\mu+\sum_{j=1}^{p} x_{ij}\beta_j \tag{5}$$

Sendo $\hat{\beta}=(\hat{\beta}_1,\hat{\beta}_2,\ldots,\hat{\beta}_p)^T$, a estimativa LASSO $(\hat{\mu},\hat{\beta})$ para problemas de classificação é definida pela função de máxima verossimilhança penalizada descrita na Equação 6:

$$l(\hat{\mu},\hat{\beta}) \;=\; \frac{1}{n}\sum_{i=1}^{n}\left[y_i\left(\mu+\sum_{j=1}^{p} x_{ij}\beta_j\right)-\log\left(1+e^{\mu+\sum_{j=1}^{p} x_{ij}\beta_j}\right)\right] \tag{6}$$

$$\text{sujeito à restrição } \sum_{j=1}^{p}|\beta_j|\le t \quad \text{para } t\ge 0\,,$$

onde *t* é um parâmetro de penalização, também representado pela letra grega λ (lambda) em outra formulação da equação, e que deve ser determinado separadamente. Normalmente, os algoritmos de implementação do LASSO fornecem o valor ótimo para tal parâmetro, utilizando uma análise por validação cruzada de um intervalo de *n* possíveis valores. Essa restrição permite que algumas estimativas dos coeficientes de regressão sejam exatamente zero, realizando simultaneamente um procedimento de encolhimento e seleção de modelos.

Observando a Figura 9, a área azul representa a restrição *t* do LASSO e os círculos em vermelho representam os coeficientes de estimativas β calculados pela maximização de verossimilhança. Assim, se *t* for suficientemente grande, mais semelhante será o método LASSO de uma regressão logística comum, ou seja, selecionará os mesmos atributos, entre os quais podem ter muitos irrelevantes. Portanto, quanto maior a restrição LASSO (menor valor de *t*), menos atributos irrelevantes entram no modelo.

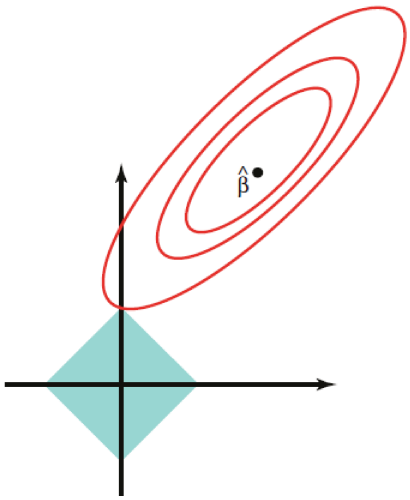

Figura 9: Área azul indicando restrição LASSO e estimativas do método de máxima verossimilhança circundados em vermelho. Quanto menor área azul, menos atributos entram no modelo.

Fonte: Adaptado de JAMES *et al.*, 2013.

3.5.3. RANDOM FOREST

Random Forest é uma técnica de classificação e regressão desenvolvida por Breiman (2001), que consiste num conjunto de árvores de decisão combinadas para solucionar problemas de classificação. Cada uma dessas árvores de decisão é construída utilizando uma amostra aleatória inicial dos dados e, a cada divisão desses dados, um subconjunto aleatório de *m* atributos é utilizado para escolha do atributo mais informativo. Assim, a técnica utiliza a estratégia denominada *bagging* (*Booststrap Aggregating*) (BREIMAN, 1996), a qual consiste numa abordagem para criar classificadores em amostras *bootstrap* dos dados na montagem da floresta, e a seleção aleatória de atributos para a construção de cada árvore de decisão. No final, Random Forest gera uma lista dos atributos mais importantes no desenvolvimento da floresta, que são determinados pela importância acumulada do atributo nas divisões dos nós de cada árvore da floresta (JAMES *et al.*, 2013).

O algoritmo básico da técnica Random Forest, num problema de classificação, segue os

seguintes passos da Figura 10:

A Figura 11 ilustra um exemplo de um conjunto de árvores de decisão desenvolvidas utilizando amostras *bootstrap* e seleção aleatória de atributos.

Dado um conjunto de dados $X = x_{1,} x_{2,} \ldots, x_j$ e $Y = y_{1,} y_{2,\ldots} y_k$.

Para $b = 1, 2, 3, \ldots B$, repita:

(a) Cria uma amostra *bootstrap* (X_b, Y_b) com *n* exemplos de (X, Y).

(b) Ajusta uma árvore de decisão f^b para o conjunto de treinamento (X_b, Y_b), utilizando *m* atributos para a escolha de cada nó.

Fim de repetição.

Gera o modelo final: $\hat{f}(x) = \sum_{b=1}^{B} f^b(x)$, que calcula os votos obtidos por cada modelo f^b, resultando uma classificação final de acordo com a votação majoritária.

Figura 10: Algoritmo básico da técnica Random Forest (BREIMAN, 2001).

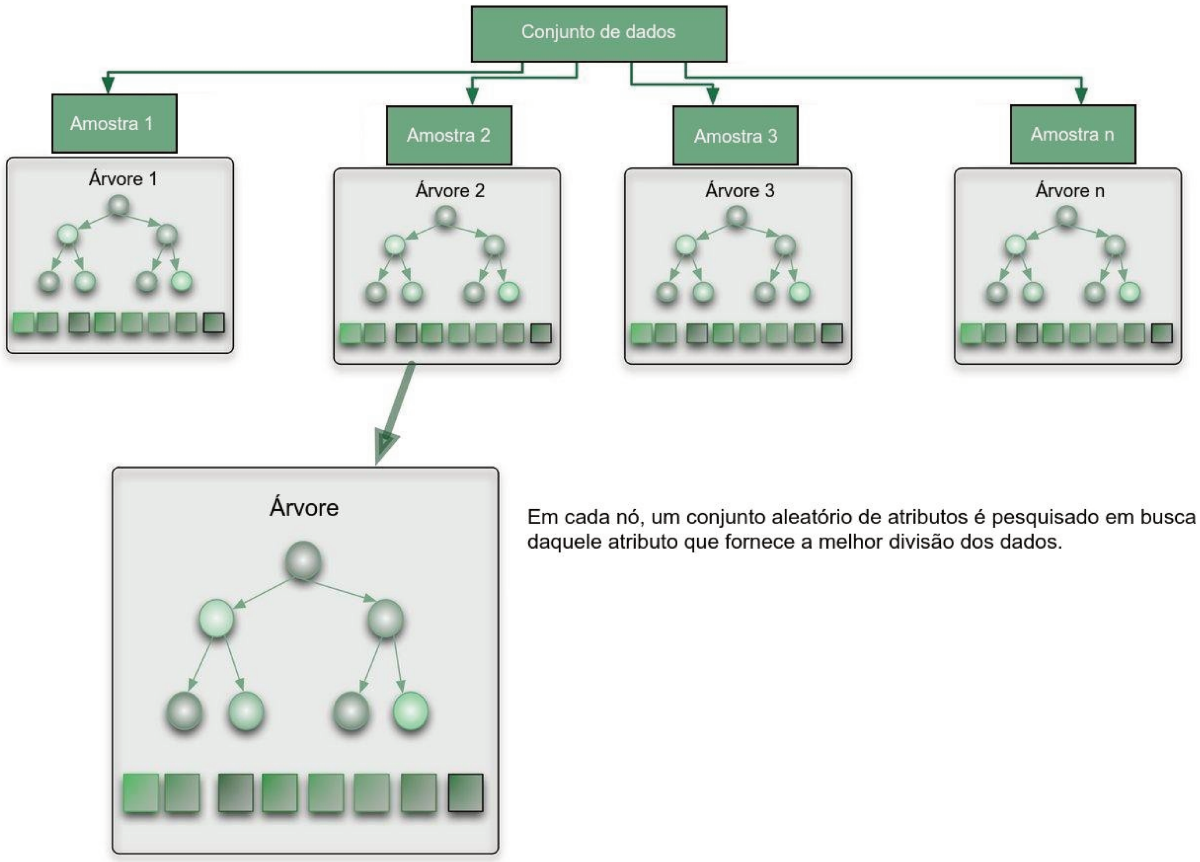

Figura 11: Exemplo de árvores de decisão construídas utilizando amostras bootstrap e seleção aleatória de atributos pela técnica Random Forest.

Fonte: http://citizennet.com/blog/2012/11/10/random-forests-ensembles-and-performance-metrics.

Random Forest possui uma performance excelente na tarefa de classificação, muitas vezes comparável às Máquinas de Vetores de Suporte. Embora seja menos utilizado que outros classificadores mais tradicionais, várias características favorecem o uso da técnica Random Forest em grandes conjuntos de dados, tais como (BREIMAN, 2001):

- Pode ser aplicada em problemas que envolvam mais atributos do que amostras (instâncias).
- Pode ser usada tanto em problemas que possuam apenas duas classes (binários) quanto

naqueles que possuem mais que duas classes (multiclasses).

- Possui um bom desempenho nas predições nas quais os conjuntos de dados possuem muitos ruídos.
- Não produz um classificador superajustado aos dados treinados, ou seja, evita o *overfitting*.
- Fornece medidas de importância de cada atributo.
- Não é necessário muitos ajustes em parâmetros para se atingir um desempenho excelente.

Ainda de acordo com Breiman (2001), dada uma floresta construída de árvores simples e seus respectivos subconjuntos de atributos, Random Forest define uma função que mede a extensão em que o número de votos para uma dada classe excede a votação para qualquer outra classe.

3.5.4. BOOSTING

Boosting nasceu dentro da comunidade de Machine Learning a partir da suposição de que, para qualquer distribuição de dados X, a existência de um classificador fraco implica na existência de um classificador forte (*strong learner*). A ideia era transformar múltiplos classificadores ruins em um único muito bom. Este problema foi resolvido por Schapire (1990), que provou ser possível obter um classificador forte a partir de um fraco.

Essa definição pode ser interpretada da seguinte maneira: um classificador pode ser considerado fraco se a probabilidade deste classificador ser construído, com base numa amostra *D*, tiver erro menor do que 50%, porém, ainda será ligeiramente melhor do que se escolher aleatoriamente uma das classes com probabilidade maior que 50%. Assim, a combinação destes múltiplos classificadores fracos podem resultar num classificador com erro próximo ou igual a zero (FREUND e SCHAPIRE, 1999).

Os métodos desta abordagem funcionam aplicando-se sequencialmente um algoritmo de classificação a versões reponderadas do conjunto de dados de treinamento, dando maior peso aos registros classificados erroneamente no passo anterior.

Assim como Random Forest, a técnica Boosting funciona perturbando a amostra de treinamento. Mas enquanto Random Forest perturba essa amostra aleatoriamente, através de re-amostragem, Boosting gera, em cada passo, uma distribuição dando maior peso às observações classificadas erroneamente no passo anterior (HASTIE *et al.*, 2011).

O algoritmo que mostra a execução básica da técnica Boosting é descrito na Figura 12.

Dado um conjunto de dados de treinamento $X = x_{1,} x_{2,} \ldots, x_j$ e $Y = y_{1,} y_{2,\ldots} y_k$.

Define $\hat{f}(x) = 0$ e $residuos_i = y_i$ para todos os registros do treinamento.

Para $b = 1, 2, 3, \ldots B$, repita:

(a) Ajusta um modelo f^b para o conjunto de treinamento $(X, residuos)$.

(b) Atualiza $\hat{f}$ com o novo modelo:

$$\hat{f}(x) = \hat{f}(x) + f^b(x).$$

(c) Atualiza os resíduos (erros na classificação):

$$residuos_i = residuos_i - f^b(x_i).$$

Fim de repetição.

Gera o modelo final: $\hat{f}(x) = \sum_{b=1}^{B} f^b(x)$, que calcula os votos obtidos por cada modelo f^b, resultando uma classificação final de acordo com a votação majoritária.

Figura 12: Algoritmo básico da técnica Boosting (JAMES *et al.*, 2013).

Considere o exemplo seguinte, ilustrado na Figura 13, onde há uma distribuição de duas classes representadas pelos sinais de positivo (+) e negativo (-). Na rodada 1, os pontos classificados erroneamente (sinais positivos circulados) têm seus pesos aumentados para a próxima etapa. Na rodada 2, devido aos pesos dos erros, a nova classificação desloca a reta vertical, gerando novos erros (sinais negativos circulados). Na terceira rodada, uma reta horizontal foi traçada, dessa vez com menos erros (sinal negativo circulado). No final, a combinação dos classificadores desenvolvidos resulta num classificador global muito superior, sendo que cada classificador tem um peso distinto na votação final do comitê.

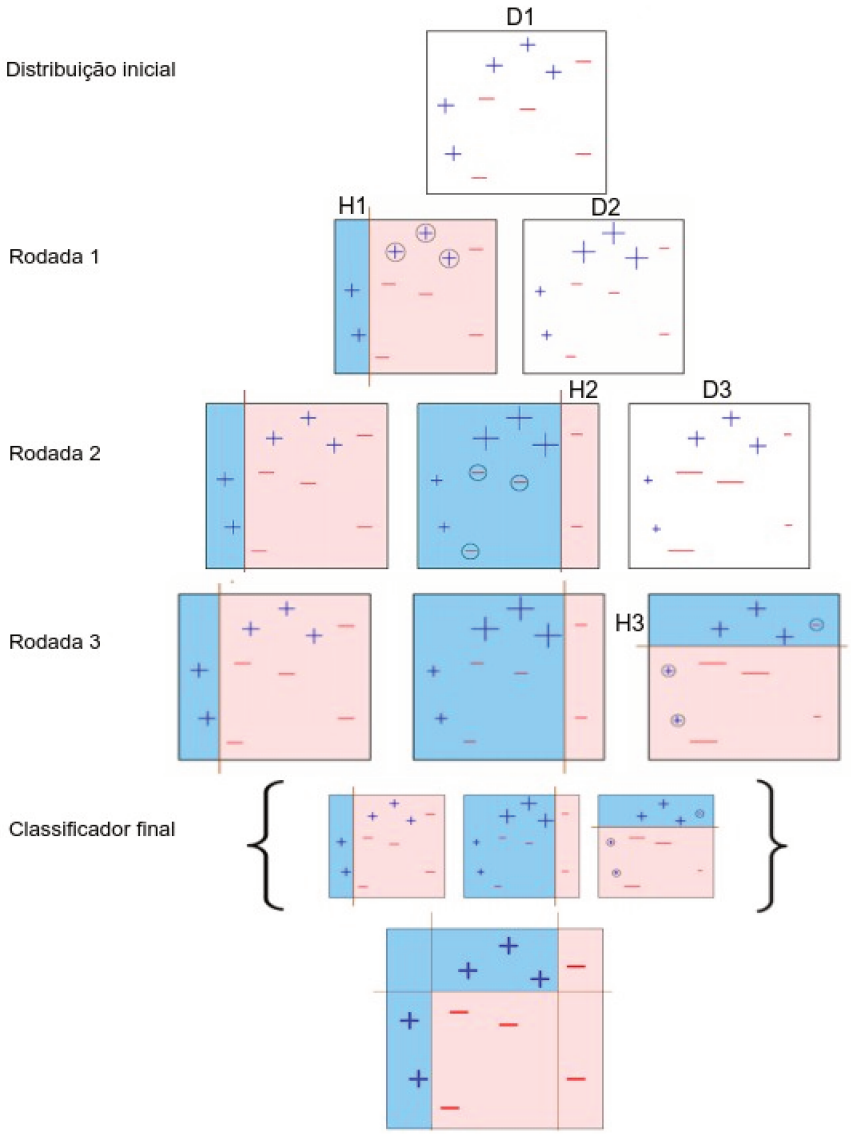

Figura 13: Dinâmica de funcionamento da técnica Boosting.

Fonte: Adaptado de FREUND e SCHAPIRE, 2014.

Um dos limites teóricos do algoritmo Boosting implica que o erro na amostra de treinamento decai exponencialmente com o número de iterações do algoritmo. Empiricamente, observa-se que, após algumas iterações, o erro na amostra de treinamento é muito menor que na amostra de teste, confirmando o resultado teórico, conforme mostra a Figura 14 (JAMES *et al.*, 2013).

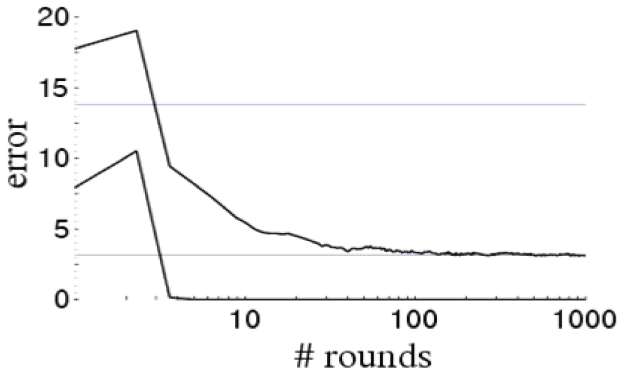

Figura 14: Curvas de erros de treinamento (abaixo) e teste (acima) indicando que quanto maior o número de iterações, menor o erro nos dados de treinamento.

Fonte: FREUND e SCHAPIRE, 1999.

Da mesma forma que Random Forest, Boosting produz uma lista das variáveis mais importantes no desenvolvimento do conjunto de classificadores, que são obtidas pela importância acumulada da variável nas divisões dos nós de cada árvore construída (JAMES *et al.*, 2013).

3.6. MODELAGEM E SELEÇÃO DE MARCADORES DE DNA

Nos últimos anos, a evolução do conhecimento sobre o conteúdo da informação genética, assim como as tecnologias disponíveis para o sequenciamento de genomas em larga escala, ocorreu de forma sem precedentes. Como consequência, um volume muito grande de informações sobre genomas de diversos organismos foi se acumulando nos bancos de dados públicos. Entre essas informações, estão os marcadores moleculares SNP, que são gerados aos milhares para cada indivíduo pelos sequenciadores atuais.

Diversos estudos já foram conduzidos na construção de modelos computacionais e estatísticos para identificação de conjuntos de SNP que possam estar relacionados com características fenotípicas interessantes em variados organismos. Mokry *et al.* (2013), por exemplo, construíram um modelo utilizando o algoritmo Random Forest para a identificação de

marcadores SNP ligados à espessura de gordura de gados Canchim, que é uma estratégia importante para o melhoramento genético de qualidade da carne. Neste processo, primeiramente selecionaram 1% dos SNP mais relevantes de cada cromossomo. Após, foi selecionado 1% dos SNP mais importantes do subconjunto anterior. Por fim, foi ajustada uma análise de regressão seguindo um processo "stepwise" de seleção dos SNP de interesse (SNP), obtendo-se, assim, um conjunto de 21 SNP relacionados com a espessura de gordura. O conjunto de SNP identificados pela abordagem Random Forest mostrou forte correlação com a espessura de gordura dos animais analisados.

Wu *et al.* (2012) apresentaram um método mais ágil de busca dos conjuntos de *m* atributos utilizados na construção das árvores de decisão de um modelo Random Forest. Para avaliação do método, utilizou dados de marcadores SNP de indivíduos (caso e controle, ou seja, indivíduos com e sem a característica avaliada, respectivamente) com Alzheimer e Parkinson. O modelo Random Forest gerado, além de ter tido um desempenho computacional melhor que os modelos utilizando os métodos de busca tradicionais de subconjuntos de atributos, obteve uma alta taxa de acertos. Além disso, forneceu um conjunto dos 25 marcadores SNP mais informativos que estão ligados a genes interessantes nos estudos de casos com Parkinson, principalmente por estarem associados a problemas de desordem neurológica.

Lewis *et al.* (2011) descreveram um modelo utilizando o método de Análise de Componentes Principais (PCA) como um dos algoritmos para a identificação dos marcadores SNP mais relevantes na rastreabilidade de um bovino. Seus experimentos demonstraram que cerca de 250 a 500 SNP, de um total de 30 mil SNP de 19 raças, foram suficientes para atingir quase 100% de acurácia na predição racial de um indivíduo.

González-Recio *et al.* (2010) fizeram a comparação de modelos bayesianos com o algoritmo L2-Boosting em dois conjuntos de marcadores SNP, sendo que um conjunto continha informações sobre o tempo de vida produtiva de 4.702 touros da raça Holandesa e outro continha as médias da taxa de conversão alimentar de 394 frangos. Comparando com os classificadores Bayesian LASSO e BayesA, o L2-Boosting foi uma alternativa competitiva para aplicações de seleção genômica, proporcionando alta precisão nas predições das classes de novos animais com um tempo computacional relativamente curto.

Ayers e Cordell (2010) utilizaram diversos algoritmos de regressão penalizada, tais como LASSO, Regressão de Ridge e Elastic Net, para modelagem dos SNP mais relevantes presentes nos genes de diversas doenças. Os resultados mostraram que os algoritmos de penalização forneceram um modelo esparso com apenas os marcadores SNP mais relevantes para identificação de cada doença analisada, não permitindo a inclusão de variáveis redundantes no modelo e demonstrando maior poder na detecção de SNP importantes do que as abordagens convencionais, gerando menos falsos positivos. Wu *et al.* (2009) desenvolveram um modelo LASSO para seleção de SNP relevantes que estivessem associados à doença Celíaca (doença que afeta o intestino delgado) de humanos, sendo que os resultados demonstraram ser possível identificar conjuntos menores de SNP com alta interação entre os mesmos.

Yi e Xu (2008) utilizaram diversos métodos bayesianos, entre eles o LASSO bayesiano, para associação entre marcadores SNP e QTL (*Quantitative Trait Locus*, que são regiões do DNA ligadas a determinados genes que expressam uma característica fenotípica quantitativa, como por exemplo, a altura de um indivíduo) em uma população de retrocruzamento de cevada. Os resultados mostraram que a abordagem bayesiana obteve modelos cujas inferências foram mais precisas do que a maioria das análises não-bayesianas, apesar de ser computacionalmente mais cara.

Existem pesquisas bem-sucedidas que desenvolveram modelos estatísticos para a seleção de SNP relevantes em DNA de outros animais domésticos. Suekawa *et al.* (2010) desenvolveram um modelo para discriminar as carnes provenientes de gados de raças locais e de raças importadas dos Estados Unidos por meio da análise de frequência alélica, encontrando cinco marcadores SNP capazes de distinguir os gados japoneses dos gados americanos. Sasazaki *et al.* (2011) construíram um modelo que identificou 11 marcadores SNP importantes para gados provenientes dos Estados Unidos, e que os diferenciavam dos gados japoneses, sendo que 4 desses marcadores estavam presentes no trabalho citado anteriormente. Ambas as pesquisas utilizaram uma matriz de 50K de marcadores SNP e foram realizadas em decorrência da preocupação dos consumidores japoneses com a segurança alimentar, principalmente após o surto de BSE (encefalopatia espongiforme bovina, vulgarmente conhecida como doença da vaca louca) que atingiu alguns países, inclusive os Estados Unidos, onde a doença foi identificada em

dezembro de 2003, de acordo com os próprios autores. No trabalho de Pant *et al.* (2012) os resultados revelaram que um conjunto de três a cinco marcadores SNP pode distinguir duas raças de bovinos americanos, Holstein e Jersey, por meio de identificação de marcadores com alelos específicos para uma ou outra raça. Porém, conforme se aumenta o número de raças, o número de marcadores para distingui-las também deverá aumentar.

Heaton *et al.* (2005), utilizando como base a distribuição das frequências alélicas e genotípicas de 165 vacas e 3 touros de diversas raças, observaram 20 marcadores SNP altamente informativos capazes de identificar, com alta acurácia, um animal dentro de um rebanho, auxiliando na rastreabilidade dos produtos de origem bovina. Já em Rohrer *et al.* (2007), em que foram utilizadas amostras de 155 suínos, foi mostrado que um total de 60 marcadores SNP foi suficiente para comprovar a identificação de qualquer animal dentro de um espaço amostral suficientemente grande.

Existem ainda outros trabalhos que apresentaram modelos para seleção de genes em dados de DNA. Por exemplo, Diaz-Uriarte e Alvarez (2006) propuseram uma abordagem que utilizou Random Forest como um dos algoritmos para um modelo que identificasse os genes mais relevantes na classificação de amostras de DNA em estudo de expressão gênica de câncer em humanos. Os resultados mostraram que o modelo conseguiu selecionar um conjunto menor de genes com alta acurácia, demonstrando ser de grande utilidade para biomédicos, bioinformatas e pesquisadores da área genômica.

4. MATERIAL E MÉTODOS

As atividades do projeto de pesquisa foram executadas nos laboratórios de Inteligência Computacional e Bioinformática Aplicada da Embrapa Informática Agropecuária.

A metodologia utilizada é composta de quatro etapas principais, a saber: entendimento do negócio, entendimento dos dados, preparação dos dados e modelagem, conforme ilustrado na Figura 15. Cada etapa da metodologia é descrita com mais detalhes nas seções subsequentes. Essa metodologia é baseada na modelo CRISP-DM (*Cross Industry Standard Process for Data*

Mining) (Chapman *et al*., 2000).

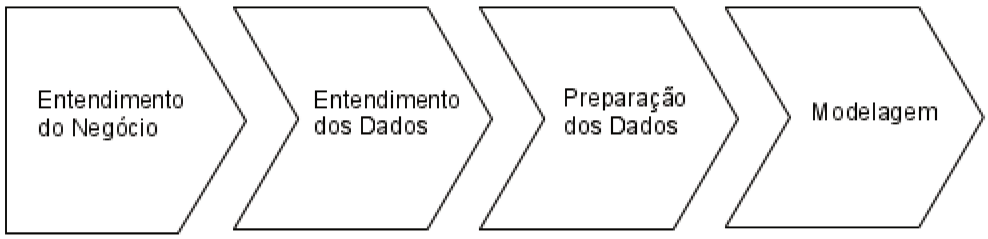

Figura 15: Fluxograma da metodologia utilizada para o trabalho.

4.1. ENTENDIMENTO DO NEGÓCIO

Na fase de entendimento do negócio, buscou-se realizar uma compreensão do domínio, procurando conhecer os objetivos e requisitos do projeto sob uma perspectiva de negócio. Em seguida, este conhecimento foi transformado em um problema de mineração de dados e num plano inicial para atingir os objetivos. Considerando que o objetivo principal desse trabalho é o desenvolvimento de modelos baseados em técnicas de mineração de dados para selecionar os principais marcadores SNP para as raças Crioula, Morada Nova e Santa Inês, foi realizada uma pesquisa contínua em busca de conhecimentos sobre as características destas raças, bem como a situação atual da ovinocultura e suas projeções no cenário nacional e internacional. Também foram estudados alguns conceitos sobre os assuntos concernentes a marcadores moleculares SNP e genética populacional, procurando compreender suas aplicações dentro do campo da genômica animal. Além disso, buscou-se entender alguns conceitos relacionados às técnicas de mineração de dados apropriadas para o problema da pesquisa. Toda essa pesquisa culminou na elaboração do Capítulo 3 (Revisão de Literatura) deste trabalho.

4.2. ENTENDIMENTO DOS DADOS

Nesta etapa do trabalho, realizou-se a aquisição inicial dos dados, buscando identificar problemas de qualidade e a detecção das informações necessárias para atingir os objetivos traçados. O conjunto de dados analisado foi obtido do Consórcio Internacional do Genoma Ovino (ISGC *et al.*, 2010) por meio da Rede Genômica Animal, projeto da Empresa Brasileira de Pesquisa Agropecuária (Embrapa).

O conjunto total dos dados inicial era composto de 3.004 animais domésticos de 71 raças de ovinos provenientes da África, Ásia, América do Sul, Europa, Oriente Médio, Austrália, Estados Unidos e Caribe. A primeira versão da genotipagem destes dados foi finalizada em 2009, e contou com a participação de diversas instituições em todo o mundo, incluindo a Embrapa como representante brasileira. Esta genotipagem usou o genoma bovino como referência para o mapeamento dos genes e marcadores nos ovinos (ISGC *et al.*, 2010). Neste processo, pequenos fragmentos de DNA de ovinos foram utilizados na montagem do genoma ovino por meio do alinhamento com o genoma bovino, conforme mostra a Figura 16.

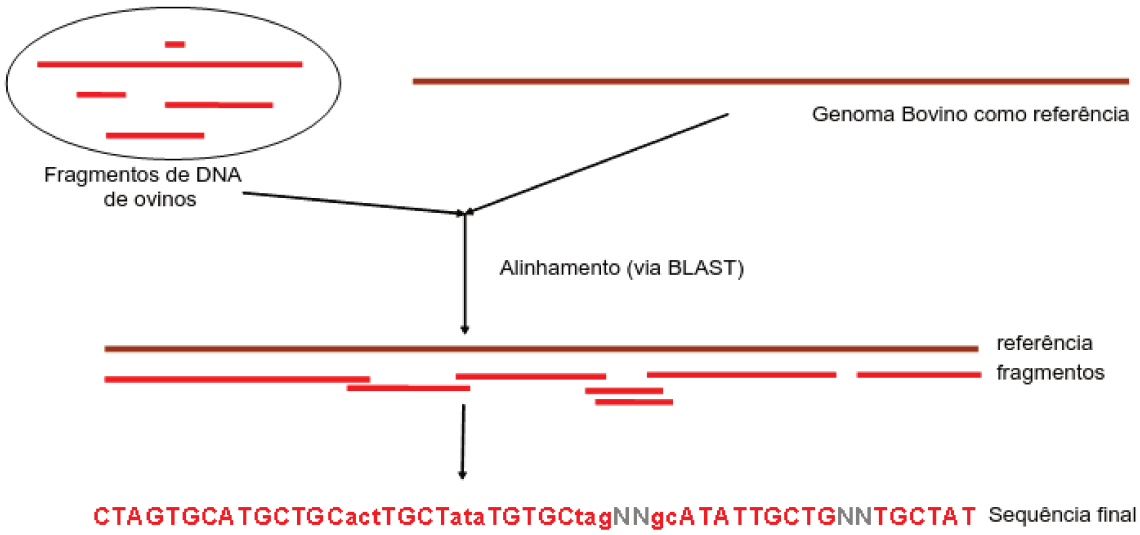

Figura 16: Esquema de montagem do genoma ovino realizada pelo Consórcio Internacional de Ovinos.

Fonte: Adaptado de http://www.sheephapmap.org.

Para este projeto, o conjunto de dados moleculares SNP utilizado continha 49.034 marcadores do conjunto total, no entanto, foram selecionados apenas os 72 animais pertencentes às raças estudadas, sendo 23 animais da raça Crioula, 22 da Morada Nova e 27 da Santa Inês. Observa-se, então, que se trata de uma matriz em que o número de marcadores (p) é muito maior que o número de instâncias (n), isto é, $p>>n$.

Cada um desses marcadores SNP possui um valor de genótipo, que é composto por dois alelos, provenientes do pai e da mãe do animal. Cada alelo pode conter uma Adenina (A) ou uma Timina (T) ou uma Citosina (C) ou uma Guanina (G). O valor de cada marcador pode variar em, no máximo, três genótipos (ex.: AA, GG, AG).

A Figura 17 ilustra o formato do conjunto de dados de ovinos em estudo.

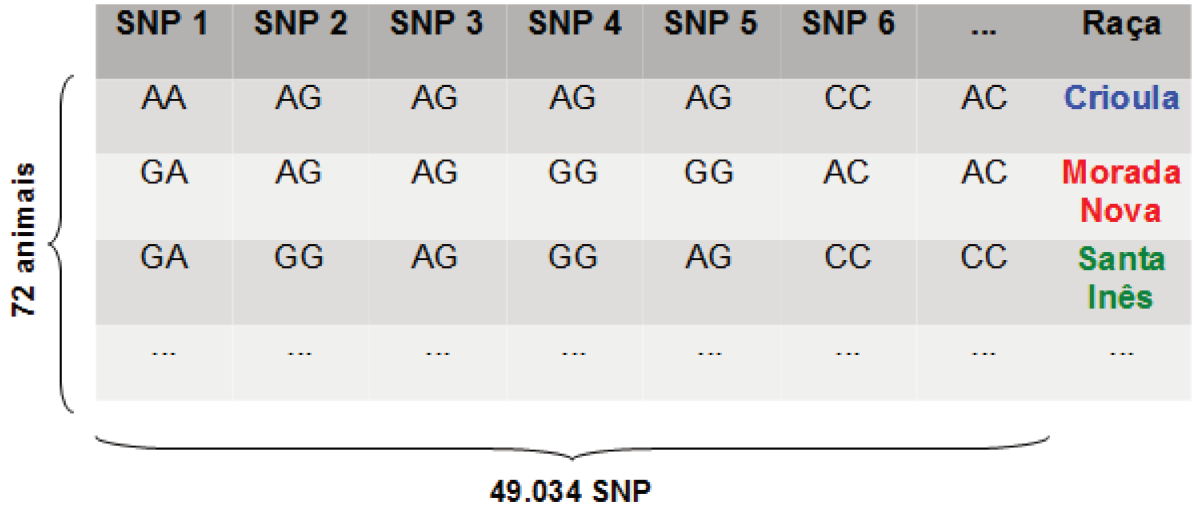

Figura 17: Formato do conjunto de dados de marcadores SNP das três raças em estudo.

4.3. PREPARAÇÃO DOS DADOS

Após a aquisição dos dados, a qualidade destes foi analisada na etapa de preparação dos dados. Como os dados obtidos para este trabalho já haviam sido tratados em relação à remoção de marcadores com valores faltantes e remoção de *outliers*, não houve a necessidade de se utilizar procedimentos para tratamento dos dados concernentes a estes aspectos. Foi realizada uma

verificação, via *script* feito em R, quanto à existência de amostras idênticas dentro do conjunto de dados. Também removeu-se os marcadores presentes no cromossomo X, pois, conforme recomendação de especialistas, este é um dos cromossomos que determinam o sexo do animal, sendo que nos animais machos, aparecerá junto ao cromossomo Y. Como não existem marcadores identificados para o cromossomo Y no conjunto de dados analisado, os valores genotípicos para o cromossomo X, em animais machos, será apenas o alelo vindo da mãe, comprometendo a identificação racial do animal. Verificou-se, por fim, a existência de marcadores SNP que tivessem apenas um valor de genótipo para todas as raças, ou seja, um par de alelos (por exemplo, "AG") idêntico em todos os animais. Também seguindo recomendação de especialistas, um valor único de genótipo em todas as raças faz do marcador um atributo irrelevante para o modelo de identificação de raça.

Após a verificação, constatou-se que não existiam amostras idênticas dentro do conjunto de dados. Entretanto, havia 384 marcadores SNP com valor único para todas as raças, os quais foram removidos do conjunto de dados final por serem totalmente irrelevantes para os objetivos do trabalho. Quanto aos marcadores do cromossomo X, removeu-se 1.502 marcadores pertencentes ao cromossomo. Desta forma, o conjunto final continha 47.148 marcadores para serem utilizados na fase da modelagem.

4.4. MODELAGEM

A partir do conjunto de dados final, técnicas de classificação foram selecionadas e aplicadas ao conjunto de dados moleculares das três raças de ovinos. Para o desenvolvimento de modelos de predição e seleção dos atributos, foram utilizadas as técnicas LASSO, Random Forest e Boosting, cujos algoritmos foram desenvolvidos utilizando o *software* R, uma linguagem de desenvolvimento integrado, para cálculos estatísticos e gráficos, amplamente utilizada por especialistas em estatística e mineração de dados.

Com o objetivo de obter modelos de predição com uma visualização mais amigável, foi utilizado o pacote caret (KUHN, 2013) como interface para as técnicas anteriores. Além de facilitar a visualização dos modelos, caret automatiza a escolha dos melhores valores para alguns

parâmetros das técnicas, como o número de árvores ideal tanto para Random Forest quanto para Boosting, assim como o número de atributos ótimo para desenvolvimento de cada árvore no Random Forest.

Foram realizados vários experimentos, utilizando cada uma das técnicas, procurando obter modelos que fornecessem os melhores resultados em termos de acurácia e menor número de marcadores selecionados. Para tanto, os principais parâmetros de cada uma das técnicas foram ajustados diversas vezes para atingir tal objetivo.

LASSO foi a primeira técnica a ser aplicada, e o único parâmetro testado foi o intervalo de possíveis valores para o coeficiente de penalização *t*, comumente representado pela letra grega λ (lambda). O número padrão deste intervalo é de 100 valores possíveis (JAMES et al, 2013; FRIEDMAN *et al.*, 2010), obtidos separadamente pelo algoritmo LASSO, via validação cruzada, sobre os dados analisados.

Por fim, os parâmetros da Tabela 3 foram configurados para a execução da técnica LASSO:

Tabela 3: Parâmetros utilizados pelo algoritmo LASSO.

Parâmetro	Descrição	Valor
x	Valores dos atributos preditivos	Matriz com genótipos dos SNP
y	Valores do atributo meta	Lista com raças dos ovinos correspondentes à matriz de genótipos
family	Tipo do atributo meta, sendo binomial para duas classes e multinomial para problemas envolvendo mais que duas classes	multinomial
alpha	Algoritmo de penalização a ser utilizado, sendo 1 indicando LASSO, 0 para Ridge Regression e 0.5 para Elastic Net	1
nfolds	Número de subconjuntos a serem utilizados para particionamento dos dados em treinamento e validação	10

Após a aplicação da técnica LASSO, utilizou-se Random Forest para a busca dos marcadores SNP mais relevantes, associados a cada uma das raças. Random Forest constrói diversas árvores de decisão utilizando reamostras *bootstrap* de dados, formando um comitê de classificadores. Para montagem de cada nó destas árvores, o melhor *split* é obtido de uma amostra aleatória de *m* atributos ao invés de todos. Como consequência da construção desta floresta, foi

possível determinar os marcadores mais importantes para o modelo. No entanto, não há uma seleção dos atributos mais importantes para cada uma das classes, a exemplo da técnica LASSO, e sim uma listagem geral ordenando os atributos mais importantes para o modelo (do atributo mais importante ao menos relevante).

Os parâmetros configurados para a utilização do algoritmo Random Forest no modelo final são apresentados na Tabela 4.

Tabela 4: Parâmetros utilizados pelo algoritmo Random Forest.

Parâmetro	Descrição	Valor
x	Valores dos atributos preditivos	Matriz com genótipos dos SNP
y	Valores do atributo meta	Lista com raças dos ovinos correspondentes à matriz de genótipos
ntree	Número de árvores a serem construídas. Neste caso, foi determinado pelo pacote caret.	1.000
mtry	Número de atributos que devem ser selecionados aleatoriamente para determinar o *split* no nó de cada árvore. Neste caso, foi determinado pelo pacote caret.	309
importance	Determina se os atributos devem ser classificados de acordo com sua importância para o modelo.	True

Assim como Random Forest, a técnica Boosting foi utilizada para fornecer um modelo com a listagem dos marcadores SNP mais importantes na identificação das raças. Os classificadores construídos pelo algoritmo Boosting, neste trabalho, são baseados em árvores de decisão, as quais são construídas em distribuições reponderadas dos dados, dando maior peso às observações classificadas erroneamente no passo anterior.

De uma forma geral, uma árvore de decisão é um modelo gráfico representado por nós e ramos, onde os nós intermediários, ou decisórios, representam os testes de atributos (variáveis independentes), enquanto que os ramos representam os resultados desses testes. O nó localizado no topo da árvore representa seu início e é denominado nó-raiz. Já o nó externo, que não possui um nó descendente, localizado na extremidade inferior, é denominado folha ou terminal, e representa o valor de predição do atributo-meta ou classe (HAN *et al.*, 2011).

.

Para a execução do algoritmo Boosting, os parâmetros apresentados na Tabela 5 foram utilizados.

Tabela 5: Parâmetros utilizados pelo algoritmo Boosting.

Parâmetro	Descrição	Valor
x	Valores dos atributos preditivos	Matriz com genótipos dos SNP
y	Valores do atributo meta	Lista com raças dos ovinos correspondentes à matriz de genótipos
n.trees	Número de árvores (ou iterações) a serem ajustadas para o modelo final. Neste caso, foi determinado pelo pacote caret.	1.000
distribution	Tipo de distribuição da variável resposta, podendo ser Bernoulli para duas classes, multinomial para mais de duas classes ou Gaussiana para outros tipos.	multinomial

Após a obtenção dos modelos e dos conjuntos de marcadores mais importantes para identificação das raças, foi realizada uma análise da frequência alélica de cada um desses marcadores dentro das raças. Esta análise foi necessária para saber a frequência de um alelo em uma raça. Quanto maior a frequência de um alelo, maior o número de homozigotos (duas cópias do mesmo alelo num *locus*) dentro da raça. Por outro lado, quanto menor a frequência deste mesmo alelo em pelo menos uma das outras duas raças, maior a possibilidade do marcador discriminar uma raça de outra.

Com a obtenção dos modelos preditivos concebidos para seleção dos marcadores mais informativos e a análise das frequências alélicas, foi selecionado um subconjunto menor de marcadores SNP, com intersecção em dois ou três modelos, com maior potencial para identificar cada uma das três raças pesquisadas.

4.4.1. AVALIAÇÃO DOS MODELOS DE CLASSIFICAÇÃO

Na avaliação de um modelo de classificação procura-se, em suma, testar o desempenho do

modelo. O processo de avaliação, de forma geral, ocorre com a divisão do conjunto de dados inicial em duas partes disjuntas, sendo que uma parte constitui o conjunto de treinamento e outra o conjunto de teste. Neste trabalho, as técnicas utilizaram dois tipos de particionamento dos dados: validação cruzada e *bootstrap*.

A validação cruzada é utilizada, principalmente, quando a quantidade de dados para a divisão em treinamento e teste é limitada (WITTEN *et al.*, 2011). Na validação cruzada, os dados são particionados em k sub-conjuntos mutuamente exclusivos (*folds*) de tamanhos aproximadamente iguais. O indutor é treinado e testado k vezes, sendo que em cada vez é testado com uma das partições e treinado com o restante. Ao final, a taxa de erro é calculada como a média dos k valores das taxas de erros.

O *bootstrap* consiste em gerar os conjuntos de treinamento e teste a partir de uma seleção randômica dos exemplos do conjunto de dados total. Geralmente, esse processo de classificação se repete por várias vezes, gerando múltiplos conjuntos de treinamento e teste. A cada ciclo, as amostragens são selecionadas com reposição, isto é, um mesmo exemplo poderá aparecer mais de uma vez no mesmo sub-conjunto (treinamento ou teste).

O próximo passo após a escolha do método de particionamento do conjunto de dados é a definição das medidas de desempenho que serão utilizadas para avaliar os classificadores. Todas essas medidas são derivadas de uma matriz que ilustra a qual classe cada exemplo pertence e também a qual classe esse exemplo foi classificado por um classificador. Essa matriz é conhecida como Matriz de Confusão ou Matriz de Erros, conforme pode ser vista na Tabela 6, ilustrando o caso de um problema de duas classes.

Tabela 6: Matriz de confusão para duas classes.

	Classe predita		
Classe Real	**Positiva**	**Negativa**	**Total**
Positiva	VP	FN	P
Negativa	FP	VN	N
Total	P'	N'	P + N

Considerando a matriz de confusão da Tabela 6, existem apenas duas classes, sendo que uma é considerada como positiva e a outra como negativa. Na coluna Total, P é o valor total de casos positivos e N é o total de casos negativos existentes no conjunto de treinamento. Já na linha Total, P' é o total de casos que o modelo classificou como casos positivos e N' é o total de casos classificados como negativos. Ainda com base na matriz da Tabela 6, outras medidas podem ser derivadas como segue:

- Verdadeiros Positivos (VP): são os exemplos que pertencem à classe positiva e foram corretamente classificados como pertencentes a essa mesma classe.
- Falsos Negativos (FN): são os exemplos que pertencem à classe positiva e foram incorretamente classificados como pertencentes à classe negativa.
- Verdadeiros Negativos (VN): são os exemplos que pertencem à classe negativa e foram corretamente classificados como pertencentes à classe negativa.
- Falsos Positivos (FP): são os exemplos que pertencem à classe negativa e foram incorretamente classificados como pertencentes à classe positiva.

Neste trabalho, os modelos foram analisados por meio dos valores da acurácia e do coeficiente Kappa.

A acurácia, ou taxa de acerto, fornece a porcentagem de observações que foram classificadas corretamente pelo classificador, sendo definida conforme a Equação 7.

$$Acurácia = \frac{VP + VN}{P + N} \tag{7}$$

A medida Kappa de Cohen (COHEN, 1960) é mais uma maneira utilizada para medir o desempenho do classificador. O Kappa mede o grau de concordância entre as classes preditas e observadas, deduzindo o número esperado de acertos (utilizando uma classificação ao acaso) do número real de acertos do classificador (WITTEN *et al.*, 2011).

O valor máximo da medida é 1, onde este valor significa total concordância. Quando o valor está próximo a 0, a medida indica nenhuma concordância, ou a concordância foi

exatamente a esperada pelo acaso. As divisões arbitrárias entre estes valores representam um nível ruim até 0,4, regular até 0,6, moderado até 0,8 e excelente acima de 0,8 (LANDIS e KOCH, 1977).

Baseando-se na matriz de confusão da Tabela 6, o coeficiente Kappa pode ser definido pela Equação 8:

$$K = \frac{P(a) - P(e)}{1 - P(e)} \tag{8}$$

Onde $P(a)$ é equivalente à proporção de acertos nas classes (Equação 9) e $P(e)$ é a probabilidade de concordância esperada (concordância ao acaso) para as mesmas classes da matriz de confusão (Equação 10).

$$P(a) = \frac{VP + VN}{P + N} \tag{9}$$

$$P(e) = \frac{(P' P) + (N' N)}{(P + N)^2} \tag{10}$$

4.4.2. SOFTWARES UTILIZADOS

Os principais *softwares* utilizados para a produção deste trabalho foram R (versão 3.0.1) e Weka (HALL *et al.*, 2009), versão 3.6.8. A plataforma na qual foi desenvolvido este trabalho foi Ubuntu, uma das distribuições do sistema operacional Linux. Esses *softwares* utilizados são livres e gratuitos, ou seja, não houve a necessidade de licenças especiais para realizar este trabalho.

O pacote estatístico R foi apontado como o segundo *software* mais utilizado para mineração de dados em uma pesquisa, cujos resultados são exibidos na Figura 18. Devido à enorme quantidade de pacotes e recursos disponíveis, foi extensamente utilizado na etapa de desenvolvimento dos modelos preditivos que resultassem na seleção dos marcadores mais

relevantes para cada uma das raças. Para isso, além dos pacotes já instalados por padrão, foram instalados os pacotes para aplicação dos algoritmos referentes às técnicas de modelagem. O pacote instalado para a execução do algoritmo LASSO foi o glmnet (FRIEDMAN *et al.*, 2010), enquanto que para o algoritmo Random Forest foi instalado o pacote randomForest (LIAW e WIENER, 2002). Por fim, para a execução da técnica Boosting foi instalado o algoritmo gbm (RIDGEWAY, 2013). Segundo James *et al.* (2013), estes algoritmos são os mais adequados para cada uma das técnicas escolhidas. Além destes pacotes, instalou-se o pacote caret (KUHN, 2013), como dito anteriormente, com vistas a obter modelos de predição com uma visualização mais amigável e também para a escolha dos melhores valores para alguns parâmetros para cada técnica aplicada.

Por sua vez, o *software* Weka (*Waikato Environment for Knowledge Analysis*) é formado por um conjunto de algoritmos de aprendizado de máquina e diversas ferramentas de análise e visualização de dados, que auxiliam no processo de mineração de dados (WITTEN *et al.*, 2011). Weka foi escolhido pela facilidade que oferece na visualização dos dados por meio de diversos gráficos, auxiliando principalmente na análise das frequências alélicas dos marcadores dentro das raças. Assim como o R, o Weka é um dos principais *softwares* livres utilizados para mineração de dados de acordo com aos resultados da pesquisa disponíveis na Figura 18.

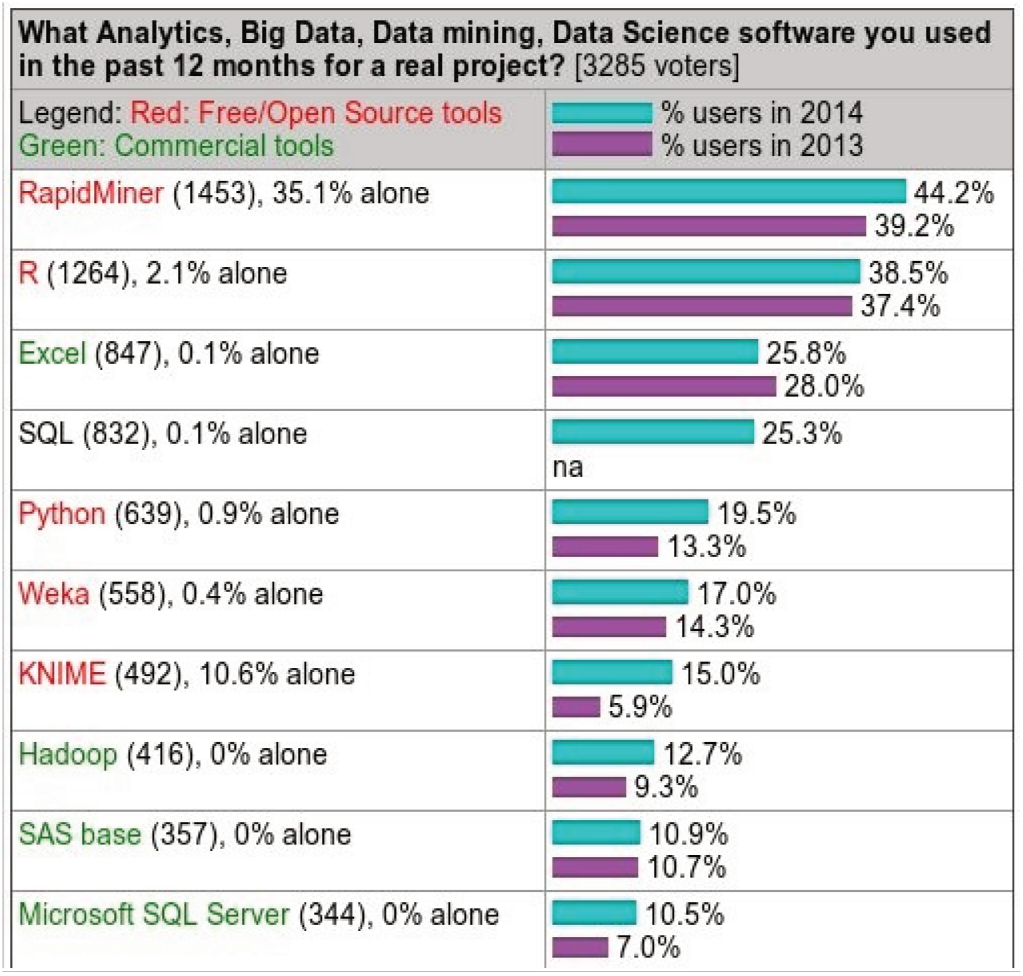

Figura 18: Levantamento publicado em maio de 2014 que apresenta os dez softwares mais utilizados em 2013 e 2014, com 3285 participantes.

Fonte: http://www.kdnuggets.com/polls/2014/analytics-big-data-mining-data-science-software.html

5. RESULTADOS E DISCUSSÃO

Neste capítulo são apresentados os resultados obtidos por meio da aplicação das técnicas LASSO, Random Forest e Boosting em um conjunto de dados compostos por marcadores SNP de ovinos, contendo observações das raças Crioula, Morada Nova e Santa Inês. Além disso, é apresentada uma análise da frequência alélica dos marcadores selecionados dentro dessas raças. Por fim, seleciona-se um subconjunto com os marcadores SNP com maior potencial de identificação das raças.

5.1. MODELO GERADO POR MEIO DA TÉCNICA LASSO

Na aplicação do algoritmo LASSO, o parâmetro de penalização *t*, comumente representado pela letra grega λ (lambda), é determinado separadamente. Em particular, o próprio algoritmo glmnet fornece o valor ótimo para tal parâmetro, utilizando uma análise por validação cruzada de um intervalo de 100 possíveis valores. Para avaliar a relação entre a acurácia e o número de marcadores selecionados, foi alterado o intervalo de 100 para 1.000 valores possíveis. Entretanto, o número de marcadores selecionados e a acurácia permaneceram inalteradas, mantendo-se, então, os 100 valores fornecidos pelo pacote caret (utilizado para automatizar a escolha dos melhores valores para alguns parâmetros do modelo).

Deste intervalo, o valor ótimo para λ obtido foi 0,0035243. Com este valor de λ, o preditor selecionou 26 marcadores SNP relevantes para o modelo, encolhendo a zero os outros marcadores restantes, considerados irrelevantes para o modelo. Dos 26 marcadores selecionados, seis se destacaram para a raça Crioula, 10 para a raça Morada Nova e 10 para a raça Santa Inês.

Um conjunto de seis marcadores foram selecionados com coeficiente diferente de zero para a raça Crioula. Os marcadores selecionados com suas respectivas informações sobre cromossomo, posição no cromossomo, alelos presentes no *locus*, alelo específico (ou predominante) da raça e frequência do alelo específico para a raça Crioula e em outras duas raças estão descritos na Tabela 7. Todas as tabelas relativas ao algoritmo LASSO estão ordenadas por

cromossomo e SNP.

Tabela 7: Frequências alélicas dos marcadores SNP selecionados pelo algoritmo LASSO para a raça Crioula em relação às outras duas raças.

SNP	Cromossomo	Posição	Alelos*	Frequência alélica**		
				Crioula	Morada Nova	Santa Inês
OAR1_268303279_X.1	1	268303279	[G/A]	0.78	0.07	0.09
OAR2_55861669.1	2	55861669	[A/C]	0.84	0	0.05
OAR3_173071993.1	3	173071993	[G/A]	0.90	0.04	0.31
OAR4_25990541.1	4	25990541	[G/A]	0.91	0.04	0.60
OAR16_39888776.1	16	39888776	[A/G]	0.89	0.11	0.15
s17456.1	19	13006214	[A/C]	0.62	0.15	0.16

** Alelo específico para a raça Crioula do lado esquerdo. ** Frequência do alelo específico na população Crioula e nas raças Morada Nova e Santa Inês.*

Observa-se que os valores das frequências alélicas para a raça crioula são bem superiores àqueles das demais raças, sendo um bom sinal de que esses SNP sejam bons discriminantes das raças, como observa-se nos trabalhos de Suekawa *et al.* (2010) e Sasazaki *et al.* (2011).

A Figura 19 ilustra as distribuições das frequências alélicas dos marcadores da Tabela 7 entre as raças.

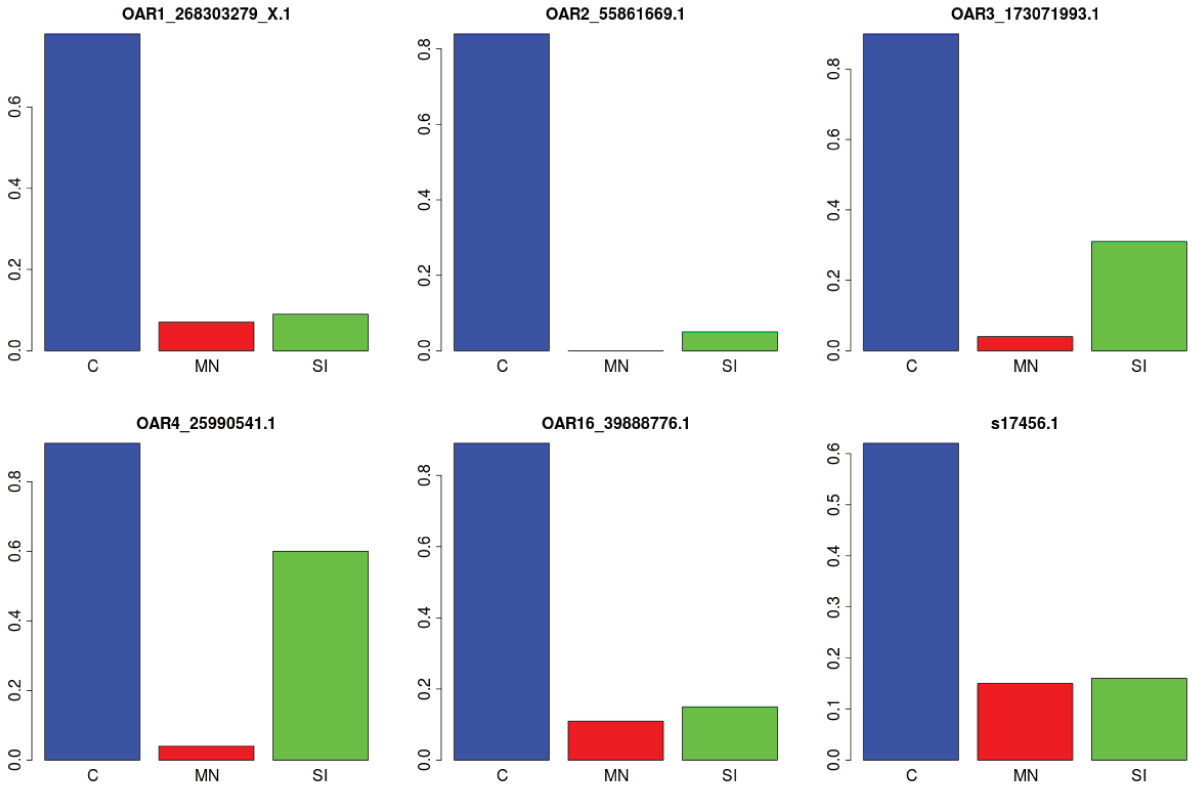

Figura 19: Frequências alélicas dos marcadores relevantes, selecionados pelo algoritmo LASSO, para as três raças em estudo.

Legenda: C – Crioula; MN – Morada Nova; SI – Santa Inês.

De forma geral, todos os marcadores mostraram alto potencial de identificação da raça Crioula. Destacam-se, entre outros, quatro deles (OAR2_55861669.1, OAR3_173071993.1, OAR4_25990541.1 e OAR16_39888776.1) com frequência alélica acima de 80%. O marcador OAR1_268303279_X.1, que pertence ao cromossomo um, possui uma frequência de 78% dentro da raça Crioula, que também pode ser considerada elevada.

Além disso, tão importante quanto a alta frequência do alelo dentro da raça, é sua baixa presença em outras duas raças. O marcador OAR2_55861669.1, por exemplo, tem o alelo "A" ausente na raça Morada Nova e 5% de presença na raça Santa Inês, sendo um bom diferenciador entre as raças. Aliás, todos os marcadores da raça Crioula possuem diferenças de frequências

alélicas muito altas em relação às raças Morada Nova e Santa Inês. Isto se deve, talvez, ao fato da mesma possuir as características físicas da raça mais distintas entre todas as três, por possuir tamanho diminuto e ser lanada (PAIVA, 2005).

Para a raça Morada Nova, o algoritmo LASSO identificou 10 marcadores relevantes, cujas informações estão listadas na Tabela 8.

Tabela 8: Frequências alélicas dos marcadores SNP, selecionados pelo algoritmo LASSO para a raça Morada Nova em relação às outras duas raças.

SNP	Cromossomo	Posição	Alelos*	Frequência alélica**		
				Morada Nova	Crioula	Santa Inês
OAR1_169522497.1	1	169522497	[A/G]	0.73	0.17	0.15
OAR1_187375309_X.1	1	187375310	[A/G]	0.86	0.02	0.31
OAR1_194627962.1	1	194627962	[G/A]	0.73	0	0.02
s06182.1	5	30787155	[A/G]	0.93	0.15	0.31
OAR6_39029427.1	6	39029427	[A/G]	0.84	0.17	0.11
OAR7_106207879.1	7	106207879	[A/G]	0.86	0	0.74
OAR10_33338187.1	10	33338187	[A/G]	0.90	0.22	0.28
OAR17_22334380.1	17	22334380	[G/A]	0.79	0.19	0.13
OAR17_8472049.1	17	8472049	[A/G]	0.95	0.22	0.37
OAR20_45964534.1	20	45964534	[G/A]	0.75	0	0.15

** Alelo específico para a raça Morada Nova do lado esquerdo. ** Frequência do alelo específico na população Morada Nova e nas raças Crioula e Santa Inês.*

Na Figura 20 é possível observar claramente a presença dos alelos específicos da raça Morada Nova nas outras raças.

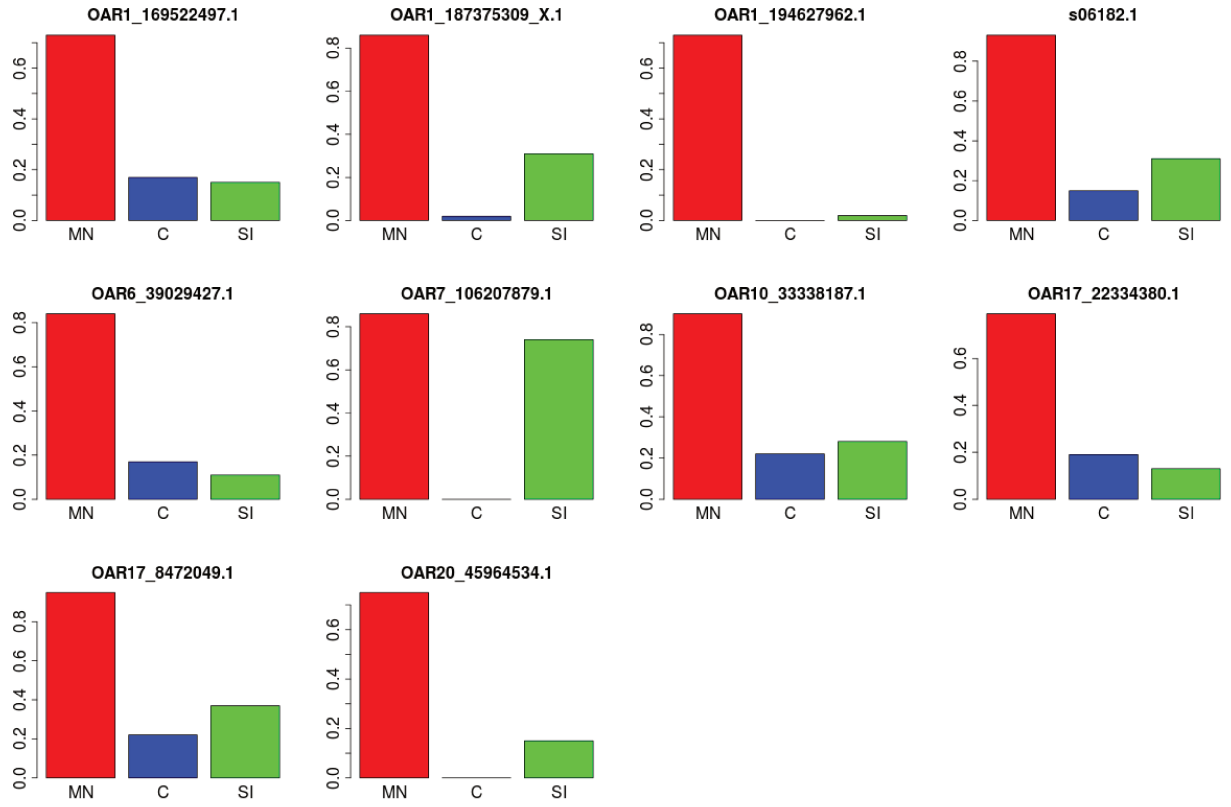

Figura 20: Frequências alélicas dos marcadores relevantes, selecionados pelo algoritmo LASSO para a raça Morada Nova e para as outras duas raças.

Legenda: C – Crioula; MN – Morada Nova; SI – Santa Inês.

Os destaques para a raça Morada Nova são três SNP (OAR1_169522497.1, OAR1_187375309_X.1 e OAR1_194627962.1) no cromossomo um (com posições muito próximas) e dois SNP (OAR17_8472049.1 e OAR17_22334380.1) no cromossomo 17. Um total de seis marcadores possuem frequência acima de 80%. Dois marcadores (OAR1_194627962.1 e OAR20_45964534.1) não possuem seus alelos na raça Crioula, constituindo bons separadores entre estas raças; um deles (OAR1_194627962.1) está quase ausente na raça Santa Inês por possuir baixa frequência.

Foi observado ainda que há uma frequência relativamente maior dos alelos dos animais Morada Nova na raça Santa Inês (um, inclusive, próximo de 75%). Isto talvez seja explicado pelo fato dos animais Santa Inês serem originários do cruzamento entre Morada Nova e outros ovinos

sem raça definida (SRD) do nordeste brasileiro, fazendo com que muitos ovinos Santa Inês preservem características genotípicas do Morada Nova (FIGUEIREDO *et al.*, 1990).

Para a raça Santa Inês foram selecionados 10 marcadores, cujas informações são apresentadas na Tabela 9.

Tabela 9: Frequências alélicas dos marcadores SNP, selecionados pelo algoritmo LASSO para a raça Santa Inês em relação às outras duas raças.

SNP	Cromossomo	Posição	Alelos	Frequência alélica**		
			*	Santa Inês	Crioula	Morada Nova
OAR1_144015756_X.1	1	144015756	[G/A]	0.78	0.37	0.15
OAR2_145195113.1	2	145195113	[A/G]	0.74	0.04	0.38
OAR2_242658985.1	2	242658985	[A/G]	0.85	0.17	0.29
s20468.1	2	56248983	[A/G]	0.76	0.15	0
OAR3_165050963.1	3	165050963	[A/G]	0.80	0.02	0.07
s16949.1	3	164901721	[G/A]	0.89	0.15	0.18
s17180.1	4	102298030	[G/A]	0.74	0.28	0.18
OAR7_21409209.1	7	21409209	[G/A]	0.61	0.02	0.11
s11241.1	7	30741909	[C/A]	0.81	0.35	0.34
s59000.1	18	45393237	[A/G]	0.87	0.30	0.38

** Alelo específico para a raça Santa Inês do lado esquerdo. ** Frequência do alelo específico na população Santa Inês e nas raças Crioula e Morada Nova.*

A Figura 21 ilustra a disposição da frequência alélica da raça Santa Inês e nas raças comparadas.

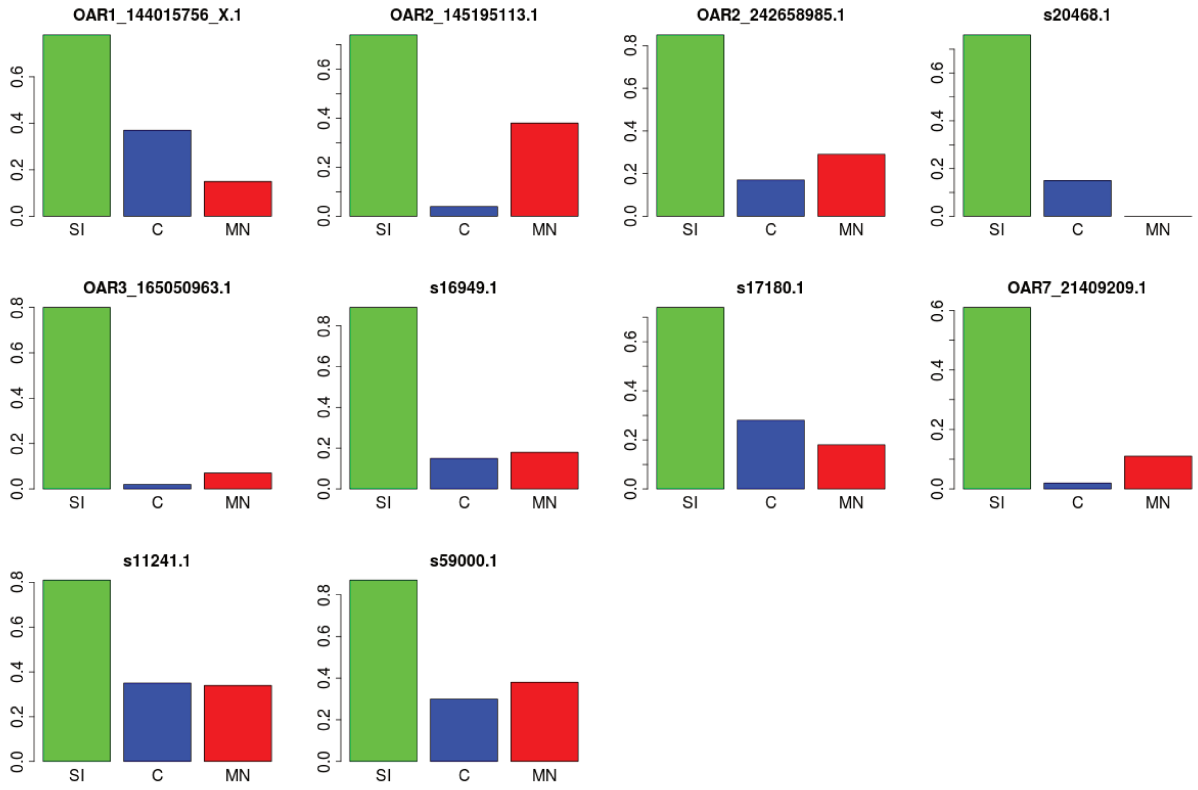

Figura 21: Frequências alélicas dos marcadores relevantes, selecionados pelo algoritmo LASSO para a raça Santa Inês e para as outras duas raças.

Legenda: C – Crioula; MN – Morada Nova; SI – Santa Inês.

Dentre os marcadores selecionados para a raça Santa Inês, três pertencem ao cromossomo dois (OAR2_145195113.1, OAR2_242658985.1 e s20468.1), dois ao cromossomo três (OAR3_165050963.1 e s16949.1) e dois ao cromossomo sete (OAR7_21409209.1 e s11241.1). Uma observação importante vem do fato que os três marcadores do cromossomo três estão em posições muito próximas um do outro, principalmente OAR3_165050963.1 e s16949.1.

De uma maneira geral, os marcadores para a raça Santa Inês têm altas frequências em seus alelos, tendo como destaque o marcador s20468.1, cujo alelo "A" está ausente na raça Morada Nova. Outro marcador, o OAR7_21409209.1, surge com 2% de presença dos alelos na raça Crioula, o que identifica sua baixa frequência.

A acurácia atingida com o conjunto de 26 marcadores SNP selecionados pelo algoritmo LASSO foi de 98% na predição de novas raças, e o índice Kappa foi igual a 0,97. Para os

conjuntos de treinamento e de teste, utilizou-se validação cruzada em 10 subconjuntos dos dados do conjunto treinamento. Segundo Tan *et al.* (2005), o método de validação cruzada com k partições, sendo $k=10$, é considerado um bom estimador do erro de classificação, número também utilizado neste trabalho. O algoritmo LASSO teve ótimo desempenho em termos de acurácia, como demonstrado em Ayers e Cordell (2010), cujos resultados também confirmaram uma boa performance de outras técnicas de regressão penalizada, como Ridge Regression e Elastic-net.

5.2. MODELO GERADO POR MEIO DA TÉCNICA RANDOM FOREST

Com a aplicação do algoritmo Random Forest, obteve-se uma listagem dos atributos (marcadores) mais importantes para o modelo de identificação das raças ovinas. Foram realizados experimentos alterando o número de árvores a serem construídas e o número de atributos selecionados aleatoriamente para determinar o melhor *split* em cada nó. Experimentou-se modelos combinando de 1.000 a 5.000 árvores, e conjuntos aleatórios de atributos variando de 20 a 47.147 atributos para divisão (*split*) dos nós. Após esses experimentos, o melhor resultado obtido foi utilizando os parâmetros fornecidos pelo pacote caret, que resultou em 1.000 árvores e 309 marcadores para *split* dos nós.

Como todos os atributos são ordenados de acordo com sua importância, utilizou-se os 24 melhores classificados, pois a partir desta posição os marcadores restantes pouco contribuíam (menos que 2%) para o modelo. Mokry *et al.* (2013) utilizaram um critério de seleção diferente, pois como o número de marcadores era muito maior que este caso (cerca de 700 mil SNP para cada animal), primeiramente selecionou-se 1% dos SNP mais relevantes de cada cromossomo. Após isso, foi selecionado 1% dos SNP mais importantes do subconjunto anterior, sendo selecionados 70 marcadores SNP pela técnica Random Forest, utilizando tal critério. No modelo final, selecionou-se 21 marcadores, dentre os 70 identificados, por meio de análise de regressão.

A Figura 22 destaca os 24 marcadores e suas importâncias no modelo Random Forest:

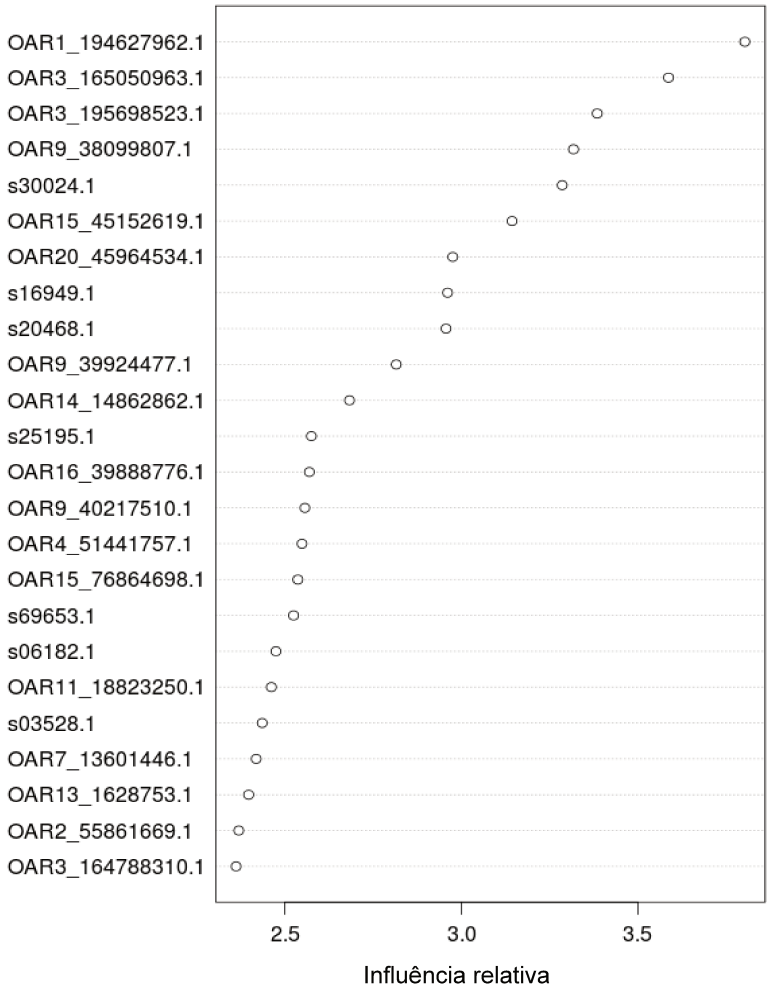

Figura 22: Os 24 marcadores melhores classificados pelo algoritmo Random Forest.

Do conjunto total de 24 marcadores, oito marcadores também foram selecionados pelo algoritmo LASSO (OAR2_55853730.1, OAR16_39888776.1, OAR1_194627962.1, s06182.1, OAR20_45964534.1, s20468.1, OAR3_165050963.1 e s16949.1). Agrupando-se os marcadores fornecidos pelo modelo Random Forest de acordo com a raça, desenvolveu-se três tabelas para análise da frequência do alelo específico de cada uma delas em relação às outras.

A Tabela 10 mostra os marcadores predominantes na raça Crioula e as frequências dos alelos específicos desta raça em relação à Morada Nova e Santa Inês.

Tabela 10: Frequências alélicas dos marcadores SNP, selecionados pelo algoritmo Random Forest para a raça Crioula em relação às outras duas raças.

SNP	Cromossomo	Posição	Alelos*	Frequência alélica**		
				Crioula	Morada Nova	Santa Inês
OAR2_55853730.1	2	55853730	[A/C]	0.85	0	0.07
OAR4_51441757.1	4	51441757	[A/G]	0.91	0.25	0.16
OAR7_13601446.1	7	13601446	[A/G]	0.91	0.11	0.85
OAR11_18815864.1	11	18815864	[A/G]	0.93	0.34	0.22
OAR14_14862862.1	14	14862862	[A/G]	0.83	0.02	0.05
OAR15_45152619.1	15	45152619	[G/A]	0.76	0.02	0.02
OAR15_76864698.1	15	76864698	[G/A]	0.65	0	0.03
OAR16_39888776.1	16	39888776	[A/G]	0.89	0.11	0.14
s25195.1	25	7203123	[G/A]	0.93	0.02	0.30
s30024.1	25	7165805	[C/A]	0.91	0.02	0.28

** Alelo específico para a raça Crioula do lado esquerdo. ** Frequência do alelo específico na população Crioula e nas raças Morada Nova e Santa Inês.*

A partir da Figura 23 é possível notar como estão distribuídos os marcadores selecionados para a raça Crioula e seus alelos nas três raças.

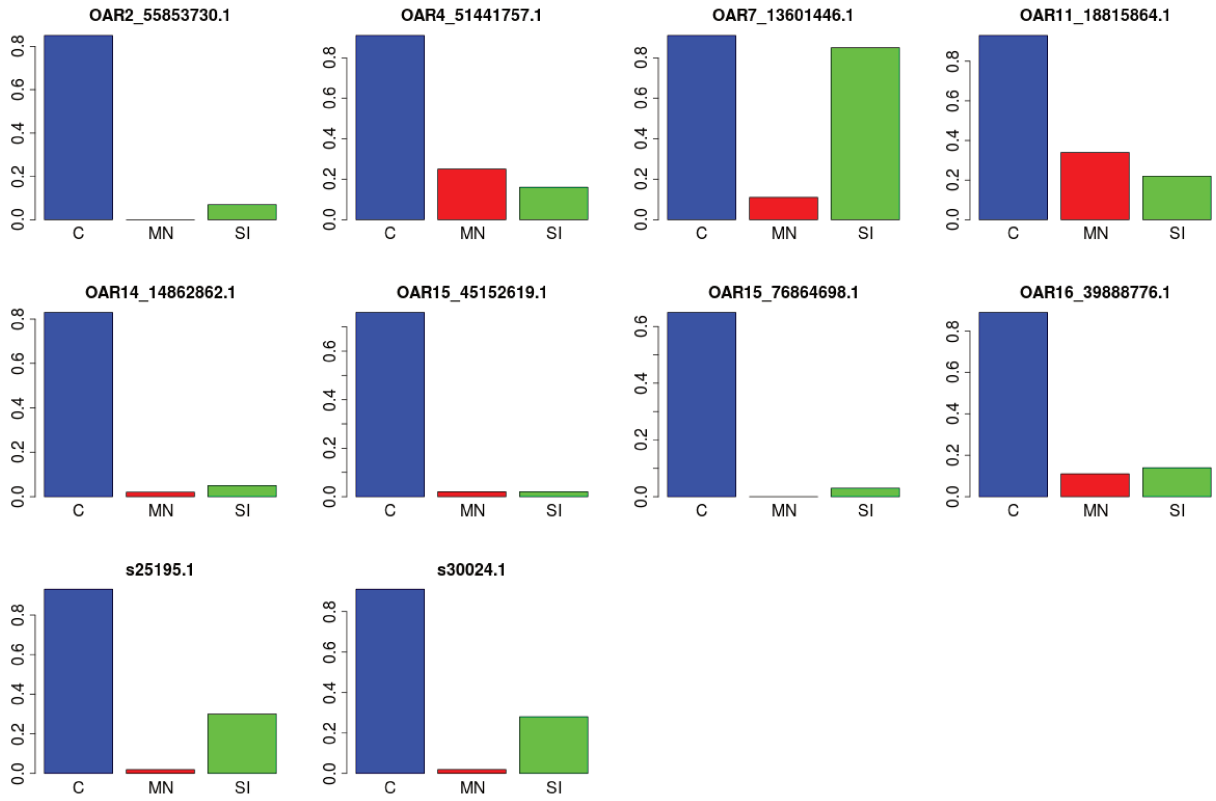

Figura 23: Frequências alélicas dos marcadores relevantes, selecionados pelo algoritmo Random Forest para a raça Crioula e para as outras duas raças.

Legenda: C – Crioula; MN – Morada Nova; SI – Santa Inês.

Do conjunto de 10 marcadores identificados pela técnica Random Forest, como importantes para a raça Crioula, dois também foram identificados pela técnica LASSO (OAR2_55853730.1 e OAR16_39888776.1). Os dois SNP do cromossomo 25 estão em posições próximas e com frequência acima de 90% dentro da raça, surgindo como bons separadores em relação às outras raças. Outro marcador de destaque é o OAR7_13601446.1, que possui uma frequência acima de 90% para a raça Crioula, porém, a frequência para a Santa Inês surge com 85%, o que é muito próximo da frequência de ovinos Crioula.

De maneira geral, todos os SNP fornecidos pelo modelo Random Forest se mostraram importantes na identificação da raça Crioula, com destaque para os quatro marcadores em intersecção com o modelo LASSO. O restante dos marcadores também mostrou uma alta diferença de frequência alélica entre a Crioula e as raças Morada Nova e Santa Inês, com

destaque para os marcadores OAR14_14862862.1, OAR15_45152619.1 e OAR15_76864698.1.

Na Tabela 11, os SNP com predominância na raça Morada Nova são comparados às raças Crioula e Santa Inês, por meio de suas frequências alélicas.

Tabela 11: Frequências alélicas dos marcadores SNP, selecionados pelo algoritmo Random Forest para a raça Morada Nova em relação às outras duas raças.

SNP	Cromossomo	Posição	Alelos*	Frequência alélica**		
				Morada Nova	Crioula	Santa Inês
OAR1_194627962.1	1	194627962	[G/A]	0.73	0	0.02
s06182.1	5	30787155	[A/G]	0.93	0.15	0.31
OAR9_39924477.1	9	39924477	[A/C]	0.95	0.17	0.33
OAR13_1628753.1	13	1628753	[G/A]	0.63	0.34	0.03
OAR20_45964534.1	20	45964534	[G/A]	0.75	0	0.15

** Alelo específico para a raça Morada Nova do lado esquerdo. ** Frequência do alelo específico na população Morada Nova e nas raças Crioula e Santa Inês.*

A Figura 24 mostra como estão distribuídos os marcadores selecionados para a raça Morada Nova e seus alelos nas três raças.

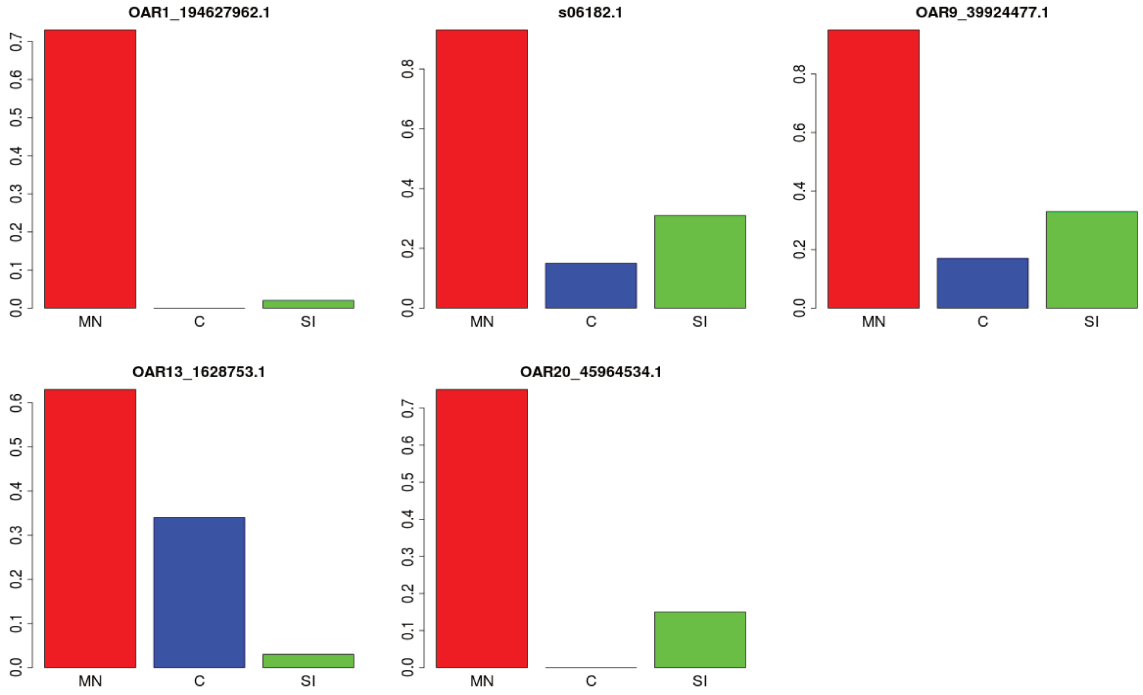

Figura 24: Frequências alélicas dos marcadores relevantes, selecionados pelo algoritmo Random Forest para a raça Morada Nova e para as outras duas raças.

Legenda: C – Crioula; MN – Morada Nova; SI – Santa Inês.

O algoritmo Random Forest indicou cinco marcadores importantes para a raça Morada Nova. Como destaque, existem três marcadores também indicados por LASSO (OAR1_194627962.1, s06182.1 e OAR20_45964534.1). Observa-se que os marcadores OAR1_194627962.1 e OAR20_45964534.1 surgem com frequência acima de 70% na Morada Nova e praticamente ausente nas outras duas raças. O marcador s06182.1 se destaca com uma frequência alélica de 93% na raça Morada Nova, apesar de sua frequência em outras duas raças ter ficado entre de 15% e 40%.

Na Tabela 12 destacam-se os SNP com alta frequência de alelo na raça Santa Inês em comparação às raças Crioula e Morada Nova.

Tabela 12: Frequências alélicas dos marcadores SNP, selecionados pelo algoritmo Random Forest para a raça Santa Inês em relação às outras duas raças.

SNP	Cromossomo	Posição	Alelos*	Frequência alélica**		
				Santa Inês	Crioula	Morada Nova
s03528.1	1	28583773	[A/G]	0.92	0.43	0.23
s20468.1	2	56248983	[A/G]	0.76	0.15	0
OAR3_164788310.1	3	164788310	[G/A]	0.89	0.22	0.18
OAR3_165050963.1	3	165050963	[A/G]	0.80	0.02	0.07
OAR3_195698523.1	3	195698523	[A/G]	0.66	0.15	0.04
s16949.1	3	164901721	[G/A]	0.89	0.15	0.18
s69653.1	3	164951744	[G/A]	0.90	0.08	0.36
OAR9_38099807.1	9	38099807	[G/A]	0.72	0.30	0.02
OAR9_40217510.1	9	40217510	[C/A]	0.54	0.08	0.02

** Alelo específico para a raça Santa Inês do lado esquerdo. ** Frequência do alelo específico na população Santa Inês e nas raças Crioula e Morada Nova.*

Na Figura 25 são exibidos os marcadores selecionados para a raça Santa Inês e as frequências de seus alelos nas três raças.

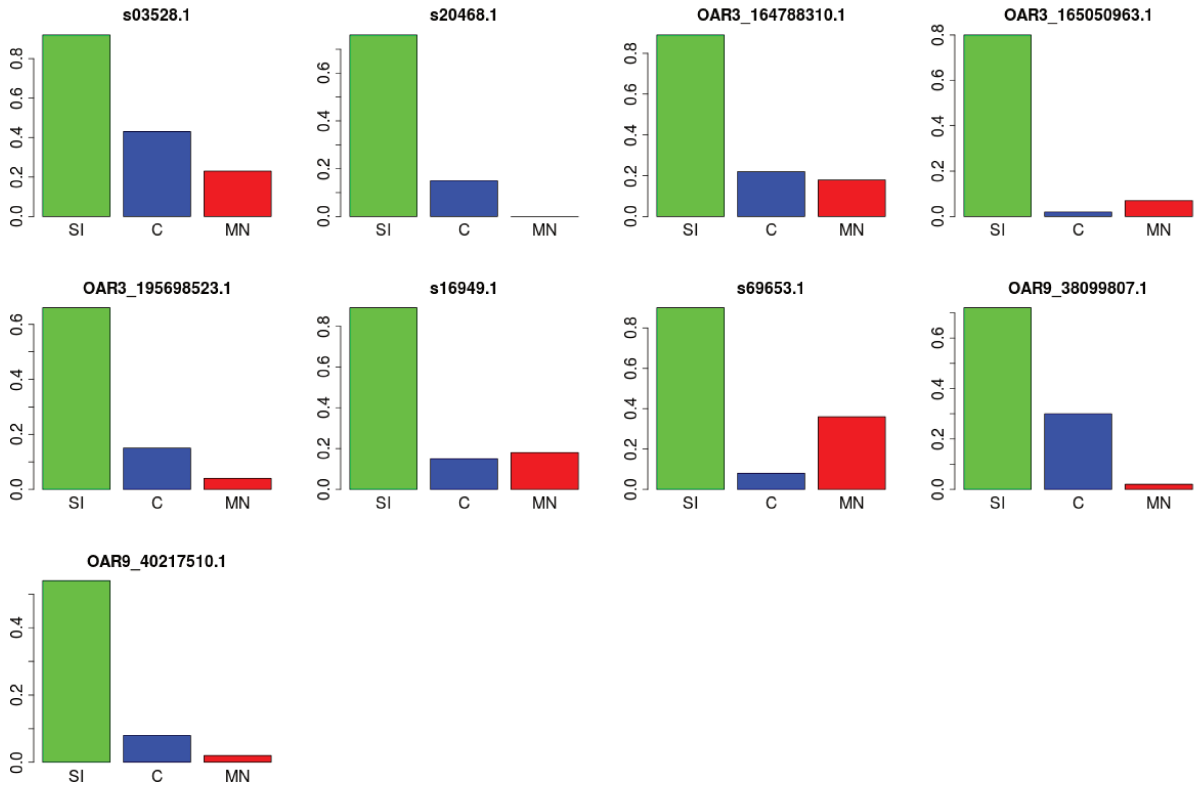

Figura 25: Frequências alélicas dos marcadores relevantes, selecionados pelo algoritmo Random Forest para a raça Santa Inês e para as outras duas raças.

Legenda: C – Crioula; MN – Morada Nova; SI – Santa Inês.

Para a raça Santa Inês, nove marcadores foram selecionados com altas frequências alélicas. Destes, três estavam presentes no modelo fornecido pelo algoritmo LASSO (s20468.1, OAR3_165050963.1 e s16949.1). Um dado interessante é que cinco marcadores são originados do cromossomo três (OAR3_164788310.1, OAR3_165050963.1, OAR3_195698523.1, s16949.1 e s69653.1), sendo que quatro deles estão praticamente um do lado do outro (OAR3_164788310.1, OAR3_165050963.1, s16949.1 e s69653.1) no cromossomo. Foi observado ainda que o marcador OAR3_195698523.1 tem frequência de 66% dentro da Santa Inês, além de frequências baixas em outras duas raças

Para treinamento e teste, utilizou-se amostragens *bootstrap* dos dados. Foram desenvolvidas e combinadas 1.000 árvores utilizando as amostras *bootstrap*, e 309 atributos para o conjunto aleatório *m* na escolha do melhor atributo para *split* dos nós. Os números de 1.000

árvores e 309 atributos para *m* foram determinados automaticamente pelo pacote caret. O comitê de classificadores que formaram a floresta obteve uma acurácia de 100% e Kappa de 1, ou seja, assim como a técnica LASSO, Random Forest também teve um excelente desempenho.

5.3. MODELO GERADO POR MEIO DA TÉCNICA BOOSTING

Assim como Random Forest, o algoritmo Boosting foi utilizado para encontrar os marcadores SNP mais importantes para as raças estudadas. Os classificadores construídos pelo algoritmo Boosting são baseados em múltiplas árvores de decisão desenvolvidas em distribuições ajustadas dos dados. Boosting possui um procedimento interno para seleção de atributos, que exclui do modelo final aquelas variáveis com nenhuma influência na predição das classes (JAMES *et al.*, 2013). Neste caso, foram selecionados 334 marcadores com influência acima de zero.

O único parâmetro testado para o algoritmo Boosting foi o número de classificadores (neste caso, árvores de decisão) a serem construídos. Avaliou-se modelos desenvolvidos com totais de 1.000, 5.000 e 10.000 árvores para a formação do comitê de classificadores, sendo que o melhor resultado, em termos de acurácia e Kappa, ocorreu com 1.000 árvores, número fornecido pelo pacote caret.

O algoritmo Boosting gera uma listagem ordenada dos melhores atributos conforme a influência de cada um deles no modelo. Neste caso, selecionou-se os 14 melhores marcadores, pois os marcadores restantes a partir desta posição pouco contribuíam (menos que 1%) para o modelo. A Figura 26 destaca os 14 marcadores com maior influência, segundo o algoritmo Boosting.

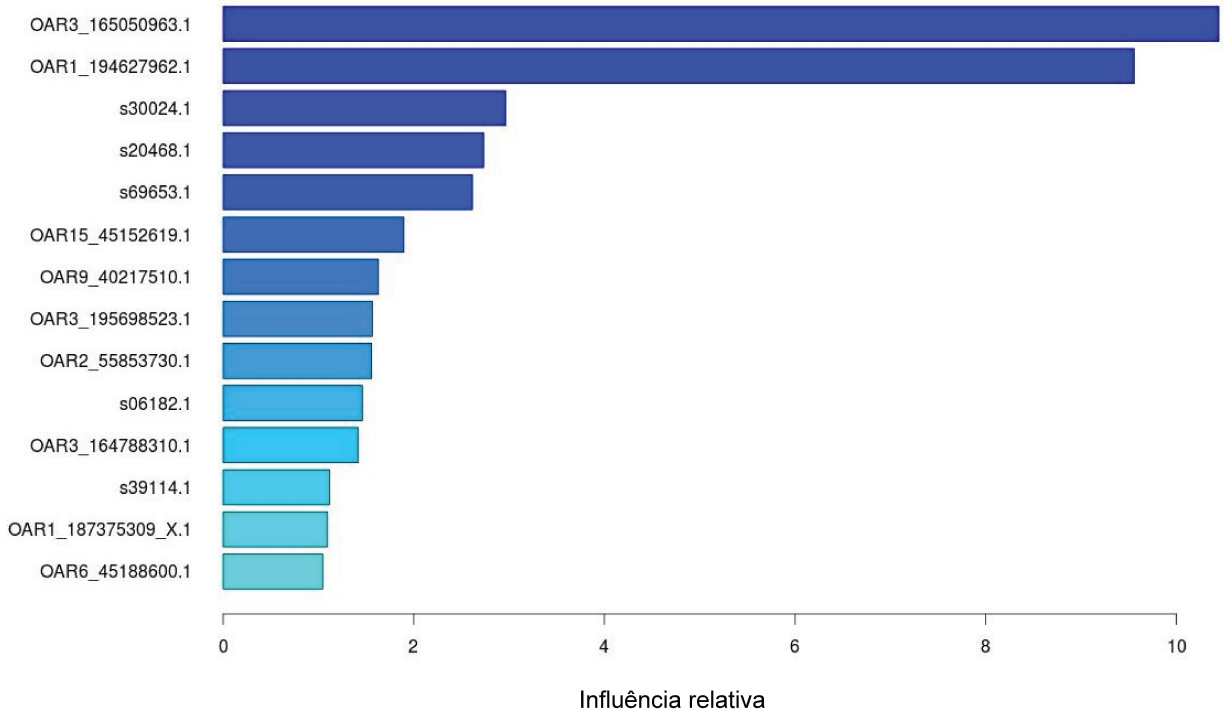

Figura 26: Os 14 marcadores mais bem classificados pelo algoritmo Boosting.

Entre os 14 marcadores ordenados pela técnica Boosting, cinco (OAR2_55853730.1, OAR3_165050963.1, OAR1_194627962.1, s20468.1 e s06182.1) estavam presentes nos modelos LASSO e Random Forest, um (OAR1_187375309_X.1) estava somente no modelo LASSO e seis (OAR15_45152619.1, s30024.1, s69653.1, OAR3_164788310.1, OAR3_195698523.1 e OAR9_40217510.1) somente no modelo Random Forest. Com isto, o algoritmo Boosting selecionou apenas dois marcadores diferentes das técnicas anteriores para seu modelo, a saber: s39114.1 e OAR6_45188600.1.

Realizando-se o agrupamento dos marcadores selecionados pelo algoritmo Boosting, de acordo com a raça do animal, foram construídas três tabelas contendo as frequências dos alelos específicos de cada raça em comparação às outras duas raças. Na Tabela 13 estão descritos os marcadores com predominância na raça Crioula e as frequências dos alelos específicos desta raça em relação à Morada Nova e Santa Inês.

Tabela 13: Frequências alélicas dos marcadores SNP, selecionados pelo algoritmo Boosting para a raça Crioula em relação às outras duas raças.

SNP	Cromossomo	Posição	Alelos*	Frequência alélica**		
				Crioula	Morada Nova	Santa Inês
OAR2_55853730.1	2	55853730	[A/C]	0.85	0	0.07
OAR15_45152619.1	15	45152619	[G/A]	0.76	0.02	0.02
s30024.1	25	7165805	[C/A]	0.91	0.02	0.28

** Alelo específico para a raça Crioula do lado esquerdo. ** Frequência do alelo específico na população Crioula e nas raças Morada Nova e Santa Inês.*

Na Figura 27 é possível observar os marcadores com alelos predominantes na raça Crioula e suas frequências nas outras raças.

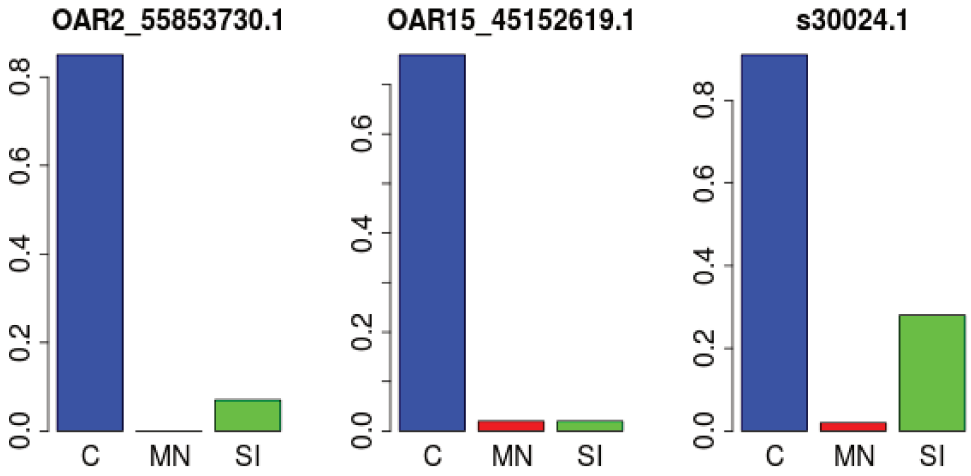

Figura 27: Frequências alélicas dos marcadores relevantes, selecionados pelo algoritmo Boosting para a raça Crioula e para as outras duas raças.

Legenda: C – Crioula; MN – Morada Nova; SI – Santa Inês.

Na lista de marcadores importantes para a raça Crioula, um deles (OAR2_55853730.1) foi indicado nos dois modelos anteriores, e outros dois marcadores (OAR15_45152619.1 e s30024.1) foram selecionados no modelo Random Forest, reforçando o alto potencial destes marcadores para identificação da raça Crioula. Outro destaque é todos possuem frequência acima de 75%

dentro da raça Crioula, e alta diferenciação alélica para as duas raças.

A Tabela 14 traz uma listagem dos marcadores com predominância na raça Morada Nova e suas frequências nas outras raças.

Tabela 14: Frequências alélicas dos marcadores SNP, selecionados pelo algoritmo Boosting para a raça Morada Nova em relação às outras duas raças.

SNP	Cromossomo	Posição	Alelos*	Frequência alélica**		
				Morada Nova	Crioula	Santa Inês
OAR1_187375309_X.1	1	187375310	[A/G]	0.86	0.02	0.31
OAR1_194627962.1	1	194627962	[G/A]	0.73	0	0.02
s06182.1	5	30787155	[A/G]	0.93	0.15	0.31

** Alelo específico para a raça Morada Nova do lado esquerdo. ** Frequência do alelo específico na população Morada Nova e nas raças Crioula e Santa Inês.*

A Figura 28 ilustra as frequências dos marcadores identificados pelo Boosting como relevantes para a raça Morada Nova nas três raças.

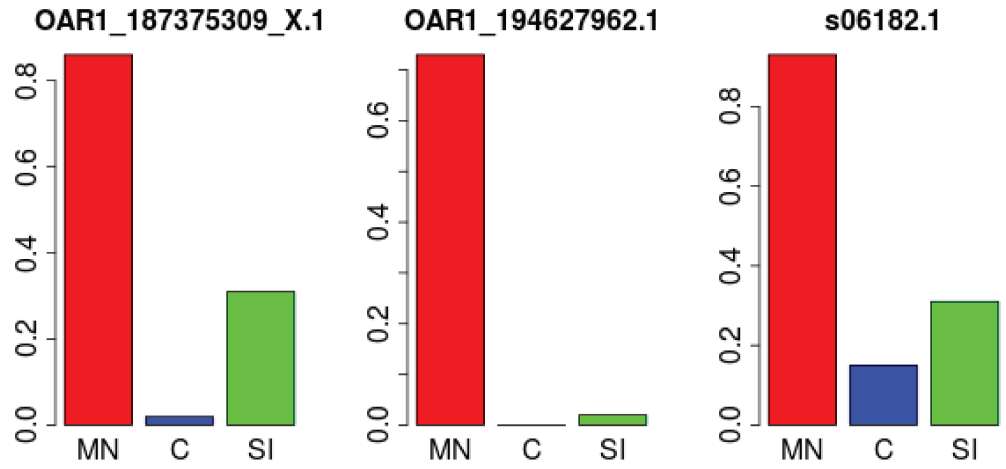

Figura 28: Frequências alélicas dos marcadores relevantes, selecionados pelo algoritmo Boosting para a raça Morada Nova e para as outras duas raças.

Legenda: C – Crioula; MN – Morada Nova; SI – Santa Inês.

O algoritmo Boosting separou três marcadores com maior frequência em Morada Nova, sendo dois deles (OAR1_194627962.1 e s06182.1) presente nos dois modelos anteriores e um (OAR1_187375309_X.1) no modelo LASSO. Nota-se que, assim como Random Forest, o algoritmo Boosting também indicou poucos SNP para a raça Morada Nova. O marcador OAR1_194627962.1, presente nos outros dois modelos, possui frequência acima de 70% em animais Morada Nova, de apenas 2% na Santa Inês e ausente na Crioula, resultado que reforça este marcador como um bom discriminante de raças. O marcador OAR1_187375309_X.1, também indicado pelo modelo LASSO, surge com frequência acima de 80% nos animais Morada Nova, o que também demonstra o bom potencial destes SNP.

A Tabela 15 descreve os marcadores associados à raça Santa Inês e suas frequências nas outras duas raças.

Tabela 15: Frequências alélicas dos marcadores SNP, selecionados pelo algoritmo Boosting para a raça Santa Inês em relação às outras duas raças.

SNP	Cromossomo	Posição	Alelos*	Frequência alélica**		
				Santa Inês	Crioula	Morada Nova
s20468.1	2	56248983	[A/G]	0.76	0.15	0
OAR3_164788310.1	3	164788310	[G/A]	0.89	0.22	0.18
OAR3_165050963.1	3	165050963	[A/G]	0.80	0.02	0.07
OAR3_195698523.1	3	195698523	[A/G]	0.66	0.15	0.04
s39114.1	3	232410568	[A/G]	0.59	0.08	0.07
s69653.1	3	164951744	[G/A]	0.90	0.08	0.36
OAR6_45188600.1	6	45188600	[G/A]	0.78	0.06	0.04
OAR9_40217510.1	9	40217510	[C/A]	0.54	0.08	0.02

** Alelo específico para a raça Santa Inês do lado esquerdo. ** Frequência do alelo específico na população Santa Inês e nas raças Crioula e Morada Nova.*

Na Figura 29 são mostrados os marcadores selecionados para a raça Santa Inês e as frequências de seus alelos nas três raças.

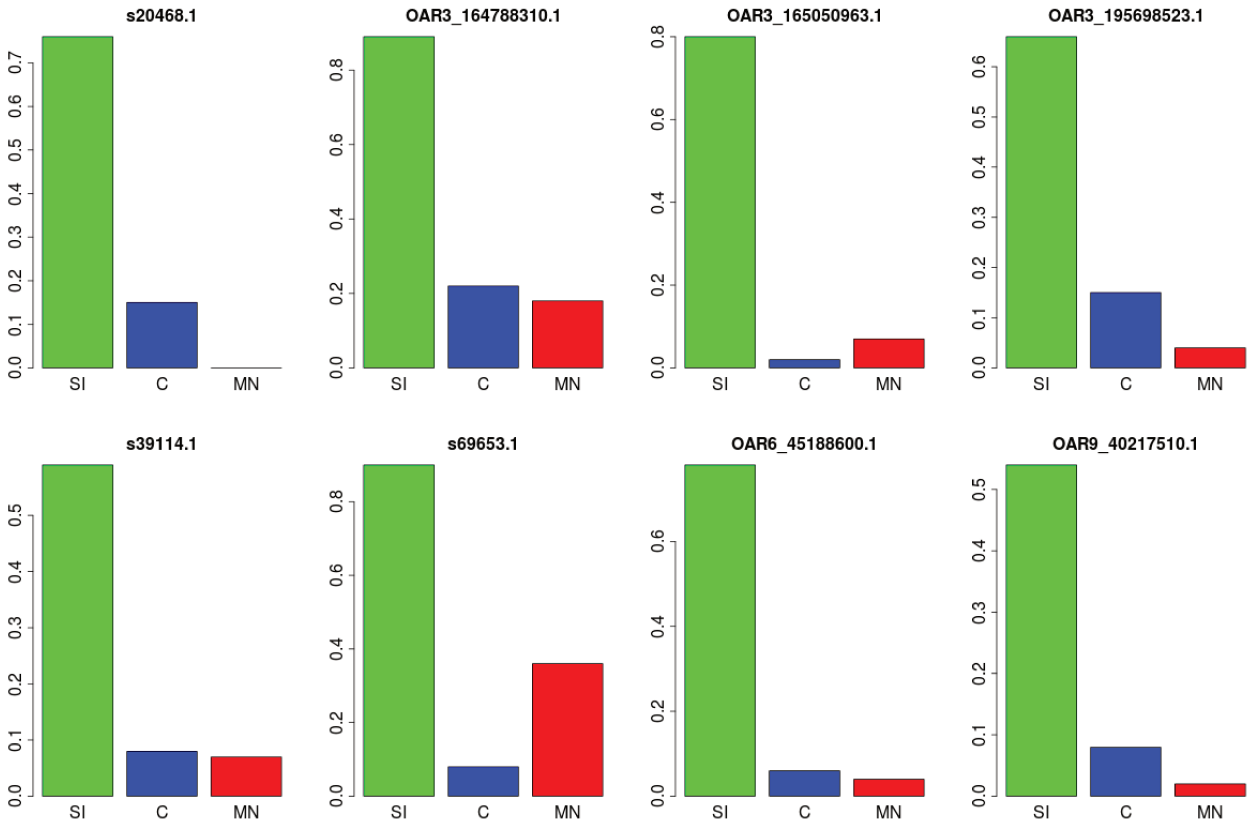

Figura 29: Frequências alélicas dos marcadores relevantes, selecionados pelo algoritmo Boosting para a raça Santa Inês e para as outras duas raças.

Legenda: C – Crioula; MN – Morada Nova; SI – Santa Inês.

Dentre os marcadores fornecidos pelo modelo Boosting com frequências altas para a raça Santa Inês, destacam-se dois deles (s20468.1 e OAR3_165050963.1) também selecionados pelas técnicas LASSO e Random Forest, e quatro (OAR3_164788310.1, OAR3_195698523.1, s69653.1 e OAR9_40217510.1) apenas pela técnica Random Forest. Além disso, dois SNP (s39114.1, OAR6_45188600.1) foram selecionados exclusivamente pela técnica Boosting. Novamente se observa que o cromossomo três abrange um grande número de marcadores relacionados à Santa Inês, e todos muito próximos.

De forma geral, grande parte dos marcadores selecionados para a raça Santa Inês apresenta alta frequência de alelo (alguns acima de 80%). Destaque para os dois marcadores (s20468.1 e OAR3_165050963.1) que também foram indicados pelos dois modelos anteriores,

atestando seu potencial de identificação da raça Santa Inês. Entre os marcadores listados no modelo Boosting e Random Forest estão os SNP OAR3_164788310.1 e s69653.1, com frequências acima de 80% na raça Santa Inês. Os dois SNP identificados somente pelo algoritmo Boosting (s39114.1 e OAR6_45188600.1) se destacam com uma frequência acima de 50% na raça Santa Inês, e com frequências abaixo de 10% tanto na Crioula quanto na Morada Nova, demonstrando que também podem ser úteis na separação das raças.

Para realização de treinamento e teste, o algoritmo Boosting foi executado por meio de validação cruzada em 10 subconjuntos de dados para suas amostras. Desta forma, para cada subconjunto, foram gerados 1.000 classificadores (do tipo árvore de decisão) utilizando amostras ajustadas dos dados de treinamento. O modelo final foi obtido por meio da média dos 10 subconjuntos. A acurácia e o Kappa obtidos pelo modelo, com a combinação dos classificadores ajustados, foi de 100% e 1, respectivamente. Observando esses resultados, pode-se pensar em indícios de *overfitting*, porém os parâmetros ajustados para a execução do algoritmo Boosting foram obtidos pelo caret de forma a evitar um super-ajuste do modelo. Pode ocorrer *overfitting* em Boosting, mesmo que de forma gradual, quando o número de árvores é muito grande para o conjunto de dados analisado (JAMES *et al.*, 2013). Por isso, o uso de uma ferramenta (como o caret) para automatizar a escolha do número de árvores se torna importante nestas situações. Este ótimo desempenho também foi obtido no trabalho de González-Recio *et al.* (2010), que utilizou o algoritmo L2-Boosting em dois conjuntos de marcadores SNP (de touros e frangos), obtendo alta precisão nas predições das classes de novos animais com um tempo computacional relativamente curto.

5.4. MARCADORES COM MAIOR POTENCIAL DE IDENTIFICAÇÃO DAS RAÇAS

Com o desenvolvimento dos três modelos preditivos e a seleção dos principais marcadores para identificação das raças, foi realizada uma análise para verificação daqueles SNP que convergiam em dois ou três modelos. Marcadores que estão nestas intersecções, provavelmente, têm maior potencial de identificação das raças.

Para tanto, foram construídos três diagramas de Venn, um para cada raça, com o objetivo

de ilustrar os marcadores selecionados pelos três modelos e mostrar onde houve intersecção entre os mesmos. O primeiro diagrama de Venn se refere à raça Crioula, e pode ser visualizado na Figura 30.

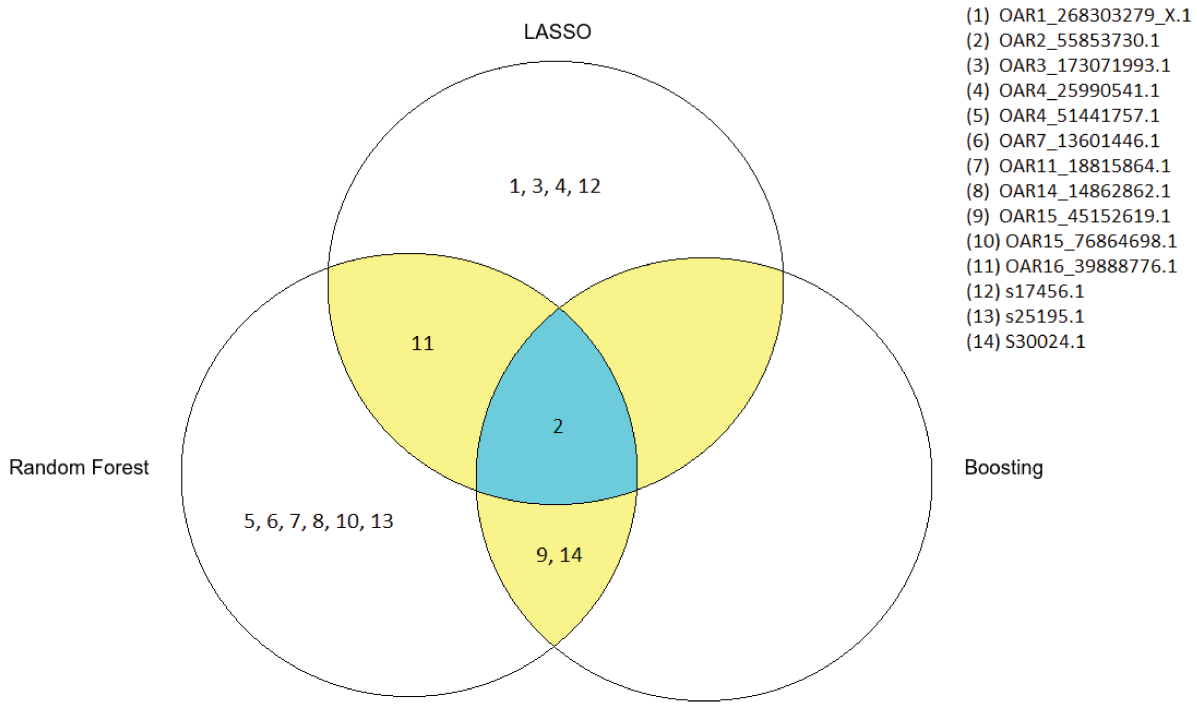

Figura 30: Diagrama de Venn para os marcadores selecionados para a raça Crioula pelos três modelos.

O diagrama de Venn para a raça Crioula mostra que o marcador OAR2_55853730.1 foi selecionado nos três modelos, mostrando uma frequência de 85% para a raça Crioula e frequências quase nulas para outras duas raças, ou seja, é um SNP com alto potencial de identificação da raça Crioula. Além disso, outros marcadores se interseccionaram em dois modelos, sendo que o marcador OAR16_39888776.1 surgiu nos resultados do modelos LASSO e Random Forest, e os marcadores OAR15_45152619.1 e s30024.1 nos modelos Boosting e Random Forest. A identificação dos mesmos marcadores por dois modelos também podem indicar que esses SNP possuem alta relevância na identificação da raça. O marcador OAR15_45152619.1, por exemplo, surge com frequência acima de 70% para a raça Crioula e

frequências abaixo de 5% para Morada Nova e Santa Inês, corroborando sua importância no modelo.

O diagrama de Venn referente à raça Morada Nova é apresentado na Figura 31.

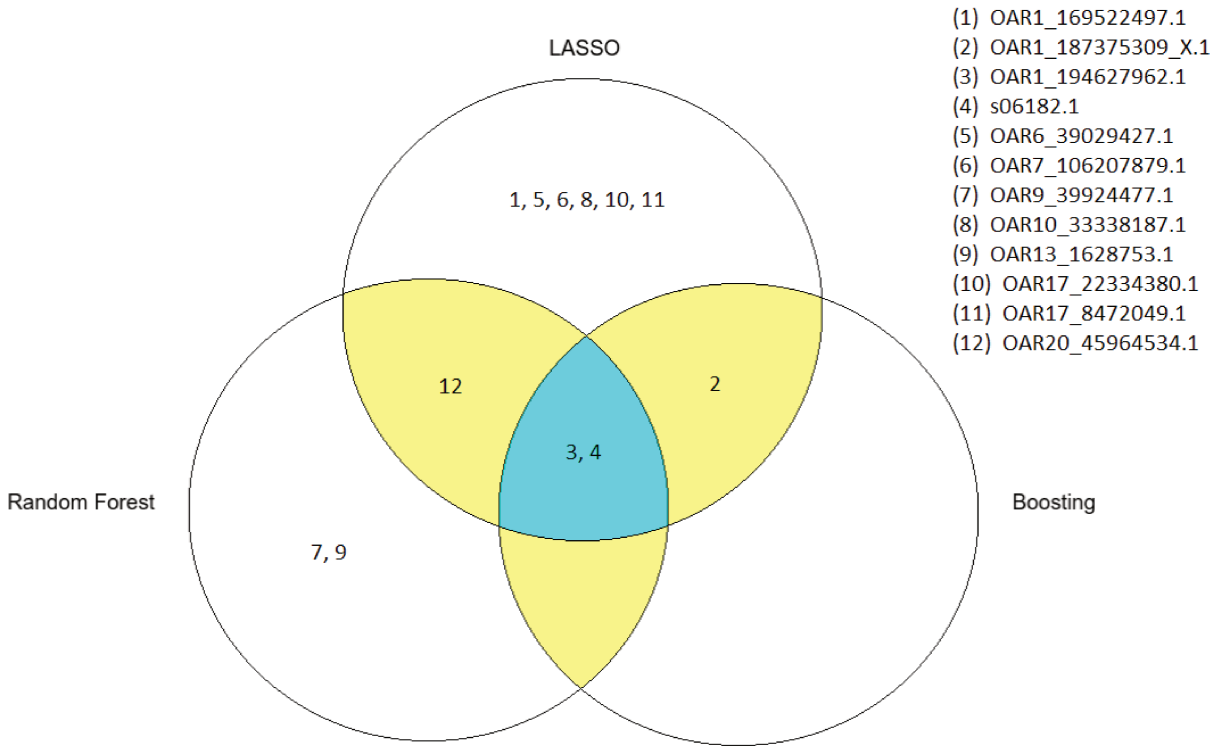

Figura 31: Diagrama de Venn para os marcadores selecionados para a raça Morada Nova pelos três modelos.

O diagrama de Venn para a raça Morada Nova exibe os marcadores OAR1_194627962.1 e s06182.1 em intersecção nos três modelos, possuindo frequência acima de 70% para a raça Morada Nova e frequências abaixo de 30% nas raças Crioula e Santa Inês, o que caracteriza esses SNP como bons discriminantes da raça Morada Nova. Os modelos LASSO e RandomForest indicaram o marcador OAR20_45964534.1 em comum, exibindo uma frequência de 75% para Morada Nova e frequências abaixo de 15% em outras duas raças, reforçando o potencial deste marcador. Por fim, os modelos LASSO e Boosting selecionaram o marcador OAR1_187375309_X.1 em comum, o qual possui frequência acima de 80% dentro da raça Morada Nova, colocando-o também como altamente relevante para a raça. De forma geral, os

marcadores encontrados em comum entre os modelos apresentam elevado potencial de discriminação da raça Morada Nova.

Por fim, o diagrama de Venn referente à raça Santa Inês é mostrado na Figura 32.

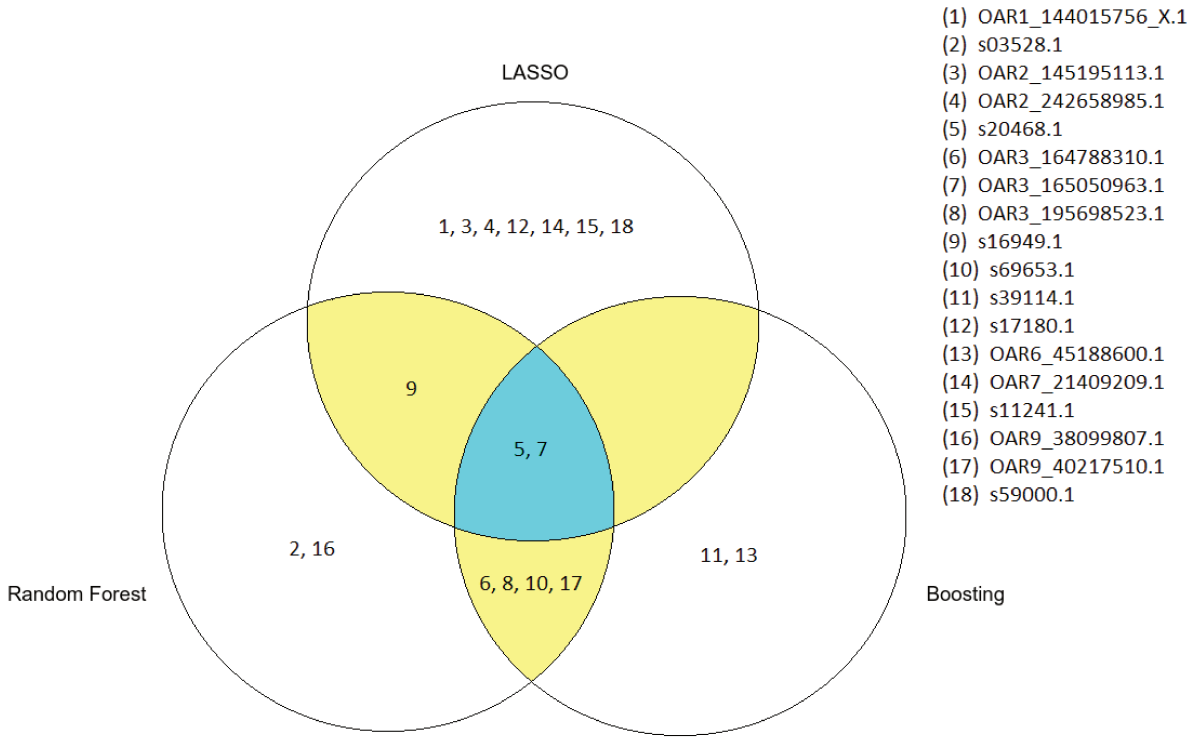

Figura 32: Diagrama de Venn para os marcadores selecionados para a raça Santa Inês pelos três modelos.

Em relação à raça Santa Inês, o diagrama de Venn mostra diversos marcadores em intersecção. Nos três modelos, nota-se a presença de dois marcadores: s20468.1 e OAR3_165050963.1. Esses marcadores encontrados pelos três modelos apresentam frequências acima de 70% em ovinos Santa Inês e frequências abaixo de 10% em outras duas raças, confirmando a alta capacidade dos mesmos na discriminação da raça. Os quatro marcadores obtidos tanto pelo modelo Random Forest quanto pelo modelo Boosting foram: OAR3_164788310.1, OAR3_195698523.1, s69653.1 e OAR9_40217510.1. O marcador OAR3_164788310.1, por exemplo, possui frequência de 89% para a raça Santa Inês, e frequências abaixo de 20% nos animais Crioula e Morada Nova, destacando-o como um

potencial identificador da raça. Apenas um marcador, o s16949.1, apareceu em comum tanto em LASSO quanto em Random Forest, exibindo uma frequência próxima de 90% dentro da raça Santa Inês e frequências abaixo de 20% nas raças Crioula e Morada Nova, também colocando-o como um potencial discriminante da raça Santa Inês.

Analisando os três diagramas referentes aos marcadores classificados como relevantes, pode-se afirmar que cinco marcadores possuem o maior potencial de identificação de cada uma das raças, pois foram selecionados nos três modelos. Estes marcadores estão descritos na Tabela 16.

Tabela 16: Marcadores SNP selecionados pelos três modelos e suas raças predominantes.

SNP	Cromossomo	Posição	Alelos*	Raça Predominante
OAR2_55853730.1	2	55853730	[A/C]	Crioula
OAR1_194627962.1	1	194627962	[G/A]	Morada Nova
s06182.1	5	30787155	[A/G]	Morada Nova
s20468.1	2	56248983	[A/G]	Santa Inês
OAR3_165050963.1	3	165050963	[A/G]	Santa Inês

* *Alelo específico para a raça predominante do lado esquerdo.*

Existem outros 10 marcadores que foram classificados como relevantes por duas das três técnicas, mostrando também um bom potencial de identificação racial, e que estão listados na Tabela 17.

Tabela 17: Marcadores SNP selecionados por dois modelos e suas raças predominantes.

SNP	Cromossomo	Posição	Alelos*	Raça Predominante
OAR15_45152619.1	15	45152619	[G/A]	Crioula
OAR16_39888776.1	16	39888776	[A/G]	Crioula
s30024.1	25	7165805	[C/A]	Crioula
OAR1_187375309_X.1	1	187375309	[A/G]	Morada Nova
OAR20_45964534.1	20	45964534	[G/A]	Morada Nova
s03528.1	1	28583773	[A/G]	Santa Inês
OAR3_164788310.1	3	164788310	[G/A]	Santa Inês
s69653.1	3	164951744	[G/A]	Santa Inês
s16949.1	3	164901721	[G/A]	Santa Inês
OAR9_40217510.1	9	40217510	[C/A]	Santa Inês

** Alelo específico para a raça predominante do lado esquerdo.*

Considerando apenas os marcadores que foram selecionados por dois e três modelos, um total de 15 marcadores demonstra ter grande potencial na identificação das raças estudadas. Deste total, quatro são específicos para a raça Crioula, quatro para a Morada Nova e sete para a Santa Inês.

Esse número de marcadores é próximo aos resultados de trabalhos relacionados à identificação racial em bovinos. Em seu modelo para discriminar as carnes originadas de gados japoneses e gados dos Estados Unidos, Suekawa *et al.* (2010) encontraram cinco marcadores por meio de análise de frequência alélica capazes de distinguir os gados japoneses dos gados americanos. Por sua vez, Sasazaki *et al.* (2011) desenvolveram um modelo no qual foram selecionados 11 marcadores SNP importantes para gados provenientes dos Estados Unidos. Em ambas as pesquisas foi utilizada uma matriz de 50K de marcadores SNP.

Em trabalhos relacionados à seleção de marcadores para outros fins, conjuntos com quantias próximas de marcadores foram selecionados. No trabalho de Heaton *et al.* (2005), por exemplo, por meio da análise de frequência alélica, foram identificados 20 SNP com alta relevância capazes de identificar, com alta precisão, um animal dentro de um rebanho, auxiliando na rastreabilidade dos produtos de origem bovina.

Para avaliar o potencial destes 15 SNP identificados nos modelos, criou-se um conjunto de dados contendo os ovinos com apenas estes 15 marcadores, e um outro conjunto de dados

contendo 20 marcadores SNP selecionados aleatoriamente do conjunto inicial de dados, e que, obviamente, não estavam nas listagens dos marcadores mais relevantes indicados pelos três modelos. Com estes dois conjuntos de dados, foram aplicadas novamente as três técnicas para se avaliar a acurácia e Kappa dos modelos gerados para cada um deles. Os resultados comprovaram que os marcadores selecionados pelos três modelos são mais eficientes, em termos de acurácia e Kappa, que um conjunto aleatório de SNP, como pode ser visto na Tabela 18.

Tabela 18: Medidas de avaliação dos modelos obtidos com os marcadores selecionados pelos modelos e com marcadores selecionados aleatoriamente.

Conjunto de dados	Acurácia			Kappa		
	LASSO	Random Forest	Boosting	LASSO	Random Forest	Boosting
Marcadores selecionados pelos modelos	96%	95%	100%	0,95	0,94	1
Marcadores selecionados aleatoriamente	33%	51%	28%	0,03	0,28	0,01

Os modelos obtidos para seleção dos marcadores SNP mais importantes poderão ser utilizados na certificação racial de animais já cadastrados nos bancos de germoplasma e de novos animais a serem inseridos nestes bancos. Além dos bancos de germoplasma, os modelos poderão ser utilizados por associações de criadores interessados em certificar seus animais, verificando o quão puro são os animais de seu rebanho. O MAPA também poderá utilizá-lo para realizar o controle de animais registrados que possivelmente apresentam alelos de outras raças, ocasionando a reclassificação ou mesmo a revogação de animais já registrados. Além dos modelos, os marcadores SNP selecionados poderão ser empregados na construção de ferramentas de genotipagem de SNP de baixa densidade, como os microarranjos, onde serão inseridos apenas os marcadores selecionados para genotipagem de novos ovinos.

6. CONCLUSÕES

Levando-se em consideração os modelos obtidos por meio da aplicação de técnicas que combinam métodos preditivos e seleção de atributos, pode-se concluir que é possível desenvolver modelos baseados em técnicas de mineração de dados para selecionar os marcadores SNP mais relevantes para as raças Crioula, Morada Nova e Santa Inês.

Pela análise de trabalhos relacionados, este é o primeiro trabalho a utilizar as três técnicas escolhidas (LASSO, Random Forest e Boosting), em um mesmo estudo, envolvendo dados de marcadores moleculares SNP, seja de ovinos ou de outros organismos.

A avaliação dos modelos com aplicação das três técnicas escolhidas revelou resultados promissores para a seleção dos marcadores SNP mais informativos relativos à identificação das raças de ovinos estudadas. Em particular, os modelos gerados pelas técnicas LASSO e Boosting obtiveram resultados melhores, em termos de acurácia e Kappa, em comparação com o modelo gerado pela técnica Random Forest.

Com relação aos resultados obtidos por meio das medidas de avaliação dos modelos, os resultados revelam que o número de marcadores identificados como relevantes, nos três modelos, ficou dentro do limite esperado, ou seja, menor que 0,2% do total de marcadores inicial, demonstrando que é viável a construção de classificadores com modelos compactos e eficazes para o problema de certificação racial de ovinos.

Na intersecção dos atributos que compõem os modelos, foram encontrados 15 marcadores com maior potencial de identificação das raças. Esses marcadores são considerados com maior potencial por terem sido selecionados por mais de um modelo, o que indica que realmente possuem alta correlação com a raça associada. Após uma nova aplicação das técnicas nesse conjunto de 15 marcadores e em outro conjunto aleatório de marcadores, os resultados comprovaram maior potencial do conjunto de 15 SNP na identificação das raças.

Como principal contribuição deste trabalho, os modelos desenvolvidos para seleção dos marcadores SNP mais eficientes na identificação das raças podem ser utilizados na certificação racial de animais já depositados nos bancos de germoplasma e de novos animais a serem inclusos nestes bancos, assim como podem ser utilizados por associações de criadores interessadas em

certificar seus animais, e pelo MAPA, no controle de animais registrados em seus bancos de dados.

Além disso, os marcadores SNP selecionados pelos modelos poderão ser empregados na geração de um produto, na forma de uma ferramenta de genotipagem de SNP de baixa densidade, como um microarranjo de SNP, constituindo-se também numa ferramenta auxiliar para identificação racial dos ovinos. É importante ressaltar que, quanto menor o número de marcadores montados no arranjo, menor o custo total da construção do produto.

Espera-se ainda que os modelos gerados produzam um impacto positivo no estado da arte para modelos com ovinos, abrindo caminho para a modelagem de outros fenótipos de interesse econômico da ovinocultura.

Durante a execução do trabalho, foram identificadas algumas possibilidades de trabalhos futuros, as quais estão descritas a seguir:

- Utilizar técnicas que combinem seleção de atributos e modelos preditivos com um conjunto de dados, incluindo outras raças de ovinos.

- Aplicar as técnicas de mineração de dados na seleção de marcadores SNP relacionados a diferentes características fenotípicas dos ovinos.

- Proceder com a validação experimental para verificar a eficiência dos modelos gerados. Essa avaliação deverá ser feita em laboratório, conduzida por especialista na área.

7. REFERÊNCIAS BIBLIOGRÁFICAS

AGRAWAL, R.; SRIKANT, R. Fast Algorithms for Mining Association Rules. **International Conference on Very Large Databases,** Santiago, Chile, 1994.

ARAÚJO, F. C.; MEDEIROS, J. X., Análise dos modos de governança da cadeia produtiva de ovinos no Distrito Federal: estudo de caso do frigorífico AICO por meio da análise multicritério. XLI Congresso Brasileiro de Economia e Sociologia Rural. 2003. **Anais...** Juiz de Fora-MG, 2003.

AYERS, K. L.; CORDELL, H. J. SNP selection in genome-wide and candidate gene studies via penalized logistic regression. **Genetic epidemiology**, v. 34, n. 8, p. 879-91, 2010.

BEIGUELMAN, B. Genética de populações humanas. Ed. SBG, Ribeirão Preto, 2008.

BREIMAN, L. Bagging predictors. **Machine Learning,** v. 24, n. 1, p. 123-140, 1996.

BREIMAN, L. Random Forests. **Machine Learning**, Boston, Netherlands, Boston, v. 45, n. 1, p. 5-32, 2001.

BRISOLA, M. V.; ESPIRITO SANTO, E. Panorama da cadeia produtiva da ovinocultura no Brasil. In: SIMPÓSIO MINEIRO DE OVINOCULTURA, III., Lavras, 2003. **Anais...**Minas Gerais: UFLA/Lavras, 2003.

CAETANO, A. R. Marcadores SNP: coneitos básicos, aplicações no manejo e no melhoramento animal e perspectiva para o futuro. **Revista Brasileira de Zootecnia**, v. 38, n. spe, p. 64-71, 2009.

CHAPMAN, P.; CLINTON, J.; KERBER, R.; KHABAZA, T.; REINARTZ, T.; SHEARER,

C.; WIRTH, R. **CRISP-DM 1.0:** step-by-step data mining guide. Illinois: SPSS, 78p. 2000.

COHEN, J. A. A coefficient of agreement of nominal scales. **Educational and Psychological Measurement**. v. 20, p. 37-46, 1960.

COSTA, N. G. **A cadeia produtiva de carne ovina no Brasil rumo às novas formas de organização da produção**. Dissertação (Mestrado) - Faculdade de Agronomia e Medicina Veterinária, Universidade de Brasília, Brasília, 2007. 182p.

DIAZ-URIARTE, R.; ALVAREZ, S. Gene selection and classification of microarray data using random forest. **BMC Bioinformatics**, v. 7, n. 3, jan. 2006.

FACÓ, O.; PAIVA, S. R.; ALVES, L, R. N.; LÔBO, R. N. B.; VILLELA, L. C. V.; **Raça Morada Nova: origem, características e perspectivas**, Sobral: Embrapa Caprinos, 2008. 43 p. - (Documentos / Embrapa Caprinos, ISSN 1676-7659; 75).

FAO - FOOD AND AGRICULTURE ORGANIZATION OF THE UNITED NATIONS. **FAO Statistical Yearbook 2012**: World Food and Agriculture. Roma, 2012. 366 p. Disponível em: <http://www.fao.org/docrep/015/i2490e/i2490e00.pdf>. Acesso em: 20 novembro de 2012.

FARAH, S.; DNA Segredos e Mistérios, ed. 2, São Paulo: Sarvier, 2007. 538 p.

FAYYAD, U.; PIATETSKY-SHAPIRO, G.; SMYTH, P. From Data Mining to Knowledge Discovery: an overview. In: **Advances In Knowledge Discovery & Data Mining. Menlo Park: American Association for Artificial Intelligence**, 1996. p. 1-34.

FERNANDES, A. A. O. **Genetic and phenotypic parameter estimates for growth, survival and reproductive traits in Morada Nova hair sheep**. 183 p. Thesis (Degree of Doctor of Philosophy) - Oklahoma State University, 1992.

FERNANDES, A. A. O.; BUCHANAN, D.; SELAIVE-VILLAROEL, A. B.. Avaliação dos fatores ambientais no desenvolvimento corporal de cordeiros desmamados da raça Morada Nova. **Revista Brasileira de Zootecnia**, v. 30, n. 5, p.1460-1465, 2001.

FIGUEIREDO, E. A. P.; OLIVEIRA, E. R.; BELLAVER, C. **Performance dos ovinos deslanados do Brasil**. Sobral: EMBRAPA-CNPC, 1980. 32 p. (EMBRAPA-CNPC. Circular Técnica, 1).

FIGUEIREDO, E. A. P.; SHELTON, M.; BARBIERI, M. E. Avaliable genetic resources: the origin and classification of the world's sheep. In: **Hair Sheep Production in Tropical and Subtropical Regions**, Davis, USA, p. 25-36, 1990.

FREUND, Y.; SCHAPIRE, R. A short introduction to boosting. **Journal of Japanese Society for Artificial Intelligence**, v. 14(5), p. 771-780, 1999.

FREUND, Y.; SCHAPIRE, R. A Tutorial on Boosting. Disponível em: <http://www.cc.gatech.edu/~thad/6601-gradAI-fall2013/boosting.pdf>. Acesso em: 14 fev 2014.

FRIEDMAN, J.; HASTIE, T.; TIBSHIRANI, R. Regularization Paths for Generalized Linear Models via Coordinate Descent. **Journal of Statistical Software**, v. 33, n. 1, p. 1-22, 2010.

GIBSON, G.; MUSE, S. V. A Primer of Genome Science. ed.3. Sinauer Associates. 2009.

GONZÁLEZ-RECIO, O.; WEIGEL, K. A.; GIANOLA, D.; NAYA, H.; ROSA, G. J. M. L2-Boosting Algorithm Applied to High-Dimensional Problems in Genomic Selection. **Genetics Research**, v. 92, n. 03, p. 227–237, 2010.

GOUVEIA, J. J. de S. **A utilização da genômica de populações na análise das principais raças de ovinos brasileiras**. Tese (Doutorado) – Universidade Federal do Ceará, Fortaleza, 2013.

98 p.

GURGEL, M. A.; SOUZA, A. A.; LIMA, F. A. M. Avaliação do feno de leucena no crescimento de cordeiros Morada Nova em confinamento. **Pesquisa Agropecuária Brasileira**, Brasília, v. 27, n. 11, p.1519-1526, 1992.

HALL, M.; FRANK, E.; HOLMES, G.; PFAHRINGER, B.; REUTEMANN, P.; WITTEN, I.H. The WEKA Data Mining Software: An Update. **SIGKDD Explorations**, v. 11, n. 1. 2009.

HAN, J.; KAMBER, M.; PEI, J. Data Mining: Concepts and Techniques. Morgan Kaufmann Publishers, ed. 3, San Francisco, CA, USA, 2011.

HARTL, D. L.; CLARCK, A. G. Principles of Population Genetics. 3ª ed. Sunderland, Massachusetts: Sinauer Associates, 1997.

HASTIE, T.; TIBSHIRANI, R.; FRIEDMAN, J. The Elements of Statistical Learning: Data Mining, Inference, and Prediction. Ed. Springer, London, 745 p., 2011.

HEATON, M. P., KEEN, J. E., CLAWSON, M. L., HARHAY, G. P., BAUER, N., SCHULTZ, C., GREEN, B. T., DURSO, L. M., CHITKO MCKOWN, C. G., LAEGREID, W. W. Use of bovine single nucleotide polymorphism markers to verify sample tracking in beef processing. **Journal of the American Veterinary Medical Association**, v. 226, n. 8, p. 1311-1314, 2005.

HILL, C. M.; MALONE, L. C.; TROCINE, L. Data Mining and Traditional Regression. In: BOZDOGAN, Hamparsum. **Statistical Data Mining and Knowledge Discovery**. Knoxville: Chapman & Hall/crc, 2003. p. 17.

IBGE. Produção da Pecuária Municipal. v. 39. Brasília, 2011. Disponível em: <ftp://ftp.ibge.gov.br/Producao_Pecuaria/Producao_da_Pecuaria_Municipal/2011/ppm2011.pdf>.

Acesso em 25 novembro de 2012.

JAMES, G.; HASTIE, T.; TIBSHIRANI, R. An Introduction to Statistical Learning: With Applications in R. Ed. Springer, London, 429 p., 2013.

KIM, S.; MISRA, A. SNP genotyping: Technologies and biomedical applications. **Annual Review of Biomedical Engineering**, v. 9, p. 289-320, 2007.

KUHN, M. caret: Classification and Regression Training. R package version 5.16-24, 2013.

LANDIS, J. R.; KOCH, G. G. The measurement of observer agreement for categorical data. **Journal of Biometrics**. Michigan, v. 33, p. 159-174, 1977.

LEWIS, J.; ABAS, Z.; DADOUSIS, C.; LYKIDIS, D.; PASCHOU, P.; DRINEAS, P. Tracing cattle breeds with principal components analysis ancestry informative SNPs. **PloS one**, v. 6, n. 4, p. e18007, 2011.

LI, R.; LI, Y.; FANG, X. SNP detection for massively parallel whole-genome resequencing. **Genome Research**, 2009.

LIAW, A.; WIENER, M. Classification and Regression by randomForest. **R News**, v. 2, n. 3, p.18-22, 2002.

MARIANTE, A. S.; CAVALCANTE, N. Animais do descobrimento: raças domésticas da História do Brasil. Brasília: Embrapa Sede. Embrapa Recursos Genéticos e Biotecnologia, 232 p., 2000.

MARIANTE, A. S.; ALBUQUERQUE, M. S. M.; EGITO, A. A.; MCMANUS, C.; LOPES, M. A.; PAIVA, S. R. Present status of the conservation of livestock genetic resources in Brazil. **Livestock Sci**., v.120, n.3, p.204-212, 2009.

MINISTÉRIO do Desenvolvimento, Indústria e Comércio Exterior – MDIC. **Estudo de mercado externo de productos derivados da ovinocaprinocultura**. Passo Fundo/RS: Méritos, 2010. 168 p.

MIRANDA, R. M. **Avaliação da influência de fatores genéticos e de meio sobre a produtividade de ovinos no cerrado** (continuação do projeto inicial). Projeto CNPq, 1990.

MOKRY, F. B.; HIGA, R. H.; MUDADU, M. de A.; LIMA, A. O. de; MEIRELLES, S. L. C.; SILVA, M. V. G. B.; CARDOSO, F. F.; OLIVEIRA, M. M. de O.; URBINATI, I.; NICIURA, S. C. M.; TULLIO, R. R.; ALENCAR, M. M. de; REGITANO, L. C. de A. Genome-wide association study for backfat thickness in Canchim beef cattle using Random Forest approach. **BMC Genetics**, London, v. 14, n. 47, 2013.

OECD-FAO Agricultural Outlook 2012-2021, OECD Publishing and FAO. Disponível em: <http://dx.doi.org/10.1787/agr_outlook-2012-en>. Acesso em: 23 novembro de 2012.

OSÓRIO, J. C. S.; OSÓRIO, M. T. M. Zootecnia de ovinos: raças, lã, morfologia, avaliação de carcaça, comportamento em pastejo. Pelotas: Universidade Federal de Pelotas, 2005. 243p.

PAIVA, S. R. **Caracterização da diversidade genética de ovinos no Brasil com quatro técnicas moleculares**. Tese (Doutorado)- Universidade Federal de Viçosa, Viçosa, 2005. 108p.

PANT, S. D.; SCHENKEL, F. S.; VERSCHOOR, C. P.; KARROW, N.A. Use of Breed-Specific Single Nucleotide Polymorphisms to Discriminate Between Holstein and Jersey Dairy Cattle Breeds. **Animal Biotechnology**, v.23, n.1, p.1-10, 2012.

REIS, F. A.; CABRAL, L. da S.; PACHECO, R. D. L.; GOMES, R. da C. Hurdles to the expansion of sheep meat supply chain in Central Brazil. In: REUNIÃO ANUAL DA SOCIEDADE BRASILEIRA DE ZOOTECNIA, 49, Brasília, DF. A produção animal no mundo

em transformação: **Anais...** Brasília, DF: SBZ, 2012.

REZENDE, S. O.; PUGLIESI, J. B.; MELANDA, E. A.; DE PAULA, M. F. Mineração de Dados. In: REZENDE, S. O. Sistemas Inteligentes: fundamentos e aplicações. 1ed. São Paulo: Manole, p. 307-336, 2003.

RIDGEWAY, G. gbm: Generalized Boosted Regression Models. **R package version 2.1**, 2013.

ROHRER, G. A.; FREKING, B. A.; NONNEMAN, D. Single nucleotide polymorphisms for pig identification and parentage exclusion. **Animal genetics**, v. 38, p. 253-258, 2007.

ROORKIWAL, M; SAWARGAONKAR, S. L.; CHITIKINENI, A.; THUDI, M.; SAXENA, R. K.; UPADHYAYA, H. D.; VALES, M. I.; RIERA-LIZARAZU, O.; VARSHNEY, R. K. Single nucleotide polymorphism genotyping for breeding and genetics applications in chickpea and pigeonpea using the BeadXpress platform.. **The Plant Genome**, v. 6, n. 2, 2013.

SANTOS, J. R. S. **Composição Física e Química dos Cortes Comerciais da Carcaça de Ovinos Santa Inês Terminados em Pastejo e Submetidos a Diferentes Níveis de Suplementação**. Patos, UFCG. 2007. 96p. (Dissertação - Mestrado em Zootecnia – Sistemas Agrosilvipastoris no Semiárido).

SASAZAKI, S.; HOSOKAWA, D.; ISHIHARA, R.; AIHARA, H.; OYAMA, K.; MANNEN, H. Development of discrimination markers between Japanese domestic and imported beef. **Animal science journal**, v. 82, n. 1, p. 67-72, 2011.

SCHAPIRE, R. The strength of weak learnabilíty. **Machine Learning**, v. 5, p. 197-227, 1990.

SILVA, R. R.da. **O Agronegócio Brasileiro de Carne Caprina e Ovina**. Salvador, 2002.

SOUZA, J. D. F. de; SOUZA, O. R. G. de; CAMPEÃO, P. Mercado e comercialização na

ovinocultura de corte no Brasil. In: CONGRESSO DA SOCIEDADE BRASILEIRA DE ECONOMIA, ADMINISTRAÇÃO E SOCIOLOGIA RURAL, 50, 2012, **Anais...** Vitória. Agricultura e desenvolvimento rural com sustentabilidade. Vitória: Sociedade Brasileira de Economia, Administração e Sociologia Rural, 2012.

SUEKAWA, Y.; AIHARA, H.; ARAKI, M.; HOSOKAWA, D.; MANNEN, H.; SASAZAKI, S. Development of breed identification markers based on a bovine 50K SNP array. **Meat science**, v. 85, n. 2, p. 285-8, jun. 2010.

TAN, P. N., STEINBACH, M., KUMAR, V. Introduction to Data Mining, (First Edition). Addison-Wesley Longman Publishing Co., Inc., Boston, MA, USA. 2005.

THE INTERNATIONAL SHEEP GENOMICS CONSORTIUM – ISGC; ARCHIBALD, A.L.; COCKETT, N.E.; DALRYMPLE, B.P.; FARAUT, T.; KIJAS, J.W.; MADDOX, J.F.; MCEWAN, J.C.; HUTTON ODDY, V.; RAADSMA, H.W.; WADE, C.; WANG, J.; WANG, W.; XUN, X.The sheep genome reference sequence: a work in progress. **Anim. Genet.**, n.41, p.449–453, 2010.

TIBSHIRANI, R. Regression Shrinkage and Selection via the Lasso, **Statistics in Medicine**, v.16, p. 385-395, 1997.

VAZ, C. M. S. L. Morfologia e Aptidão da Ovelha Crioula Lanada. Bagé: EMBRAPA Pecuária Sul. (Documentos, 22), 2000. 20 p.

VIANA, J. G. A. Panorama Geral da Ovinocultura no Mundo e no Brasil; **Revista Ovinos**, v. 12, n. 4, 2008.

WITTEN, I. H.; FRANK, E.; HALL, M. A. Data Mining: Practical Machine Learning Tools and Techniques. Morgan Kaufmann Publishers, San Francisco, CA, USA, 2011.

WU, Q.; YE, Y.; LIU, Y.; NG, M. K. SNP selection and classification of genome-wide SNP data

using stratified sampling random forests. **IEEE Transactions on Nanobioscience**, v.11, p.216–227, 2012.

WU, T. T.; CHEN, Y .F.; HASTIE, T.; SOBEL, E.; LANGE, K. Genome-wide association analysis by lasso penalized logistic regression. **Bioinformatics**, v. 25: p. 714–721, 2009.

YI, N.; XU, S. Bayesian LASSO for quantitative trait loci mapping. **Genetics**, v. 179, p. 1045–1055, 2008.