Universidade Federal de São Carlos
Centro de Ciências Exatas e de tecnologia
Departamento de Estatística

Seleção de marcadores SNP: uma aplicação com diferentes metodologias

Mariana Pavan Ióca

Trabalho de Conclusão de Curso

Universidade Federal de São Carlos
Centro de Ciências Exatas e de Tecnologia
Departamento de Estatística

Seleção de marcadores SNP: uma aplicação com diferentes metodologias

Mariana Pavan Ióca
Profª.Drª. Daiane Ap. Zuanetti

Trabalho de Conclusão de Curso a ser apresentado como parte dos requisitos para obtenção do título de Bacharel em Estatística.

São Carlos
8 de Dezembro de 2020

Mariana Pavan Ióca

Seleção de marcadores SNP: uma aplicação com diferentes metodologias

Este exemplar corresponde à redação final do trabalho de conclusão de curso devidamente corrigido e defendido por Mariana Pavan Ióca e aprovado pela banca examinadora.

São Carlos, 2020.

Banca Examinadora

- Prof[a].Dr[a]. Daiane Ap. Zuanetti
- Prof[a].Dr[a]. Maria Silvia de Assis Moura
- Prof.Dr. Ricardo Ferreira

Dedicatória

Dedico este trabalho para Gabriele Oliveira. Nós conseguimos.

Agradecimentos

Primeiramente gostaria de agradecer à Profª Daiane Zuanetti, pela parceria e amizade construídas ao longo desse trabalho. Ao seu apoio, conselhos, horas de reuniões e risadas, muito obrigada.

Aos professores da banca Maria Silvia de Assis Moura e Ricardo Ferreira, pelas correções, dicas e tempo dedicado para me ajudar na construção desse trabalho.

Agradeço à minha família Laura, Sueli e José, pelos meses ouvindo que eu estava fazendo meu TCC, ouvindo sobre o trabalho para me deixar menos ansiosa e pelo apoio. Ao Jack, pelas horas de companhia enquanto eu escrevia e ao seu amor incondicional.

Aos integrantes do PET, pela amizade, aprendizados e suporte. Em especial ao Pedro Ferreira Filho, pelas horas e horas de conversas e conselhos.

Agradeço também à Gabriela Massoni, Lais Sebastiany, Luiz Eduardo Amaral, Mariana Oliveira, Rafaela Marques e Renan Vinícius Rodrigues, pelo apoio imenso e por me acalmarem ao longo dessa jornada.

Resumo

A quantidade e a complexidade dos dados gerados devido ao avanço nas tecnologias de sequenciamento genético fez da análise estatística uma ferramenta essencial para a interpretação correta de resultados. No entanto, ainda não há um consenso sobre quais metodologias são mais adequadas para esses dados. Além disso, os dados genéticos apresentam uma grande quantidade de variáveis (marcadores, genótipos, etc) e poucas observações, logo, a utilização de algumas metodologias estatísticas tornam-se inviáveis. Os objetivos desse trabalho de conclusão de curso são: *(i.)* estudar duas metodologias de seleção de variáveis - Florestas Aleatórias e LASSO, *(ii.)* aplicá-las em dados genéticos para selecionar marcadores SNP (do inglês *Single Nucleotide Polymorphism*) presentes nos indivíduos que caracterizam a presença ou não de uma doença e *(iii.)* comparar suas performances.

Palavras-chave: *SNP, LASSO, Florestas Aleatórias, Seleção de Variáveis.*

Sumário

Lista de Tabelas

Lista de Figuras

Capítulo 1

Introdução

A ciência genética estuda a presença, a variação e a transmissão de características através de gerações. Os primeiros estudos datam de 1860, nos quais, o monge austríaco Gregor Mendel desvendou os princípios da hereditariedade a partir da realização de cruzamentos de linhagens de ervilhas. Entretanto, foi apenas em meados da década de 1950 que Francis Crick, James Watson e Maurice Wilkins descobriram o DNA (ácido desoxirribonucleico) e como as informações genéticas eram armazenadas. O DNA é formado por duas fitas de polinucleotídeos compondo uma dupla hélice que contém as bases nitrogenadas adenina (A), timina (T), citosina (C) e guanina (G), as quais se unem nas duplas AT e CG a partir de uma ponte de hidrogênio.

Com o avanço nas tecnologias de sequenciamento, diminuindo o custo e aumentando a velocidade de obtenção de dados, surgiram os primeiros trabalhos sobre marcadores moleculares (sequências de DNA capazes de apresentar o polimorfismo dos indivíduos em estudo). A primeira estratégia para o estudo de marcadores moleculares acontece no início da década de 1980, com a caracterização de marcadores RFLP (do inglês *Restriction Fragment Length Polymorphism*) em suínos (Chardon *et al.*, 1985) e bovinos (Beckmann *et al.*, 1986).

Os primeiros trabalhos analisam dados usando metodologias limitadas e trabalhosas, que foram evoluídas a partir do desenvolvimento tecnológico e propiciaram o desdobramento de metodologias com maior precisão. Atualmente RFLP não é mais utilizada e foi substituída pela técnica de SNPs (do inglês *Single Nucleotide Polymorphism*). Marcadores SNPs constituem-se pela variação de apenas um nucleotídeo (A,T,C,G) em um determinado gene, podendo ou não ocasionar uma alteração do fenótipo (característica que pode ser observada). Às vezes, um SNP não é a causa de uma doença, todavia, sua identi-

ficação ajuda a estabelecer localizações no genoma de fatores genéticos que contribuem à variabilidade, ou seja, na presença da doença.

Apesar dos avanços tecnológicos, ainda há poucos estudos acerca de dados genéticos, mais especificamente aqueles que estudam marcadores SNPs, pois tal situação apresenta problemas que barram a utilização das ferramentas estatísticas mais tradicionais. Normalmente, encontra-se dados com um número de variáveis superior ao de observações, variáveis muito correlacionadas ou eventos raros. Em busca de superar essa dificuldade, uma das soluções encontradas foi utilizar a seleção de variáveis para que sejam identificados os SNPs mais associados ao fenótipo de interesse. Os métodos mais utilizados são baseados na estimação do modelo de regressão linear simples entre o genótipo de cada SNP e o fenótipo em estudo, nos quais os SNPs mais significantes são escolhidos via testes de hipóteses. Esta abordagem apresenta vantagens, tais como: baixo tempo de processamento computacional, facilidade de uso e de interpretação dos resultados. Entretanto, Zeng *et al.* (2015); Oliveira *et al.* (2015); Feng *et al.* (2012) alertam para as deficiências dessa abordagem que não considera a estrutura de associação entre os SNPs, não permite que a interação entre os efeitos de dois ou mais SNPs sobre o fenótipo seja estimada e normalmente possui baixo poder.

Com o objetivo de superar as deficiências apresentadas, algumas metodologias passaram a ser estudadas, dentre elas, algumas classificadas como metodologias de aprendizagem de máquina. Pode-se destacar: Florestas Aleatórias (Mokry *et al.*, 2013; Oliveira *et al.*, 2015; Breiman, 2001), Componentes Principais (Lewis *et al.*, 2011), análise de frequência alélica (Suekawa *et al.*, 2010; Sasazaki *et al.*, 2011), Algoritmos Genéticos (Goldberg, 1989; Oliveira *et al.*, 2015), métodos de regressão por mínimos quadrados parciais (SPLS - do inglês *Sparse Partial Least Squares*, Chun e Keleş, 2010) e LASSO (*Least Absolute Shrinkage and Selection Operator*) (Park e Casella, 2008; Oliveira *et al.*, 2015).

Os objetivos do presente trabalho, portanto, são estudar e aplicar duas dessas metodologias: Florestas Aleatórias e LASSO em dados genéticos para selecionar marcadores SNPs presentes nos indivíduos que caracterizam a presença ou não de uma doença e comparar as suas performances.

Esse relatório de TCC está organizado como a seguir. No Capítulo 2 é feita uma breve descrição sobre fundamentos genéticos importantes para entender o problema em questão. Os Capítulos 3 e 4 apresentam as metodologias LASSO e Florestas Aleatórias, respectivamente. No Capítulo 5 é feita a descrição do banco de dados que será utilizado

nesse estudo e a análise descritiva dos SNPs utilizados para a simulação da resposta. O Capítulo 6 apresenta os procedimentos para a aplicação do LASSO e das Florestas Aleatórias no banco de dados em estudo, bem como os resultados encontrados.

Capítulo 2

Fundamentos Genéticos

Neste Capítulo apresentaremos alguns fundamentos genéticos essenciais para a compreensão do problema que vamos estudar. Eles serão apresentados em lista para facilitar a compreensão. São eles:

1. *Fenótipos:* características observadas em um indivíduo. Exemplo: cor dos olhos, cor da pele, presença de uma doença.

2. *Genótipo:* constituição genética de um indivíduo.

3. *DNA:* o ácido desoxirribonucleico, é uma molécula presente em todas as células dos seres vivos. No DNA estão contidas todas as informações genéticas do organismo. Ele é composto por uma dupla hélice que contém as bases nitrogenadas adenina (A), timina (T), citosina (C) e guanina (G).

4. *Bases Nitrogenadas:* fazem parte da composição do DNA e podem ser divididas em dois grupos:

   - Bases púricas ou purinas: adenina e guanina;
   - Bases pirimídicas ou pirimidinas: citosina, timina e uracila.

   No DNA humano essas bases se unem a partir de pontes de hidrogênio nos seguintes pares: adenina (A) e timina (T), citosina (C) e guanina (G). O DNA humano, por exemplo, é composto por mais de 3 bilhões dessas bases organizadas em sequência.

5. *Gene:* é uma sequência do DNA responsável por uma característica específica herdada geneticamente, formado por sequência de alelos.

6. *Alelo:* as partes que compõem um gene. Os genes são diferentes partes do DNA que determinam as características do indivíduo, já os alelos são variações específicas do gene que determinam a forma como as características irão se expressar.

    - Dominante: determina uma característica.
    - Recessivo: só se expressa quando em dose dupla, pois na presença de um dominante, ele se torna inativo.

7. *Indivíduo Heterozigoto:* indivíduo cujos alelos para determinada característica são diferentes.

8. *Indivíduo Homozigoto:* indivíduo cujos alelos para determinada característica são iguais.

9. *Cromossomo:* longa sequência de DNA, que contém vários genes.

    - Diploide: são espécies que apresentam fitas duplas de cromossomos, os quais denominamos homólogos por apresentarem, em geral, sequências de DNA iguais, podendo exibir pequenas variações. Via de regra espécies que se reproduzem sexualmente são diploides.

    Na espécie humana, temos 23 pares de cromossomos, sendo 22 pares autossômicos, que definem as características gerais de uma espécie e um par de cromossomos sexuais, responsáveis pela determinação do sexo do indivíduo.

10. *Leis de Mendel:* criadas no século XIX por Gregor Mendel, o objetivo dessas leis é explicar como as características de um organismo são passadas para seus descendentes.

    - 1ª Lei de Mendel: Lei da Segregação dos Fatores. Consiste em: a característica de um indivíduo é determinada pela união de dois genes, cada um proveniente de um "pai", em que se estabelece a relação de dominância. O indivíduo filho herdará de cada pai o gene M (característica dominante) ou m (característica recessiva), podendo ter a característica determinada por MM (dominante), Mm (dominante) ou mm (recessiva).
    - 2ª Lei de Mendel: Lei da Segregação Independente. Consiste em: as características são herdadas de forma independente. Por consequência, temos que as informações presentes em cada cromossomo não interfere nos demais.

11. *Doença complexa:* doença que é associada a diversos genes. Também são consideradas as condições ambientais para o desenvolvimento desse tipo de doença.

12. *Marcadores genéticos:* é uma sequência curta do DNA (podendo chegar ao nível de ser a informação de uma base nitrogenada) com localização conhecida no cromossomo que pode ser usada para diferenciar indivíduos e espécies.

13. *Marcadores SNP:* do inglês *Single Nucleotide Polymorphism*, são mutações em bases nitrogenadas (adenina, timina, citosina e guanina) em um determinado gene que pode ou não ocasionar uma alteração no fenótipo. A mutação mais comum é a troca de bases pertencentes ao mesmo grupo, ou seja, a troca entre purinas (A por G ou G por A) e a troca entre pirimidinas (C por T ou T por C). Também pode ocorrer entre bases de grupos diferentes, mas esses casos são menos comuns. Em espécies não endogâmicas (espécies em que não há a união entre indivíduos geneticamente semelhantes) os SNPs são abundantes (Caetano, 2009).

Capítulo 3

LASSO

Neste Capítulo será apresentada a metodologia LASSO para seleção de variáveis e estimação de parâmetros. Na Seção 3.1 será explicada a metodologia e o parâmetro de penalização λ que nela é utilizada. Na Seção 3.2 serão discutidas formas de se obter o valor de λ de modo que sejam obtidos os melhores estimadores a partir do LASSO. A Seção 3.3 traz uma discussão sobre alguns problemas encontrados na metodologia. Apresentamos os Modelos Lineares Generalizados na Seção 3.4 como uma solução aos problemas encontrados no LASSO. E, finalmente, a Seção 3.5 aborda a implementação do LASSO no *software* R.

3.1 Metodologia

O LASSO (*Least Absolute Shrinkage and Selection Operator*), criado por Tibshirani (1996), é uma metodologia de seleção de variáveis, cujo objetivo é encontrar um estimador de Mínimos Quadrados mais parcimonioso. Considere o modelo linear,

$$y_k = \sum_{i=0}^{d} (\beta_i x_{ki}) + \epsilon_k, \tag{3.1}$$

$k = 1, 2, \ldots, N$ com $d \in \mathbb{N}$, em que y_k é o valor real da variável resposta do k-ésimo indivíduo, $\beta_i s$ são os parâmetros reais desconhecidos, x_{ki} é o valor real da i-ésima covariável para o k-ésimo indivíduo e ϵ_k é o erro aleatório considerado para o k-ésimo indivíduo.

De acordo com Izbicki e Santos (2018) a ideia principal do LASSO é acrescentar a restrição $\sum_{i=1}^{d} |\beta_i| \leq c$, em que $c = c(\lambda)$, à fórmula tradicional de mínimos quadrados utilizada para estimar os valores dos $\beta_i' s$. Com esse acréscimo as estimativas de alguns $\beta_i' s$

serão aproximadamente zero e, logo, o método seleciona para o modelo aquelas covariáveis que apresentam $\beta_i \neq 0, i = 1, ..., d$. Chen e Gopalakrishnan (1998) afirmam que o LASSO também pode ser chamado de uma metodologia de busca de base (do inglês *basis pursuit*).

Tal metodologia também pode ser utilizada para definir a complexidade do modelo: quanto maior o número de covariáveis presentes no modelo, mais complexo ele é. Geralmente menores valores de c representam modelos com menor grau de complexidade, pois usualmente levam à seleção de menores quantidades de covariáveis. Assim, a metodologia LASSO procura por:

$$\hat{\boldsymbol{\beta}} = arg \min_{\boldsymbol{\beta} \in \mathbb{R}^{d+1}} \sum_{k=1}^{n} \left(y_k - \beta_0 - \sum_{i=1}^{d} \beta_i x_{ki} \right)^2 + \lambda \sum_{i=1}^{d} |\beta_i| \,, \tag{3.2}$$

em que $\lambda \geq 0$ é um parâmetro de regularização. É válido ressaltar que a penalidade utilizada no LASSO é de rápida implementação e frequentemente apresenta boa performance computacional. De acordo com Friedman *et al.* (2001), se $\sum_{i=1}^{d} \left| \hat{\beta}_i \right| \leq c(\lambda)$, em que $\hat{\beta}_i, i = 1, ..., d$, são as estimativas de Mínimos Quadrados para o modelo linear (3.2), então as estimativas dos coeficientes do LASSO são os $\hat{\beta}_i$'s.

O LASSO pode ser estendido para uma grande variedade de funções objetivos que não apenas uma regressão linear, tais como para Modelos Lineares Generalizados, que serão um pouco mais detalhados na Seção 3.4. Na aplicação desse trabalho utilizaremos a versão do LASSO que prediz uma probabilidade de sucesso (como uma Regressão Logística) e não um valor na reta real.

3.2 A Escolha do λ

Na busca por um λ suficientemente pequeno, as estimações de alguns coeficientes são zero devida à natureza da restrição utilizada no LASSO. De acordo com Friedman *et al.* (2001) se $t_0 = \sum_{i=1}^{d} \left| \hat{\beta}_i \right| \leq c(\lambda)$, em que $\hat{\beta}_i$s são as estimativas de Mínimos Quadrados para o modelo linear (3.2) então as estimativas do LASSO são $\hat{\beta}_i$.

Note que para cada valor de λ obtém-se um conjunto diferente de $\hat{\beta}_i$ e se $\lambda = 0$ o LASSO torna-se idêntico ao estimador de Mínimos Quadrados, ou seja, com todos os coeficientes diferentes de zero. A principal vantagem dessa metodologia é que nos permite escolher o valor de λ, nos fornecendo um modelo com boa predição e também que selecione só as covariávies mais significativas. No caso do Mínimos Quadrados sem a restrição, todas as

covaríaveis seriam colocadas no modelo. Pode-se destacar duas características positivas a respeito do LASSO:

1. Possibilita estimar o modelo para muitos valores de λ simultaneamente e;

2. Devido às diversas estimativas iguais a zero, o modelo resultante é de fácil interpretação e através dela identificamos as variáveis significativas (Izbicki e Santos, 2018).

3.2.1 Validação Cruzada

Para que possamos encontrar o melhor modelo que o LASSO pode estimar para o nosso banco de dados é fundamental que a escolha do λ seja feita da melhor maneira. Em geral, segundo Izbicki e Santos (2018), a escolha de λ é feita a partir da validação cruzada de diferentes modelos estimados. Vale ressaltar que o LASSO não realiza a validação cruzada, pois nessa metodologia é encontrado um modelo para o λ fornecido a ele.

A validação cruzada é uma técnica que avalia como os resultados de uma análise estatística serão generalizados para um conjunto de dados independentes. O método mais utilizado de validação cruzada é o *k-fold*. Hastie *et al.* (2017) define o algoritmo da validação cruzada da seguinte forma:

- De forma aleatória, o banco de dados é dividido em k subconjuntos aproximadamente do mesmo tamanho e mutuamente exclusivos. Para efeito de ilustração vamos supor um banco de dados com 500 observações e $k = 5$, temos então a divisão nos subconjuntos a, b, c, d e e dada na Tabela 3.1:

Tabela 3.1: Exemplo de divisão dos dados com $k = 5$ para validação cruzada.

a	b	c	d	e
n =100	n =100	n =100	n =100	n =100

- Em seguida é feita uma combinação de 4 desses 5 grupos e estimados modelos para diversos λ. Em nosso exemplo, vamos supor $\lambda = 1$, 2 e 3. O modelo será estimado com os subconjuntos a, b, c e d, para ilustração, e será testado na e - ésima parte.

Tabela 3.2: Exemplo de combinação dos dados com $k = 5$ para estimação do modelo na validação cruzada.

a	b	c	d	e
n =100	n =100	n =100	n =100	n =100

Na Tabela 3.2 os elementos em azul compõe a parcela do banco de dados a partir do qual o modelo será estimado considerando $\lambda = 1, 2, 3$. Obtemos assim, 3 modelos para essa combinação de dados. Registramos o Erro de Predição (EP) de cada modelo obtido que é calculado sobre o subconjunto e, não utilizado na estimação do modelo. Nesse caso o EP é definido como sendo a soma dos resíduos ao quadrado da base e.

O mesmo é feito para as demais combinações de divisões do banco de dados. Ou seja, em nosso exemplo teremos 3 modelos para cada uma das seguintes combinações da Tabela 3.3.

Tabela 3.3: Combinações para $k = 5$ do banco de dados em validação cruzada.

Treinamento	Teste
$b + c + d + e$	a
$a + c + d + e$	b
$a + b + d + e$	c
$a + b + c + e$	d
$a + b + c + d$	e

Com os EP's de cada modelo em mãos, é feita uma média desses valores para cada λ. Aquele que apresentar menor média de EP é escolhido como o λ que nos dará o melhor modelo.

3.3 Problemas Encontrados

Durante os estudos da metodologia foram observados alguns problemas. Como se sabe de estudos anteriores, as estimativas dos $\beta_i's$ são viciadas e elas podem mudar significativamente dependendo da amostra treino e teste utilizada na estimação, chegando ao ponto de $\beta_i's$ que em uma amostra tiveram estimativa próxima de zero, em outra apresentar um alto valor.

Uma alternativa para resolver esse problema é a utilização de Modelos Lineares Generalizados (MLGen) para o cálculo das estimativas dos $\beta_i's$. Nesse caso, o LASSO seria

utilizado para a seleção de variáveis e o modelo final seria obtido a partir do ajuste de um MLGen.

3.4 Modelos Lineares Generalizados

Os Modelos Lineares Generalizados (MLGen, do inglês *Generalized linear models - GLM*) foram propostos por Nelder e Wedderburn (1972). Até sua proposição tentava-se ajustar modelos normais lineares para quase qualquer tipo de fenômeno aleatório, mas em inúmeros casos era necessário realizar transformações nas variáveis para que tal modelo fosse considerado adequado.

Com as limitações do modelo normal, os MLGen foram criados com base na ideia de abrir um leque de opções para a distribuição da variável resposta (Paula, 2004), pois esses modelos permitem que a distribuição de $Y|X$ pertença à família exponencial linear e não apenas à família normal. Dentre as distribuições que pertencem a essa família podemos citar: Poisson, Normal Inversa, Normal, Gama e Binomial, da qual falaremos mais adiante.

Um MLGen possui três componentes (Paula, 2004):

- Componente aleatório: conjunto de variáveis aleatórias independentes com distribuição pertencente à família exponencial linear;
- Componente sistemático: que engloba o vetor de parâmetros $\boldsymbol{\beta}$ e as covariáveis $\mathbf{x}$;
- Função de ligação (F.L.): uma função estritamente monótona e duplamente diferencial que relaciona o componente aleatório ao componente sistemático.

Algumas distribuições apresentam uma função de ligação canônica. Seu uso apresenta duas vantagens:

1. Garante a unicidade do Estimador de Máxima Verossimilhança de $\boldsymbol{\beta}$ e
2. Simplifica o algoritmo de estimação de $\boldsymbol{\beta}$.

Tais vantagens tornam o ajuste via MLGen uma alternativa para solucionar o problema que encontramos no LASSO, o qual não apresenta estimativa única para os β_i. Como o banco de dados utilizado nesse trabalho apresenta variável resposta binária, a distribuição utilizada para o ajuste do MLGen será a Binomial com $n = 1$.

3.4.1 Distribuição Binomial

A distribuição Binomial é utilizada no ajuste de MLGen quando o interesse é a probabilidade de ocorrência de um dos valores de uma variável binária em relação ao outro. No caso do presente trabalho temos que o número de ensaios associado a cada indivíduo da amostra é 1, o que caracteriza um caso especial da distribuição Binomial: uma Bernoulli.

A função de ligação canônica da distribuição Binomial é a logito, dada por:

$$g(p_k) = log\left(\frac{p_k}{1-p_k}\right) = \beta_0 + \sum_{i=1}^{d} \beta_i x_{ki},$$

em que p_k é a probabilidade de sucesso do k-ésimo indivíduo. Além das vantagens por ser F.L. canônica, a função usada para respostas binárias leva a um modelo com parâmetros interpretáveis e valores ajustados que pertencem ao espaço paramétrico de p. Um MLGen com resposta binária e F.L. logito é denominado Regressão Logística.

Como já dito no final da Seção 3.1, o LASSO também é muito utilizado para selecionar variáveis, sendo a estimação dos coeficientes de regressão feita a partir de MLGen. Seguiremos esse método para a aplicação da metodologia LASSO nesse estudo.

3.5 LASSO no *software* R

O LASSO já está implementado em pacotes do R, sendo um deles o *glmnet* (Friedman *et al.*, 2010). Criado por Jerome Friedman, Trevor Hastie e Rob Tibshirani, nesse pacote é feita a estimação não somente dos β_i como também do melhor λ a partir da validação cruzada.

Capítulo 4

Florestas Aleatórias

Neste Capítulo será apresentada a metodologia Florestas Aleatórias. Na Seção 4.1 será discutido um breve histórico de metodologias que serviram de base para a formulação das Florestas Aleatórias: Árvores de Decisão, *Boosting* e *Bagging*, pois uma das motivações para a criação das Florestas Aleatórias foi solucionar alguns problemas encontrados nessas metodologias. Na Seção 4.2 será apresentada a metodologia Florestas Aleatórias, dando enfoque para sua utilização em regressão. Na Seção 4.3 será discutida a determinação da quantidade m das d covariáveis que devem ser utilizadas na estimação do modelo. Na Seção 4.4 serão apresentadas as amostras *Out-of-Bag*, que são utilizadas para a validação cruzada do modelo e estimação dos erros. A Seção 4.5 aborda a implementação dessa metodologia no *software* R.

4.1 Histórico

As Árvores de Decisão, *Boosting* e *Bagging* são metodologias essenciais para o desenvolvimento das Florestas Aleatórias. Discutiremos brevemente a respeito dessas três técnicas para melhor compreensão da metodologia que utilizaremos nesse estudo.

4.1.1 Árvores de Decisão

Árvore de Decisão ou de regressão é uma metodologia de aprendizado de máquina supervisionado. De forma simples, uma árvore é construída a partir da significância de cada covariável e um nó é criado a partir de cada covariável significativa. Assim, procuramos a covariável preditora que apresenta maior efeito na variável resposta e a partir dela cria-se um nó. Por exemplo, se a idade do paciente é superior a 50 anos e

a probabilidade do mesmo desenvolver problemas cardíacos aumenta consideravelmente a partir dessa idade, a variável idade se transforma em um nó para predizer o risco de desenvolver problemas cardíacos e temos um ramo para pacientes maiores de 50 anos e um para menores, como pode ser observado na Figura 4.1.

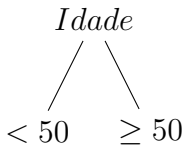

Figura 4.1: Árvore de Decisão com o primeiro nó e seus ramos.

Em seguida procura-se identificar qual desses dois ramos deve ser particionado a partir de uma nova covariável. Suponha, por exemplo que pacientes com idade superior a 50 anos e com antecedente familiar possuam maior probabilidade de desenvolver problemas cardíacos do que pacientes sem histórico familiar e pacientes com menos de 50 anos. O segundo ramo seria criado e pode ser observado na Figura 4.2.

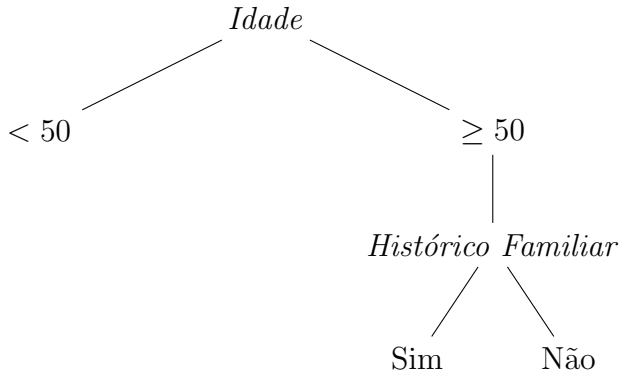

Figura 4.2: Árvore de Decisão com os dois primeiros nós e seus ramos.

Segundo Izbicki e Santos (2018) a Árvore de Decisão é obtida a partir de dois passos:

- ***I.*** *Criação de uma árvore completa*: procura-se encontrar as partições nas quais a variável resposta (Y) apareça de forma homogênea nas folhas.

- ***II.*** *Poda da árvore:* cada nó é retirado, um por vez, e observa-se o efeito causado por essa retirada no Erro de Predição para os dados de validação. A partir desses valores, decide-se quais nós permanecerão na árvore. Essa etapa é essencial para evitar o super ajuste do modelo aos dados de treinamento e para a redução da complexidade da árvore.

4.1.2 *Bagging* e *Boosting*

O *Bagging* (Breiman, 1996) é uma técnica cujo objetivo é reduzir a variância apresentada em um modelo de previsão. Essa metodologia consiste em ajustar a mesma árvore de regressão para diversas amostras *bootstrap*, as quais são independentes entre si. Ao final, cada árvore tem o peso de um voto e a decisão é tomada pela maioria simples destes, ou seja, 50% + 1 do total de votos. Para dados com grande variabilidade e baixo viés, o *Bagging* é uma metodologia que normalmente apresenta bons resultados, ou seja, boas previsões individuais.

Diferentemente do *Bagging*, no *Boosting* (Schapire *et al.*, 1998) o voto final é dado de forma ponderada, pois conforme são estimadas, as árvores vão dando maior peso para as observações que foram anteriormente previstas de forma errônea. Outra diferença entre as metodologias é: enquanto o *Bagging* preocupa-se com manter a variância pequena da predição, no *Boosting* o foco está na redução do viés com melhores predições, ou seja, a técnica busca reduzi-lo. Em geral, a performance do *Boosting* é superior ao do *Bagging*, logo, na maioria dos casos opta-se pelo *Boosting* (Hastie *et al.*, 2017).

4.2 Florestas Aleatórias

A metodologia de Florestas Aleatórias (Breiman, 2001) é a combinação de diversas Árvores de Decisão diferentes e não correlacionadas para a tomada de decisão. Sua performance é similar ao do *Boosting* (Hastie *et al.*, 2017). Considerando um vetor aleatório Θ_ℓ gerado independentemente dos vetores aleatórios anteriores $(\Theta_1, ..., \Theta_{\ell-1})$, mas com mesma distribuição e $\mathbf{x} = (x_1, ..., x_d) \in \mathbb{R}^d$ é um vetor com as d covariáveis utilizadas na estimação das Florestas Aleatórias para um específico indivíduo, Breiman (2001) define as Florestas Aleatórias da seguinte forma:

Definição 1.1: Uma Floresta Aleatória é um classificador que consiste em uma coleção de classificadores estruturados em árvore $\{h(\mathbf{x}, \Theta_\ell), \ell = 1, ...\eta\}$ em que Θ_ℓ são vetores aleatórios independentes e identicamente distribuídos e cada árvore emite um voto único na categoria mais frequente para o valor de $\mathbf{x}$ observado.

A ideia principal das Florestas Aleatórias é reduzir a correlação entre as árvores sem aumentar muito a variância da predição. Ou seja, a metodologia procura encontrar um equilíbrio entre o *Boosting* e o *Bagging*. Podemos definir o algoritmo para o crescimento das florestas da seguinte forma:

1. Montar η amostras *bootstrap* dos dados originais;

2. Para cada uma das η amostras, crescer uma árvore (T_ℓ), em que $\ell = 1, ..., \eta$, seguindo os seguintes passos para cada árvore:
   i. Selecionar aleatoriamente m das d covariáveis;
   ii. Escolher a covariável mais significante dentre as m sorteadas;
   iii. Dividir o nó em dois "nós filhos";

   - Realizar passos ii. e iii. até crescer a árvore toda e a poda ser realizada;

3. Realizar a predição.

No caso da regressão via Florestas Aleatórias, a predição final é feita a partir da média das predições, ou seja, a partir da Equação (4.1) que representa a predição para a Floresta composta por η árvores:

$$\hat{f}^{\eta} = \frac{1}{\eta} \sum_{l=1}^{\eta} T_\ell(\mathbf{x}) \tag{4.1}$$

em que $T_\ell(\mathbf{x})$ é o valor predito para o valor $\mathbf{x}$ na ℓ-ésima árvore estimada.

4.3 Determinação de m

Breiman e sua colaboradora Adele Cutler recomendam que para a regressão via Florestas Aleatórias o sugerido da quantidade de variáveis a ser usada em cada árvore (m) é $m = \frac{d}{3}$ e pelo menos 5 nós. Na prática os melhores valores para esses parâmetros podem ser definidos a partir dos dados que estão sendo analisados, ou seja, podem ser estimados.

4.4 Amostras *Out-of-Bag* e Validação Cruzada

Uma das características das Florestas Aleatórias é o uso de amostras *Out-of-Bag* (OOB), uma metodologia de validação cruzada, para crescer e podar as árvores e estimar a melhor floresta. Hastie *et al.* (2017) define as OOB's da seguinte forma:

1. Para cada observação y_k é feita a predição de seu valor utilizando apenas as árvores estimadas a partir das amostras *bootstrap* que não contém y_k;

2. Calcula-se o erro total das predições feitas a partir das amostras OOB's.

Para efeito de ilustração, vamos utilizar um exemplo simples. Suponha que construímos uma floresta com 5 árvores, ou seja, 5 amostras *bootstrap* (a, b, c, d, e) em um conjunto de 5 observações $(1, 2, 3, 4, 5)$. As amostras são dadas na Tabela 4.1.

Tabela 4.1: Exemplo de amostras *boostrap* para OOB's.

Amostra	Observações Utilizadas
a	1,2,3,1,2
b	1,3,5,1,3
c	2,4,5,2,4
d	2,3,4,2,3
e	1,4,5,1,4

A partir de cada amostra presente na Tabela 4.1 é encontrada uma Árvore de Decisão. As predições via OOB's são feitas como mostrado na Tabela 4.2.

Tabela 4.2: Exemplo de estimação via amostras OOB.

Obs. estimada	Árvore referente as amostras utilizadas na predição
1	c, d
2	b, e
3	c, e
4	a, b
5	a, d

Podemos observar na Tabela 4.2 que a predição para cada observação será feita a partir de florestas compostas apenas por árvores que não contêm a respectiva observação na sua estimação. Com as estimativas em mãos, calculamos o erro da predição de forma similar ao do método *k-fold* explicado na Seção 3.2.1. Na estimação via Florestas Aleatórias a fase de teste é encerrada quando o erro das OOB's é estabilizado.

4.4.1 Importância das Variáveis

As amostras OOB's são utilizadas para o cálculo da importância das variáveis, a qual é dada pela mudança no Erro de Predição quando a i-ésima covariável é excluídas das árvores, mantendo-se as demais constantes. Assim, para encontrar a importância de cada variável seguimos o seguinte algoritmo:

1. Calcula-se o Erro de Predição das OOB's das η árvores;

2. Para cada uma das m variáveis de cada árvore retira-se os nós relativos a ela e calcula-se o novo erro;

3. Calcula-se a Importância Parcial (IP) da i-ésima variável na ℓ-ésima Árvore ($IP_{i\ell}$) como sendo a diferença relativa (em porcentagem) entre o Erro da ℓ-ésima Árvore sem a i-ésima covariável e o Erro da ℓ-ésima Árvore completa, dada pela Equação (4.2) como:
$$IP_{il} = \left(\frac{EP_i - EP_c}{EP_c}\right) * 100; \tag{4.2}$$

4. A importância da i-ésima variável é a média das η's Importâncias Parciais, ou seja:
$$I_i = \frac{1}{\eta}\sum_{\ell=1}^{\eta} IP_{i\ell}.$$

4.5 Florestas Aleatórias no *software* R

As Florestas Aleatórias estão implementadas no R no pacote *randomForest* (Liaw e Wiener, 2002), o qual implementa a metodologia criada por Breiman. Os valores padrões dos parâmetros presentes no pacote são as recomendações do autor. A partir dele podemos encontrar Florestas Aleatórias para classificação e regressão, sendo o último caso o que usaremos nesse estudo.

Capítulo 5

Banco de Dados GAW17

Nesse Capítulo será apresentado o banco de dados escolhido para o atual estudo (Seção 5.1). Na Seção 5.2 apresentamos uma breve análise descritiva das variáveis destacadas pelo autor do banco como importantes na simulação dos dados.

5.1 Sobre o Banco de Dados

O banco de dados escolhido para esse estudo se chama *Genetic Analysis Workshop 17* (GAW17) e contém dados simulados e reais de 697 indivíduos sem parentesco, sendo 327 homens e 370 mulheres. A simulação de uma doença complexa e fatores de risco foi realizada baseada nos dados reais contidos no *1000 Genomes Project*. Os dados *1000 Genomes Project* são projetados para pesquisar a variação genética em vários grupos de populações humanas, sendo elas: Europa, Leste Asiático, do sul da Ásia, África Ocidental e dos índios americanos. Nessa simulação foram obtidos 24487 SNPs.

A simulação foi feita para uma doença comum e complexa que apresenta uma prevalência de 30% na população. Em conjunto com a doença, foram simulados três outros fenótipos quantitativos contínuos: Q1, Q2 e Q4 que não serão explorados nesse trabalho, além do estado de tabagismo. Os marcadores simulados para a composição do banco de dados são autossômicos, ou seja, não há marcadores presentes nos cromossomos sexuais. Para mais detalhes sobre a simulação desses dados ver Almasy *et al.* (2011).

Originalmente as informações contidas no banco de dados estavam com as bases nitrogenadas A,T,C,G em 16 pares possíveis: A/A, T/T, C/C, G/G, A/C, A/G, A/T, C/A, C/G, C/T, G/A, G/C, G/T, T/A, T/C, T/G. Para esse estudo classificamos as observações da seguinte forma:

- A/A ou T/T: homozigoto dominante = 1;
- C/C ou G/G: homozigoto recessivo = -1; e
- A/C, A/G, A/T, C/A, C/G, C/T, G/A, G/C, G/T, T/A, T/C, T/G: heterozigoto = 0.

Em nosso banco de dados temos a informação de 24487 SNPs divididos em 22 cromossomos.

5.2 Análise Descritiva

Em Almasy *et al.* (2011) temos o destaque para 51 SNPs, os quais foram usados pelo autor na determinação da presença ou não da doença nos indivíduos em estudo. Eles estão distribuídos da seguinte forma: 30 SNPs no cromossomo 1, 3 no cromossomo 2, 5 no cromossomo 8, 6 no cromossomo 14, 1 no cromossomo 16 e 2 nos cromossomos 17,18 e 19. Na Tabela 5.1 podemos observar a quantidade de indivíduos em cada categoria ($-1, 0$ ou 1) para cada um desses 51 SNPs destacados.

Podemos observar na Tabela 5.1 que aproximadamente 30% da base de dados apresenta a doença em estudo, portanto não estamos lidando com uma doença rara. Também notamos que 30 dos SNPs destacados em Almasy *et al.* (2011) apresentam apenas uma observação em uma categoria diferente dos demais 696. Esses resultados mostram que os SNPs que determinam a doença são incomuns, o que faz do nosso estudo um caso de pouquíssima variabilidade nas covariáveis e, consequentemente, difícil seleção destas como variáveis significativas.

Tabela 5.1: Análise Descritiva dos indivíduos nos SNPs destacados, em que o número logo após a letra C no nome do SNP identifica de qual cromossomo ele é proveniente.

cromossomo 1				**cromossomo 2**			
SNP	*-1*	*0*	*1*	*SNP*	*-1*	*0*	*1*
C1S9391	0	1	696	C2S2286	696	1	0
C1S9423	696	1	0	C2S2288	693	4	0
C1S9432	683	13	1	C2S2307	0	1	696
C1S9445	696	1	0	**cromossomo 8**			
C1S9446	696	1	0	C8S4825	696	1	0
C1S9449	696	1	0	C8S4839	696	1	0
C1S9455	693	4	0	C8S886	696	1	0
C1S9457	696	1	0	C8S900	695	2	0
C1S7061	689	7	1	C8S909	695	2	0
C1S11396	696	1	0	**cromossomo 14**			
C1S3181	696	1	0	C14S1381	696	1	0
C1S3182	696	1	0	C14S1382	0	5	692
C1S5748	0	1	696	C14S3630	0	1	696
C1S9164	695	2	0	C14S3695	696	1	0
C1S9165	0	1	696	C14S3704	0	5	692
C1S9172	691	6	0	C14S3706	0	246	451
C1S9173	0	2	695	**cromossomo 16**			
C1S9174	696	1	0	C16S1894	0	1	696
C1S9189	688	9	0	**cromossomo 17**			
C1S9200	696	1	0	C17S4578	39	154	504
C1S9222	0	1	696	C17S4581	0	1	696
C1S9250	695	2	0	**cromossomo 18**			
C1S9266	693	4	0	C18S2475	696	1	0
C1S9267	694	3	0	C18S2492	0	24	673
C1S9306	696	1	0	**cromossomo 19**			
C1S9320	696	1	0	C19S4929	695	2	0
C1S9333	696	1	0	C19S4937	695	2	0
C1S9346	696	1	0	**Doença**			
C1S9373	696	1	0	*0*		*1*	
C1S2919	696	1	0	488		209	

A principal consequência de termos a maioria dos SNPs importantes sendo raros, é o aumento da complexidade na seleção de variáveis que realmente impactam e determinam a propensão à doença. Em nosso estudo, a presença de cada um dos 51 SNPs destacados em Almasy *et al.* (2011) representa um aumento na probabilidade do indivíduo apresentar a doença. Sendo assim, uma pessoa que tem os 51 SNPs é a que apresenta maior probabilidade de ter a doença em estudo. Como temos casos raros, não temos uma quantidade alta de indivíduos que apresentam um grande número dos 51 SNPs destacados, logo, teremos probabilidades não muito elevadas para indivíduos doentes e não doentes, o que aumenta a chance de erro.

Uma forma de facilitar a identificação de covariáveis raras seria aumentar a amostra. Todavia, por mais que recentemente estudos genéticos vêm ganhando expressividade, os SNPs ainda continuarão a ter baixa frequência. Além disso, encontramos mais SNPs dentre os 24487 totais que apresentam características semelhantes a esses 51, ou seja, que apresentam 1 caso em uma categoria e 696 em outra, o que poderá causar a seleção de variáveis diferentes nas nossas metodologias.

Outro fator de complexidade desses dados é que para a modelagem dividimos a base entre Treino (contendo 70% das observações) e Teste (contendo 30% das observações). Sendo assim, os SNPs que apresentam apenas 1 observação diferente serão complexos de serem avaliados e modelados, pois em alguns casos essa única observação estará no treino e em outros na validação. Como consequência, a seleção ou não dos SNPs poderá ser afetada, bem como o aumento do Erro de Predição.

Capítulo 6

Resultados

Neste Capítulo iremos apresentar os resultados obtidos com o ajuste de modelos a partir das metodologias LASSO e Florestas Aleatórias. Na Seção 6.1 comentaremos sobre o procedimento padronizado feito para o ajuste dos modelos. A Seção 6.2 apresenta os resultados sobre a estabilidade das metodologias na seleção das variáveis em diferentes amostras. Na Seção 6.3 comentaremos sobre os resultados obtidos aplicando o LASSO e sua combinação com MLGen. As Seções 6.4 e 6.5 contêm os resultados obtidos via Florestas Aleatórias e de sua combinação com MLGen, respectivamente. Por último, na Seção 6.6 temos a comparação entre os desempenhos dos modelos encontrados. Os códigos utilizados para a análise estão disponíveis em: https://github.com/Mariana3112/TCC.git

6.1 Procedimento

De acordo com a 2ª Lei de Mendel, as informações contidas em cada par de cromossomos são independentes das demais. Assim, para a seleção dos SNPs que influenciam a presença da doença as metodologias foram aplicadas separadamente em cada um dos 22 cromossomos autossômicos.

Inicialmente a base foi dividida aleatoriamente em duas partes: a primeira com 70% da base (489 observações) que chamaremos de Treino e a segunda com os demais 30% da base (208 observações), que chamaremos de Teste. Para a estimação dos modelos, inclusive determinação do parâmetro de regularização do LASSO (λ), utilizaremos apenas a base de Treino, em seguida, os modelos obtidos serão aplicados na base de Teste. Esse método possibilita a comparação do desempenho dos modelos em uma base diferente daquela utilizada em sua construção, nos mostrando como eles se comportariam em dados

completamente independentes.

No Capítulo 3 vimos que o LASSO pode selecionar variáveis diferentes e apresentar distintas e viciadas estimativas para os β_i, dependendo da divisão da base Treino para realização da validação cruzada. Levando em consideração esses fatos e que um dos objetivos do presente trabalho é comparar as duas metodologias, optamos pela realização de 21 ajustes diferentes (com sementes distintas) de cada uma delas. Assim, poderemos compará-las em três aspectos:

1. Erro na classificação das observações;

2. Precisão na seleção dos SNPs destacados em Almasy *et al.* (2011);

3. Estabilidade na seleção das covariáveis.

Vimos nos Capítulos 3 e 4 que ambas as metodologias utilizam alguma forma de validação cruzada para a estimação completa dos modelos, as quais separam a base de Treino em duas: Treino Efetivo e Validação. Em cada uma das 21 vezes que ajustamos os modelos, alteramos a semente que determina como essa divisão será feita. Foram registradas as covariáveis selecionadas em cada rodada de cada um dos 22 cromossomos em ambas as metodologias. Para o LASSO, registramos todas os SNPs que apresentam coeficiente de regressão diferente de zero e para as Florestas Aleatórias os 30 SNPs mais importantes.

6.2 Estabilidade na Seleção de Variáveis

Notamos uma variação na seleção das covariáveis em ambas as metodologias entre os 21 ajustes realizados, sendo esta mais acentuada no LASSO. Por essa razão, para finalmente selecionarmos os SNPs importantes em cada metodologia, decidimos analisar a frequência com que cada SNP apareceu nos 21 ajustes. Para cada ajuste feito, foram registradas as covariáveis selecionadas pelo LASSO e as 30 covariáveis com maior importância nas Florestas Aleatórias. Ao final dos 21 ajustes, contabilizamos em quantos deles cada um dos SNPs foi selecionado, em seguida, montamos alguns cenários de corte. Esses cenários são a quantidade de covariáveis que foram selecionadas em pelo menos r, em que $r = 1, ..., 21$, ajustes. Os resultados obtidos podem ser observados na Tabela 6.1.

Tabela 6.1: Número de SNPs que apareceram em cada cenários.

Cenários	**LASSO**	**Florestas Aleatórias**
21	0	123
≥17	0	257
≥14	1	337
≥12	3	401
≥11	5	444
≥7	66	654
≥1	961	2935

Analisando a Tabela 6.1 temos que se definirmos um ponto de corte para a seleção das variáveis importantes como sendo de aparições em pelo menos 17 rodadas (80%), o LASSO não seleciona nenhuma covariável e no outro extremo, aparições em pelo menos um dos ajustes, ambas as metodologias selecionam muitos SNPs. Tais resultados deixam explícito que, para o conjunto de dados utilizado nesse trabalho, o LASSO apresenta grande instabilidade na seleção de covariáveis e que as Florestas Aleatórias não são constantes, porém são mais estáveis que o LASSO.

Considerando os valores apresentados na Tabela 6.1 e que uma solução para a estimação dos coeficientes de regressão no LASSO seria a combinação dele com Modelos Lineares Generalizados, nos quais precisamos ter um número de covariáveis inferior ao de observações do Treino Efetivo do modelo, optamos pelo cenário de corte em 11 aparições. Ou seja, consideramos que as metodologias selecionaram como SNPs importantes os que apareceram em pelo menos 11 dos 21 ajustes.

6.3 LASSO

Na Seção 6.2 vimos que 5 covariáveis foram selecionadas em pelo menos 11 ajustes do LASSO. Na Tabela 6.2 podemos ver quais são os SNPs selecionados e em quais categorias as observações estão, tanto na base toda, quanto no Treino.

Tabela 6.2: Análise Descritiva dos SNPs selecionados pelo LASSO, em que o número logo após a letra C no nome do SNP identifica de qual cromossomo ele é proveniente.

	Base Completa			**Treino**		
SNPs	*-1*	*0*	*1*	*-1*	*0*	*1*
C3S5389	685	12	0	482	7	0
C3S5742	683	12	2	479	9	1
C3S4611	693	4	0	486	3	0
C15S774	0	4	693	0	4	485
C18S2320	693	4	0	485	4	0

Primeiramente temos que apenas 1 dos SNPs selecionados está em algum dos cromossomos que contém SNPs importantes na determinação da doença: C18S2320, presente no cromossomo 18. Porém, ele não está presente na lista apresentada na Tabela 5.1 e o LASSO não obteve sucesso na seleção de nenhuma das 51 covariáveis citadas em Almasy *et al.* (2011).

Destacamos também que as observações heterozigóticas dos SNPs C15S774 e C18S2320 estão todas na base Treino e, consequentemente, nenhuma na base Teste. Essas covariáveis se enquadram, portanto, em um caso complexo de avaliação, pois sua raridade dificulta a modelagem no sentido que sua real importância não será confirmada na base de dados independente.

Apesar de não selecionar nenhuma das covariáveis esperadas, é possível notar que o LASSO conseguiu identificar SNPs semelhantes aqueles usados na simulação. Observando as Tabelas 5.1 e 6.2 vemos que C3S5389 apresenta características semelhantes das observações a C1S9189, porém, na Tabela 6.3 vemos que os indivíduos que apresentam 0 nesses SNPs são diferentes, logo, um não poderia substituir o outro.

Tabela 6.3: Comparação entre os SNPs C1S9189 e C3S5389.

	C1S9189		
		-1	0
C3S5389	-1	676	9
	0	12	0

O mesmo acontece com o SNP C3S5742, que poderia ser visto como um substituto ao C1S9432, mas não apresenta as mesmas características em todos os indivíduos da base,

como podemos observar na Tabela 6.4. O C15S774 se assemelha a C14S1382 e C14S3704, mas como vemos na Tabela 6.5 também não apresentam características iguais em toda a amostra.

Tabela 6.4: Comparação entre os SNPs C1S9432 e C3S5742.

	C1S9432			
		-1	0	1
C3S5742	-1	669	13	1
	0	12	0	0
	1	2	0	0

Tabela 6.5: Comparação entre os SNPs C15S774, C14S1382 e C14S3704.

	C14S1382			C14S3704	
		0	1	0	1
C15S774	0	0	4	0	4
	1	5	688	5	688

O SNP C18S2320 é um caso diferente dos demais pois encontramos dois motivos para sua seleção. Assim como as outras covariáveis selecionadas ele apresenta características semelhantes a um SNP importante de outro cromossomo, nesse caso, tanto ele quanto C3S4611 poderiam ter sido selecionados no lugar de C1S9455, C1S9266 ou C2S2288, mas não é o que ocorre, como podemos observar nas Tabelas 6.6 e 6.7. Um outro motivo é justamente o fato dele pertencer ao cromossomo 18. Pelas Leis de Mendel temos informações presentes no mesmo gene estão relacionadas, porém, o SNP C18S2320 não pertence ao mesmo gene que C18S2475, assim, não podemos dizer que um substituiu o outro por trazer a mesma informação genética.

Tabela 6.6: Comparação entre os SNPs C18S2320, C1S9455, C1S9266 e C2S2288.

	C1S9455			C1S9266		C2S2288	
		-1	0	-1	0	-1	0
C18S2320	-1	689	4	689	4	689	4
	0	4	0	4	0	4	0

Tabela 6.7: Comparação entre os SNPs C3S4611, C1S9455, C1S9266 e C2S2288.

	C1S9455			C1S9266		C2S2288	
		-1	0	-1	0	-1	0
C3S4611	-1	689	4	689	4	689	3
	0	4	0	4	0	3	1

No Capítulo 3 vimos que as estimativas para os β_i podem ser viciadas quando calculadas pelo LASSO. Uma possível solução para esse problema seria a utilização da metodologia para a seleção de variáveis e o ajuste final do modelo ser feito a partir de uma Regressão Logística. Assim, obtivemos o modelo da Equação (6.1)

$$log\left(\frac{\hat{p}}{1-\hat{p}}\right) = 41.706 + (2.906 * C3S5389) + (1.875 * C3S5742) + \\ (12.68 * C3S4611) - (12.68 * C15S774) + (12.68 * C18S2320). \tag{6.1}$$

Em seguida aplicamos o modelo na base de Teste e fizemos o cálculo da Erro de Predição:

$$EP_{LASSO} = 51,640.$$

Essa medida será importante mais à frente do estudo em que faremos a comparação entre os modelos resultantes de cada metodologia usada.

6.4 Florestas Aleatórias

Na Seção 6.2 vimos que no caso das Florestas Aleatórias, 444 SNPs foram selecionados em pelo menos 11 das 21 rodadas. Por ser um grande número de covariáveis, não iremos fazer uma análise detalhada de suas características como no caso do LASSO.

Apesar de ser mais estável que o LASSO na seleção de variáveis, notamos que as Florestas Aleatórias também selecionam diferentes variáveis dependendo das amostras OOB's geradas. Com isso em mente, optamos por rodar a metodologia mais 5 vezes com diferentes sementes para a separação da base de Treino e Validação, e somente com os 444 SNPs mais importantes. A escolha do modelo final foi feita pelo menor Erro de Predição(EP_{FA}).

$$EP_{FA} = 50,075$$

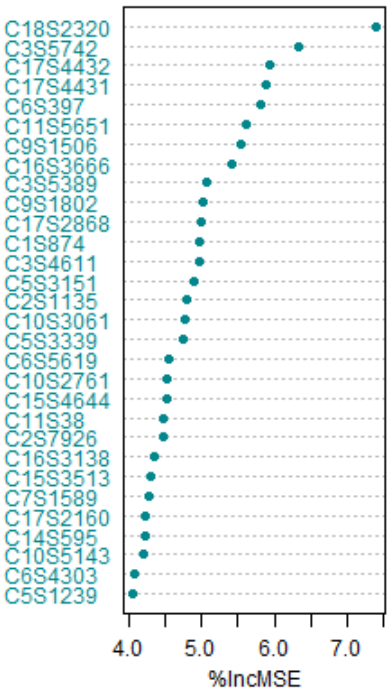

Figura 6.1: Importância das variáveis via Florestas Aleatórias, em que o número antes da letra S no nome do SNP representa de qual cromossomo ele é proveniente.

Na Figura 6.1 podemos observar as covariáveis que apresentam maior importância no modelo final. Notamos que quase todos os cromossomos presentes na Tabela 5.1 estão representados, com exceção dos cromossomos 8 e 19. Todavia, novamente nenhum SNP destacado por Almasy *et al.* (2011) foi selecionado.

Podemos observar que os dois SNPs mais importantes segundo as Florestas Aleatórias são: C18S2320 e C3S5742, também selecionados pelo LASSO. Os SNPs C3S5389 e C3S4611, identificados pelo LASSO são, respectivamente, a 9 e 13 variáveis mais importantes nas florestas.

Assim como no LASSO, podemos interpretar os resultados de duas formas. A seleção de SNPs de cromossomos diferentes daqueles presentes na Tabela 5.1 possivelmente ocorreu por conta de uma maior concentração de informação em covariáveis cujas características se assemelham àquelas utilizadas na simulação. Já a seleção de outros SNPs presentes nos cromossomos esperados provavelmente ocorreu pela relação da informação de SNPs do mesmo gene.

6.5 Florestas Aleatórias e Regressão Logística

Como o método das Florestas Aleatórias ainda apresenta um grande número de SNPs selecionados, muito maior do que o real número de SNPs significativos, resolvemos estimar um terceiro modelo como sendo a combinação dos métodos de Florestas Aleatórias e Regressão Logística. A ideia desse terceiro método é utilizar as Florestas Aleatórias para uma pré-seleção de variáveis e depois estimar, via regressão logística, o modelo final. Esse modelo final foi ajustado de duas maneiras:

1. Com as 444 variáveis selecionadas;

2. Aplicando uma nova seleção de variáveis, agora via Stepwise.

Com as 444 variáveis selecionadas anteriormente, foi feito o ajuste de uma Regressão Logística. Por se tratar de um modelo muito extenso, traremos apenas o EP para que possamos compará-lo com os demais casos estudados:

$$EP_{FARLT} = 79,375.$$

Em busca de um modelo mais parcimonioso, decidimos usar a seleção de variáveis via Stepwise em conjunto com a Regressão Logística. O Stepwise é uma metodologia na qual as variáveis são adicionadas e retiradas do modelo em cada passo a partir de um critério pré estabelecido, no caso o critério utilizado foi o AIC. Nessa métrica, quanto menor o valor encontrado, melhor o modelo. E as variáveis são incluídas ou excluídas do modelo se o valor do AIC diminuir com a ação que foi testada.

Após a realização de 1000 iterações, encontramos um modelo com 75 covariáveis. Novamente traremos apenas a medida de desempenho por se tratar de um modelo extenso. Assim, temos o EP resultante dado por:

$$EP_{FARLS} = 76,915.$$

6.6 Comparação de Desempenho

Na Tabela 6.8 temos as Somas de Quadrados dos Erros dos quatro modelos finais que foram encontrados. Observamos que ambos os casos em que combinamos as metodologias Florestas Aleatórias e MLGen apresentam os piores resultados, pois são os modelos com

maiores erros. Os modelos que combinam a seleção via LASSO com ajuste via MLGen e a seleção e modelagem via Florestas Aleatórias apresentam erros bem semelhantes, porém o modelo utilizando apenas Florestas Aleatórias tem um erro inferior aos demais.

Tabela 6.8: Medida de desempenho dos modelos resultantes.

Seleção	**Modelagem**	**EP**
LASSO	MLGen	51,640
Florestas Aleatórias	Florestas Aleatórias	50,075
Florestas Aleatórias	MLGen	79,375
Florestas Aleatórias + Stepwise	MLGen	76,915

Temos assim, a partir dos resultados observados na Seção 6.2 e considerando o conjunto de dados que foi analisado nesse trabalho, que a metodologia Florestas Aleatórias é mais estável na seleção das variáveis quando a base de Treino é alterada se comparada com a seleção feita pelo LASSO. Além disso, a partir da Tabela 6.8 concluímos que o erro na previsão de novos casos utilizando apenas as Florestas Aleatórias é inferior aos demais testados, sendo que o LASSO apresenta um erro muito próximo. Em relação à complexidade do modelo, o ajuste pelo LASSO é bem mais simples sendo composto por apenas 5 covariáveis enquanto o ajuste com florestas apresenta 444. Nenhuma das metodologias estudadas apresentou ótima performance em selecionar os 51 SNPs importantes. Mas isso já era esperado se considerarmos a complexidade dos dados analisados.

Dessa forma, podemos concluir que para o banco de dados em estudo e com base na métrica utilizada para comparação, a melhor opção seria usar a seleção e o ajuste via Florestas Aleatórias. Porém, é válido ressaltar que o modelo em que as covariáveis foram selecionadas via LASSO e o ajuste foi feito usando Regressão Logística obteve um desempenho semelhante e é mais parcimonioso.

Capítulo 7

Conclusões e Estudos Futuros

Com o desafio de encontrar metodologias que possibilitem o estudo de cenários com mais covariáveis do que observações, como por exemplo estudos de genética, a literatura nos indicou que modelos de *Machine Learning* poderiam ser uma solução para essa problemática. A partir dessa ideia, escolhemos duas metodologias para realizar o estudo de seleção de marcadores SNPs que impactam a propensão de doença ou não: LASSO e Florestas Aleatórias.

Ao explorar o banco de dados GAW17, escolhido para a análise, nos deparamos com o desafio de selecionar covariáveis raras ou com pouquíssima variabilidade, o que dificulta a seleção correta de características que são determinantes na resposta.

Comparamos a estabilidade do LASSO e das Florestas Aleatórias a partir da frequência em que as mesmas covariáveis eram selecionadas ao modificarmos a base de Treino Efetivo e observamos se os marcadores realmente utilizados na simulação dos dados foram ou não selecionados pelas metodologias. Seguimos para a análise do erro na classificação das observações, no qual procuramos a metodologia que apresentasse o menor Erro de Predição.

Após as análises concluímos que, a partir das métricas escolhidas, do ponto de vista preditivo encontramos que realizar a seleção de variáveis e o ajuste do modelo via Florestas Aleatórias apresentou o menor erro de predição. Todavia, destacamos que o modelo usando o LASSO para a seleção de variáveis e o ajuste via Regressão Logística é mais parcimonioso e apresente erro semelhante.

Nenhuma das metodologias estudadas foi capaz de selecionar corretamente covariáveis significativas e isso talvez se deva ao fato de que grande parte delas apresenta pouquíssima variabilidade na amostra considerada. E essa é uma situação muito frequente em dados

genéticos, pois mutações acontecem em um número muito pequeno de pessoas.

A partir dos resultados obtidos nesse estudo, alguns pontos ainda devem ser explorados em estudos futuros. Entre eles, se destacam o estudo e aplicação de metodologias que selecionem de forma mais efetiva covariáveis raras importantes e que considerem associação familiar entre indivíduos da amostra. Em Zeng *et al.* (2015) podemos observar o uso de Análise de Componentes Principais, Modelos Mistos, além da sugestão da uma modelagem em dois estágios com diferentes abordagens.

Referências Bibliográficas

Almasy, L., Dyer, T. D., Peralta, J. M., Kent, J. W., Charlesworth, J. C., Curran, J. E. e Blangero, J. (2011). Genetic analysis workshop 17 mini-exome simulation. In *BMC Proceedings*, volume 5, page S2. BioMed Central.

Beckmann, J., Kashi, Y., Hallerman, E., Nave, A. e Soller, M. (1986). Restriction fragment length polymorphism among israeli holstein-friesian dairy bulls. *Animal Genetics*, **17**(1), 25–38.

Breiman, L. (1996). Bagging predictors. *Machine learning*, **24**(2), 123–140.

Breiman, L. (2001). Random forests. *Machine Learning*, **45**(1), 5–32.

Caetano, A. R. (2009). Marcadores SNP: conceitos básicos, aplicações no manejo e no melhoramento animal e perspectivas para o futuro. *Revista Brasileira de Zootecnia*, **38**(8), 64–71.

Chardon, P., Kirszenbaum, M., Cullen, P. R., Geffrotin, C., Auffray, C., Strominger, J. L., Cohen, D. e Vaiman, M. (1985). Analysis of the sheep MHC using HLA class I, II, and C4 cDNA probes. *Immunogenetics*, **22**(4), 349–358.

Chen, S. S. e Gopalakrishnan, P. S. (1998). Clustering via the Bayesian information criterion with applications in speech recognition. In *Proceedings of the 1998 IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP'98 (Cat. No. 98CH36181)*, volume 2, pages 645–648. IEEE.

Chun, H. e Keleş, S. (2010). Sparse partial least squares regression for simultaneous dimension reduction and variable selection. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, **72**(1), 3–25.

Feng, Z. Z., Yang, X., Subedi, S. e McNicholas, P. D. (2012). The LASSO and sparse least squares regression methods for SNP selection in predicting quantitative traits.

IEEE/ACM Transactions on Computational Biology and Bioinformatics (TCBB), **9**(2), 629–636.

Friedman, J., Hastie, T. e Tibshirani, R. (2001). *The elements of statistical learning*, volume 1. Springer series in statistics New York.

Friedman, J., Hastie, T. e Tibshirani, R. (2010). Regularization paths for generalized linear models via coordinate descent. *Journal of Statistical Software*, **33**(1), 1–22.

Goldberg, D. E. (1989). Genetic algorithms in search, optimization, and machine learning. *Summary the Applications of GA-Genetic Algorithim for Dearling with some Optimal Calculations in Economics*.

Hastie, T., Tibshirani, R. e Jerome, F. (2017). *The Elements of Statistical Learning Data (2nd edition)*. Springer. https://doi. org/10.1007/b94608.

Izbicki, R. e Santos, T. (2018). Machine learning sob a ótica estatística: Uma abordagem preditivista para estatística com exemplos em R. *Notes*.

Lewis, J., Abas, Z., Dadousis, C., Lykidis, D., Paschou, P. e Drineas, P. (2011). Tracing cattle breeds with principal components analysis ancestry informative snps. *PloS one*, **6**(4), e18007.

Liaw, A. e Wiener, M. (2002). Classification and regression by randomForest. *R News*, **2**(3), 18–22.

Mokry, F. B., Higa, R. H., de Alvarenga Mudadu, M., de Lima, A. O., Meirelles, S. L. C., da Silva, M. V. G. B., Cardoso, F. F., de Oliveira, M. M., Urbinati, I., Niciura, S. C. M. *et al.* (2013). Genome-wide association study for backfat thickness in canchim beef cattle using random forest approach. *BMC genetics*, **14**(1), 47.

Nelder, J. A. e Wedderburn, R. W. (1972). Generalized linear models. *Journal of the Royal Statistical Society: Series A (General)*, **135**(3), 370–384.

Oliveira, F. C. d. *et al.* (2015). Um método para seleção de atributos em dados genômicos - tese de doutorado da ufjf.

Park, T. e Casella, G. (2008). The Bayesian LASSO. *Journal of the American Statistical Association*, **103**(482), 681–686.

Paula, G. A. (2004). *Modelos de regressão: com apoio computacional*. IME-USP.

Sasazaki, S., Hosokawa, D., Ishihara, R., Aihara, H., Oyama, K. e Mannen, H. (2011). Development of discrimination markers between japanese domestic and imported beef. *Animal Science Journal*, **82**(1), 67–72.

Schapire, R. E., Freund, Y., Bartlett, P., Lee, W. S. *et al.* (1998). Boosting the margin: A new explanation for the effectiveness of voting methods. *The Annals of Statistics*, **26**(5), 1651–1686.

Suekawa, Y., Aihara, H., Araki, M., Hosokawa, D., Mannen, H. e Sasazaki, S. (2010). Development of breed identification markers based on a bovine 50k snp array. *Meat Science*, **85**(2), 285–288.

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society: Series B (Methodological)*, **58**(1), 267–288.

Zeng, P., Zhao, Y., Qian, C., Zhang, L., Zhang, R., Gou, J., Liu, J., Liu, L. e Chen, F. (2015). Statistical analysis for genome-wide association study. *Journal of biomedical research*, **29**, 285–97.