Capítulo 14

Mineração de dados: conceitos e um estudo de caso sobre certificação racial de ovinos

Fábio Danilo Vieira
Stanley Robson de Medeiros Oliveira

1 Introdução

1.1 Descoberta de conhecimento em banco de dados

Nos últimos anos, observa-se que uma grande quantidade de dados cresce de forma acelerada em diversos campos de conhecimento, fato que dificulta a sua interpretação, pois o 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 a transformar os dados em conhecimento (HAN et al., 2011).

Grande parte dessas técnicas e ferramentas podem ser encontradas no processo de Descoberta de Conhecimento em Bases de Dados, da sigla em inglês *Knowledge Discovery in Databases* (KDD). 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, como *Online analytical processing* (Olap), mas, sim, uma forma de se aperfeiçoar os resultados obtidos por meio das explorações realizadas pelas ferramentas atuais (REZENDE et al., 2003).

As aplicações das técnicas estão presentes em praticamente todos os setores do conhecimento humano. Na área de negócios, por exemplo, utilizam-se técnicas em detecção de fraudes em cartões, na criação de perfis de clientes de acordo com suas compras, entre outros. Na agricultura, podem ser utilizadas, em previsão de geadas, sistemas de alerta para a ferrugem do cafeeiro, sistemas de alerta para a ferrugem asiática da soja, etc. Na medicina, pode-se identificar terapias médicas de sucesso para diversas doenças. Na bioinformática, para se buscar padrões em sequências de DNA, 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 1, com muitas decisões sendo feitas pelo especialista do domínio dos dados.

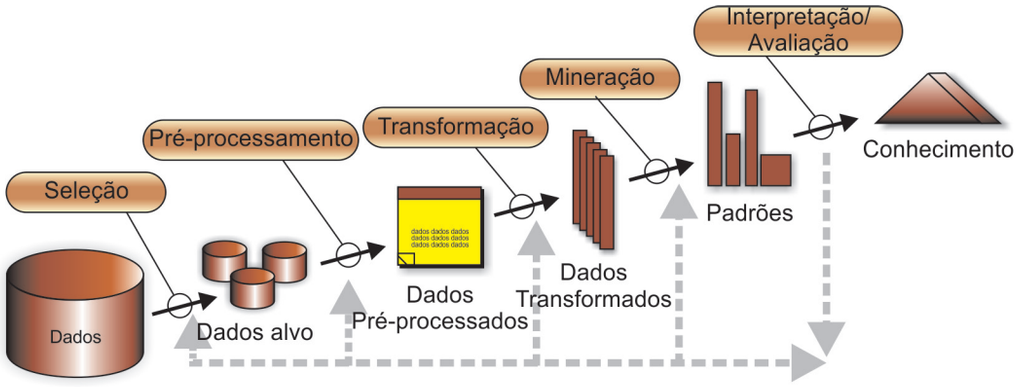

Figura 1. As fases do processo KDD.
Fonte: adaptado de Fayyad et al. (1996).

Os passos do processo KDD consistem em:

1) **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.
2) **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.
3) **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.
4) **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.
5) **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.
6) **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.
7) **Implantação do conhecimento descoberto (Conhecimento)**: incorporação deste conhecimento à performance do sistema ou, simplesmente, documentá-lo e reportá-lo às partes interessadas.

1.2 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), como ilustradas na Figura 2.

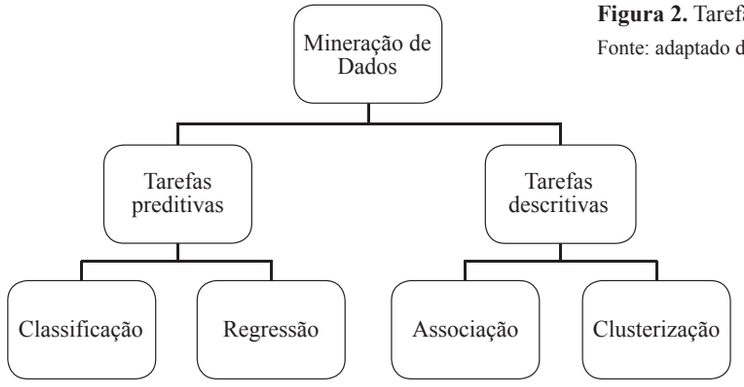

Figura 2. Tarefas de Mineração de Dados.
Fonte: adaptado de Rezende et al., 2003.

As tarefas preditivas têm como objetivo principal a construção de modelos que possam predizer a classe de um novo evento 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 classe determinada. A escolha de uma ou mais tarefas irá depender do problema a ser solucionado. As tarefas tradicionais de mineração de dados representadas na Figura 2 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: 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; 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 $l \cap r = \emptyset$, 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.

1.3 Modelo do processo de descoberta de conhecimento em bases de dados

Com o objetivo de padronizar o processo de descoberta de conhecimento, em 1996 foi criado o modelo de processo *CRoss-Industry Standart Process for Data Mining* (Crisp-DM), que divide o ciclo de vida de um projeto de mineração de dados em seis fases, a saber: compreensão do domínio, entendimento dos dados, preparação dos dados, modelagem, avaliação e distribuição (CHAPMAN et al., 2000).

As fases do modelo do processo estão ilustradas na Figura 3. O círculo externo traduz o aspecto cíclico de um projeto de mineração de dados, uma vez que após encontrar uma solução, o projeto não é necessariamente finalizado, e a partir de novos conhecimentos adquiridos podem ocorrer

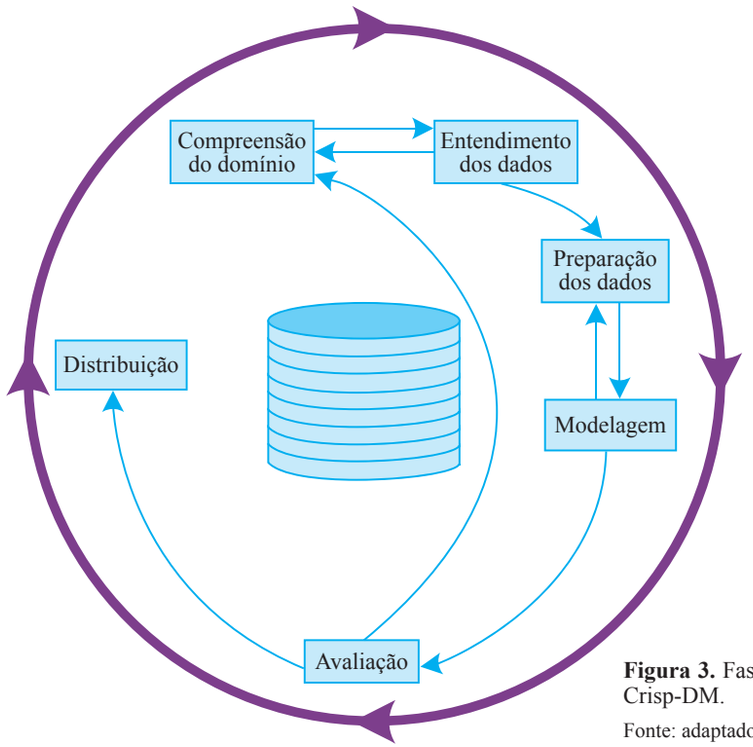

Figura 3. Fases do modelo de processo Crisp-DM.

Fonte: adaptado de Chapman et al. (2000).

novos questionamentos que levam a ações mais específicas. As setas internas ilustram as relações entre as fases, que indicam que a sequência entre elas não é rígida, sendo comum haver a necessidade de voltar ou avançar entre as fases.

A seguir, encontra-se uma breve descrição de cada fase do processo:

- Compreensão do domínio: compreender os objetivos e requisitos do projeto e transformar esse conhecimento em um problema de mineração de dados. Definir um plano preliminar para atingir esses objetivos.
- Entendimento dos dados: inicia-se com uma coleção de dados inicial e prossegue com atividades de exploração de dados, para se familiarizar, identificar problemas de qualidade, fazer as primeiras hipóteses e identificar possíveis subconjuntos que possam abrigar informações ocultas sobre esses dados.
- Preparação dos dados: a fase de preparação dos dados contempla todas as atividades necessárias para a construção do conjunto de dados final, no qual serão aplicadas as técnicas de modelagem. As atividades incluem, por exemplo, limpeza de dados, seleção e transformação de atributos, entre outras.
- Modelagem: nessa fase são escolhidas e aplicadas as técnicas de mineração de dados, e seus parâmetros são calibrados. Diversas técnicas podem ser aplicadas ao mesmo problema, embora cada técnica necessite de formatos específicos e necessite voltar para a fase de preparação de dados.
- Avaliação: nesse estágio, tem-se o modelo (ou modelos) com boa qualidade. Os resultados são comparados e interpretados conforme a área de aplicação. É importante reavaliar todas as etapas do processo para se ter a certeza de que o modelo atende às necessidades e aos objetivos do projeto.
- Distribuição: a criação de modelos geralmente não finaliza um projeto. O conhecimento obtido deve ser documentado, organizado e apresentado para os usuários, para que estes possam saber quais ações devem ser realizadas para aproveitar os modelos criados.

2 Estudo de caso

2.1 Modelagem para certificação racial de ovinos

O Brasil possui diversas raças de ovinos que foram desenvolvidas 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 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 Empresa Brasileira de Pesquisa Agropecuária (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 *deoxyribonucleic acid* (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 *Single Nucleotide Polymorphism* (SNP) mostraram ser mais efetivos no auxílio da certificação racial de animais domésticos (PANT et al., 2012; SASAZAKI et al., 2011; SUEKAWA et al., 2010). Os marcadores SNP 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. 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) (). 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 Ministério Agricultura Pecuária e Abastecimento (Mapa), 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 (KIM; MISRA, 2007; ROORKIWAL et al., 2013). 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).

2.2 Metodologia

O exemplo utilizado para demonstrar o processo de aplicação de técnicas de mineração de dados aborda o desenvolvimento de modelos para selecionar os principais marcadores SNP na identificação racial de animais pertencentes às raças Crioula, Morada Nova e Santa Inês. Para dar suporte aos procedimentos realizados no exemplo, optou-se por seguir o modelo de processo Crisp-DM, já explicado anteriormente. Cada uma das seis fases do processo de análise do Crisp-DM, para o estudo de caso analisado, estão descritas a seguir.

Fase 1 - Compreensão do domínio: Considerando que o objetivo principal do exemplo utilizado 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.

Fase 2 - Entendimento dos dados: O conjunto de dados analisado no exemplo foi obtido do Consórcio Internacional do Genoma Ovino (ARCHIBALD et al., 2010) por meio da Rede Genômica Animal, projeto da Embrapa. Este conjunto era composto por 72 animais das raças estudadas (23 animais da raça Crioula, 22 da Morada Nova e 27 da Santa Inês), sendo que cada animal possuía 49.034 marcadores SNP. Observa-se, então, que o conjunto de dados é 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, sendo que cada alelo pode conter uma Adenina (A) ou uma Timina (T) ou uma Citosina (C) ou uma Guanina (G). A Figura 4 ilustra o formato do conjunto de dados de ovinos em estudo.

	SNP 1	SNP 2	SNP 3	SNP 4	SNP 5	SNP 6	...	Raça
72 animais	AA	AG	AG	AG	AG	CC	AC	Crioula
	GA	AG	AG	GG	GG	AC	AC	Morada Nova
	GA	GG	AG	GG	AG	CC	CC	Santa Inês
	...	...	...	...	...	...	...	...

49.034 SNP

Figura 4. Formato do conjunto de dados de marcadores SNP das três raças em estudo.

Fase 3 - Preparação dos dados: Na etapa de preparação dos dados, realizou-se uma verificação quanto à existência de amostras idênticas dentro do conjunto de dados e de marcadores SNP que tivessem um valor único de genótipo para todas as raças. Após a verificação, constatou-se que não existiam amostras idênticas. Entretanto, existiam 384 marcadores SNP com valor único para todas as raças, os quais foram removidos do conjunto de dados final.

Fase 4 - Modelagem: Na etapa da modelagem do exemplo utilizado, foram aplicadas técnicas que combinam seleção de atributos e desenvolvimento de modelos preditivos para identificar os marcadores SNP mais relevantes para três raças de ovinos. Entretanto, devido ao elevado número de atributos (marcadores SNP) e o baixo número de registros (animais), técnicas capazes de lidar com esta situação foram empregadas, a saber: *Least Absolute Shrinkage and Selection Operator* (Lasso), Random Forest e Boosting.

Lasso é um método de regressão penalizada utilizado para reduzir os efeitos dos atributos que não contribuem para identificação da classe (atributo-meta ou variável resposta), encolhendo seus coeficientes para zero e excluindo-os do modelo (TIBSHIRANI, 1997). O método é usado normalmente para estimar os parâmetros de regressão $\beta = (\beta_1, \beta_2, ..., \beta_p)^T$ no modelo da Equação 1:

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

onde, y_i é a raça do i-ésimo animal (i = 1, 2, ..., 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, ..., p) do animal i; o coeficiente β_j representa o efeito do marcador j na raça; e_i é o erro residual. 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).

Sendo $\hat{\beta} = (\hat{\beta}_1, \hat{\beta}_2, ..., \hat{\beta}_p)^T$, a estimativa Lasso para problemas de classificação é definida pela função de máxima verossimilhança penalizada descrita na Equação 2:

$$l(\hat{\mu}, \hat{\beta}) = \frac{1}{n}\sum_{i=1}^{n}[y_i(+\sum_{j=1}^{p} x_{ij}\beta_j) - \log(1 + e^{\mu + \sum_{j=1}^{p} x_{ij}\beta_j})] \quad (2)$$

$$\text{Sujeito à restrição } \sum_{j=1}^{p} |\beta_j| \leq t \text{ para } t \geq 0,$$

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.

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 árvore 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 dos atributos mais informativos. 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). Os principais passos do algoritmo Random Forest podem ser vistos na Figura 5.

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 pelos modelos f^b, resultando uma classificação final de acordo com a votação majoritária.

Figura 5. Algoritmo básico da técnica Random Forest.
Fonte: Breiman (2001).

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 evitar *overfitting*, foi utilizada a abordagem Random Forest que, em geral, lida melhor com o problema de sobreajuste nos modelos (MEGETO et al., 2014).

A ideia principal da técnica Boosting é transformar múltiplos classificadores ruins em um único muito bom (FREUND; 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. O algoritmo que mostra a execução básica da técnica Boosting é descrito na Figura 6.

Para aplicação das técnicas de modelagem, escolheu-se o software R (versão 3.0.1). O pacote instalado para o algoritmo Lasso foi o glmnet (FRIEDMAN et al., 2010), para Random Forest foi instalado o pacote randomForest (LIAW; WIENER, 2002) e, para Boosting, foi instalado o algoritmo gbm (RIDGEWAY, 2013). Além destes, instalou-se o pacote caret (KUHN, 2013), utilizado para a escolha dos melhores valores para alguns parâmetros de cada técnica aplicada.

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 pelos modelos f^b, resultando uma classificação final de acordo com a votação majoritária.

Figura 6. Algoritmo básico do algoritmo Boosting.

Fonte: James et al. (2013).

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, antes da utilização do pacote caret, 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 λ (lambda). O número padrão deste intervalo é de 100 valores possíveis (FRIEDMAN et al., 2010; JAMES et al., 2013), obtidos separadamente pelo algoritmo Lasso, via validação cruzada, sobre os dados analisados. Após a aplicação da técnica Lasso, utilizou-se Random Forest para a busca dos marcadores SNP mais relevantes, associados a cada uma das raças. Os parâmetros avaliados para Random Forest foram o número de árvores a serem construídas e o número de atributos selecionados para determinar o *split* em cada nó das árvores. Com a construção desta floresta, foi possível determinar os marcadores mais importantes para o modelo (do atributo mais importante ao menos relevante). Assim como Random Forest, Boosting foi utilizado para fornecer um modelo com a listagem dos marcadores mais importantes na identificação das raças. O único parâmetro testado para Boosting foi o número de classificadores a serem desenvolvidos para o modelo final. Os classificadores construídos pela técnica Boosting foram baseados em árvores de decisão, as quais foram construídas em distribuições reponderadas dos dados.

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, a fim de verificar o quanto um alelo estava presente em uma raça e ausente em outras duas. Por fim, foi selecionado um subconjunto menor de marcadores SNP com maior potencial de identificação das três raças pesquisadas.

Fase 5 - Avaliação: nesta etapa do processo, os modelos já foram desenvolvidos. Antes de passar a fase final de desenvolvimento dos modelos finais, é importante revisar todos os passos executados, para verificar se os objetivos foram alcançados. No fim dessa etapa, uma decisão a respeito do uso dos resultados da análise deve ser tomada.

Para avaliar o desempenho dos modelos, dividiu-se o conjunto de dados inicial em duas partes disjuntas, sendo que uma parte constitui o conjunto de treinamento e outra o conjunto de teste. As técnicas utilizaram dois tipos de particionamento dos dados: validação cruzada e *bootstrap*. Na validação cruzada, os dados são particionados em *k* sub-conjuntos de tamanhos aproximadamente iguais, e o indutor é treinado e testado *k* vezes. Para cada uma das vezes, o indutor é testado com uma das partições e treinado com o restante. O bootstrap consiste em gerar os conjuntos de treinamento e teste a partir de uma amostragem randômica dos dados, repetindo esse processo de classificação por várias vezes. A cada ciclo, as amostragens são selecionadas com reposição, isto é, um mesmo exemplo poderá aparecer mais de uma vez no mesmo sub-conjunto.

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, enquanto o Kappa (COHEN, 1960) 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).

Fase 6 - Distribuição: a construção dos modelos geralmente não é a fase final de um processo de mineração de dados. O conhecimento obtido deve ser organizado e apresentado para os usuários, para que estes possam saber quais ações devem ser realizadas para aproveitar os modelos criados. Dependendo do projeto, a distribuição pode ser simplesmente a geração de um relatório ou, em outras vezes, pode ser uma tarefa mais complexa. Em muitos casos, a tarefa da distribuição é responsabilidade do usuário não analista dos dados (não especialista), por isso a importância que estes tenham algum conhecimento da forma como podem utilizar os modelos.

A fase 5 será abordada com mais detalhes no decorrer da próxima seção (Resultados obtidos). Em relação à fase 6 do exemplo utilizado, esta será realizada em trabalhos futuros, quando os modelos obtidos serão repassados para especialistas em genômica animal, que poderão aplicá-los no desenvolvimento de ferramentas de genotipagem de baixa densidade, entre outras aplicações.

2.3 Resultados obtidos

Na aplicação do algoritmo Lasso, para obtenção do melhor valor de λ, avaliaram-se intervalos de 100 e de 1.000 valores possíveis. Entretanto, o número de marcadores selecionados e a acurácia permaneceram inalterados, mantendo-se, então, os 100 valores fornecidos por caret. Com o valor ótimo de λ, o algoritmo Lasso selecionou 29 marcadores relevantes, dos quais, cinco se destacaram para a raça Crioula, 12 para Morada Nova e 12 para Santa Inês. Os cinco marcadores que se destacaram para Crioula e suas respectivas informações estão descritas na Tabela 1. Cabe ressaltar que, 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.

De forma geral, todos os marcadores mostraram alto potencial de identificação da raça Crioula, destacando-se, entre outros, quatro marcadores (OARX_121724022.1, s56924.1, OARX_78903642.1 e OARX_29830880.1) pertencentes ao cromossomo X. Foi observado que todos os marcadores da raça Crioula possuem altas diferenças de frequências em relação às outras raças, o que se deve, provavelmente, ao fato de ela possuir as características físicas mais distintas entre elas, como possuir tamanho diminuto e ser lanada (PAIVA, 2005).

Tabela 1. Frequências alélicas dos marcadores SNP selecionados pelo algoritmo LASSO para a raça Crioula.

SNP	Cromossomo	Posição	Alelos*	Frequência alélica**		
				Crioula	Morada Nova	Santa Inês
OARX_121724022.1	X	121724022	[C/A]	0.98	0.02	0.05
OARX_29830880.1	X	29830880	[A/G]	0.80	0.00	0.05
OARX_78903642.1	X	78903642	[A/G]	0.95	0.07	0.09
s56924.1	X	53358543	[A/G]	0.98	0.13	0.15
OAR1_268303279_X.1	1	268303280	[G/A]	0.78	0.07	0.09

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

Para a raça Morada Nova, LASSO identificou os 12 marcadores listados na Tabela 2.

Tabela 2. Frequências alélicas dos marcadores SNP, selecionados pelo algoritmo LASSO para a raça Morada Nova.

SNP	Cromossomo	Posição	Alelos*	Frequência alélica**		
				Morada Nova	Crioula	Santa Inês
s05480.1	X	52592630	[G/A]	0.93	0.15	0.22
OAR1_187375309_X.1	1	187375310	[A/G]	0.86	0.02	0.31
OAR1_194627962.1	1	194627962	[G/A]	0.73	0.00	0.02
DU373896_534.1	3	139464759	[A/C]	0.82	0.35	0.15
s32131.1	4	22382506	[A/G]	0.98	0.32	0.42
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
OAR9_39924477.1	9	39924477	[A/C]	0.95	0.17	0.33
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	0.00	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.

Os destaques para a raça Morada Nova são dois SNP (OAR1_187375309_X.1 e OAR1_194627962.1) no cromossomo um e dois SNP (OAR17_8472049.1 e OAR17_22334380.1) no cromossomo 17, além do total de seis marcadores com frequência acima de 90%. Foi observado ainda que há uma frequência relativamente maior dos alelos dos animais Morada Nova na raça Santa Inês. Isto talvez seja explicado pelo fato de os animais Santa Inês serem originários do cruzamento entre Morada Nova e outros ovinos sem raça definida do nordeste brasileiro, fazendo com que muitos ovinos Santa Inês preservem características genotípicas do Morada Nova (PAIVA, 2005).

Para a Santa Inês, foram selecionados os 12 marcadores apresentados na Tabela 3.

Dentre os marcadores selecionados para a raça Santa Inês, três pertencem ao cromossomo dois (OAR2_145195113.1, OAR2_242658985.1 e s20468.1), três ao cromossomo três (OAR3_153703374.1, OAR3_165050963.1 e s16949.1) e três ao cromossomo sete

Tabela 3. Frequências alélicas dos marcadores SNP, selecionados pelo algoritmo Lasso para a raça Santa Inês.

SNP	Cromossomo	Posição	Alelos*	Frequência alélica**		
				Santa Inês	Crioula	Morada Nova
OARX_53305527.1	X	53305527	[A/G]	0.72	0.00	0.09
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	0.00
OAR3_153703374.1	3	153703374	[A/G]	0.76	0.41	0.13
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
OAR5_93120389.1	5	93120389	[G/A]	0.89	0.19	0.38
OAR7_21409209.1	7	21409209	[G/A]	0.61	0.02	0.11
OAR7_94733688.1	7	94733688	[G/A]	0.98	0.37	0.59
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.

(OAR7_21409209.1, OAR7_94733688.1 e s11241.1). Uma observação importante vem do fato de que dos três marcadores do cromossomo, três estão em posições muito próximas. De maneira geral, os marcadores para a raça Santa Inês têm altas diferenças de frequência alélica em relação às outras raças, tendo como destaques os marcadores OARX_53305527.1 e s20468.1.

A acurácia atingida com o conjunto de 29 marcadores SNP selecionados pelo algoritmo Lasso foi de 100% na predição de novas raças, e o índice Kappa foi igual a 1. O algoritmo Lasso teve ótimo desempenho, tanto em termos de acurácia quanto computacionalmente (tempo de execução baixo, em torno de uma hora e 57 minutos), 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.

Random Forest gerou uma listagem dos marcadores mais importantes para o modelo de identificação das raças ovinas. Experimentou-se modelos combinando de 1.000 a 5.000 árvores, e conjuntos aleatórios de atributos variando de 20 a 49.033 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 313 marcadores para *split*. Selecionou-se, então, os 27 melhores SNP classificados, pois, a partir desta posição os SNP restantes contribuíam com menos que 2% para o modelo. Em Mokry et al. (2013), utilizou-se um critério de seleção diferente, no qual, primeiramente selecionou-se 1% dos SNP mais relevantes de cada cromossomo e, em seguida, foi selecionado 1% dos SNP mais importantes do subconjunto anterior, sendo selecionados 70 marcadores SNP pela técnica Random Forest, utilizando tal critério.

Do conjunto total de 27 marcadores, nove marcadores também foram selecionados pelo algoritmo Lasso. 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 4 mostra os marcadores predominantes na raça Crioula e as frequências dos alelos específicos desta raça em relação à Morada Nova e Santa Inês.

Tabela 4. Frequências alélicas dos marcadores SNP, selecionados pelo algoritmo Random Forest para a raça Crioula.

SNP	Cromossomo	Posição	Alelos*	Frequência alélica**		
				Crioula	Morada Nova	Santa Inês
OARX_121724022.1	X	121724022	[C/A]	0.98	0.02	0.05
OARX_29830880.1	X	29830880	[A/G]	0.80	0.00	0.05
OARX_78903642.1	X	78903642	[A/G]	0.95	0.07	0.09
s56924.1	X	53358543	[A/G]	0.98	0.13	0.15
OAR1_23724877.1	1	23724877	[G/A]	0.50	0.00	0.04
OAR2_212548956.1	2	212548956	[G/A]	0.80	0.04	0.18
OAR2_55853730.1	2	55853730	[A/C]	0.85	0.00	0.07
OAR11_18815864.1	11	18815864	[A/G]	0.93	0.34	0.22
s71482.1	14	41937578	[G/A]	0.91	0.18	0.50
OAR15_45152619.1	15	45152619	[G/A]	0.76	0.02	0.02
OAR16_39888776.1	16	39888776	[A/G]	0.89	0.11	0.15
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.

Do conjunto de 13 marcadores identificados por Random Forest para a raça Crioula, quatro também foram identificados por LASSO (OARX_121724022.1, OARX_29830880.1 e OARX_78903642.1, s56924.1). Os dois SNP do cromossomo 25 estão em posições próximas e com frequência acima de 90% dentro da raça, surgindo como bons separadores raciais. De forma geral, os SNP fornecidos por Random Forest se mostraram importantes na identificação da raça Crioula.

Na Tabela 5, os SNP com predominância na raça Morada Nova são listados.

Tabela 5. Frequências alélicas dos marcadores SNP, selecionados pelo algoritmo Random Forest para a raça Morada Nova.

SNP	Cromossomo	Posição	Alelos*	Frequência alélica**		
				Morada Nova	Crioula	Santa Inês
OAR1_194627962.1	1	194627962	[G/A]	0.73	0.00	0.02
OAR2_54691204.1	2	54691204	[G/A]	0.57	0.04	0.00
OAR18_65638912.1	18	65638912	[G/A]	1.00	0.56	0.41

* 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.

O algoritmo Random Forest indicou três marcadores importantes para a raça Morada Nova. Como destaque, observa-se os marcadores OAR1_194627962.1, indicado também pelo modelo Lasso, e OAR2_54691204.1, com frequência acima de 50% na Morada Nova e praticamente ausente nas outras duas raças. O marcador OAR18_65638912.1 se destaca com frequência de 100% na raça Morada Nova, apesar de sua frequência em outras duas raças ter ficado entre 40% e 60%.

Na Tabela 6 pode-se observar os SNP com alta frequência na raça Santa Inês.

Tabela 6. Frequências alélicas dos marcadores SNP, selecionados pelo algoritmo Random Forest para a raça Santa Inês.

SNP	Cromossomo	Posição	Alelos*	Frequência alélica**		
				Santa Inês	Crioula	Morada Nova
OARX_53305527.1	X	53305527	[A/G]	0.72	0.00	0.09
s61697.1	-	-	[C/A]	0.68	0.06	0.04
OAR1_175474366.1	1	175474366	[G/A]	0.55	0.24	0.00
s03528.1	1	28583773	[A/G]	0.92	0.43	0.23
s20468.1	2	56248983	[A/G]	0.76	0.15	0.00
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_76802154.1	9	76802154	[A/G]	0.96	0.32	0.50

* 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.

Para a raça Santa Inês, 11 marcadores foram selecionados com altas frequências alélicas. Destes, quatro estavam presentes no modelo fornecido pelo algoritmo LASSO (OARX_53305527.1, 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). O marcador s61697.1 também se destaca com alta frequência na raça Santa Inês e com frequências abaixo de 7% na raça Crioula e Morada Nova.

Para treinamento e teste, foram desenvolvidas e combinadas 1.000 árvores utilizando as amostras bootstrap. O comitê de classificadores que formaram a floresta obteve uma acurácia de 99% e Kappa de 0,98. Em relação ao desempenho, o modelo Random Forest foi gerado em duas horas e 44 minutos, ou seja, um tempo que pode ser considerado aceitável levando-se em consideração o elevado número de atributos do conjunto de dados.

Na aplicação da técnica Boosting, o único parâmetro testado foi o número de classificadores (neste caso, árvores de decisão) a serem construídos. Avaliou-se modelos desenvolvidos com totais entre 1.000 e 10.000 árvores, sendo que o melhor resultado, em termos de acurácia e Kappa, ocorreu com 1.000 árvores, número fornecido pelo pacote caret. Selecionou-se os 20 melhores marcadores, pois os SNP a partir desta posição pouco contribuíam (menos que 1%) para o modelo. Entre os 20 marcadores ordenados por Boosting, seis estavam presentes nos modelos Lasso e Random Forest, dois estavam somente em Lasso e sete somente no modelo Random Forest. Com isto, Boosting selecionou apenas cinco marcadores diferentes das técnicas anteriores. Na Tabela 7 estão descritos os SNP predominantes na raça Crioula e suas frequências.

Na lista de marcadores importantes para a raça Crioula, dois deles (OARX_121724022.1 e s56924.1) foram indicados nos dois modelos anteriores, e outros dois (OAR2_55853730.1 e OAR15_45152619.1) foram selecionados no modelo Random Forest, demonstrando o alto potencial destes marcadores. Os marcadores indicados apenas no modelo Boosting (OAR4_51441757.1, OAR6_110447914.1 e s30024.1) também mostraram ser potenciais discriminantes de raças.

Tabela 7. Frequências alélicas dos marcadores SNP, selecionados pelo algoritmo Boosting para a raça Crioula.

SNP	Cromossomo	Posição	Alelos*	Frequência alélica**		
				Crioula	Morada Nova	Santa Inês
OARX_121724022.1	X	121724022	[C/A]	0.98	0.02	0.05
s56924.1	X	53358543	[A/G]	0.98	0.13	0.15
OAR2_55853730.1	2	55853730	[A/C]	0.85	0.00	0.07
OAR4_51441757.1	4	51441757	[A/G]	0.91	0.25	0.16
OAR6_110447914.1	6	110447914	[G/A]	0.67	0.04	0.02
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.

A Tabela 8 traz uma listagem dos marcadores com predominância na raça Morada Nova.

Tabela 8. Frequências alélicas dos marcadores SNP, selecionados pelo algoritmo Boosting para a raça Morada Nova.

SNP	Cromossomo	Posição	Alelos*	Frequência alélica**		
				Morada Nova	Crioula	Santa Inês
OAR1_194627962.1	1	194627962	[G/A]	0.73	0.00	0.02
s32131.1	4	22382506	[A/G]	0.98	0.32	0.42
s06182.1	5	30787155	[A/G]	0.93	0.15	0.31
s10365.1	10	21720029	[G/A]	0.45	0.00	0.00

* 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.

O algoritmo Boosting separou cinco marcadores com maior frequência em Morada Nova, sendo um deles (OAR1_194627962.1) presente nos dois modelos anteriores e dois (s32131.1, s06182.1) no modelo LASSO. O marcador OAR1_194627962.1 possui frequência de apenas 2% na Santa Inês e ausente na Crioula, resultado que o confirma como um bom discriminante de raças. Os marcadores s32131.1 e s06182.1 surgem com frequência acima de 90% nos animais Morada Nova, o que também demonstra o bom potencial destes SNP.

A Tabela 9 apresenta os marcadores associados a raça Santa Inês.

Dentre os marcadores fornecidos pelo modelo Boosting para a raça Santa Inês, destacam-se três deles (OARX_53305527.1, s20468.1, OAR3_165050963.1) também selecionados pelas técnicas Lasso e Random Forest. Além disso, dois SNP (s39114.1, OAR9_40217510.1) foram selecionados exclusivamente por Boosting. De forma geral, a maioria dos marcadores selecionados para a raça Santa Inês apresenta alta frequência de alelo. Destaque para os três SNP também indicados pelos dois modelos anteriores, atestando seu potencial de identificação da raça Santa Inês.

Para realização de treinamento e teste, o algoritmo Boosting foi executado por meio de validação cruzada em 10 subconjuntos de dados, sendo que 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-

Tabela 9: Frequências alélicas dos marcadores SNP, selecionados pelo algoritmo Boosting para a raça Santa Inês.

SNP	Cromossomo	Posição	Alelos*	Frequência alélica**		
				Santa Inês	Crioula	Morada Nova
OARX_53305527.1	X	53305527	[A/G]	0.72	0.00	0.09
s61697.1	-	-	[C/A]	0.68	0.06	0.04
s03528.1	1	28583773	[A/G]	0.92	0.43	0.23
s20468.1	2	56248983	[A/G]	0.76	0.15	0.00
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
s39114.1	3	232410568	[A/G]	0.59	0.08	0.07
s69653.1	3	164951744	[G/A]	0.90	0.08	0.36
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.

-se acreditar que há indícios de *overfitting*, porém os parâmetros ajustados para a execução do algoritmo foram obtidos pelo caret de forma a evitar um super-ajuste do modelo. O tempo de execução do algoritmo Boosting foi o menor entre os três modelos, sendo finalizado em uma hora e nove minutos. Este bom 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 com um tempo computacional relativamente curto.

Com o desenvolvimento dos três modelos e a seleção dos principais marcadores para identificação das raças, foi realizada uma análise daqueles SNP que convergiam em dois ou três modelos. A intersecção dos modelos envolvendo a raça Crioula mostra que os marcadores OARX_121724022.1 e o s56924.1 foram selecionados nos três modelos, demonstrando alta relevância na identificação da raça Crioula. O marcador OARX_ 121724022.1, em especial, possui uma frequência de 98%, ou seja, demonstra ser um SNP com alto potencial de identificação da raça.

A intersecção presente nos três modelos, relativa à raça Morada Nova, exibe o marcador OAR1_194627962.1 com frequência de 73% para a raça Morada Nova e frequências praticamente nulas nas outras raças, o que caracteriza esse SNP como bom discriminante da raça. Os modelos Lasso e Boosting selecionaram os SNP s32131.1 e o s06182.1, os quais possuem frequências acima de 90% na raça Morada Nova, colocando-os também como altamente relevantes para a raça.

Em relação à raça Santa Inês, a intersecção mostra que, nos três modelos, há a presença de três SNP (OARX_53305527.1, s20468.1 e OAR3_165050963.1) que apresentam frequências acima de 70% em ovinos Santa Inês e abaixo de 10% em outras raças, confirmando alta capacidade na discriminação racial. Entre os marcadores obtidos por Random Forest e Boosting, destaca-se o s61697.1, com frequência de 68%, posicionando-o como um potencial identificador da raça.

Considerando apenas os marcadores selecionados por dois e três modelos, um total de 18 marcadores demonstra ter grande potencial na identificação das raças estudadas. Esse número de marcadores é próximo aos resultados de trabalhos relacionados à identificação racial em bovinos,

como em Suekawa et al. (2010), onde foram encontrados cinco marcadores por meio de análise de frequência alélica capaz de distinguir gados japoneses e americanos. Por sua vez, Sasazaki et al. (2011) desenvolveram um modelo no qual foram selecionados 11 SNP importantes para gados provenientes de rebanhos dos Estados Unidos.

2.4 Considerações finais

A avaliação dos modelos com aplicação das três técnicas de mineração de dados escolhidas revelou resultados promissores para a seleção dos marcadores SNP mais informativos, que identificam as raças 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 Random Forest. Considerando que o conjunto de dados utilizado possui um elevado número de atributos, as técnicas utilizadas reduziram o número de SNP para menos de 0,2%. Na intersecção dos marcadores que compõem os modelos, foram encontrados 18 SNP com maior potencial de identificação das raças, indicando que realmente os marcadores selecionados possuem alta correlação com a raça associada. Os modelos desenvolvidos podem ser utilizados na certificação racial de animais já depositados em bancos de germoplasma e de novos animais a serem inclusos nestes bancos, assim como poderão ser utilizados por diversos segmentos ligados à ovinocultura, como por exemplo, associações de criadores interessadas em certificar seus animais, e pelo Ministério da Agricultura, Pecuária e Abastecimento(Mapa), no controle de animais registrados que apresentam alelos de outras raças, possibilitando a reclassificação desses animais. Adicionalmente, a metodologia proposta poderá ser extendida para toda e qualquer espécie animal de produção.

3 Referências

AGRAWAL, R.; SRIKANT, R. Fast algorithms for mining association rules. In: INTERNATIONAL CONFERENCE ON VERY LARGE DATABASES, 1994, Santiago. **Proceedings**... Santiago: Morgan Kaufmann, 1994. p. 1-32.

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. **Animal Genetics**, Malden, n. 41, n. 5, p. 449-453, Oct. 2010. DOI: 10.1111/j.1365-2052.2010.02100.x.

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

BREIMAN, L. Random forests. **Machine Learning**, Boston, v. 45, n. 1, p. 5-32, Oct. 2001. DOI: 10.1023/A:1010933404324.

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

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, 2000. 78 p.

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

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.

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

FRIEDMAN, J.; HASTIE, T.; TIBSHIRANI, R. Regularization paths for generalized linear models via coordinate descent. **Journal of Statistical Software**, Los Angeles, v. 33, n. 1, p. 1-22, 2010.

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**, New York, v. 92, n. 3, p. 227-237, 2010.

GOUVEIA, J. J. de S. **A utilização da genômica de populações na análise das principais raças de ovinos brasileiras**. 2013. 98 p. Tese (Doutorado) - Universidade Federal do Ceará, Fortaleza.

HAN, J.; KAMBER, M.; PEI, J. **Data mining**: concepts and techniques. 2nd. ed. San Francisco: Morgan Kaufmann, 2011. 529 p.

HASTIE, T.; TIBSHIRANI, R.; FRIEDMAN, J. **The elements of statistical learning**: data mining, inference, and prediction. London: Springer, 2011. 745 p.

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, 2003. p. 17.

JAMES, G.; HASTIE, T.; TIBSHIRANI, R. **An introduction to statistical learning**: with applications in R. London: Springer, 2013. 429 p.

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

KUHN, M. **Caret**: classification and regression training. R package version 5.16-24, 2013.

LIAW, A.; WIENER, M. Classification and regression by random forest. **R News**, Austria, v. 2, n. 3, p. 18-22, 2002.

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 Science**, Amsterdam, v. 120, n. 3, p. 204-212, 2009.

MEGETO, G. A. S.; OLIVEIRA, S. R. de M.; PONTE, E. D.; MEIRA, C. A. A. Árvore de decisão para classificação de ocorrências de ferrugem asiática em lavouras comerciais com base em variáveis meteorológicas. **Engenharia Agrícola**, Jaboticabal, v. 34, n. 3, jun. 2014.

MOKRY, F. B.; HIGA, R. H.; MUDADU, M. A.; LIMA, A. O.; MEIRELLES, S. L. C.; SILVA, M. V. G. B.; CARDOSO, F. F.; OLIVEIRA, M. M. O.; URBINATI, I.; NICIURA, S. C. M.; TULLIO, R. R.; ALENCAR, M. M.; REGITANO, L. C. Genome-wide association study for backfat thickness in Canchim beef cattle using Random Forest approach. **BMC Genetics**, London, v. 14, n. 47, p. 1-11, 2013. DOI:10.1186/1471-2156-14-47.

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

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**, New York, v. 23, n. 1, p. 1-10, 2012.

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

RIDGEWAY, G. **Generalized boosted regression models**. Version 2.1. [S.l.: s.n], 2013. 34 p.

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**, Madison, v. 6, n. 2, p. 1-10, Aug. 2013.

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**, Tokio, v. 82, n. 1, p. 67-72, 2011.

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**, Barking, v. 85, n. 2, p. 285-8, June 2010.

TAN, P.-N.; STEINBACH, M.; KUMAR, V. **Introduction to data mining**. Boston: Addison Wesley, 2006. 769 p.

TIBSHIRANI, R. Regression shrinkage and selection via the lasso, **Statistics in Medicine**, New York, v. 16, p. 385-395, 1997.

WITTEN, I. H.; FRANK, E.; HALL, M. A. **Data mining**: practical machine learning tools and techniques. 3rd. ed. San Francisco: Morgan Kaufmann , 2011. 664 p.