# Uso de técnicas de reconhecimento de padrões e genetica para a predição de transtornos psiquiatricos da infância

Walkiria Karla Resende

Dissertação apresentada
ao
Instituto de Matemática e Estatística
da
Universidade de São Paulo
para
obtenção do título
de
Mestre

Programa: Programa Interunidades em Bioinformática

Orientador: Prof. Dra. Helena Brentani

Coorientador: Prof. Dr. Fabricío Martins Lopes

São Paulo, Maio de 2017

# Uso de técnicas de reconhecimento de padrões e dados genéticos para a predição de transtornos psiquiátricos da infância

Esta é a versão original da dissertação elaborada pela
candidata Walkiria Karla Resende, tal como
submetida à Comissão Julgadora.

# Agradecimentos

À vida pelas oportunidades que tive.

Agradeço à minha família que sempre me apoiou em todas as minhas decisões! Obrigada mãe, pai, minhas irmãs Wânia e Ana Izabel. Obrigada a família do meu marido, sem o apoio de vocês tudo teria sido muito mais difícil, vocês também são a minha família.

Henrique obrigada por cada instante compartilhado! Pelo carinho e dedicação que sempre teve por mim! Vencemos juntos.

Obrigada Dra. Helena pela disponibilidade e por tanto conhecimento compartilhado. Foi uma oportunidade incrível trabalhar com você.

Obrigada Fabrício pelos vários conselhos e incentivo ao decorrer deste trabalho.

Obrigada ao Grupo de Genética da Dra. Helena pelo conhecimento compartilhado.

Obrigada Looplex pelo apoio e pelos horários flexíveis.

Obrigada à todos os mestres que tive durante a vida acadêmica!

"A ciência é uma disposição de aceitar os fatos mesmo quando eles são opostos aos desejos". (Burrhus Frederic Skinner)

# Resumo

Resende, W. K. **Uso de técnicas de reconhecimento de padrões e genetica para a predição de transtornos psiquiátricos da infância**. 2017. Dissertação - Instituto de Matemática e Estatística, Universidade de São Paulo, São Paulo, 2017.

Os transtornos psiquiátricos são o resultado da interação entre diversos genes e o ambiente. Além disso durante a infância o diagnóstico de tais desordens não é uma tarefa trivial, uma vez que diferentes transtornos compartilham os mesmos sintomas. Apesar desta dificuldade, é muito importante que ele seja feito o mais cedo possível, o que fará com que o indivíduo obtenha melhor resposta a intervenção.

O fato de trabalharmos com desordens poligênicas e multifatoriais reafirma a dificuldade do diagnóstico. Um gene pode estar relacionado com diferentes transtornos e diferentes transtornos compartilham vários genes, ou seja genes não codificam doenças e sim endofenótipos comportamentais.

Este trabalho consiste de 3 etapas. Na primeira, fizemos a genotipagem dos dados e garantimos consistência dos resultados de acordo com análises de Equilíbrio de Hardy e Weinberg. Na segunda mapeamos os SNPs de acordo com o conhecimento biológico a priori. Procurando evitar que ocorram representações insignificantes ou ainda um *overfitting* será feita uma seleção de SNPs baseados em conhecimento a priori e integrar o conhecimento biológico com os algoritmos de seleção. Na última etapa propomos uma colaboração à predição at risk de transtornos da infância, elaborando um classificador de alta sensibilidade e especificidade.

Considerando que a seleção de características esteja adequado ao problema, o desafio a ser solucionado será a escolha e parametrização do classificador. Serão utilizadas 723 amostras, sendo 503 controles e 221 casos. Estas amostras são provenientes do projeto do Instituto Nacional de Psiquiatria do Desenvolvimento (INPD).

**Palavras-chave:** reconhecimento de padrões, classificação, doenças complexas, psiquiatria.

# Sumário

**Lista de Abreviaturas** vii

**Lista de Figuras** ix

**Lista de Tabelas** xi

**1 Introdução** 1

1.1 Justificativa . . . 3

1.2 Objetivos . . . 4

**2 Conceitos** 7

2.1 SNPs e Risco Poligênico . . . 7

2.2 Reconhecimento de Padrões . . . 10

2.2.1 Seleção de Características . . . 14

2.2.2 Dimensão da amostra e das características . . . 18

2.2.3 Balanceamento de Classes . . . 18

2.2.4 Escolha do modelo para classificação . . . 19

2.2.5 SNPs e aprendizado de máquinas . . . 19

**3 Metodologia** 23

3.1 Amostras . . . 24

3.1.1 Genotipagem e Controle de Qualidade . . . 25

3.1.2 Integração com conhecimento Biológico . . . 26

3.1.3 Seleção de Características . . . 27

3.1.4 Construção do classificador . . . 27

**4 Resultados** 31

4.1 Amostras . . . 31

4.2 Seleção de Características . . . 34

4.2.1 Escore de Fisher . . . 34

4.2.2 Regularização Lasso . . . 34

4.2.3 SFFS . . . 34

4.2.4 FBST . . . 34

4.2.5 Integrando Critérios Biológicos . . . 35

**5 Conclusões** **43**
5.1 Discussão . . . . . 43
5.2 Conclusões . . . . . 44

**A Modelo Relacional** **45**

**Referências Bibliográficas** **47**

# Lista de Abreviaturas

| | |
|---|---|
| AMD | (*Age-related Macular Degeneration*) |
| DI | Deficiência Intelectual |
| DNA | Ácido Desoxirribonucleico ( *Deoxyribonucleic Acid*) |
| FBST | Teste Baisiano Completo ( *Full Bayesian Significance Test for Coefficients of Variation*) |
| FS | Escore de Fisher (*Fisher Score*) |
| GWAS | Estudo de Associação Ampla do Genoma (*Genome Wide Association*) |
| LD | Desequilibrio de Ligação (*Linkage Disequilibrium*) |
| MAF | Alelo de menor frequência (*Minor Allele Frequency*) |
| PRS | Escore de Risco Poligênico (*Polygenic Risk Score*) |
| PCA | Análise de Componentes Principais (*Principal Component Analysis*) |
| RBF | Função de Base Radial (*Radial Basis Function*) |
| RNA | Redes Neurais Artificiais |
| RNA | Acido Ribonucleico (*Ribonucleic Acid*) |
| SNPs | Polimorfismos de Nucleotídeo Único ( *Single Nucleotide Polymorphisms)*) |
| SVM | Máquinas de Vetores de Suporte (*Support Vector Machine*) |
| SWFS | (*Sliding Window Sequential Forward Feature Selection*) |
| TDAH | Transtorno de Déficit de Atenção e Hiperatividade |
| TEA | Transtorno do Espectro Autista |
| TOC | Transtorno Obsessivo Compulsivo |
| TAE | Teoria do Aprendizado Estatístico |

# Lista de Figuras


1.1 Espaço de sintomas do DSM-IV. Figura adaptada de "The small world of psychopathology", BORSBOOM et al., 2011 . . . 1
1.2 Proporção de variantes com diferentes propriedades funcionais nas regiões gênicas. Extraído de: "GWASdb: a database for human genetic variants identified by genome-wide association studies", 2012 . . . 4

2.1 Fator genético em cada transtorno de acordo com estudos em gêmeos. Extraído de: . 8
2.2 Descrição da figura mostrada. . . . 11
2.3 Principais etapas durante o processo de aprendizado de máquina. Figura adaptada de "Supervised machine learning: A review of classification techniques", 2007 . . . 12
2.4 Um conjunto de características é a entrada de um processo que seleciona um subconjunto de caracteristicas que serão avaliadas e, de acordo com um critério determinado, serão usadas para a classificação (validação) ou serão descartadas. . . . 14
2.5 O primeiro passo é aplicar o SFS, o qual adiciona uma característica por vez baseado na função objetivo. O SFFS adiciona outro passo, que é a exclusão de uma característica por vez , passo do SBS, do subconjunto obtido no primeiro passo e avalia o novo subconjunto. Se a exclusão leva a uma melhor performance, então aquela característica é removida e volta para o primeiro passo . O processo acontece até que o número desejado de features seja atingido (Chandrashekar e Sahin, 2014). Figura extraída de: "Redes complexas de expressão gênica: síntese, identificação, análise e aplicações, 2011" . . . 18
2.6 Figura extraída de Python machine learning, 2015 . . . 21

3.1 Fluxo de desenvolvimento do trabalho . . . 24
3.2 Ilustração do Banco de dados criado para facilitar a obtenção e integração de informações . . . 27
3.3 Fluxo de seleção de características, integrando informações biólogicas e algoritmos . 28
3.4 Exemplo de validação cruzada com 5 folds. . . . 29
3.5 Matriz de Confusão - Adaptada de Sokolova, 2009 . . . 30

4.1 PCA feito com os dados após o controle de qualidade para verificar homogeneidade da população. . . . 33
4.2 Distribuição dos scores de Fisher em todos os SNPs da lâmina para amostras caso e controle. . . . 34

4.3 Diagrama de Venn dos SNPs associados a psiquitria, os que estão em genes e selecionados por FBST . . . . . 37
4.4 Diagrama de Venn dos SNPs associados a psiquiatria, os que estão em região regulatória e selecionados por FBST . . . . . 37

A.1 Modelo Relacional do Banco de Dados para Integração de informações biológicas . . 46

# Lista de Tabelas


2.1 Uso de Risco Poligênico em Psiquiatria. Referências: a - Frank et al. (2012); b - Collins et al (2012); Whalley et al. (2012);c - Service et al. (2012);d - Demirkan et al. (2011);e - Luciano et al. (2012);f - Otowa et al. (2012);g - ISC (2009); Derks, Vorstman, Ripke, Kahn and Ophoff (2012); Ikeda et al. (2011); Levinson et al. (2012); Ripke, O'Dushlaine, et al. (2013);h - Hamshere, Langley, et al. (2013);i - Anney et al. (2012);j - Vrieze, McGue, and Iacono (2012); . . . . . 10

2.2 Resumo dos métodos de particionamento. Extraído de (Monard e Baranauskas, 2003) 13

2.3 Métodos de validação e quando usar. Adaptado de http://leitang.net/papers/ency-cross-validation.pdf . . . . . 13

4.1 Número de casos e controles separados por sexo . . . . . 31

4.2 Número de casos de acordo com transtornos psiquiátricos caracterizados pelo DSM-5 31

4.3 Número de comorbidades por indivíduo . . . . . 32

4.4 Exposição ao estresse em caso e controle do sexo feminino . . . . . 32

4.5 Exposição ao estresse em caso e controle do sexo masculino . . . . . 33

4.6 Critérios de remoção utilizados neste trabalho e número de SNPs removidos . . . . . 33

4.7 Quantidade de SNPs selecionados variando $\alpha$ . . . . . 34

4.8 Número de SNPs selecionados de acordo com a variação do e-valor . . . . . 35

4.9 Resultado da classificação dos Top 5 SNPs selecionados pelo método FBST em seguida por Fisher . . . . . 35

4.10 Resultado da classificação dos Top 5 SNPs selecionados pelo método FBST em seguida por Lasso . . . . . 35

4.11 Quantidade de SNPs em cada critério biológico . . . . . 35

4.12 Parâmetros de avaliação da classificação realizada com filtros de características biológicas e os algoritmos utilizados - Fisher . . . . . 36

4.13 Parâmetros de avaliação da classificação realizada com filtros de características biológicas e os algoritmos utilizados - LASSO . . . . . 36

4.14 Resultado da Intersecção Biológicos - Fisher . . . . . 37

4.15 Resultado da Intersecção Biológicos - LASSO . . . . . 38

4.16 Melhores resultados . . . . . 38

4.17 TDAH com SNPs psiquiatria . . . . . 39

4.18 SNPS em Psiquiatria e Região regulatória . . . . . 39

4.19 SNPS em FBST e Região Regulatória . . . . . 40

4.20 SNPS em FBST e Genes . . . . . 40

4.21 SNPS em Psiquiatria e Genes . . . . . 41
4.22 Resultado da melhor AUC em cada método para meninos e meninas . . . . . 41

# Capítulo 1

# Introdução

As doenças psiquiátricas são, em geral, caracterizadas por grande heterogeneidade clínica (Fanous e Kendler, 2005; Fornito e Bullmore, 2012), ou seja, a apresentação clínica é extremamente variável, várias combinações de diferentes sintomas podem resultar em um mesmo diagnóstico. Por outro lado, um sintoma pode ser compartilhado entre os diferentes transtornos psiquiátricos, tais como o prejuízo de cognição social, compartilhado por transtorno do espectro autista (TEA), esquizofrenia (Korkmaz, 2011), transtorno bipolar e depressão maior (Domschke, 2013; Owoeye *et al.*, 2013), entre outros. Além da heterogeneidade fenotípica, altas taxas de comorbidade são detectadas, ou seja, o diagnóstico de mais de um transtorno psiquiátrico em um mesmo paciente é muito comum. Por exemplo, (Simonoff *et al.*, 2008) relataram que cerca de 70% das crianças com diagnóstico de TEA apresentaram pelo menos uma comorbidade e 48% apresentaram duas ou mais.

Esse compartilhamento pode ser visualizado na Figura 1.1: cada sintoma do DSM-IV (*Diagnostic and Statistical Manual of Mental Disorders IV*) é um nó e existe uma aresta entre os sintomas se eles são critério para um mesmo transtorno. Dito de outro modo, pode-se afirmar que os sintomas estão conectados quando são comuns a um mesmo transtorno. Com esse estudo foi possível visualizar a alta comunalidade de sintomas entre os diagnósticos psiquiátricos, assim como entender padrões de comorbidade (Borsboom *et al.*, 2011).

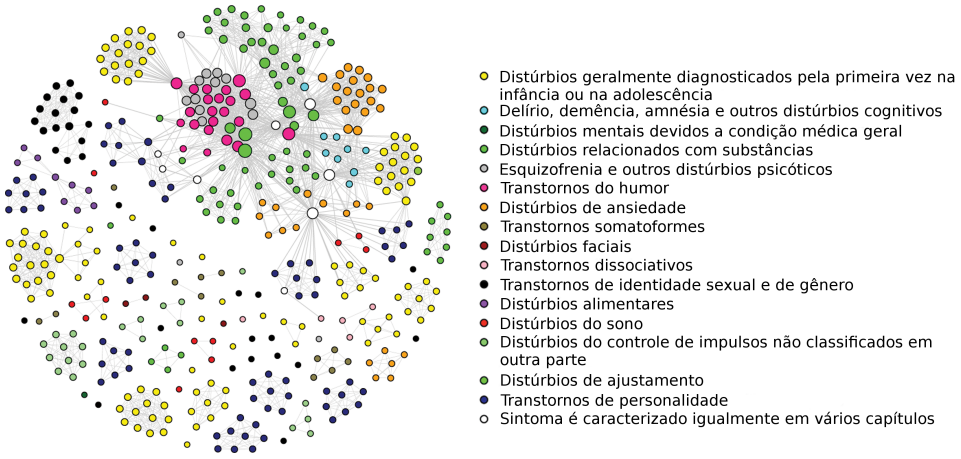


**Figura 1.1:** *Espaço de sintomas do DSM-IV. Figura adaptada de "The small world of psychopathology", BORSBOOM et al., 2011*

Por outro lado, hoje em dia já está claro que os comportamentos humanos são resultado de interações genéticas e influências do ambiente (Plomin e Kosslyn, 2001; Plomin *et al.*, 2001). Neles, os efeitos dos alelos gênicos envolvidos variam e espera-se que poucos alelos tenham um grande efeito, sendo que a maioria exerce um efeito pequeno (Marian, 2012). A contribuição individual de variantes pode não ser suficiente para alterar a suscetibilidade para um fenótipo complexo, sendo necessária a contribuição de vários alelos de pequeno efeito agindo ao mesmo tempo (Marian, 2012). Pode-se dizer então que o fator genético atua de forma quantitativa nesses fenótipos (Fanous e Kendler,

2005; Plomin *et al.*, 2009). Dessa maneira, os transtornos passam a ser pensados como traços quantitativos extremos e o foco está na dimensão destes traços (Burmeister *et al.*, 2008; Goh *et al.*, 2007; Plomin *et al.*, 2009) e na busca por marcadores biológicos de diagnóstico.

O genoma humano contém aproximadamente 3,2 bilhões de pares de bases, os quais estão arranjados em 22 pares de cromossomos somáticos e 2 cromossomos sexuais. O genoma é composto de cerca de 20000 genes codificantes contendo aproximadamente 180.000 exons que, coletivamente, envolvem 35 milhões de pares de bases ou cerca de 1,1% do genoma. A transcrição do genoma também produz mais de 50.000 RNAs não codificantes, incluindo mais de 2000 microRNAs. Além disso, cada gene transcreve múltiplos variantes de *splice* de mRNA com um ou dois variantes de *splice* dominantes. O genoma também contêm 100.000 elementos regulatórios que modulam a expressão do gene e, consequentemente, o fenótipo.

Humanos têm alta diversidade genética. Eles diferem em aproximadamente 0.1% de seus genomas. O dbSNP lista mais de 37 milhões de variantes entre os humanos. Com a exceção de gêmeos idênticos, dois humanos não contêm genomas idênticos. Cada genoma contém cerca de 4 milhões de variações em sua sequência de DNA, que coletivamente afeta metade dos genes em cada genoma. Grande parte das variações de DNA no genoma são SNPs (do inglês: *Single Nucleotide Polymorphisms - SNPs*), mas existem variações estruturais que afetam grande parte dos nucleotídeos no genoma. Essas variações incluem deleções, inserções, duplicações e rearranjos de grandes segmentos de DNA e, assim, podem aumentar ou diminuir o número de cópias de genes das duas cópias naturais. Tais variações são chamadas de variações do número de cópias, do inglês *Copy Number Variants (CNVs)* Entre os aproximadamente 3.5 milhões de SNPs em cada genoma, cerca de 10.000 SNPs são não-sinônimas (nSNPs). Além disso, cada genoma contêm em média entre 50 e 100 variantes associadas a doenças herdadas e cerca de 30 variantes de novo (Consortium *et al.*, 2010).

Para a busca dos marcadores biológicos, o foco em estudos de traços complexos está inicialmente nas variantes do DNA entre os indivíduos, as quais podem contribuir para a suscetibilidade da doença (Frazer *et al.*, 2009; Hirschhorn e Daly, 2005). Grande parte da variação genética humana é representada por SNPs e acredita-se que eles contribuam com a maior carga para explicar diferenças fenotípicas entre os indivíduos (Ramensky *et al.*, 2002). Os SNPs são variações de base única como, por exemplo, a substituição de uma citosina (C) por uma timina (T) em uma posição específica do DNA (Aguiar-Pulido *et al.*, 2010; Zhou e Wang, 2007). As características dessa substituição são (Zhou e Wang, 2007): (a) Muito comuns no genoma humano, podem ocorrer a cada 100 - 300 pares de base; (b) dois de cada três SNPs são variações de citosina para timina; (c) estáveis entre gerações.

Do ponto de vista biológico, vale ressaltar que, dependendo de onde a variante ocorre no genoma e de qual é o nível de impacto nas proteínas e/ou nos processos regulatórios, podemos observar diferentes consequências em nível fenotípico — e, quando pensamos em um traço fenotípico quantitativo dependente da contribuição de vários alelos de pequeno efeito, mesmo SNPs fora de regiões codificadoras, ou ainda aqueles em região codificadora sinônimos, que não alteram o aminoácido gerado, podem ser importantes (Malhotra *et al.*, 2004; Shastry, 2009).

Atualmente a forma mais comum de se estudar as variações é utilizando a técnica de associação genômica em larga escala (GWAS), em que vários SNPs são investigados ao mesmo tempo em casos e controles. Com esses estudos, mais de 1000 regiões já foram associadas a aproximadamente 165 doenças, entre elas: diabetes, câncer e artrite reumatoide (McCarthy *et al.*, 2008). No entanto, em transtornos psiquiátricos foram poucos os resultados positivos. Inicialmente acreditava-se que este fato estava relacionado apenas ao número amostral, mas posteriormente foi observado que, se traços complexos são explicados pela ação conjunta de vários alelos, seria necessário analisar e associar os SNPs de forma combinada e não individualmente como estava sendo feito (McCarthy *et al.*, 2008). Este problema foi parcialmente resolvido com o cálculo de risco por escores poligênicos (Dudbridge, 2013). A investigação dos vários SNPs associados nos fornece um escore que, associado a um peso, fornece o risco poligênico. No entanto, alguns pontos precisam ser levantados quando o objetivo é buscar SNPs para o cálculo do risco:

- Este modelo deve ser usado quando as amostras são independentes, com fenótipos caracteri-

zados da mesma forma e grandes o suficiente para executar dois GWAS, um para treinamento e outro para replicação (Dudbridge, 2013).

- Se já existe um GWAS prévio que foi usado como amostra de treinamento para seleção dos alelos para o escore poligênico, teoricamente podemos realizar apenas o GWAS de replicação; no entanto, nesse caso a questão da origem das populações comparadas precisa ser levada em consideração, um problema importante no caso da população brasileira (Domingue *et al.*, 2014).

- Parâmetros técnicos da análise de GWAS, como controle de qualidade e imputação de dados, podem afetar a comparação final de duas bases de dados diferentes.

- Modelos da relação entre a combinação dos SNPs, outras variantes genéticas, exposição ao meio e susceptibilidade a doenças não são levados em consideração.

- Existe uma dificuldade na interpretação biológica dos escores de risco poligênico, uma vez que a seleção de alelos é feita baseada no ranqueamento de seus p-valores sem nenhum conhecimento biológico agregado.

Assim, em situações como a deste estudo no qual existe uma amostra pequena de casos psiquiátricos e controles, vinda de uma população Brasileira, e sem nenhum estudo feito anteriormente, seria muito difícil a construção de um escore de risco poligênico para classificação. Por outro lado, abordagens com aprendizado de máquina vêm sendo sugeridas para classificação em outras doenças complexas. Um problema em aberto nessa área é a seleção de características, que no caso são os SNPs a serem usados, diante de amostras pequenas (De La Vega *et al.*, 2005). É interessante notar que, apesar dos estudos de GWAS evidenciarem que SNPs associados a doenças não apresentam uma distribuição aleatória no genoma, mas que grande parte deles está localizada em regiões importantes para mecanismos de controle da expressão gênica, o uso desta informação não é levado em conta na seleção de características (Maurano *et al.*, 2012). Dessa forma, este trabalho visa ao estudo de 2 pontos:

1. Diferentes abordagens para seleção de características (SNPs) levando em conta ou não conhecimento biológico a priori;

2. Utilizando SNPs, selecionados pelas diferentes abordagens de seleção de características, avaliar a performance de classificação para transtornos psiquiátricos e controles com técnicas de reconhecimento de padrões/aprendizado de máquina;

## 1.1 Justificativa

Levando-se em conta que:

- Transtornos psiquiátricos podem ser vistos como dimensionais, ou seja, representam a composição de diferentes traços normais do comportamento humano que, em dado momento, ultrapassam certos limiares, levando a variações caracterizadas por um conjunto de sintomas que ocorrem por tempo maior que o definido como aceitável e que traz prejuízo à funcionalidade do indivíduo. Sendo assim, um mesmo sintoma pode aparecer em diferentes transtornos e existe alta comorbidade entre os transtornos até então definidos como categóricos.

- Dados de genética e imagem vêm mostrando que vários transtornos compartilham a mesma base biológica e eles são considerados doenças complexas, ou seja, poligênicos e multifatoriais.

- Genes não codificam doenças e sim comportamentos, e vários genes associados a várias doenças psiquiátricas são melhor vistos hoje como associados a comportamentos e não a doenças específicas.

- Estudos de GWAS têm contribuído com escores poligênicos para classificação e predição de risco. No entanto, alguns pontos como necessidade de estudos anteriores, ou necessidade de estudos com número amostral muito grande para seu cálculo, problemas relacionados ao uso das mesmas variantes em populações de origem diferente, precisam ser levados em consideração. Para a população Brasileira o uso de escores de GWAS europeus ou americanos pode ser um sério problema, pois somos uma população miscigenada, normalmente muito diferente em estudos genéticos de outras populações.

Faz-se necessário buscar marcadores para classificação de transtornos psiquiátricos que:

- Usem vários SNPs ao mesmo tempo;
- Possam ser doença inespecífico ou entendam as mesmas como dimensões do comportamento;
- Usem técnicas de classificação não linear;
- Usem um número de características adequadas, diante de amostras pequenas, para que o classificador possa generalizar bem;

Além disso, um ponto interessante quanto aos resultados de GWAS é que para diferentes doenças, mas certamente fundamental para os transtornos psiquiátricos, a distribuição genômica dos SNPs implicados não é aleatória. Li *et al.* (2012) calcularam a distribuição das variantes genéticas resultantes de estudos de GWAS em diferentes regiões genômicas (Figura 1.2), 43,5% de todas as variantes estão em regiões gênicas e os outros 56,5% estão localizadas em regiões intergênicas, que são áreas que contêm enhancers, elementos promotores e outros reguladores, e que podem estar envolvidos na regulação dos genes e em redes regulatórias (Li *et al.*, 2012). Dessa forma, usar características biológicas para concentrar esforços em SNPs localizados em regiões genômicas com funções importantes pode contribuir para a seleção de características (SNPs) para a classificação de transtornos psiquiátricos.

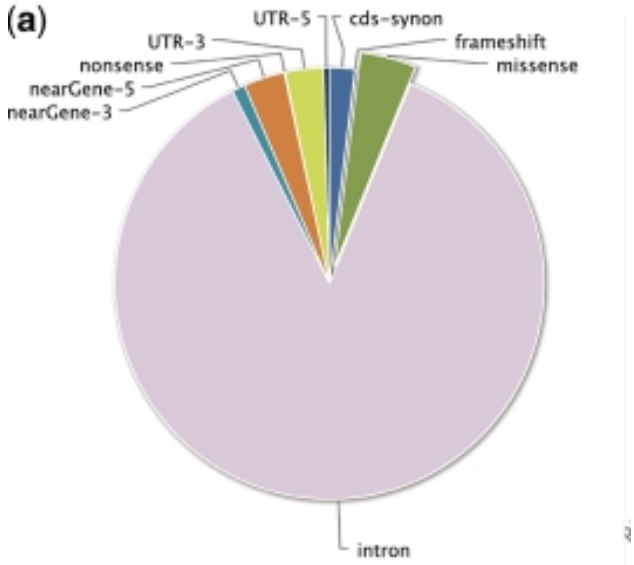


**Figura 1.2:** *Proporção de variantes com diferentes propriedades funcionais nas regiões gênicas. Extraído de: "GWASdb: a database for human genetic variants identified by genome-wide association studies", 2012*

## 1.2 Objetivos

Usar técnicas de reconhecimento de padrões, classificando as amostras da população em saudável ou doente para transtornos psiquiátricos a partir de um conjunto de SNPs como características dessas amostras.

**Objetivos Específicos**

- Avaliar os achados realizados por técnicas de reconhecimento de padrões para seleção de características com as anotações biológica e funcional em um conjunto de SNPs pré-selecionados para construir um classificador.
- Definir critérios para a identificação dos SNPs no conjunto de dados selecionado;
- Estudar técnicas de reconhecimento de padrões para a classificação das amostras;
- Selecionar as características (SNPs) do conjunto amostral;
- Integrar as características;
- Escolher um modelo para a classificação;
- Realizar os treinamento e os testes de classificação;
- Comparar os métodos de seleção de características utilizados e os métodos de classificação.

# Capítulo 2

# Conceitos

O objetivo desta dissertação é a busca de SNPs para classificação de transtornos psiquiátricos e, assim sendo, a revisão bibliográfica será apresentada em 3 tópicos:

- SNPs e risco poligênico;
- Reconhecimento de padrões e seleção de características;
- SNPs e técnicas de classificação.

As próximas Seções apresentam referenciais teóricos sobre esses temas.

## 2.1 SNPs e Risco Poligênico

Estudos com gêmeos e com famílias têm por objetivo compreender a composição genética e ambiental na patologia de certa doença. No Quadro 2.1 podemos observar como esses estudos estabelecem essas contribuições (Lima).

$h^2 = \frac{(C_m z - C_d z)}{(1 - C_d z)}$ em que $h^2$ é a herdabilidade (que mede a fração da variação em razão dos genes); $C_m z$ é a taxa de concordância entre co-gêmeos $MZ$ e $C_d z$ entre co-gêmeos $DZ$. Em um estudo de esquizofrenia, por exemplo, foram encontradas concordâncias de 58% e 13% para gêmeos $MZ$ e $DZ$, respectivamente. Logo, a herdabilidade $h^2$ para esquizofrenia neste estudo é de $(0,58 - 0,13)/(1 - 0,13) = 0,52$ (aproximadamente), indicando que a esquizofrenia tem cerca de metade da sua variação fenotípica dependendo da variação genotípica.

Diferentes estudos (Figura 2.1) sobre diversos transtornos psiquiátricos reportaram uma contribuição significante de fatores genéticos com herdabilidade, estimada em torno de 40-80% (Bulik *et al.*, 2000; Sullivan *et al.*, 2000). Esses estudos mostram que a herança de variantes genéticas não explica completamente a etiopatogenia, mas é importante na etiologia desses transtornos. Dessa forma, a identificação de variantes genéticas específicas associadas a transtornos psiquiátricos tem sido objeto de estudo por décadas Committee *et al.* (2009); Visscher *et al.* (2012). A seleção de SNPs para testes de associação é um problema relativamente antigo e não solucionado mesmo para os estudos com apenas um gene candidato (Balding, 2006; Cousin *et al.*, 2003; Phuong *et al.*, 2005).

Existem várias abordagens para encontrar as variantes relacionadas aos transtornos. Uma possibilidade é selecionar SNPs baseados em suas posições no gene ou baseado na natureza de alteração do DNA (Cousin *et al.*, 2003). No entanto, não se pode dizer que esses sejam os critérios mais eficientes. Outra possibilidade para a seleção de SNPs é procurar por padrões de Desequilíbrio de Ligação (do inglês:*Linkage disequilibrium - LD*) entre marcadores (Cichon *et al.*, 2009). O LD ocorre quando o genótipo de duas regiões fixas do cromossomo (*loci*) não são independentes uma da outra. Normalmente, os *loci* que estão fisicamente perto possuem um LD maior do que aqueles que estão afastados (Ardlie *et al.*, 2002). Contudo, alguns trabalhos (como, por exemplo, o de

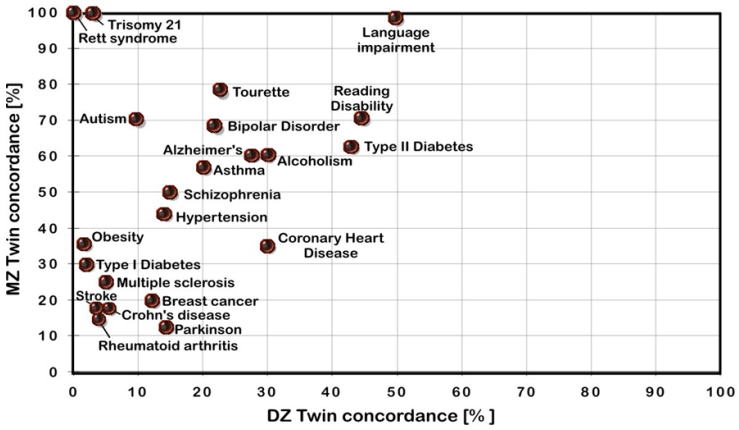


**Figura 2.1:** *Fator genético em cada transtorno de acordo com estudos em gêmeos. Extraído de:*

Pritchard e Przeworski (2001)) argumentam que a definição do que seja perto ou afastado não é clara, uma vez que o cálculo do LD não obedece a uma linearidade, e a regra para excluir um SNP se torna subjetiva. Além disso, o LD é fortemente influenciado por fatores demográficos, tais como etnia, mistura de populações e os padrões de LD dentro de genes candidatos são, portanto, difíceis de prever e podem variar entre populações (Pritchard e Przeworski, 2001).

É cada vez mais claro que doenças complexas são doenças poligênicas e multifatoriais, de forma que o fenótipo apresentado é um produto da ação de vários genes ao mesmo tempo e que um conjunto de variações comuns (definidas como alelos com frequência em dada população maior que 1%) explica melhor tais fenótipos. Nesse sentido, avanços recentes têm permitido testar sistematicamente as variantes genéticas no genoma para associação com traços medidos em indivíduos não relacionados sem nenhuma hipótese a priori, não havendo a necessidade da escolha prévia nem de genes nem de SNPs candidatos (Gratten *et al.*, 2014; Nurnberger Jr *et al.*, 2016; Shih *et al.*, 2004).

Os estudos de associação ampla visam a identificar e compreender variantes que influenciam em doenças comuns (Gibson, 2012; McCarthy *et al.*, 2008), usando tecnologias de genotipagem de alta eficiência (em inglês: *high-throughput*) (Bush e Moore, 2012) para analisar centenas de milhares de SNPs e relacioná-los a condições clínicas e/ou características mensuráveis. Essas características são chamadas endofenótipos (Gibson, 2012; McCarthy *et al.*, 2008), que são traços mensuráveis quantitativos herdáveis (Castellanos e Tannock, 2002). Mesmo com essa estratégia de genotipagem mais eficiente, não é necessário genotipar todas as variantes dos DNA: dada a estrutura de haplótipos do genoma humano, apenas certa quantidade de SNPs precisa ser selecionada para representar o genoma.

Isso se tornou viável com o Projeto International HapMap, um consórcio entre vários países organizado para produzir um mapa haplotípico de várias populações e, então, facilitar a escolha de SNPs (Manolio e Collins, 2009). O mapa foi construído com base no estudo de Desequilíbrio de Ligação entre marcadores (Committee *et al.*, 2009). A distância genômica na qual a taxa de LD decai indica quantos marcadores genéticos são necessários para representar blocos do genoma que foram herdados juntos (isso é um haplótipo) (Wray *et al.*, 2008). Sendo assim, podemos estabelecer um conjunto de tag SNPs que representam o genoma inteiro e podem ser utilizados em lâminas, ou microarranjos de SNPs (de Bakker *et al.*, 2005; Strachan e Read, 2016).

Dado um conjunto de genótipos, os pesquisadores buscam SNPs que tenham frequência alélica significantemente diferente entre dois grupos (Manor e Segal, 2013; Wu *et al.*, 2010).

Até o momento, com base nos estudos de GWAS foi possível concluir que (Korte e Farlow, 2013; Stranger *et al.*, 2011; Visscher *et al.*, 2012):

- Muitos *loci* contribuem para a variação de traços complexos;
- Em um *loci* existem múltiplos alelos associados com doenças em frequências variadas;
- Existem evidências de que as mesmas variantes estão associadas a múltiplos traços;

- Algumas variantes estão associadas a traços complexos em etnias diferentes;
- Quando todos os SNPs são considerados simultaneamente é encontrada uma grande proporção de variação genética aditiva.

Portanto, devido às evidências de que os SNPs estão envolvidos e agem conjuntamente, deu-se início a análises integradas de conjuntos de SNPs ao invés do estudo individual. A forma de associação dos SNPs com transtornos foi parcialmente resolvida com o cálculo de risco poligênico. Esse termo refere-se ao conjunto de múltiplas variantes associadas a um transtorno e tem sido usado para entender a contribuição poligênica em doenças comuns e predizer o risco de um indivíduo para uma doença (Dudbridge, 2013). O resultado desse cálculo é um escore de risco (PRS), que é feito somando os alelos associados à característica, normalmente ponderados pelo tamanho dos efeitos estimados a partir de um estudo de GWAS (Pirooznia *et al.*, 2012), como apresentado na Equação 2.1 (Dudbridge, 2013):

$$\sum_{m}^{i=1} \hat{\beta}_{i1} G_i \tag{2.1}$$

A aplicação do PRS tem crescido nos últimos anos devido à sua utilidade para a detectar a etiologia genética compartilhada entre traços, colaborar para o entendimento da arquitetura de traços fenotípicos, usar os SNPs como biomarcadores, prover conhecimento sobre a evolução, entre outros (Euesden *et al.*, 2014). Essa é uma forma de explicar os fatores genéticos em casos nos quais os marcadores não alcançam significado individualmente em estudos de GWAS (Committee *et al.*, 2009; Glessner e Hakonarson, 2009; Manor e Segal, 2013; Patnala *et al.*, 2013).

O primeiro trabalho a utilizar o PRS foi realizado pelo Consórcio Internacional de Esquizofrenia. Dado que poucos marcadores individuais eram significantes, e a hipótese doença comum variante comum estava em aberto, foi especulada a possibilidade da influência de milhares de marcadores de pequeno efeito. Com uma amostra caso e uma controle, utilizando SNPs de alta qualidade (filtrados de acordo com o MAF, taxa de genotipagem, desequilíbrio de ligação), e com um limiar permissivo, muitos SNPs foram selecionados. Ao realizar o teste, foi verificado um enriquecimento na amostra caso, evidenciando a influência de vários SNPs no fenótipo (Purcell *et al.*, 2009).

Em síntese, um estudo de GWAS é conduzido em um conjunto de treinamento e então os marcadores são ordenados por suas evidências para a associação, geralmente por seus p-valores. A partir daí, uma amostra de replicação independente é analisada para construir uma pontuação poligênica que consiste na soma ponderada de seus alelos associados à característica por algum subconjunto de marcadores no topo da ordenação. Duas aplicações relacionadas, mas distintas desta pontuação são, então, possíveis. Em primeiro lugar, testando para a associação entre o escore e o traço fenotípico na amostra de replicação, pode-se determinar se marcadores associados residem dentro daqueles que contribuem para a pontuação. Em segundo lugar, e talvez mais útil, a pontuação poligênica pode ser utilizada para prever valores de traço individuais ou risco de doença (Plomin *et al.*, 2009), potencialmente indicando um preditor com melhores propriedades de discriminação do que uma discriminação baseada apenas em marcadores estabelecidos (Dudbridge, 2013).

Na Tabela 2.1, as referências, adaptadas do trabalho de Wray *et al.* (2014), mostram estudos de cálculo de risco poligênico para vários transtornos psiquiátricos.

| | Fenótipo | Referência |
|---|---|---|
| 1 | Dependência de álcool | a |
| 2 | Transtorno Bipolar | b |
| 3 | Escalas de temperamento de Cloninger: agressividade, evitação, busca de novidade, dependência da recompensa, persistência | c |
| 4 | Extroversão depressiva grave | d |
| 5 | Neuroticismo e extroversão | e |
| 6 | Transtorno do Pânico | f |
| 7 | Esquizofrenia | g |
| 8 | Transtorno do déficit de atenção e hiperatividade | h |
| 9 | Transtorno do espectro autista | i |
| 10 | Desinibição comportamental, uso de álcool, uso e dependência de drogas | j |

**Tabela 2.1:** *Uso de Risco Poligênico em Psiquiatria. Referências: a - Frank et al. (2012); b - Collins et al (2012); Whalley et al. (2012);c - Service et al. (2012);d - Demirkan et al. (2011);e - Luciano et al. (2012);f - Otowa et al. (2012);g - ISC (2009); Derks, Vorstman, Ripke, Kahn and Ophoff (2012); Ikeda et al. (2011); Levinson et al. (2012); Ripke, O'Dushlaine, et al. (2013);h - Hamshere, Langley, et al. (2013);i - Anney et al. (2012);j - Vrieze, McGue, and Iacono (2012);*

Estes trabalhos concluem que há evidências de risco poligênico em várias doenças. Apenas (Service *et al.*, 2012) não encontrou resultados significativos de associação em seu estudo.

## 2.2 Reconhecimento de Padrões

Um padrão é um objeto abstrato, tal que um conjunto de medidas descreve um objeto físico. O reconhecimento de padrões consiste em descobrir as formas dos dados utilizando algoritmos (Bishop, 2007; Duda *et al.*, 2012). É um problema de estimar a densidade das funções em um espaço dimensional e dividi-lo em regiões de classes ou categorias (Fukunaga, 2013).

Em reconhecimento de padrões, os dados podem ser representados pelo conjunto $X = (x1), (x2), x(xn)$ e $Y(y1, y2, ym)$, em que $X$ é uma matriz de informações (características) que representam uma amostra que, por sua vez, pode ser de documentos, valores de imóveis, e-mails, sequências de DNA, etc. Os valores $(x1), (x2), x(xn)$ podem ser de dois tipos: **Categóricos** ou **Contínuos**. No primeiro tipo, o atributo assume valores em um conjunto finito, valores esses que podem ser subdivididos em Nominais, nos quais não há ordem entre os valores (por exemplo, cores), e Ordinais, nos quais há ordem (por exemplo, intensidades baixa, média e alta). Os dados Contínuos são um conjunto de valores reais em que existe diferença mensurável entre os valores (Categorical, 1998).

O $Y$ denota os rótulos das amostras que podem ou não estar presentes no conjunto de dados em estudo. De acordo com a existência de $Y$, existem duas categorias principais para os algoritmos de reconhecimento de padrões. No caso dos valores de $Y$ serem desconhecidos, dizemos que se trata de algoritmos não supervisionados e o objetivo é organizar os itens. Geralmente as tarefas típicas desta categoria envolvem clusterização dos dados (em que o objetivo é particionar itens em regiões homogêneas); detecção de outliers (ruídos) (em que é verificado se a nova amostra $x$ é significativamente diferente das demais); e redução de dimensionalidade (transformando $x$ em um espaço de dimensão mais baixa).

Geralmente quando se observa os pares $(x1, y1), (x2, y2)...(xn, yn)$, trata-se de casos de algoritmos supervisionados em que o objetivo é predizer o rótulo $y$ para qualquer nova entrada $x$. Um algoritmo supervisionado é chamado de "regressão", quando $Y$ assume valores contínuos, e de "classificação", quando $Y$ assume valores discretos (Zhu *et al.*, 2005). No primeiro caso, o objetivo é predizer um valor real para cada item de um conjunto. Alguns exemplos são: predição de valores de estoque ou variação na economia. Neste tipo de problema, a penalidade para um valor incorreto depende da magnitude da diferença entre o valor real e o valor predito, em contraste com um pro-

blema de classificação, em que normalmente não existe a noção de proximidade entre as categorias. No segundo caso, associa -se uma categoria para cada item de um conjunto (*classificar*). Por exemplo, em classificação de documentos podem ser associados itens com categorias, como, por exemplo, política, negócios, esportes ou tempo. Analogamente, pode-se associar fenótipos a indivíduos, como presença ou ausência de um transtorno.

Os rótulos $Y$ podem ser binários, indicando, por exemplo, a presença ou ausência de um determinado fenótipo. Existem também aplicações em que o $Y$ tem $k > 2$ rótulos e esses são denominados problemas multiclasses, como, por exemplo, as classes de enzimas ou estágios de um câncer (Duda *et al.*, 2012). A Figura 2.2 apresenta as amostras T nas linhas, sendo as m colunas as características e a coluna $Y$ os rótulos de cada amostra. Cada tupla $Ti = (x_i1, x_i2, x_i3, x_im, y_i) = (x_i, y_i)$ representa uma amostra, também denotada por $(x_i, y_i)$, onde fica subentendido que $x_i$ e $y_i$ são vetores. A última coluna, $y_i = f(x_i)$, é a função que se tenta predizer a partir dos atributos $x_i$ (Monard e Baranauskas, 2003).

| | $X_1$ | $X_2$ | $\cdots$ | $X_m$ | $Y$ |
|---|---|---|---|---|---|
| $T_1$ | $x_{11}$ | $x_{12}$ | $\cdots$ | $x_{1m}$ | $y_1$ |
| $T_2$ | $x_{21}$ | $x_{22}$ | $\cdots$ | $x_{2m}$ | $y_2$ |
| $\vdots$ | $\vdots$ | $\vdots$ | $\ddots$ | $\vdots$ | $\vdots$ |
| $T_n$ | $x_{n1}$ | $x_{n2}$ | $\ldots$ | $x_{nm}$ | $y_n$ |

**Figura 2.2:** *Representação de amostras e classes em um problema tipico de reconhecimento de padrões*

Uma das tarefas de reconhecimento de padrões é o aprendizado de máquina, em que os algoritmos são treinados para classificar novas amostras. No caso supervisionado, o classificador é uma função $f(x)$ que mapeia um conjunto de características de um exemplo para uma classe. O objetivo do processo de aprendizado é encontrar uma função f que prediz corretamente a classe $y = f(X)$, de novos exemplos $X$ (Dietterich, 2002). Isso é realizado através da pesquisa de algum classificador $f(x)$ que seja adequado ao espaço de característica formado pelos exemplos contidos em $X$ sem que exista um super ajuste (*overfitting*) (Dietterich, 2002). O superajuste em algoritmos supervisionados acontece quando o classificador $f(x)$ se torna muito especializado ao conjunto de dados utilizado durante o seu treinamento, gerando uma generalização pobre quando aplicado a um conjunto de dados desconhecido (teste) (Hawkins, 2004; Walder, 2006). Nesses casos, o algoritmo irá apresentar bons resultados para o conjunto de treinamento, mas o mesmo não acontece com os dados de teste Di Deco *et al.* (2013); Sebastiani (2002).

O aprendizado ocorre com o que chamamos de treinamento. Aprender é adquirir e melhorar o desempenho através da experiência. Em computação, escreve-se um programa parametrizado e o processo de aprendizado consiste em encontrar o conjunto de parâmetros que melhor se aproxima da função ou do comportamento desejado (Duda *et al.*, 2012).

No entanto, um classificador $f(x)$ com boa generalização em um determinado problema não garante sucesso em qualquer tarefa que possa ser aprendida. Isso acontece porque, geralmente, um algoritmo tem um escopo limitado ou um viés e não pode ser um aprendiz universal. Não existe um algoritmo de aprendizado universalmente eficiente (Ben-David *et al.*, 2011).

Algumas das principais etapas do processo de reconhecimento de padrões com algoritmos supervisionados estão sintetizados na Figura 2.3.

1. Definir o problema: deixar evidente o que se deseja compreender sobre o domínio;
2. Adquirir e armazenar: Aquisição de um subconjunto do domínio a ser estudado;
3. Pré-Processar: etapa que envolve grande quantidade de conhecimento sobre o domínio. Tratamento de valores faltantes, remoção de ruídos;

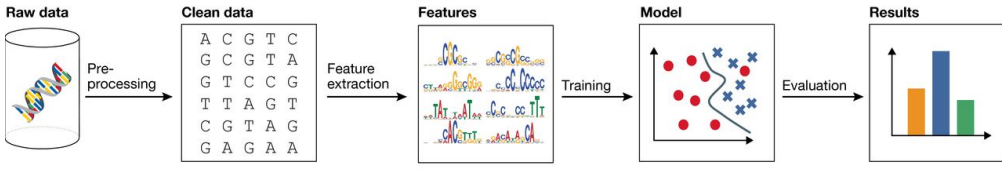


**Figura 2.3:** *Principais etapas durante o processo de aprendizado de máquina. Figura adaptada de "Supervised machine learning: A review of classification techniques", 2007*

4. Seleção ou extração de características: muitas vezes é necessário trabalhar com a redução da dimensionalidade, selecionando ou extraindo as características. Essa etapa é importante pois será a partir deste conjunto de valores que o algoritmo irá generalizar o domínio;
5. Transformação dos dados: discretização, normalização;
6. Separar em conjuntos de treinamento e teste;
7. Treinar e Validar;
8. Analisar e interpretar os resultados.

Geralmente não é possível encontrar uma classificação livre de erros, mas o objetivo é produzir um classificador que generalize bem os dados. Alguns dos fatores que impedem o sucesso do modelo proposto são dificuldades inerentes do problema, como sobreposição das classes, pequeno número de amostras, alto número de características, complexidade da classificação (quadrática versus linear), presença de ruídos e parâmetros inadequados ao algoritmo de classificação selecionado (Dua e Du, 2016).

Um problema de aprendizado é frequentemente um problema de grande dimensão com um conjunto de treinamento limitado, como ocorre frequentemente em bioinformática: existem poucas amostras disponíveis e muitas características (Saeys *et al.*, 2007). O problema é que nessas circunstâncias, os classificadores podem generalizar pobremente. Uma forma de contornar isso é usando algoritmos para seleção de características, reduzindo a dimensionalidade do problema e encontrando os melhores conjuntos de características para classificação ou regressão (Guyon e Elisseeff, 2003).

Além disso, é importante utilizar técnicas de validação com conjuntos de treino e teste para evitar resultados superotimistas. Como na maioria das vezes os dados que temos disponíveis são limitados, procuramos formas de dividi-los em um conjunto para treinamento e um conjunto para teste. Dessa forma, o classificador será testado com dados que ele não usou durante o aprendizado e pode-se evitar o superajuste (Arlot *et al.*, 2010). Esse tipo de abordagem baseia-se no fato de que os dados reais têm uma distribuição $D$; ao selecionamos um conjunto para classificação, temos uma distribuição $D'$ e este conjunto será dividido em treinamento e teste que, por sua vez, também seguem a mesma distribuição. Assim, simula-se o processo de amostragem que ocorre no mundo real, assumindo que $D'$ representa o mundo real (Monard e Baranauskas, 2003).

Os tipos mais comuns de validação são (Monard e Baranauskas, 2003):

**Holdout:** reserva um conjunto para teste (p= 1/3), e os demais para treinamento (p = 2/3). Não há sobreposição de amostras nos conjuntos de teste e treino. O problema desse método é que não se usa o conjunto todo para fazer o treinamento e o resultado é dependente da escolha do particionamento em treino/teste. As instâncias escolhidas para teste podem ser muito fáceis ou muito difíceis de classificar. Além disso, seria interessante usar as amostras do teste para o treinamento.

**Aleatório:** um conjunto de amostras é induzido (conjunto de treinamento) e o erro final é calculado como sendo a média dos erros de todas as hipóteses induzidas e calculados em conjuntos de teste independentes e extraídos aleatoriamente.

**Leave-one-out:** é um tipo de validação cruzada em que o conjunto de treinamento é n-1 e o processo é repetido n vezes, cada vez com o elemento que não foi utilizado no treinamento. Indicado para amostras pequenas.

**k-validação-cruzada:** separa o dado em k partes(chamadas de folds) de mesmo tamanho (ou quase). Uma parte é para treino e as outras são para teste, o processo é repetido k vezes. Geralmente o k tem tamanho 10.

**k-validação estratificada:** é semelhante a k-validação cruzada, mas leva em consideração a distribuição das classes para gerar os folds (os folds mantêm a proporção das classes). Dessa forma, se a classe positiva está presente em 50% dos casos, isso será mantido nos folds gerados.

**Bootstrap:** usa amostragem com reposição para formar o conjunto de treinamento. Os que não são usados no treino serão usados como teste.

A Tabela 2.2 resume os métodos de particionamento:

**Tabela 2.2:** *Resumo dos métodos de particionamento. Extraído de (Monard e Baranauskas, 2003)*

| | Holdout | Aleatório | LOO | k.Fold.CV | k.Fold.Stratified.CV | Bootstrap |
|---|---|---|---|---|---|---|
| Treinamento | pn | t | n-1 | n(k-1)/k | n(k-1)/k | n |
| Teste | (1-p)n | n-t | 1 | n/k | n/k | n-t |
| Iterações | 1 | aprox. 20 | n | k | k | aprox. 200 |
| Reposição | não | não | não | não | não | sim |
| Prevalência de Classe | não | não | não | não | sim | sim/não |

| Método | Quando usar | Prós | Contras |
|---|---|---|---|
| Holdout | quando houver dados suficientes para que um subconjunto possa ser mantido fora, e esse subconjunto é grande o suficiente para garantir estimativas estatísticas confiáveis | conjuntos de teste e treinamento independentes | reduz o número de amostras para o teste e treinamento |
| Leave one out | quando há pouquíssimas amostras | performance estimada sem viés | alta variância |
| Kfold | É útil quando o conjunto de dados de treinamento é tão pequeno que não se pode dar ao luxo de manter parte dos dados apenas para fins de validação. | estimação acurada da performance | pequenas amostras de estimação da performance; sobreposição dos dados de treinamento; elevados erros do tipo I; sub estimada variância da performance ou super estimado os graus de liberdade para comparacao |
| kfold estratificado | quando as classes são desbalanceadas | remove viés | diminui o n |
| bootstrap | quando o conjunto de treinamento é pequeno [(aproxima da distribuição) estimar um intervalo de confiança] | remove overfitting | |
| Ressubstituição | | simples | super ajuste |

**Tabela 2.3:** *Métodos de validação e quando usar. Adaptado de http://leitang.net/papers/ency-cross-validation.pdf*

### 2.2.1 Seleção de Características

Além do particionamento utilizado para validar o modelo, outras técnicas colaboram para a construção de um modelo de classificação mais acurado. A redução de dimensionalidade é conhecida por ajudar a remover ruídos e informações redundantes, facilitando a visualização e o entendimento sobre o problema e reduzindo a quantidade de recursos para armazenamento e processamento (Guyon e Elisseeff, 2003). Além disso, saber quais características são relevantes reduz o trabalho de coleta das mesmas (Hira e Gillies, 2015; Sutha e Tamilselvi, 2015).

As técnicas utilizadas podem ser divididas em extração e seleção de características. A primeira reduz a dimensionalidade criando novas características, geralmente combinando as originais; exemplos destas técnicas incluem Análise de Componentes Principais (PCA), Análise Discriminante Linear (LDA) e Análise de Correlação Canônica (CCA). Por outro lado, as abordagens de seleção de características visam a selecionar um pequeno subconjunto de recursos que minimizam a redundância e maximizam a relevância para o objetivo, como os rótulos de classe na classificação.

Resumidamente, os procedimentos em seleção de características consistem em gerar um subconjunto a partir do conjunto inicial de caraterísticas, avaliar para incluir ou descartar características de acordo com algum critério. Esse fluxo está representado na Figura 2.4 (Dash e Liu, 1997; Kumar e Minz, 2014).

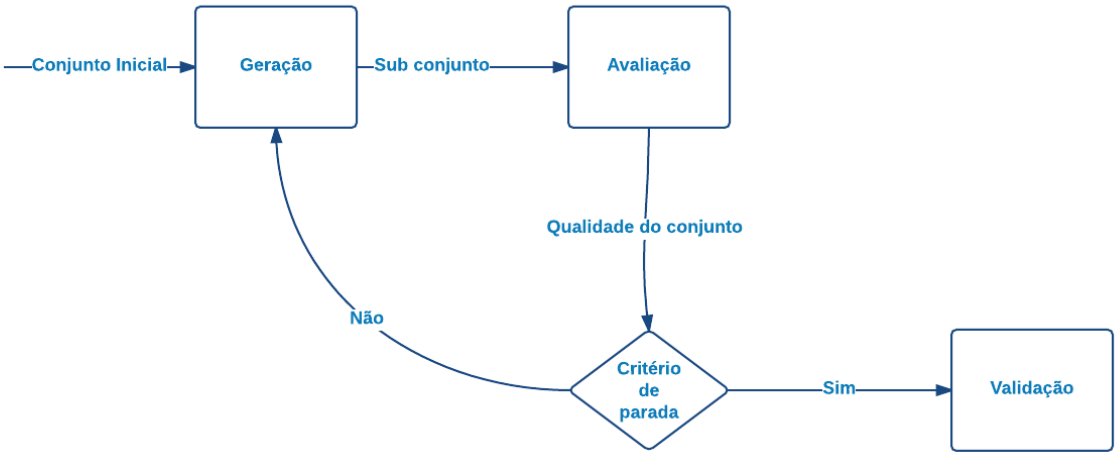


**Figura 2.4:** *Um conjunto de características é a entrada de um processo que seleciona um subconjunto de caracteristicas que serão avaliadas e, de acordo com um critério determinado, serão usadas para a classificação (validação) ou serão descartadas.*

As técnicas de redução de dimensionalidade, em geral, melhoram a performance dos algoritmos de aprendizado de máquina, fazendo com que o modelo generalize melhor, reduzam a complexidade computacional e diminuam a quantidade de dados a serem armazenados (Baharudin *et al.*, 2010). A melhora da performance se dá porque a acurácia preditiva dos algoritmos de aprendizado é reduzida na presença de características irrelevantes (Wang *et al.*, 2016). Os autores Koller e Sahami (1996) mostraram que a distribuição de características relevantes para a tarefa são ofuscadas por características irrelevantes ou redundantes. Não usar essas características pode aumentar a acurácia e diminuir a complexidade do classificador. Podem existir vários conjuntos de características que são igualmente bons, o objetivo é escolher o conjunto de características altamente discriminantes, ou seja, as que melhor descrevem aqueles dados. A relevância das características é avaliada no modo em que elas distinguem as classes dos dados.

A seleção de características encontra um subconjunto de características do conjunto original sem nenhuma transformação, e mantém os significados físicos delas. Nesse sentido, a seleção de características é superior em termos de legibilidade e interpretabilidade, o que é muito útil em aplicações como encontrar genes relevantes para uma doença específica ou uma análise de sentimento

léxica. Dado que o foco deste trabalho é seleção de características, vamos dar mais atenção a essa técnica. Os principais desafios em Seleção de Características são Silva (2011):

- Espaço de busca: Para encontrar o melhor conjunto ou os melhores conjuntos de características, é necessário testar exaustivamente todos os possíveis M subconjuntos das N características totais, esta explosão combinatorial para uma busca exaustiva leva a uma carga computacional que cresce exponencialmente com o aumento do número de características. Em termos práticos, se há mais de 30 características é inviável uma busca exaustiva (Wang *et al.*, 2016).

- Eliminação de características irrelevantes: em um grande conjunto de dados é possível que existam informações que são irrelevantes. A existência dessas características pode comprometer a representação de características relevantes

- Eliminação de redundâncias: encontrar as características redundantes irá diminuir a complexidade do modelo e o tempo de computação.

- Manutenção das características interagentes: existem características que são inúteis quando avaliadas individualmente, mas quando se considera a interação delas com outras a sua remoção pode levar à perda do significado de um conjunto. Isso se chama irredutibilidade, e significa que não se deve avaliar subconjuntos de características interagentes por partes.

O mecanismo para encontrar o subconjunto de características pode ser descrito como uma busca em um espaço de hipóteses (conjunto de soluções) (Molina *et al.*, 2002). Os algoritmos podem ser agrupados dependendo do tipo de busca que executam; são elas:

**Exponencial:** garante a melhor solução, porém, dado que testa todas as possibilidades, muitas vezes é inviável.

**Sequencial:** consiste em adicionar ou remover características usando um procedimento de busca local com complexidade polinomial. Os estudos mais comuns são a seleção sequencial para frente ou para trás (FSS e BSS) Naghibi *et al.* (2015); Tsymbal *et al.* (2003).

**Aleatória:** usa a aleatoriedade para selecionar características. A ideia é evitar que o algoritmo fique preso em ótimos locais.

Uma categorização muito utilizada para os métodos de seleção de características supervisionados são: **Filtro**, **Wrapper** e **Embutidos** (Jović *et al.*, 2015).

O Filtro separa o passo de seleção de características do passo de aprendizado do classificador, assim o bias do aprendizado não interage com o bias do algoritmo de seleção de características. Esse tipo de algoritmo depende de medidas gerais dos dados, como distância, consistência, dependência, informação e correlação. Ele é normalmente dividido em duas etapas: na primeira é feito o ranqueamento de cada característica e na segunda etapa as características com melhor rank são utilizadas em um classificador. A maior desvantagem dessa técnica é ignorar os efeitos das características selecionadas na performance do algoritmo de classificação. Relief, Fisher Score e métodos baseados em Ganho de Informação são os mais representativos do modelo Filtro.

O Wrapper usa a acurácia de um algoritmo de aprendizado para determinar a qualidade das características selecionadas. Desse modo, estes métodos são caros para rodar quando existe um grande número de características. No entanto, obtém uma acurácia melhor do que os modelos Filtro, uma vez que seleciona características que maximizam a qualidade do classificador .

O Embutido foi proposto para preencher o gap entre o Filtro e o Wrapper. Primeiro ele incorpora critérios estatísticos, assim como o modelo Filtro, para selecionar vários candidatos com uma dada cardinalidade. Depois, ele escolhe o subconjunto com a acurácia mais alta. Assim, o Embutido aproveita a boa acurácia do Wrapper e a eficiência do Filtro. Ele avalia a seleção em tempo de aprendizado; dito de outro modo, ele está ajustando o modelo e selecionando as características simultaneamente.

Neste trabalho utilizamos os seguintes métodos de seleção de características: FBST, Escore de Fisher, Regularização Lasso e SFFS.

### FBST

Os testes de significância estatística servem como uma regra decisória para aceitar ou rejeitar uma hipótese. Ao aceitar uma hipótese acredita-se que aquele resultado não ocorreu por acaso. As hipóteses são sempre em relação às amostras, mais especificamente, aos parâmetros da população como média e desvio padrão.

Quando se formula uma hipótese em relação a uma determinada característica de uma população, a amostra pode: pertencer à população de origem, portanto as diferenças observadas são decorrentes de flutuações biológicas normais ou não pertencer a essa população e as diferenças encontradas representam um efeito real, não podendo ser atribuídas ao acaso. No primeiro caso, diz-se que os valores encontrados "não são estatisticamente significativos"e no segundo "são estatisticamente significativos". Para se testar algo é necessário estabelecer uma hipótese nula e uma hipótese alternativa, sendo ambas antagônicas. A hipótese nula (comumente designada por H0) é uma hipótese tida como verdadeira até que provas estatísticas indiquem o contrário. A hipótese alternativa (comumente designada por H1 ) deve ser contrária à hipótese nula.

A metodologia FBST baseia-se na avaliação de uma evidência bayesiana ($ev(H|X)$ ou e-value) a favor da hipótese precisa H0 sob a obtenção da amostra X. No caso do FBST, o método é considerado totalmente Bayesiano por requerer apenas o conhecimento do espaço paramétrico representado pela sua distribuição a posteriori (Pereira e Stern, 2001).

Enquanto o p-valor é a probabilidade de cometer um erro, o e-valor é o número esperado de vezes que dado escore aparece aleatoriamente em um banco de dados de determinado tamanho. Assim, quanto menor o e-valor, melhor é o resultado. Uma maneira simples de calcular o e-valor é multiplicar o p-valor vezes o tamanho do banco de dados.

Utilizamos o FBST para testar a independência entre os SNPs e as classes caso e controle. As hipóteses a serem testadas são: H0: as variáveis são independentes e H1: as variáveis não são independentes (Pereira *et al.*, 2008).

### FISHER

Este método, também chamado de F-score, calcula a distância entre as médias das distribuições de duas classes em relação às suas variâncias. Quanto maior o F-score, mais discriminativa e relevante é a característica (Villela, 2011). A ideia chave é encontrar um subconjunto, tal que no espaço de características gerado, a distância entre os pontos da mesma classe seja a menor possível. Em particular, dado o conjunto selecionado de m características, a matriz de dados de entrada X $\in Rdxn$ seja reduzida para $Z \in Rmxn$ (Gu *et al.*, 2012).

O escore de Fisher é dado pela Fórmula :

$$F(x^j) = \frac{\sum_k^c n_k(\mu_k^j - \mu^j)^2}{\sum_k^c n_k(\sigma_k^j)^2} \tag{2.2}$$

Em que $\mu_k$ e $\sigma_k$ são a média e a variância da i ésima característica da classe $c$ e $n$ é o número de instâncias da classe $c$

O escore de Fisher (F-score) é um método muito utilizado dada a eficiência e a rapidez com que pode ser computado. No entanto, ele avalia as características individualmente, de forma que não é capaz de verificar a redundância das características, resultando em um subconjunto subótimo.

### Regularização Lasso

Usualmente, as técnicas de regularização utilizam alguma informação obtida a priori sobre o problema. Essa informação pode ser dada, por exemplo, por uma imposição de suavidade na solução ou pode ser obtida por meio da informação estrutural dos dados a serem tratados no problema.

O objetivo da regularização é reduzir o sobreajuste adicionando uma penalidade à complexidade para a função de perda (Bickel *et al.*, 2006).

A Regularização L1 para regressão linear com erros Gaussianos independentes (e uma verossimilhança com erros quadrado) é o tipo mais popular do Lasso (least absolute shrinkage and selection operator). É muito útil quando o número de entrada é maior que o número de amostras (p >N), pois é um método utilizado para reduzir os efeitos dos atributos que não contribuem para a identificação do atributo - meta (ou variável resposta), reduzindo seus coeficientes para zero e excluindo-os do modelo (Tibshirani *et al.*, 1997).

Um valor relativamente pequeno de x leva a uma solução que está próxima do estimador de mínimos quadrados. À medida em que se aumenta o valor de x, um coeficiente por vez é levado a zero, ou seja, o parâmetro x controla os graus de liberdade da estimação. Assim, aumentando o valor de x, pode-se controlar o número de variáveis que serão incluídas no modelo (Vidaurre *et al.*, 2013).

A Regularização L2, também chamada de Ridge evita superajuste ao reduzir (ao seja, impor uma penalidade) os parâmetros. Ele reduz todos os parâmetros pelas mesmas proporções, mas não elimina nenhum deles e não é um método de seleção de variáveis. A Ridge Regression é um método de regularização do modelo que tem como principal objetivo suavizar atributos que sejam relacionados uns aos outros e que aumentam o ruído no modelo ( multicolinearidade). Com a retirada de determinados atributos do modelo, o mesmo converge para um resultado muito mais estável em que, com a redução desses atributos, a redução em termos de acurácia do modelo se mantém inalterada. O mecanismo algorítmico que faz isso é através de um mecanismo de penalização que coloca um viés e que vai reduzindo os valores beta até não zero. Os atributos que contribuem menos para o poder preditivo do modelo são levados para a irrelevância usando esse mecanismo de penalização do viés (Tibshirani, 2011; Zou e Hastie, 2005).

Já a Regularização Lasso tem o mesmo mecanismo de penalização dos coeficientes com um alto grau de correlação entre si, mas que usa o mecanismo de penalizar os coeficientes de acordo com o seu valor absoluto (soma dos valores dos estimadores), usando o mecanismo de minimizar o erro quadrático. Isso é feito através da penalização do coeficiente até que o mesmo convirja para zero, o que naturalmente vai eliminar o atributo e reduzir a dimensionalidade do modelo (Tibshirani, 2011; Zou e Hastie, 2005).

### SFFS

A estratégia *Forward* começa sem nenhuma variável no modelo e adiciona variáveis a cada passo, a estratégia *Backward* faz o caminho oposto; incorpora inicialmente todas as variáveis e depois, por etapas, cada uma pode ser ou não eliminada. Essas duas abordagens são a base para os algoritmos de seleção SFS (*Sequential Forward Selection*) e SBS (*Sequential Backward Selection*)

O algoritmo SFFS (*(Sequential Forward Floating Selection*) é a melhoria das estratégias SFS e do SBS. O algoritmo SFS é uma abordagem que se inicia com um conjunto vazio e, conforme o algoritmo é executado, o melhor atributo $x$ é inserido no subconjunto, resultante do máximo valor da função critério, quando combinado com os atributos que já foram selecionados. Esse método tem menor custo computacional quando se deseja obter conjuntos pequenos em relação ao total de atributos (Jain e Zongker, 1997). Uma vez que um atributo tenha sido selecionado, ele não pode ser descartado do subconjunto, o que pode provocar o chamado efeito *nesting*.

O algoritmo SBS, assim como o método SFS, pode sofre do efeito *nesting*. Esse algoritmo inicia com um conjunto de atributos completo e, nas iterações do algoritmo, remove-se o atributo que maximiza a função critério. O método SBS tem menor custo computacional quando se deseja obter conjuntos grandes em relação ao total de atributos (de Andrade Oliveira *et al.*).

O SFFS é a junção dos dois algoritmos anteriores. Ele inicia com um subconjunto vazio, itera procurando características mais significativas no conjunto de características. Adiciona uma característica ao conjunto vazio e então repetidamente deleta a característica menos significativa. Depois de cada iteração os resultados são comparados aos resultados do passo anterior. Se a performance for melhor, a característica é adicionada ao subconjunto (Lopes, 2011; Shirbani e Soltanian Zadeh, 2013).

Na Figura 2.5 podemos entender melhor a execução do algoritmo:

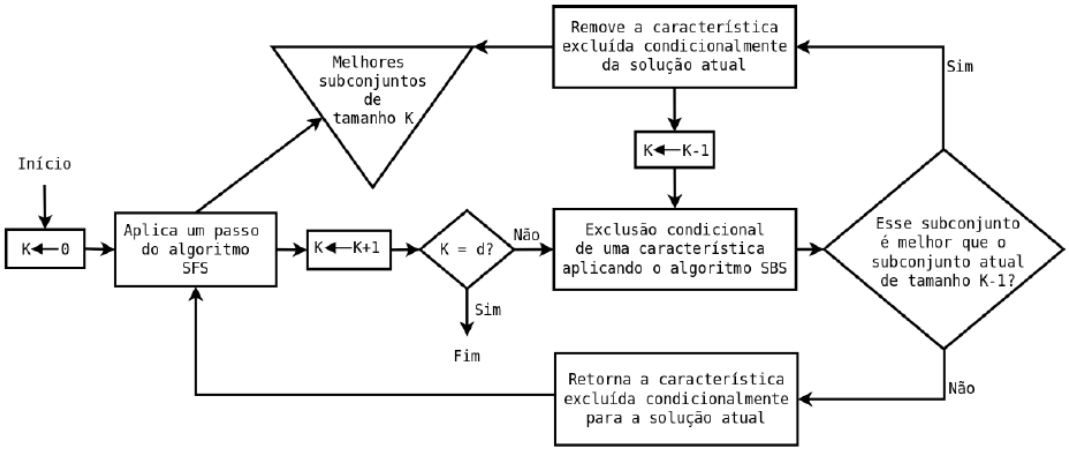


**Figura 2.5:** *O primeiro passo é aplicar o SFS, o qual adiciona uma característica por vez baseado na função objetivo. O SFFS adiciona outro passo, que é a exclusão de uma característica por vez , passo do SBS, do subconjunto obtido no primeiro passo e avalia o novo subconjunto. Se a exclusão leva a uma melhor performance, então aquela característica é removida e volta para o primeiro passo . O processo acontece até que o número desejado de features seja atingido (Chandrashekar e Sahin, 2014). Figura extraída de: "Redes complexas de expressão gênica: síntese, identificação, análise e aplicações, 2011"*

A função objetivo utilizada nesta implementação é a acurácia de um classificador KNN. O algoritmo utilizado está disponível em DimReduction.

### 2.2.2 Dimensão da amostra e das características

O tamanho da amostra é outro fator que tem influência na acurácia dos classificadores, principalmente em relação ao número de características que poderão ser utilizadas. Dessa forma, procuramos um número ótimo de características. Isso pode ser feito utilizando um banco de dados de "optimal-feature-size curves"como referência para escolher o número de features adequado. Alguns algoritmos, como SVM, são mais robustos a um grande número de features, diferente de um classificador como LDA, que não tem essa robustez (Hua *et al.*, 2005).

Para um conjunto de dados finito com pouca ou nenhuma informação a priori, a proporção entre o tamanho da amostra e a dimensionalidade deve ser o maior possível para suprimir as avaliações otimistas do desempenho do classificador. Para um dado tamanho de amostra usado no treinamento de um classificador, existe um tamanho ótimo de característica e complexidade de quantização (este último refere-se ao número de intervalos em que uma dimensão é dividida). Esse resultado é verdadeiro tanto para problemas de duas classes quanto para problemas de várias classes. A proporção do tamanho da amostra para a dimensionalidade varia de forma inversamente proporcional à quantidade de conhecimento disponível sobre as densidades condicionais da classe (Murthy e Salzberg, 1995).

Determinar o número ótimo de características é complicado pelo fato de que, se temos D potenciais características, então existem C(D,d) conjuntos de características de tamanho d e todos estes precisam ser considerados para assegurar o conjunto ótimo. (Cover e Van Campenhout, 1977; Hua *et al.*, 2005).

Pode-se afirmar que o tamanho ótimo do conjunto de características é relativo ao classificador e a distribuição dos dados e o número de características pode influenciar a performance do classificador. Dessa forma, não é aconselhável utilizar regras prontas para definir o número de características ou amostras para um problema específico.

### 2.2.3 Balanceamento de Classes

Além da seleção de características relevantes e do número de amostras disponíveis, outro fator que influencia a performance dos métodos de aprendizado de máquina é o desbalanceamento de

classes, no qual os exemplos de treinamento de uma das classes estão em maior número do que na outra. Nesta situação os sistemas de aprendizado têm dificuldades para aprender os conceitos relacionados com a classe minoritária (Bekkar e Alitouche, 2013; Chawla, 2005).

A comunidade científica concorda em relação ao fato de que o problema de desbalanceamento entre classes é o maior obstáculo no processo de aprendizado. Mas existem alguns domínios em que os algoritmos são capazes de aprender mesmo com classes desbalanceadas, o que mostra que este problema não é o único responsável pela perda de desempenho dos algoritmos (He *et al.*, 2009).

As metodologias para balanceamento mais utilizadas são: superamostragem e subamostragem. No primeiro caso, o número de amostras da classe minoritária é aumentado e no segundo caso o número de amostras da classe majoritária é reduzido (Anyfantis *et al.*, 2007). A maneira mais simples de alterar o número de amostras é de maneira aleatória. No entanto, isso pode acarretar a remoção de exemplos importantes ou levar ao superajuste. Algumas técnicas que podem substituir o modo aleatório são: Links Tomek, Condensed Nearest Neighbor (CNN) , Neighborhood Cleaning Rule (NCL) (Chawla, 2005).

Alguns algoritmos conseguem ter um bom desempenho apesar do desbalanceamento, como é o caso do SVM. No caso do SVM acontece que apenas os SV (vetores de suporte) são usados para classificação e a maioria das amostras está longe da superfície de decisão e pode ser removida sem afetar a classificação. No entanto, o classificador pode ser sensível a classes altamente desbalanceadas, fazendo com que a performance do algoritmo decaía e gere um alto número de falsos negativos (Tang *et al.*, 2009).

No caso do Random Forest, existe uma grande probabilidade de que um conjunto do bootstrap contenha poucas ou nenhuma informação da classe minoritária, o que leva a árvore de decisão a fazer uma predição errada (Bekkar e Alitouche, 2013).

Portanto é necessário um tempo dedicado à elaboração de metodologias para o pré-processamento e transformação dos dados, o que resulta em uma representação do dado que pode efetivamente colaborar com o método classificador (Bengio *et al.*, 2013; Liu, 2004).

### 2.2.4 Escolha do modelo para classificação

O modelo de classificação é uma estrutura e interpretação correspondente que resume um conjunto de dados para descrição ou predição.

A seleção do modelo consiste em revolver o problema de viés versus variância: quando se escolhe um modelo muito simples será produzido um erro alto devido um ajuste pobre (*underfitting* ) e quando o modelo escolhido é muito complexo, pode acontecer o sobreajuste (*overfitting*).

Uma maneira de se escolher o melhor modelo é comparar vários modelos com validação cruzada e então decidir o que melhor se adequa ao problema. O princípio da navalha de Ocam tem sido um guia para a seleção do modelo. Ele propõe a seleção de modelos que melhor se ajustem aos dados entre complexidades concorrentes. É o princípio da parcimônia: um modelo deve ser simples o suficiente para a computação eficiente, e complexo o suficiente para ser capaz de capturar dados específicos. Além disso, argumenta-se que o princípio da Navalha de Occam ajuda a escolher um modelo com boa generalização: um modelo complexo é susceptível de sobreajustar os dados, enquanto um modelo mais simples pode suavizar as características ruidosas da distribuição fonte. Ou seja, esse princípio propõe a escolha do modelo que melhor se ajusta aos dados e que seja o mais simples possível (Koller e Sahami, 1996).

### 2.2.5 SNPs e aprendizado de máquinas

O uso de aprendizado de máquina vem sendo testado no estudo de associação poligênica e os resultados têm sido promissores (Uppu *et al.*, 2016). Uma das vantagens de usar AM é o fato de que métodos não paramétricos lidam melhor com muitos SNPs e outras variáveis, sendo avaliadas em amostras pequenas. Outra vantagem é lidarem com interações de ordem mais altas (Ziegler *et al.*, 2008). Os algoritmos mais usados em aprendizado de máquina e análise de SNPs são Random Forest, SVM, Naive Bayes e Regressão.

### Uso de Random Forest

O *Random Forest* consiste da combinação de um conjunto de árvores de decisão (*ensemble*), em que cada árvore realiza uma classificação e cada classificação recebe um voto (*bagging*), a classe da amostra é dada de acordo com o maior número de votos. No caso de uma regressão, ao invés do voto, é calculada a média das saídas das diferentes árvores (Breiman, 2001).

O *Random Forest* funciona da seguinte maneira: para cada árvore da floresta, é criado um conjunto aleatório de treinamento proveniente do conjunto de amostras com reposição. Esse conjunto de treinamento pode ser chamado de bootstrap e contém cerca de 2/3 das amostras totais. O restante das amostras é chamado de *Out-of-bag* . A árvore cresce por particionamento recursivo. Para cada árvore as variáveis são selecionadas aleatoriamente do conjunto de todas as variáveis disponíveis e são avaliadas de acordo com a capacidade de separar os dados. A variável que resultar no maior decrescimento de impureza é escolhida para separar as amostras de cada nó pai, começando do nó raiz. A impureza é medida pelo índice Gini. Após o treinamento com o método bootstrap, o processo de separação é repetido e termina quando os nós finais são puros (contêm apenas amostras pertencentes à mesma classe) ou contêm um número específico de amostras. A árvore cresce sem poda até que tenha crescido inteiramente, chegando ao final do processo de treinamento. A performance da árvore será calculada de acordo com as amostras out-of-bag (Touw *et al.*, 2012).

Jiang *et al.* (2009) aplicaram *Random Forest* combinado com SWFS (sliding window sequential forward feature selection) em dados de GWAS de Degeneração Macular para obter um conjunto de SNPs que reduzem os erros de classificação. O foco foi obter a contribuição de cada SNP para a classificação. Com base nessa contribuição foi utilizado o SWFS, proposto para determinar um subconjunto de SNPs que estejam mais associados à doença. Segundo os autores, este tipo de estudo complementa as técnicas estatísticas existentes e colabora para a compreensão sobre o fenótipo.

Schwarz *et al.* (2010) desenvolveram o Random Jungle (RJ), que é um pacote que analisa eficientemente dados de grande dimensão. RJ elimina as variáveis utilizando o método *backward*. Nesta abordagem, é associado um escore a cada SNP e a partir desse escore os SNPs são selecionados. A principal contribuição desse trabalho é a eficiência em analisar dados de grande dimensão como GWAS.

Goldstein *et al.* (2011) investigaram um conjunto de 326 SNPs em 1343 casos de esclerose múltipla e 1379 controles. Usando *Random Forest* eles encontraram 8 SNPs intrônicos preditores de esclerose múltipla. A codificação dos SNPs foi baseada no número de alelos de menor frequência (MAF). O método foi validado com testes de associação. Assim como esses, vários outros trabalhos vêm utilizando RF com sucesso.

### Uso de SVM

O SVM (do inglês: *Support Vector Machine*) foi embasado na Teoria do Aprendizado Estatístico (TAE). Segundo essa teoria, pode-se dizer o seguinte sobre a distribuição conjunta ($P$) (Xuegong (2000)):

- Não é feita nenhuma suposição sobre a distribuição de probabilidades $P$, logo pode ser qualquer distribuição.
- Os rótulos do conjunto de treinamento podem ser não determinísticos devido a ruídos e sobreposição de classes, portanto devemos encontrar probabilidades condicionais. No caso de um pequeno ruído nos rótulos, essas probabilidades condicionais são próximas de 0 ou de 1; no caso de um grande ruído, essas probabilidades podem se aproximar de 0.5, dificultando o aprendizado.
- Assume-se que as amostras são selecionadas de maneira aleatória.
- A distribuição de $P$ é fixa. Não há alteração da distribuição $P$ ao longo do tempo
- A distribuição $P$ é desconhecida no momento do aprendizado, caso contrário o aprendizado seria trivial, reduzido a uma regressão.

A margem, por sua vez, é outro conceito fundamental. Ela é definida como a menor distância existente entre qualquer ponto do conjunto de treinamento e a reta. O SVM é um problema de otimização convexa que consiste em maximizar a margem, ou seja, encontrar o melhor hiperplano que separa as classes (Gaspar *et al.*, 2012).

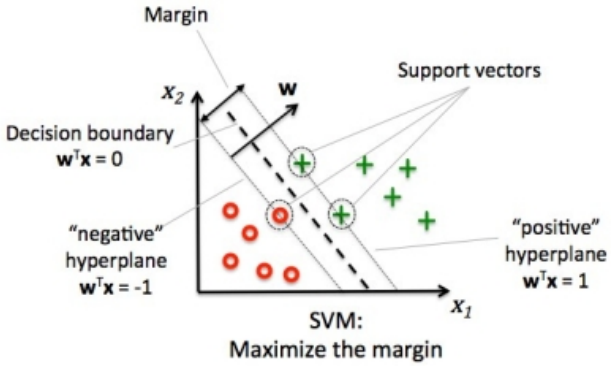


**Figura 2.6:** *Figura extraída de Python machine learning, 2015*

A Figura 2.6 ilustra o conceito de margem e de vetores de suporte.

Özgür *et al.* (2008) propuseram uma metodologia combinando mineração de texto e análise de redes para extrair os SNPs relacionados a doenças. Foi coletada uma lista com genes já relacionados à doença e construída uma rede de interação em que foi usado SVM para extrair as interações entre os genes e seus vizinhos. Para ordenar os genes da rede de acordo com sua relevância para a doença foram utilizadas medidas de topologia de redes. A abordagem foi avaliada para câncer de próstata e 95% dos 20 genes com melhor pontuação foram validados. Shen *et al.* (2012) propuseram um modelo para detectar SNPs associados com doenças, feito em duas etapas. Na primeira, os candidatos são selecionados usando SVM com penalidade L1. Na segunda etapa, são calculados p-valores para os SNPs candidatos identificados na primeira fase, usando uma regressão. A abordagem foi utilizada em dados de pacientes com Parkinson, com 540 casos e controles, com 408,000 SNPs. SNPs com p-valor maiores que 0.001 foram excluídos. Um subconjunto de 401 SNPs foi obtido e, destes, a interação entre dois obteve p-valor menor que 0.05.

Wei *et al.* (2009), aplicaram SVM em um conjunto de dados de GWAS para casos com diabetes tipo 1. O modelo foi testado em 1008 casos e 1000 controles, genotipados na plataforma Illumina e em 1529 casos e 1458 controles genotipados na plataforma Affymetrix, resultado em um AUC de 0.84% em ambos os conjuntos de teste. O estudo conclui que uma melhor predição de risco pode ser feita utilizando algoritmos que levem em consideração a interação entre um grande conjunto de marcadores. Chen *et al.* (2008), propuseram uma abordagem combinando SVM com otimização combinatória para quatro modelos: Recursive Feature Elimination (SVM-RFE), Recursive Feature Addition (SVM-RFA), (SVM-local) e Algoritmo Genético (SVM-GA). Mesmo que o custo computacional do método seja elevado, os resultados demonstraram uma forte capacidade de identificação de interações de SNPs, com uma menor ocorrência de excesso de ajuste.

### Regressão

O principal objetivo dessa técnica é obter uma equação que explique satisfatoriamente a relação entre uma variável resposta e uma ou mais variáveis explicativas, possibilitando fazer predição de valores da variável de interesse (Golberg e Cho, 2004).

A regressão consiste em aproximar uma função contínua a partir de um conjunto de exemplos composto de pontos. Tal método pode ser aplicado na previsão de preços de um produto, valores de ações no mercado de ações e prever a ocorrência de um determinado evento (regressão logística) (Golberg e Cho, 2004).

Purcell *et al.* (2007) desenvolveu uma ferramenta para análises de associação genômica que provê testes de regressão logística assumindo um modelo alélico para os efeitos principais e intera-

ções. Schwender e Ickstadt (2008) desenvolveram uma técnica que combina regressão logística com bootstrap, melhorando a seleção de características e a interpretação dos SNPs selecionados.

Listgarten *et al.* (2004) compara algoritmos de classificação. Eles analisaram um conjunto de 98 SNPs distribuídos em 45 genes com potencial relevância para o câncer de mama. Os autores usaram SVM, árvores de decisão e Naive Bayes e identificaram um conjunto de três SNPs como discriminantes do câncer de mama. Tanto SVM, quanto Naive Bayes e árvores de decisão tiveram um resultado similar em torno de 69% de força preditiva (Listgarten *et al.*, 2004).

Guo *et al.* (2016) analisaram dados de 3940 casos de anorexia nervosa e 9266 controles, utilizando regressão logística com penalização Lasso e obtiveram uma AUC de 0.693, enquanto com SVM e Gradient Boosted Trees tiveram uma AUC de 0.691 e 0.623 respectivamente. O estudo mostra que utilizar conjuntos de dados maiores são requisitos para otimizar os modelos de aprendizado de máquina e, portanto, obter um alto desempenho. Esse é um dos poucos estudos na área de psiquiatria (Guo *et al.*, 2016).

# Capítulo 3

# Metodologia

Este trabalho consiste nas seguintes etapas:

1. Controle de Qualidade das amostras
2. Mapeamento dos SNPs sequenciados:

   Estão em qual região do genoma?

   Estão em genes?

   Em regiões reguladoras?

   Relacionados a elementos regulatórios?

   Qual o estado da cromatina?

   Já foram associados a algum transtorno psiquiátrico?
3. Realizar a seleção de características:

   dadas as informações biológicas e as técnicas matemáticas, escolher subconjuntos de SNPs para caracterizar as amostras
4. Encontrar um modelo e classificar

A Figura 3.1 resume o processo que foi realizado neste trabalho:

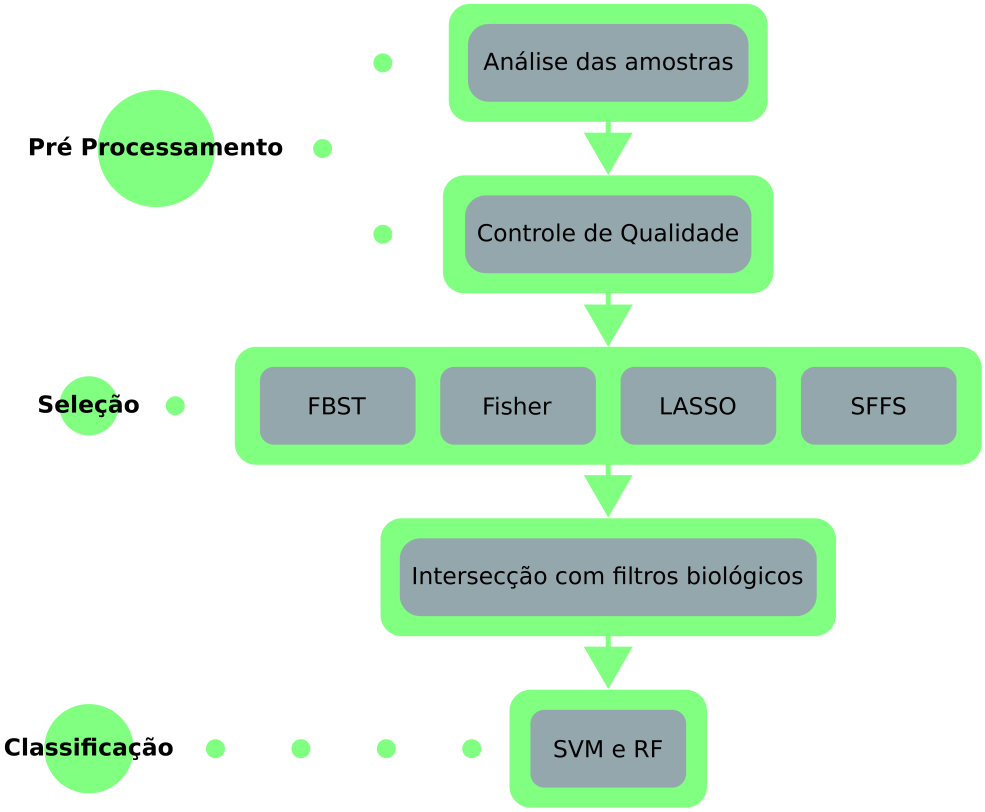


**Figura 3.1:** *Fluxo de desenvolvimento do trabalho*

## 3.1 Amostras

Este trabalho é parte do projeto do INPD, sub projeto: "Coorte de Escolares de Alto Risco para o Desenvolvimento de Psicopatologia e Resiliência na Infância - o estudo PREVENÇÃO. O projeto contou com 4 etapas específicas (triagem, etapa domiciliar, etapa escolar e etapa de neuroimagem) em inicialmente 3 fases (linha de base, seguimento de 3 anos e seguimento de 6 anos). Na etapa de triagem foram selecionados casos de alto risco, baseados na psicopatologia familiar e sub-sindrômica das crianças; na etapa domiciliar foram feitos os diagnósticos psiquiátricos da criança, pais e coleta de fatores de risco e saliva; Na etapa de avaliação escolar foram feitas avaliações neuropsicológicas e detalhamento clínico de fatores de risco para transtornos psiquiátricos.

Um total de 9.937 crianças foram incluídas na primeira fase. As entrevistas foram realizadas predominantemente com a mãe biológica (87.5%), levando em consideração sintomas de 45.394 familiares. A média de idade foi de 9 anos (DP=1,9), com uma leve predominância de sujeitos do sexo masculino 52,1% (n=5179). Dentro do universo de 9.937 sujeitos de pesquisa avaliados quanto à história familiar de transtornos psiquiátrico e quanto à apresentação sintomática da criança, os 1500 sujeitos de mais alto risco para os cinco principais transtornos de interesse para este projeto (TDAH, Ansiedade, TOC, Psicose e Aprendizagem) e uma amostragem aleatória de 1000 sujeitos foram convidados para participarem das etapas domiciliar e escolar. Definiu-se que os indivíduos de maior risco para os 5 transtornos seriam as crianças que tivessem a avaliação de sintomas no FHS para cada um dos transtornos positiva e que tivessem a maior densidade de sintomas da mesma ordem na família, através do mesmo instrumento.

Os 2.500 indivíduos selecionados foram avaliados no domicílio com uma entrevista domiciliar diagnóstica. O protocolo de avaliação domiciliar foi composto pelo DAWBA (Development and Well-Being Assessment - DAWBA), Child Behavior Checklist (CBCL), Mini International Neuropsychiatric Interview (M.I.N.I), uma avaliação de fatores de risco psiquiátricos gerais que incluem: (a) fatores gestacionais e perinatais (como tabagismo e álcool na gestação, etc.), (b) estressores na

infância (abuso, maus tratos, Bullying, rede de apoio, etc.); (c) doenças clínicas; (d) qualidade de vida geral; (e) desfechos escolares (faltas, suspensões, expulsões), uma avaliação de tratamentos e uso de serviços através de uma avaliação semiestruturada, avaliação de funcionamento familiar através da Family Environmental Scale (FES) e coleta de material genético dos pais biológicos ou irmão biológico mais velho disponível (na ausência de um dos pais biológicos).

### 3.1.1 Genotipagem e Controle de Qualidade

A genotipagem foi realizada pelo Grupo de Pesquisa do INPD. Utilizou-se a tecnologia *HumanCore - 12 v1.0 BeadChip (Illumina, Inc.)* , com 298.930 SNPs, dos quais 250421 são considerados tagSNPs. As populações representadas são: Europeia, Africana, Chinesa e Hispânica. Outras informações podem ser obtidas em Illumina.

A fase seguinte à genotipagem diz respeito ao controle de qualidade dos dados. Sabe-se que a remoção de SNPs e amostras problemáticas reduz a quantidade de testes de associação a serem realizados e, por conseguinte, reduz esforços computacionais (Turner *et al.*, 2011). As análises de qualidade foram realizadas utilizando-se o software Plink (Purcell *et al.*, 2007), e foram utilizados os seguintes critérios:

**Controle por Amostra**

- IBS (Identidade por Estado): indivíduos com mais de 95% de semelhança são descartados do conjunto de dados a ser analisado;
- *Call Rate (CRIND)*: amostras que não contenham pelo menos 90% de genótipos determinados pelo painel de genotipagem serão desconsideradas para análise.

**Controle por Variante** Para excluir os marcadores e amostras potencialmente problemáticos, os seguintes critérios foram considerados para remoção dos SNP:

- MAF (Frequência do Menor Alelo): manter os marcadores com frequência alélica menor ou igual a 5%;
- *Call Rate (CRSNP)*: exclusão de marcadores que não estejam presentes em pelo menos 98% da população;
- HWE (Equilíbrio de Hardy-Weinberg): marcadores com $p < 10^5$ para o teste exato de Fisher, ou seja, com desvios extremos de HWE, sugerem potencial erro de genotipagem.

Os seguintes comandos foram utilizados no software Plink:

```
plink --bfile INPD_SB --genome --min 0.3 --out calculo_pihat --noweb

plink --bfile INPD_SB --maf 0.05 --mind 0.01 --geno 0.01 --hwe 0.001 --remove
excluir_pihat.txt --make-bed --out inpdQCpass
```

O controle de qualidade consiste em filtrar os SNPs de baixa qualidade e amostras que possuem problemas na genotipagem. Dados genéticos podem conter erros provenientes de diversas fontes, entre elas, inconsistências em estudos de herança, manejo incorreto das amostras e erros cometidos no processo de genotipagem (Sobel *et al.*, 2002). A baixa qualidade ou baixa quantidade do material coletado são conhecidas por promover erros de genotipagem, bem como um baixo número de moléculas de DNA alvo em resultados de extração de extrema diluição de DNA ou de degradação deixa apenas algumas moléculas intactas, favorecendo o surgimento de alelos *dropouts*, isto é, amplificação de apenas um dos dois alelos presentes num lócus heterozigoto.

Para evitar erros de associação nas análises, recomenda-se o limiar de 95 a 99% de eficiência de genotipagem (*Call Rate*) para os SNPs (Turner *et al.*, 2011). É recomendado também remover os SNPs que não obedecem à frequência para estarem em Equilíbrio de HW, dado que o desequilíbrio

pode significar erro na genotipagem (Xu *et al.*, 2002). Em um organismo diploide (os cromossomos aparecem aos pares), com alelos A e a em um locus, existem três possíveis genótipos: AA, Aa, aa. Se a frequência de A for representada por p e a frequência de a por q, então as frequências genotípicas são calculadas assim:

$$p^2 + 2pq + q^2 = 1 \tag{3.1}$$

Um SNP está em desequilíbrio de HW quando suas frequências alélicas diferem das mostradas na Fórmula 3.1.1.

Ainda em relação à frequência alélica, sabe-se que a maioria dos algoritmos de chamada de genotipagem baseados em clusterização não obtém bons resultados em SNPs raros; dessa forma, é aconselhado remover aqueles que contêm um baixo MAF, pois são poucos informativos e a probabilidade de que eles rejeitem a hipótese nula H0 é baixa (statistical power).

### 3.1.2 Integração com conhecimento Biológico

Com o intuito de levarmos em consideração o efeito do SNP de acordo com suas características, construímos um banco de dados para anotar os SNPs genotipados. As características observadas estão relacionadas a SNPs em regiões regulatórias e associação destes SNPs com doenças psiquiátricas.

As informações foram obtidas através da integração dos bancos rVarBase, Biomart, GwasDB, ilustrado pela Figura 3.2. O *rVarBase* fornece informações sobre as anotações regulatórias para variantes humanas. O banco descreve as características regulatórias em três campos: estado da cromatina, sobreposição de elementos regulatórios e potenciais genes alvo. Além disso, há outras anotações, como informação sobre desequilíbrio de ligação, SNPs que estão em região regulatória e doenças ou expressão de traços quantitativos associados às variantes (Guo *et al.*, 2015). O *Biomart* é uma interface unificada para bancos de dados biomédicos distribuídos pelo mundo. Ele inclui aproximadamente 800 bancos de dados diferentes (Smedley *et al.*, 2015). Já o *GWASdb* provê informações sobre SNPs associados a traços e transtornos. Os dados foram coletados de 40 bancos de dados para anotar todos esses SNPs. As informações incluem informação baseada em gene, informação baseada em conhecimento, funções biológicas, sinais evolucionários e evidências de relação com doenças (Li *et al.*, 2011).

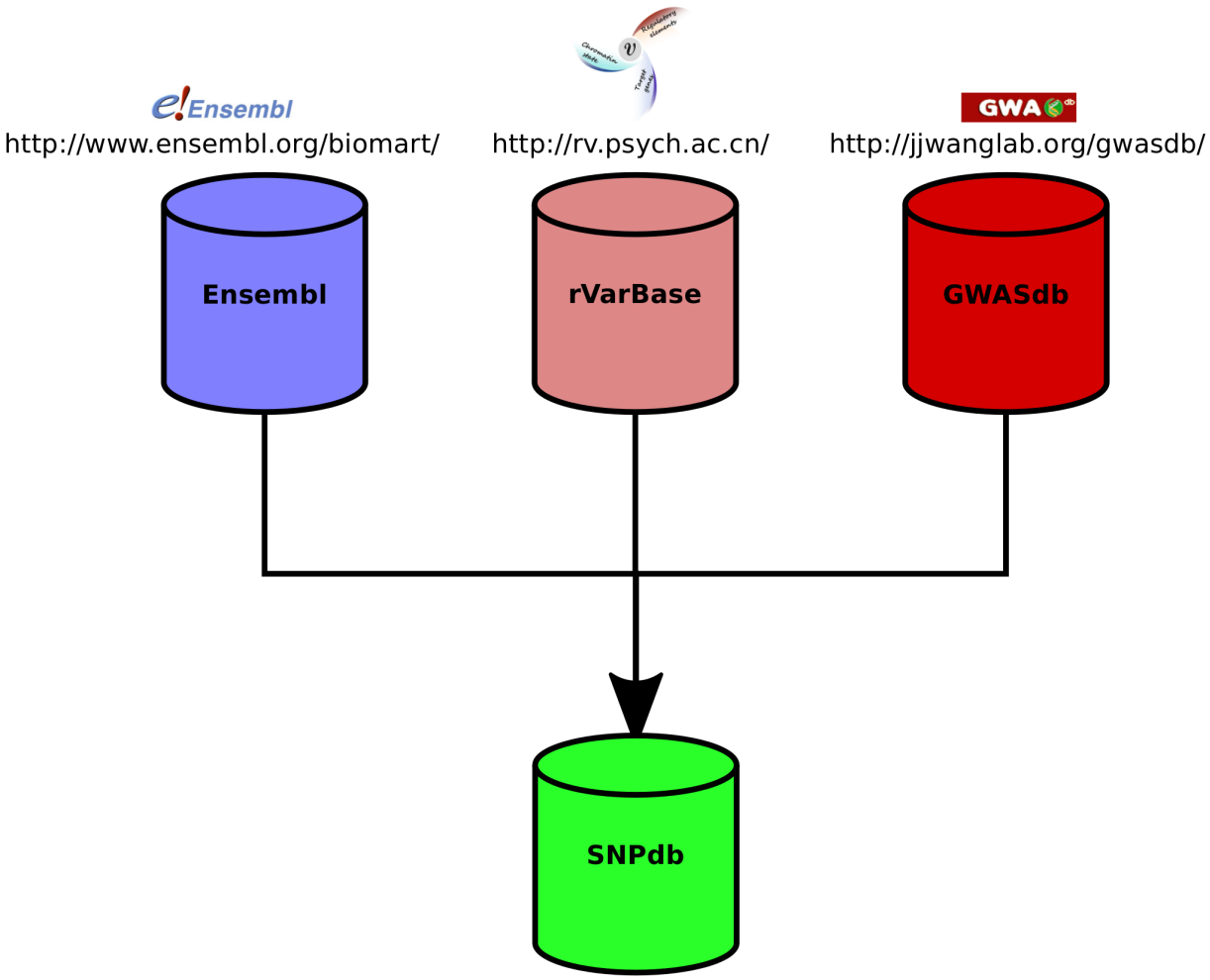


**Figura 3.2:** *Ilustração do Banco de dados criado para facilitar a obtenção e integração de informações*

Com esses resultados foram selecionados SNPs associados a qualquer estudo de GWAS, aqueles associados a transtornos psiquiátricos e também SNPs que estão em regiões regulatórias e em genes.

O Modelo Relacional do banco pode ser visto no Apêndice deste documento.

### 3.1.3 Seleção de Características

Dada a grande quantidade de dimensões em dados de Gwas, procuramos um método capaz de encontrar os melhores conjuntos de SNPs para caracterizar as amostras e então classificar novas amostras. O processo de seleção de características deste trabalho foi realizado através da utilização de três métodos Filter (FBST, Escore de Fisher e Regularização Lasso) e um método Wrapper (SFFS). Além disso, combinamos essas abordagens com as características biológicas que consideramos relevantes para serem utilizadas no classificador. Os resultados desses testes foram comparados e serão discutidos no Capítulo 4.

A Figura 3.3 apresenta o novo fluxo de trabalho. Utilizamos uma abordagem estatística e uma biológica, usamos Fisher, Lasso e SFFS e então testamos um algoritmo de classificação.

### 3.1.4 Construção do classificador

Dentro do conjunto de hipóteses H, o objetivo de um algoritmo de aprendizado de máquina é encontrar o melhor h, que é chamado de hipótese final g, próxima da função alvo f. Para que isso seja possível é necessário definir um algoritmo de aprendizado A, que inclui a função objetivo (função a ser otimizada) e métodos de otimização. O conjunto de hipóteses e a função objetivo modelam em

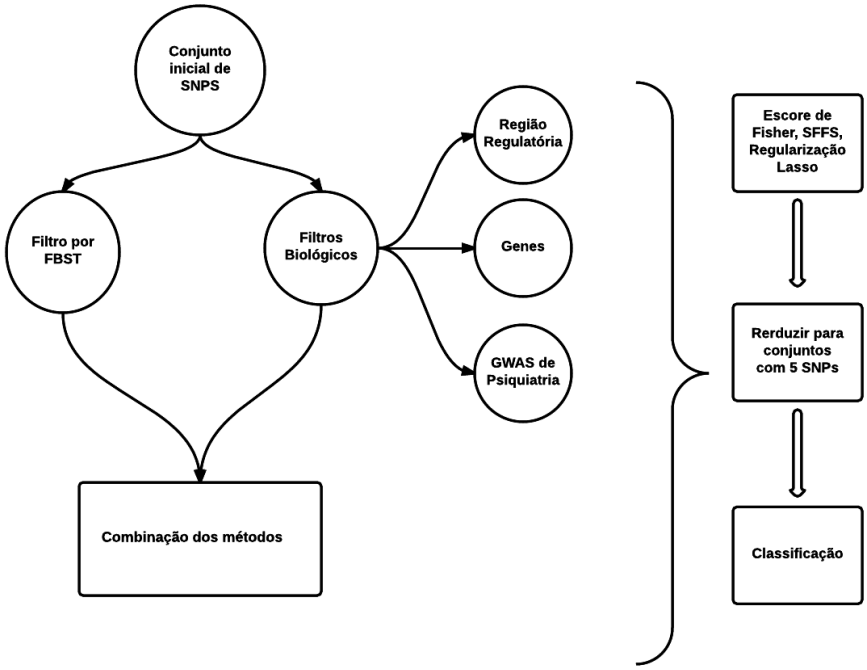


**Figura 3.3:** *Fluxo de seleção de características, integrando informações biólogicas e algoritmos*

conjunto a propriedade que não há almoço grátis. É esperado que a hipótese final g seja próxima de f e possa ser usada para predição. Os algoritmos testados foram: SVM e Random Forest.

### Parâmetros do SVM

Os algoritmos Greedy encontram boas soluções para problemas de otimização (Viswanathan *et al.*, 2011). Os testes de melhores parâmetros foram realizados usando validação cruzada com 10 folds. O algoritmo utilizado foi o SVM, com uma função RBF e os parâmetros C e Gamma foram ajustados pelo Greedy para cada conjunto específico de treinamento.

### Codificação

A codificação é o processo de padronização do conjunto de atributos para entrada no classificador, em que dado um conjunto de atributos categóricos é necessária uma codificação em valores numéricos, uma vez que um classificador SVM só trabalha com este tipo de atributo. Além disso, é necessário que todos os vetores de atributos sejam do mesmo tamanho para todas as entradas. Uma possível codificação para dados iguais aos deste trabalho foi apresentada por (Hirschhorn e Daly, 2005), em que os SNPs foram codificados em 0, 1 e 2, sendo 0 para homozigoto do alelo de maior frequência, 1 para heterozigoto e 2 para homozigoto do alelo de menor frequência ou da mesma forma como 0, 1 e 3 (Skafidas *et al.*, 2014).

### Balanceamento

O desbalanceamento acontece quando existe uma grande desproporção entre as classes, fazendo com que a classe minoritária seja erroneamente classificada. Nós balanceamos as classes fazendo com que a classe com maior número de amostras fosse reduzida ao tamanho da classe com menor número de amostras. As amostras removidas foram selecionadas aleatoriamente, outras formas devem ser exploradas em trabalhos futuros.

### Validação

Os métodos de validação foram explicados na Seção 2.2 . Neste trabalho a validação foi feita através dos métodos de validação cruzada. Um dos mais comuns é o k-validação cruzada, que consiste em dividir o conjunto de dados em k subconjuntos com aproximadamente o mesmo tamanho (Kohavi *et al.*, 1995). O k considerado normalmente é 10. Neste trabalho, subdividimos o conjunto total de dados em 10 conjuntos de treino e 10 de teste, e repetimos o procedimento de treinamento e teste por dez vezes (Efron e Tibshirani, 1995). Cada arquivo de teste possui 10% do total de dados da base, enquanto que o conjunto de treino possui 90% do conjunto total.

A Figura 3.4 mostra como o processo de validação é feito com 5 folds. Então suponhamos que possuímos 200 amostras. Vamos dividir estas 200 amostras em 5 partes, o que dá 40 amostras por parte. Pegamos uma parte e usamos como teste e as outras 4 usamos como treinamento. Esse processo é repetido k vezes, ou seja, 5 vezes. Dessa forma, todas as partes serão usadas como treino. Nesse exemplo estamos usando 80% para treinamento e 20% para teste, resultado de 100% dividido por 5.

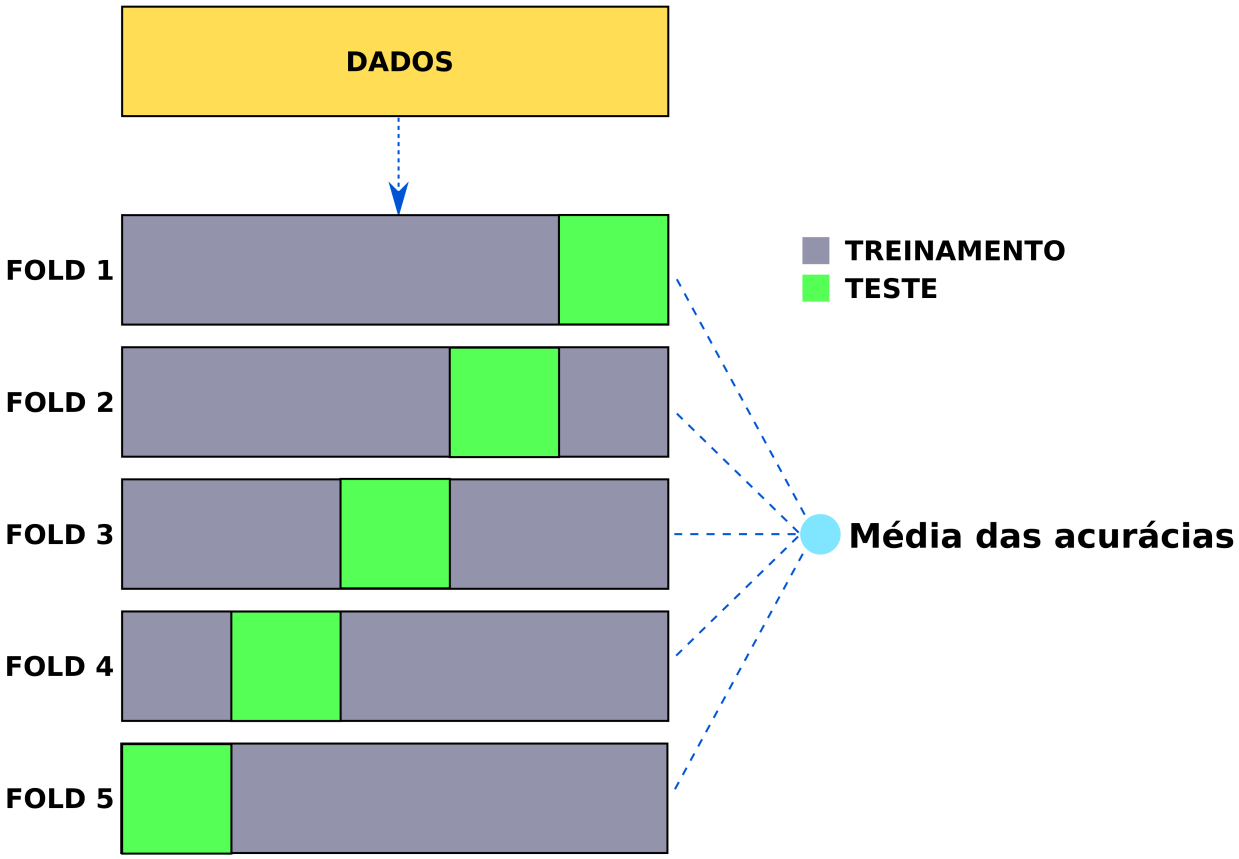


**Figura 3.4:** *Exemplo de validação cruzada com 5 folds.*

### Métricas de Avaliação do Classificador

As métricas de avaliação dizem respeito à performance de um classificador. Existem também as métricas utilizadas para a comparação entre a performance dos classificadores (Chen e Blostein, 2007).

Considerando um classificador binário e uma instância qualquer, existem quatro possíveis resultados para a classificação. Se a instância é positiva e é classificada como positiva, ela é contada como um Verdadeiro Positivo; se ela é classificada como negativa, então é contada como Falso Negativa. Se a instância é negativa e é classificada como negativa, ela é contada como uma Verdadeira Negativa, e se a instância é classificada como positiva, então é contada como Falsa Positiva. Dado

um classificador e um conjunto de instância podemos construir uma matriz de confusão (Figura 3.5), que é a base para muitas métricas.

| | Classificado como Positivo | Classificado como Negativo |
|---|---|---|
| Positivo | VP | FN |
| Negativo | FP | VN |

**Figura 3.5:** *Matriz de Confusão - Adaptada de Sokolova, 2009*

Algumas das métricas derivadas desta Tabela são:

- Acurácia: proporção de predições corretas (soma de verdadeiros positivos e verdadeiros negativos) $\frac{VP+VN}{P+N}$
- Sensibilidade ou Recall: é a capacidade que o teste diagnóstico/triagem apresenta de detectar os indivíduos verdadeiramente positivos, ou seja, de diagnosticar corretamente os doentes $\frac{VP}{VP+FN}$
- Especificidade: é a capacidade que o teste diagnóstico/triagem tem de detectar os verdadeiros negativos, isto é, de diagnosticar corretamente os indivíduos sadios $\frac{VN}{VN+FP}$
- Precisão: porcentagem de amostras positivas classificadas corretamente sobre o total de amostras classificadas como positivas $\frac{VP}{VP+FP}$
- Taxa de falsos positivos $1 - Especificidade$
- Curva ROC ( *Receiver operating characteristic*): Gráficos ROC são gráficos de duas dimensões nos quais a taxa de falso positivo é plotada no eixo Y e a taxa de falso positivo é plotada no eixo X. Um gráfico ROC descreve o tradeoff entre os benefícios (TP) e os custos(FP) (Fawcett, 2006).

A área sob a curva ROC nos diz o quão bom é o classificador, de acordo com Thomas G. Tape, os intervalos vão de Falha a Excelente, resultados entre 0.90 e 1 são considerados Excelentes, entre 0.80 e 0.90 são Bons, 0.70 e 0.80 são Aceitáveis, 0.60 e 0.70 são Pobres e de 0.50 e 0.60 são considerados Falhos (Tape, (accessado em Dezembro de 2016).

# Capítulo 4

# Resultados

## 4.1 Amostras

O número total de amostras é de 723 indivíduos. A Tabela 4.1 apresenta o número de casos e controles separados por sexo. Existe uma quantidade maior de controles, eles representam 69.5% do total de indivíduos. A porcentagem de indivíduos do sexo masculino é de 54.2% e do sexo feminino de 45.8%, sendo que do total de casos 55.6% são do sexo masculino e 44.4% são do sexo feminino. Não existe diferença de sexos entre casos e controles (chi-square = 0.2931 e p-value =0,588254. alpha=5%).

**Tabela 4.1:** *Número de casos e controles separados por sexo*

| | Female | Male |
|---|---|---|
| Controle | 234 | 269 |
| Caso | 98 | 123 |

A Tabela 4.1 apresenta a distribuição dos indivíduos pelos transtornos existentes na base. Alguns transtornos possuem mais representatividade do que outros, como é o caso de Transtorno de Depressão e Transtorno Obsessivo Compulsivo, onde um tem um número muito maior de portadores do que o outro.

**Tabela 4.2:** *Número de casos de acordo com transtornos psiquiátricos caracterizados pelo DSM-5*

| Transtornos | Feminino | Masculino | Total |
|---|---|---|---|
| Qualquer Transtorno (QT) | 98 (44.34%) | 123 (55.65%) | 221 |
| Transtorno de Depressão (TD) | 62(49.20%) | 64(50.79%) | 126 |
| Transtorno de Ansiedade (TA) | 49(54.44%) | 50 (50.50%) | 99 |
| Transtornos do neurodesenvolvimento-transtorno de deficit de atenção e hiperatividade (TND-TDAH) | 32 (35.55% ) | 58 (64.44%) | 90 |
| Transtornos Disruptivos do Controle dos Impulsos e da Conduta (TDCIC) | 30(38.96%) | 47(61.03%) | 77 |
| Transtornos Relacionados a Trauma e a Estressores (TTE) | 5(41.66%) | 7(58.33%) | 12 |
| Transtornos Alimentares (TAl) | 3 (60%) | 2 (40%) | 5 |
| Transtornos do Neurodesenvolvimento-transtorno do Espectro autista (TND-TEA) | 1 (25%) | 3 (75%) | 4 |
| Transtornos do neurodesenvolvimento-Transtorno de Tique (TND-TT) | 2 (50%) | 2(50%) | 4 |
| Transtorno Bipolar (TB) | 3 (100%) | 0 | 3 |
| Transtorno Obsessivo Compulsivo (TOC) | 1 (50%) | 1 (50%) | 2 |

A alta taxa de comorbidade entre os transtornos psiquiátricos das nossas amostras pode ser observada na Tabela 4.3. Essa Tabela reafirma o que foi dito anteriormente sobre doenças psiquiátricas e comorbidade, em que a maioria dos indivíduos apresentam dois transtornos concomitantes.

**Tabela 4.3:** *Número de comorbidades por indivíduo*

| | 1 | 2 | 3 | 4 | 5 |
|---|---|---|---|---|---|
| Meninas | 32 | 48 | 11 | 6 | 1 |
| Meninos | 46 | 54 | 17 | 3 | 3 |
| Total | 78(35.29%) | 102(46.15%) | 28(12.66%) | 9(4.07%) | 4(1.80%) |

As Tabelas 4.4 e 4.5 mostram se o individuo foi exposto a algum tipo de estresse e que estresse é este, além disso indica quantos desses indivíduos são casos ou controles. A porcentagem é em relação ao total de indivíduos que sofreram algum tipo de abuso. As Tabelas estão separadas por sexo.

**Tabela 4.4:** *Exposição ao estresse em caso e controle do sexo feminino*

| Abuso.Físico – p.valor.0 | Caso | Controle | Total |
|---|---|---|---|
| Sim | 28 (48.3%) | 30 (51.7%) | 58 |
| Não | 70 (25.6%) | 203 (74.3%) | 273 |
| Não Sabe | 0 | 1 (100%) | 1 |
| Abuso Sexual – p valor 0 | Caso | Controle | Total |
| Sim | 9 (75%) | 3 (25%) | 12 |
| Não | 88 (27.7%) | 229 (72.23%) | 317 |
| Não Sabe | 1 (33.3%) | 2 (66.6%) | 3 |
| Abuso Emocional – p valor 0 | Caso | Controle | Total |
| Sim | 62 (42.26%) | 84 (57.53) | 146 |
| Não | 36 (19.3%) | 150 (80.6%) | 186 |
| Não Sabe | 0 | 0 | 0 |
| Negligência Física – p valor 0 | Caso | Controle | Total |
| Sim | 20 (50%) | 20 (50%) | 40 |
| Não | 78 (26.7%) | 214 (73.28%) | 292 |
| Não Sabe | 0 | 0 | 0 |
| Stress Bullyng – p valor 0.66 | Caso | Controle | Total |
| Sim | 40 (40.8%) | 58 (59.18%) | 98 |
| Não | 3 (33.33%) | 6 (66.66%) | 9 |

**Tabela 4.5:** *Exposição ao estresse em caso e controle do sexo masculino*

| Abuso.Físico – p valor 0 | Caso | Controle | Total |
|---|---|---|---|
| Sim | 27 (51.9%) | 25 (48.07) | 52 |
| Não | 95 (28.2%) | 244 (72.6%) | 336 |
| Não Sabe | 1 (100%) | 0 | 1 |
| Abuso Sexual – p valor 0,42 | Caso | Controle | Total |
| Sim | 3 (30%) | 7 (70%) | 10 |
| Não | 118 (31.13%) | 261 (68.86%) | 379 |
| Não Sabe | 2 (66.66%) | 1 (33.33%) | 3 |
| Abuso Emocional – p valor 0 | Caso | Controle | Total |
| Sim | 77 (37.74%) | 127 (62.25%) | 204 |
| Não | 46 (23.23%) | 152 (76.76%) | 198 |
| Não Sabe | 0 | 0 | 0 |
| Negligência Física – p valor 0 | Caso | Controle | Total |
| Sim | 19 (57.57%) | 14 (42.42%) | 33 |
| Não | 104 (28.96%) | 255 (71.03%) | 359 |
| Não Sabe | 0 | 0 | 0 |
| Stress devido ao Bullyng – p valor 0,36 | Caso | Controle | Total |
| Sim | 46 (45.54%) | 55 (54.45%) | 101 |
| Não | 0 | 1 (100%) | 1 |

Os critérios e o número de SNPs removidos são apresentados na Tabela 4.6:
Qualidade dos SNPs

**Tabela 4.6:** *Critérios de remoção utilizados neste trabalho e número de SNPs removidos*

| HWE p.v. 0.001 | GENO 0.01 | MAF 0.05 | Total Restante |
|---|---|---|---|
| 8478 | 6763 | 47619 | 237815 |

Após a etapa de avaliação da qualidade foi feito um PCA 4.1 para verificar a homogeneidade das amostras em relação a casos e controles.

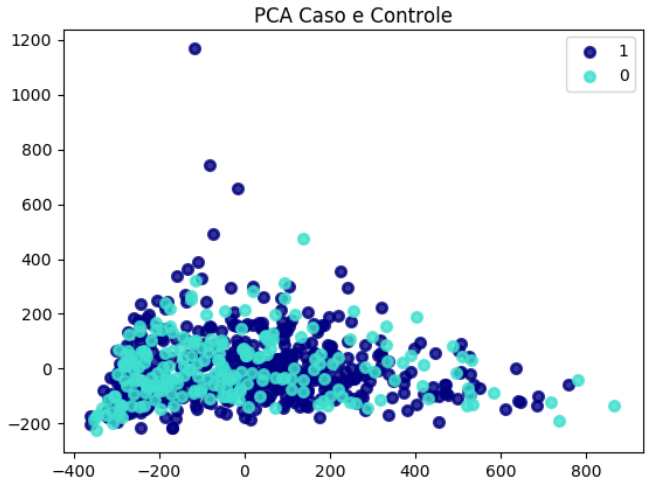


**Figura 4.1:** *PCA feito com os dados após o controle de qualidade para verificar homogeneidade da população.*

## 4.2 Seleção de Características

Os métodos utilizados para selecionar os SNPs foram FBST, Fisher, SFFS e Lasso, integrados às informações biológicas.

### 4.2.1 Escore de Fisher

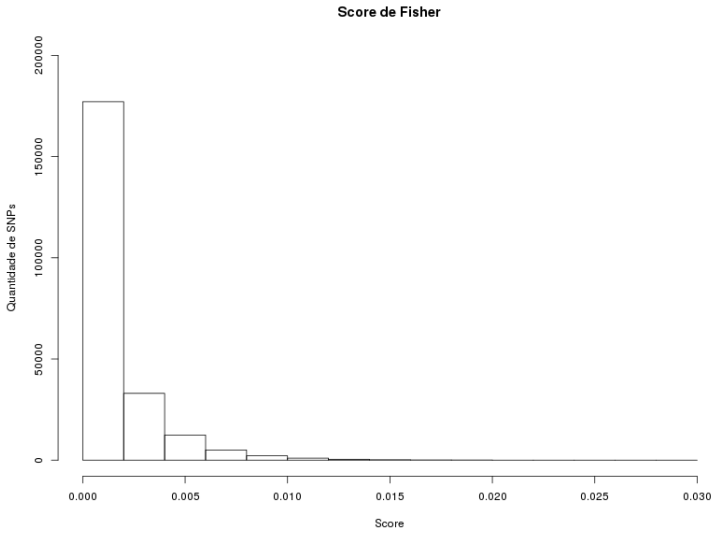


**Figura 4.2:** *Distribuição dos scores de Fisher em todos os SNPs da lâmina para amostras caso e controle.*

Podemos perceber pela Figura4.2 que a maior parte dos SNPs têm escores baixos. Como o número máximo de SNPs que usaremos no classificador é de 5, selecionaremos os 5 SNPs de maior escore para realizar a classificação.

### 4.2.2 Regularização Lasso

De acordo com a variação do parâmetro $\alpha$, foram selecionadas as seguintes quantidades de SNPs (Tabela 4.7):

**Tabela 4.7:** *Quantidade de SNPs selecionados variando* $\alpha$

| $\alpha$ 0.1 | $\alpha$ 0.3 | $\alpha$ 0.5 |
|---|---|---|
| 406 | 27 | 0 |

### 4.2.3 SFFS

Não foi viável testar o SFFS para o conjunto total de SNPs. Dado o número de SNPs, o número de combinações é muito alto e, devido a isso, optamos por outras abordagens que realizassem filtros nos SNPs antes de usar o SFFS.

### 4.2.4 FBST

Independente do algoritmo para a seleção de características, pode-se observar que a quantidade de SNPs selecionada ainda é muito alta. O número de características foi determinado dado o nosso N amostral. O nosso alfabeto tem tamanho 4 (00, 11, 12, 22). Sendo nosso N = 723, devemos descobrir quanto é $723 = 4^x$. Se fizermos $4^4$ = temos 256 indivíduos, se fizermos $4^5 = 1024$ indivíduos. Logo, devemos utilizar 4 e no máximo 5 características, para que todas as possibilidades - ou quase todas - sejam representadas.

Desse modo, duas abordagens diferentes foram utilizadas para realizar uma nova seleção de características: primeiramente, os SNPs foram selecionados com o algoritmo FBST, com uma abordagem estatística e, em um segundo momento, com características biológicas.

Os limiares considerados para o FBST foram:

**Tabela 4.8:** *Número de SNPs selecionados de acordo com a variação do e-valor*

| 0.01 | 0.05 | 0.1 |
|---|---|---|
| 128 SNPs | 1058 SNPs | 2600 SNPs |

Decidimos trabalhar com o conjunto de 1058 SNPs. As Tabelas 4.9, 4.10 mostram os resultados obtidos após aplicação dos classificadores SVM e *Random Forest* com os dados previamente filtrados por FBST e em seguida por Fisher, Lasso e SFFS. Apenas os cinco melhores SNPs foram utilizados para a classificação. Os melhores valores de C e Gamma também são mostrados nas Tabelas.

**Tabela 4.9:** *Resultado da classificação dos Top 5 SNPs selecionados pelo método FBST em seguida por Fisher*

| | Método | AUC | Especificidade | Sensibilidade | Precisão | Acc |
|---|---|---|---|---|---|---|
| Balanceado | SVM {g: 0.0001, C: 1} | 0.64 | 0.64 | 0.65 | 0.64 | 0.63 |
| | RF | 0.59 | 0.56 | 0.62 | 0.59 | 0.60 |
| Desbalanceado | SVM {g: 0.9, C: 0.5} | 0.56 | 0.95 | 0.18 | 0.66 | 0.71 |
| | RF | | | | | |

**Tabela 4.10:** *Resultado da classificação dos Top 5 SNPs selecionados pelo método FBST em seguida por Lasso*

| | Método | AUC | Especificidade | Sensibilidade | Precisão | Acc |
|---|---|---|---|---|---|---|
| Balanceado | SVM {g: 0.0001, C: 1} | 0.63 | 0.71 | 0.56 | 0.65 | 0.64 |
| | RF | 0.58 | 0.63 | 0.53 | 0.60 | 0.59 |
| Desbalanceado | SVM {g: 0.0001, C: 0.8} | 0.61 | 0.60 | 0.62 | 0.63 | 0.58 |
| | RF | 0.58 | 0.59 | 0.57 | 0.59 | 0.58 |

### 4.2.5 Integrando Critérios Biológicos

As informações biológicas foram resultados de um banco de dados construído para armazenar as características de todos os SNPs genotipados. Para cada SNP buscamos no banco de dados genótipo, alelo de menor frequência segundo o 1000 genomas, alelo ancestral, sua localização genômica, relação e distância do gene mais próximo, tipo de transcrito, impacto da variação, predição por ferramentas computacionais para explorar se o SNP teria efeito deletério e sua localização genômica de acordo com as definições. Com esse banco filtramos os SNPs de acordo com as seguintes informações (Tabela 4.11)

**Tabela 4.11:** *Quantidade de SNPs em cada critério biológico*

| | Em Região Regulatória | Em estudos de GWAS | Em Psiquiatria | Em genes |
|---|---|---|---|---|
| 1 | 188860 | 8201 | 322 | 137840 |

As Tabelas 4.12 e 4.13 mostram os resultados obtidos após aplicação do classificador SVM com os dados previamente filtrados com atributos biológicos e seleção de características por Fisher, Lasso e SFFS. Apenas os cinco melhores SNPs foram utilizados para a classificação.

**Tabela 4.12:** *Parâmetros de avaliação da classificação realizada com filtros de características biológicas e os algoritmos utilizados - Fisher*

| Filtro Biológico | Distr. | Algorit. | AUC | Espec. | Sens. | Prec. | Acur. |
|---|---|---|---|---|---|---|---|
| Região regulatória | D | SVM {g: 0.9, C: 0.5} | 0.55 | 0.95 | 0.15 | 0.64 | 0.71 |
| | | RF | - | - | - | - | - |
| | B | SVM {g: 1e-04, C: 20} | 0.67 | 0.62 | 0.72 | 0.65 | 0.67 |
| | | RF | 0.66 | 0.64 | 0.68 | 0.65 | 0.65 |
| Gwas | D | SVM {g: 1e-3, C: 1000} | 0.54 | 0.96 | 0.12 | 0.69 | 0.71 |
| | | RF | 0.55 | 0.88 | 0.21 | 0.47 | 0.68 |
| | B | SVM {g: 1e-3, C: 0.8} | 0.62 | 0.57 | 0.66 | 0.61 | 0.62 |
| | | RF | 0.54 | 0.54 | 0.53 | 0.54 | 0.55 |
| Psiquiatria | D | SVM {g: 1e-3, C: 10} | 0.50 | 0.99 | 0.00 | 0.05 | 0.69 |
| | | RF | 0.50 | 0.86 | 0.14 | 0.31 | 0.64 |
| | B | SVM {g: 1e-3, C: 20} | 0.56 | 0.42 | 0.71 | 0.55 | 0.57 |
| | | RF | 0.55 | 0.51 | 0.58 | 0.54 | 0.54 |
| Genes | D | SVM {g: 1e-3, C: 10} | 0.53 | 0.96 | 0.09 | 0.48 | 0.69 |
| | | RF | 0.54 | 0.86 | 0.23 | 0.43 | 0.67 |
| | B | SVM {g: 1e-04, C: 1} | 0.62 | 0.55 | 0.69 | 0.60 | 0.62 |
| | | RF | 0.60 | 0.56 | 0.64 | 0.59 | 0.60 |

**Tabela 4.13:** *Parâmetros de avaliação da classificação realizada com filtros de características biológicas e os algoritmos utilizados - LASSO*

| Filtro Biológico | Distr. | Algorit. | AUC | Espec. | Sens. | Prec. | Acur. |
|---|---|---|---|---|---|---|---|
| Região regulatória | D | SVM {g:0.9 C:10 } | 0.52 | 0.87 | 0.17 | 0.38 | 0.66 |
| | | RF | 0.53 | 0.87 | 0.19 | 0.40 | 0.66 |
| | B | SVM {g: 1e-3, C: 1000} | 0.58 | 0.54 | 0.61 | 0.57 | 0.57 |
| | | RF | 0.52 | 0.49 | 0.56 | 0.52 | 0.52 |
| Gwas | D | SVM {g:0.9 C:0.5} | 0.50 | 1.00 | 0.00 | 0.00 | 0.69 |
| | | RF | 0.52 | 0.84 | 0.20 | 0.37 | 0.64 |
| | B | SVM {g: 1e-04, C: 100} | 0.56 | 0.54 | 0.58 | 0.56 | 0.55 |
| | | RF | 0.53 | 0.50 | 0.57 | 0.53 | 0.52 |
| Psiquiatria | D | SVM {g: 0.9, C: 0.001} | 0.50 | 1.00 | 0.00 | 0.00 | 0.69 |
| | | RF | 0.49 | 0.96 | 0.02 | 0.18 | 0.67 |
| | B | SVM | 0.51 | 0.44 | 0.57 | 0.46 | 0.44 |
| | | RF | 0.52 | 0.52 | 0.52 | 0.52 | 0.52 |
| Genes | D | SVM {g:0.1 C:10} | 0.53 | 0.91 | 0.15 | 0.46 | 0.68 |
| | | RF | 0.57 | 0.85 | 0.29 | 0.48 | 0.68 |
| | B | SVM {g:0.001 C:0.5 } | 0.61 | 0.64 | 0.58 | 0.61 | 0.61 |
| | | RF | 0.57 | 0.64 | 0.51 | 0.58 | 0.56 |

O SFFS não encontrou SNPs o suficiente para que a classificação pudesse ser realizada. Além disso, os SNPs selecionados tiveram uma alta entropia.

Em todos os experimentos de classificação, podemos perceber que a Especificidade é muito alta, enquanto que a Sensibilidade é muito baixa. Isso indica que a classe positiva não foi predita, enquanto que a classe negativa foi. Como a classe negativa é majoritária, poderia sugerir que os resultados se deviam ao desbalanceamento. No entanto, apesar do balanceamento, os resultados não se modificaram.

Nosso próximo passo foi trabalhar com conjuntos de SNPs selecionados em pelo menos duas características de acordo com os Diagramas de Venn abaixo.

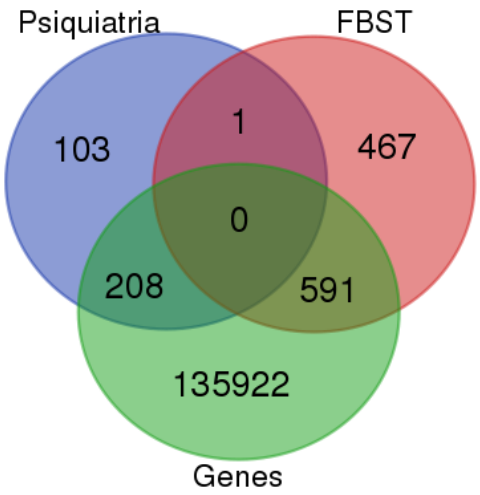


**Figura 4.3:** *Diagrama de Venn dos SNPs associados a psiquitria, os que estão em genes e selecionados por FBST*

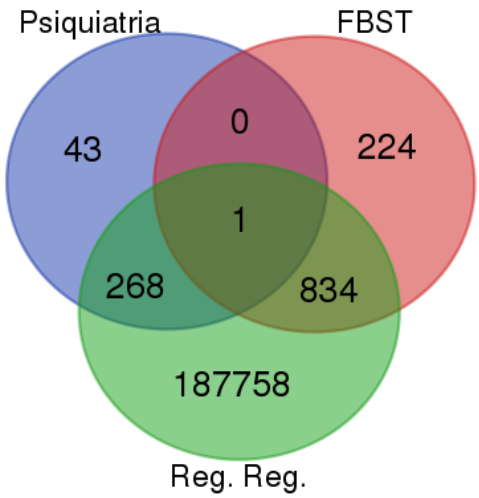


**Figura 4.4:** *Diagrama de Venn dos SNPs associados a psiquiatria, os que estão em região regulatória e selecionados por FBST*

Os resultados da classificação dos SNPs selecionados por mais de um critério estão nas Tabelas 4.14 e 4.15.

**Tabela 4.14:** *Resultado da Intersecção Biológicos - Fisher*

| Filtro Biológico | Distr. | Algorit. | AUC | Espec. | Sens. | Prec. | Acur. |
|---|---|---|---|---|---|---|---|
| Psiq x Região Reg. | D | SVM {g: 1e-3, C: 10} | 0.50 | 0.99 | 0.00 | 0.05 | 0.69 |
| | | RF | 0.48 | 0.85 | 0.11 | 0.27 | 0.63 |
| | B | SVM {g: 1e-3, C: 1} | 0.58 | 0.53 | 0.63 | 0.57 | 0.58 |
| | | RF | 0.54 | 0.54 | 0.54 | 0.53 | 0.54 |
| FBST x Região Reg. | D | SVM {g: 0.9, C: 0.5} | 0.55 | 0.95 | 0.15 | 0.64 | 0.71 |
| | | RF | 0.55 | 0.87 | 0.24 | 0.46 | 0.67 |
| | B | SVM {g: 1e-3, C: 0.5} | 0.68 | 0.68 | 0.67 | 0.68 | 0.69 |
| | | RF | 0.64 | 0.67 | 0.60 | 0.65 | 0.64 |
| FBST x Genes | D | SVM {g: 1e-3, C: 10} | 0.53 | 0.96 | 0.09 | 0.48 | 0.69 |
| | | RF | 0.54 | 0.86 | 0.22 | 0.43 | 0.67 |
| | B | SVM {g: 1e-04, C: 1} | 0.64 | 0.63 | 0.65 | 0.63 | 0.64 |
| | | RF | 0.57 | 0.50 | 0.63 | 0.56 | 0.56 |
| Psiq x Genes | D | SVM {g: 0.1, C: 0.8, } | 0.51 | 0.97 | 0.04 | 0.55 | 0.69 |
| | | RF | 0.48 | 0.85 | 0.11 | 0.30 | 0.63 |
| | B | SVM {g: 1e-04, C: 20} | 0.57 | 0.62 | 0.52 | 0.59 | 0.55 |
| | | RF | 0.57 | 0.59 | 0.54 | 0.56 | 0.56 |

**Tabela 4.15:** *Resultado da Intersecção Biológicos - LASSO*

| Filtro Biológico | Distr. | Algorit. | AUC | Espec. | Sens. | Prec. | Acur. |
|---|---|---|---|---|---|---|---|
| Psiq x Região Reg. | D | SVM {g: 0.1, C: 10} | 0.49 | 0.99 | 0.00 | 0.00 | 0.69 |
| | | RF | 0.48 | 0.95 | 0.02 | 0.15 | 0.66 |
| | B | SVM g: 1e-04, C: 100} | 0.55 | 0.58 | 0.52 | 0.55 | 0.54 |
| | | RF | 0.53 | 0.51 | 0.55 | 0.52 | 0.53 |
| FBST x Região Reg. | D | SVM {g: 0.1, C: 0.5} | 0.55 | 0.95 | 0.15 | 0.62 | 0.71 |
| | | RF | 0.54 | 0.89 | 0.19 | 0.43 | 0.68 |
| | B | SVM {g: 0.1, C: 0.5} | 0.55 | 0.55 | 0.55 | 0.55 | 0.55 |
| | | RF | 0.56 | 0.58 | 0.55 | 0.56 | 0.57 |
| FBST x Genes | D | SVM {g: 1e-3, C: 15} | 0.53 | 0.96 | 0.10 | 0.50 | 0.69 |
| | | RF | 0.56 | 0.89 | 0.22 | 0.47 | 0.69 |
| | B | SVM {g: 0.9, C: 0.8} | 0.63 | 0.56 | 0.70 | 0.61 | 0.63 |
| | | RF | 0.61 | 0.53 | 0.69 | 0.59 | 0.61 |
| Psiq x Genes | D | SVM {g: 0.1, C: 10} | 0.49 | 0.99 | 0.00 | 0.00 | 0.69 |
| | | RF | 0.48 | 0.95 | 0.02 | 0.15 | 0.66 |
| | B | SVM {g: 1e-3, C: 100} | 0.56 | 0.60 | 0.51 | 0.56 | 0.54 |
| | | RF | 0.54 | 0.54 | 0.54 | 0.54 | 0.53 |

O método Fisher teve mais sucesso que o Lasso e o SFFS. Além disso, os melhores resultados foram encontrados quando as classes foram balanceadas. Em relação ao classificador, os melhores resultados foram encontrados pelo SVM. É interessante notar que o FBTS e região regulatória, de forma geral e, principalmente, quando combinadas, alcançam os melhores resultados.

**Tabela 4.16:** *Melhores resultados*

| | | | AUC | Espec | Sens | Prec | Acur |
|---|---|---|---|---|---|---|---|
| FBST e FISHER | SVM | B | 0.64 | 64 | 65 | 64 | 63 |
| Fisher e REg Reg | SVM | B | 0.67 | 62 | 72 | 65 | 67 |
| Fisher - FBST er Reg Reg | SVM | B | 0.68 | 68 | 67 | 68 | 69 |

De acordo com os resultados apresentados, não conseguimos um classificador adequado para casos e controles. Um dos possíveis problemas seria o número amostral pequeno para atingir toda a variabilidade fenotípica incluída nos casos, uma vez que não usamos nenhum endofenótipo específico. Portanto, resolvemos testar a classificação de apenas um transtorno e levando em conta o sexo, uma vez que a prevalência de TDAH é muito maior em meninos.

**Tabela 4.17:** *TDAH com SNPs psiquiatria*

| Mét. | Algorit. | AUC | Espec. | Sens. | Prec. | Acur. | |
|---|---|---|---|---|---|---|---|
| Meninos | Fisher | SVM {g: 1e-04, C: 20} | 0.62 | 0.65 | 0.60 | 0.65 | 0.62 |
| | | RF | 0.48 | 0.51 | 0.45 | 0.51 | 0.49 |
| | Lasso | SVM {g: 1e-03, C: 10} | 0.53 | 0.55 | 0.50 | 0.56 | 0.50 |
| | | RF | 0.47 | 0.46 | 0.48 | 0.48 | 0.44 |
| | SFFS | SVM | - | - | - | - | - |
| | | RF | - | - | - | - | - |
| Meninas | Fisher | SVM {g: 1e-04, C: 15} | 0.77 | 0.70 | 0.83 | 0.76 | 0.76 |
| | | RF | | | | | |
| | Lasso | SVM {g: 0.1, C: 0.8} | 0.52 | 0.61 | 0.45 | 0.39 | 0.51 |
| | | RF | 0.57 | 0.61 | 0.54 | 0.50 | 0.58 |
| | SFFS | SVM | - | - | - | - | - |
| | | RF | - | - | - | - | - |
| Meninos e Meninas | Fisher | SVM {g: 1e-03, C: 0.5} | 0.67 | 0.66 | 0.69 | 0.68 | 0.66 |
| | | RF | 0.63 | 0.57 | 0.68 | 0.61 | 0.61 |
| | Lasso | SVM {g: 1e-04, C: 15} | 0.59 | 0.58 | 0.61 | 0.59 | 0.60 |
| | | RF | 0.63 | 0.63 | 0.64 | 0.62 | 0.62 |
| | SFFS | SVM | - | - | - | - | - |
| | | RF | - | - | - | - | - |

**Tabela 4.18:** *SNPS em Psiquiatria e Região regulatória*

| | Mét. | Algorit. | AUC | Espec. | Sens. | Prec. | Acur. |
|---|---|---|---|---|---|---|---|
| Meninos | Fisher | SVM {g: 1e-4, C: 15} | 0.63 | 0.65 | 0.61 | 0.62 | 0.62 |
| | | RF | 0.48 | 0.51 | 0.45 | 0.44 | 0.49 |
| | Lasso | SVM {g: 1e-3, C: 0.8} | 0.58 | 0.57 | 0.59 | 0.60 | 0.58 |
| | | RF | 0.55 | 0.62 | 0.48 | 0.55 | 0.58 |
| | SFFS | SVM | - | - | - | - | - |
| | | RF | - | - | - | - | - |
| Meninas | Fisher | SVM {g: 1e-3, C: 0.8} | 0.70 | 0.61 | 0.78 | 0.65 | 0.68 |
| | | RF | 0.73 | 0.69 | 0.76 | 0.71 | 0.71 |
| | Lasso | SVM | 0.62 | 0.44 | 0.79 | 0.59 | 0.63 |
| | | RF | 0.52 | 0.51 | 0.54 | 0.51 | 0.54 |
| | SFFS | SVM | - | - | - | - | - |
| | | RF | - | - | - | - | - |
| Meninos e Meninas | Fisher | SVM {g: 1e-04, C: 1000} | 0.60 | 0.50 | 0.69 | 0.59 | 0.57 |
| | | RF | 0.63 | 0.59 | 0.68 | 0.61 | 0.61 |
| | Lasso | SVM {g: 1e-04, C: 20} | 0.56 | 0.60 | 0.51 | 0.56 | 0.55 |
| | | RF | 0.54 | 0.56 | 0.53 | 0.55 | 0.53 |
| | SFFS | SVM | - | - | - | - | - |
| | | RF | - | - | - | - | - |

**Tabela 4.19:** *SNPS em FBST e Região Regulatória*

| Mét. | Algorit. | AUC | Espec. | Sens. | Prec. | Acur. | |
|---|---|---|---|---|---|---|---|
| Meninos | Fisher | SVM {g: 1e-4, C: 20} | 0.76 CHECAR | 0.73 | 0.79 | 0.75 | 0.78 |
| | | RF | 0.70 | 0.72 | 0.69 | 0.74 | 0.72 |
| | Lasso | SVM {g: 0.1, C: 0.5} | 0.61 | 0.78 | 0.44 | 0.64 | 0.65 |
| | | RF | 0.62 | 0.65 | 0.59 | 0.62 | 0.63 |
| | SFFS | SVM | - | - | - | - | - |
| | | RF | - | - | - | - | - |
| Meninas | Fisher | SVM {g: 0.9, C: 0.8} | 0.74 | 0.77 | 0.72 | 0.74 | 0.75 |
| | | RF | 0.73 | 0.70 | 0.75 | 0.70 | 0.72 |
| | Lasso | SVM {g: 1e-3, C: 100} | 0.62 | 0.63 | 0.61 | 0.63 | 0.62 |
| | | RF | 0.72 | 0.71 | 0.73 | 0.71 | 0.70 |
| | SFFS | SVM | - | - | - | - | - |
| | | RF | - | - | - | - | - |
| Meninos e Meninas | Fisher | SVM {g: 1e-4, C: 20} | 0.74 | 0.68 | 0.80 | 0.73 | 0.75 |
| | | RF | 0.75 | 0.72 | 0.79 | 0.72 | 0.76 |
| | Lasso | SVM {g: 1e-3, C: 20} | 0.68 | 0.57 | 0.79 | 0.65 | 0.68 |
| | | RF | 0.66 | 0.65 | 0.67 | 0.65 | 0.66 |
| | SFFS | SVM | - | - | - | - | - |
| | | RF | - | - | - | - | - |

**Tabela 4.20:** *SNPS em FBST e Genes*

| Mét. | Algorit. | AUC | Espec. | Sens. | Prec. | Acur. | |
|---|---|---|---|---|---|---|---|
| Meninos | Fisher | SVM {g: 1e-3, C: 0.5} | 0.68 | 0.69 | 0.68 | 0.70 | 0.70 |
| | | RF | 0.70 | 0.69 | 0.71 | 0.77 | 0.72 |
| | Lasso | SVM {g: 1e-4, C: 15} | 0.57 | 0.56 | 0.59 | 0.57 | 0.57 |
| | | RF | 0.61 | 0.62 | 0.60 | 0.63 | 0.62 |
| | SFFS | SVM | - | - | - | - | - |
| | | RF | - | - | - | - | - |
| Meninas | Fisher | SVM {g: 1e-3, C: 1} | 0.74 | 0.69 | 0.78 | 0.71 | 0.75 |
| | | RF | 0.61 | 0.66 | 0.56 | 0.62 | 0.63 |
| | Lasso | SVM {g: 1e-4, C: 20} | 0.66 | 0.75 | 0.56 | 0.66 | 0.66 |
| | | RF | 0.63 | 0.66 | 0.59 | 0.61 | 0.64 |
| | SFFS | SVM | - | - | - | - | - |
| | | RF | - | - | - | - | - |
| Meninos e Meninas | Fisher | SVM {g: 1e-3, C: 10} | 0.70 | 0.71 | 0.70 | 0.74 | 0.71 |
| | | RF | 0.70 | 0.71 | 0.69 | 0.73 | 0.71 |
| | Lasso | SVM {g: 1e-3, C: 1000} | 0.56 | 0.48 | 0.65 | 0.55 | 0.55 |
| | | RF | 0.58 | 0.57 | 0.58 | 0.57 | 0.57 |
| | SFFS | SVM | - | - | - | - | - |
| | | RF | - | - | - | - | - |

**Tabela 4.21:** *SNPS em Psiquiatria e Genes*

| Mét. | Algorit. | AUC | Espec. | Sens. | Prec. | Acur. | |
|---|---|---|---|---|---|---|---|
| Meninos | Fisher | SVM {g: 1e-4, C: 1000} | 0.65 | 0.76 | 0.54 | 0.68 | 0.66 |
| | | RF | 0.62 | 0.60 | 0.64 | 0.60 | 0.60 |
| | Lasso | SVM {g: 1e-3, C: 1} | 0.55 | 0.61 | 0.49 | 0.51 | 0.52 |
| | | RF | 0.50 | 0.50 | 0.49 | 0.50 | 0.49 |
| | SFFS | SVM | - | - | - | - | - |
| | | RF | - | - | - | - | - |
| Meninas | Fisher | SVM {g: 1e-3, C: 0.8} | 0.64 | 0.56 | 0.72 | 0.61 | 0.64 |
| | | RF | 0.64 | 0.66 | 0.61 | 0.63 | 0.62 |
| | Lasso | SVM {g: 1e-3, C: 1000} | 0.66 | 0.54 | 0.77 | 0.63 | 0.67 |
| | | RF | 0.59 | 0.62 | 0.57 | 0.62 | 0.62 |
| | SFFS | SVM | - | - | - | - | - |
| | | RF | - | - | - | - | - |
| Meninos e Meninas | Fisher | SVM {g: 1e-3, C: 20} | 0.67 | 0.61 | 0.72 | 0.65 | 0.67 |
| | | RF | 0.63 | 0.60 | 0.67 | 0.63 | 0.62 |
| | Lasso | SVM {g: 1e-3, C: 1000} | 0.53 | 0.63 | 0.42 | 0.53 | 0.53 |
| | | RF | 0.53 | 0.55 | 0.51 | 0.56 | 0.52 |
| | SFFS | SVM | - | - | - | - | - |
| | | RF | - | - | - | - | - |

Resumindo os melhores resultados para TDAH são (Tabela 4.22):

**Tabela 4.22:** *Resultado da melhor AUC em cada método para meninos e meninas*

| | Menino | Menina | Menino e Menina |
|---|---|---|---|
| PSIQ x Fisher | SVM: 0.62 | SVM: 0.77 | SVM: 0.67 RF: 0.63 |
| PSIQ x Lasso | - | - | SVM: 0.63 RF: 0.63 |
| PSIQ x REG.REG x FISHER | SVM: 0.63 | SVM: 0.70 RF: 0.73 | SVM: 0.60 RF: 0.63 |
| PSIQ x REG.REG x lasso | - | SVM: 0.62 | - |
| FBST x REG.REG x fisher | SVM: 0.75 RF: 0.70 | SVM: 0.74 RF: 0.73 | SVM: 0.74 RF: 0.75 |
| FBST x REG.REG x lasso | SVM 0.61 RF: 0.62 | SVM 0.62 RF: 0.72 | SVM: 0.68 RF:0.66 |
| FBST x GENES x fisher | SVM: 0.68 RF 0.70 | SVM: 0.74 RF: 0.61 | SVM: 0.70 RF: 0.70 |
| FBST x GENES x lasso | - | SVM: 0.66 RF: 0.63 | - |
| PSIQ x GENES x fisher | SVM: 0.65 RF: 0.62 | SVM: 0.64 RF 0.64 | SVM: 0.67 RF: 0.63 |
| PSIQ x GENES x lasso | - | SVM: 0.66 | - |

Na Tabela 4.22 podemos confirmar o que vimos anteriormente, os melhores resultados são provenientes dos SNPs selecionados pela intersecção dos métodos FBST e Região Regulatória e, posteriormente, os que foram ordenados pelo Escore de Fisher.

# Capítulo 5

# Conclusões

## 5.1 Discussão

Um dos grandes desafios nos estudos de associação genômica ampla é selecionar os SNPs adequados, sabendo que a interação entres eles deve ser levada em consideração. Dado o grande número de SNPs, uma busca exaustiva entre todas as combinações possíveis é inviável. A importância de um método de fácil replicação é a sua utilidade na prática clínica, uma vez que existe dificuldade no diagnóstico de doenças psiquiátricas, principalmente em crianças. Levando isso em consideração, estudamos a possibilidade de usar técnicas heurísticas para encontrar os conjuntos que melhor caracterizam as amostras caso e controle. As técnicas utilizadas para seleção de características foram Escore de Fisher, Lasso e SFFS. No entanto, o número de características selecionadas ainda era maior que o número estimado para essa amostra. Optamos, então, por usar uma primeira seleção por teste estatístico e por características biológicas relevantes para, depois, usar Fisher, Lasso, SFFS e SVM ou RF. Pudemos observar que quando usamos FBST e/ou regiões regulatórias, houve uma melhora no classificador, sugerindo que a seleção de SNPs deve levar em conta as mesmas. A combinação de outras características biológicas sem a presença de uma destas obteve piores resultados de classificação. No entanto, ainda assim, os testes mostraram que os conjuntos de SNPs encontrados não são capazes de discriminar casos e controles na nossa amostra. Algumas hipóteses sobre a falha em nossos testes são que:

- O N amostral era muito pequeno;
- O número de SNPs selecionados, dado o tamanho da amostra, não foi suficiente para caracterizar toda a psicopatologia estudada;
- Outros fatores fundamentais, como diferenças entre sexos, diferenças em fatores de exposição ao stress, presença de fatores protetores, não foram levados em consideração;
- O acúmulo de ruídos do processo;
- Não foi feita nenhuma otimização no classificador.

Estudos mostram a importância de um N amostral grande em problemas de aprendizado de máquina. Para trabalharmos com, por exemplo, uma centena de SNPs, seria necessário um N amostral muito maior do que o que está disponível. Além do mais, no caso da interação entre SNPs, pode ser que sejam necessários mais de 5 SNPs para representar a heterogeneidade clínica englobada por qualquer transtorno psiquiátrico. Assim, apesar dos transtornos psiquiátricos apresentarem altas taxas de comorbidade e compartilharem uma base poligênica, principalmente diante do nosso número amostral, uma melhor abordagem seria o uso de um endofenótipo, ou seja, um traço comportamental observável e quantitativo sabidamente associado à psicopatologia de alguns transtornos mentais, ou a união dos transtornos que apresentam altas taxas de comorbidade e não qualquer transtorno.

Outro ponto fundamental é que os transtornos psiquiátricos são poligênicos, mas também multifatoriais, ou seja, diferenças entre sexos e outros fatores de exposição ao stress não podem ser excluídos. Existem também fatores protetores que impedem que certos sintomas sejam elevados a transtornos. Para considerar estes fatores, são necessários estudos sobre quais são eles (Borsboom *et al.*, 2011). Além desses fatores, é importante lembrar que os dados aqui usados vieram de uma plataforma de GWAS e, portanto, só foram estudados SNPs comuns, e que por mais que estes SNPs expliquem grande parte da variabilidade genômica, outras variantes muito raras, bem como estruturais, são importantes na etiopatogenia, principalmente dos transtornos psiquiátricos.

Ainda, a quantidade de ruídos que se acumula durante o processo deve ser levada em conta. Os ruídos podem ser provenientes de vários fatores: mecanismos estocásticos inerentes ao nível molecular, variação na abundância de moléculas, heterogeneidade, sensibilidade do ensaio biológico ou artefatos de medição prevalentes especialmente em configurações de *high-throughput* (Natarajan *et al.*, 2013; **?**). Existem também fatores protetores que impedem que certos sintomas sejam elevados a transtornos. Para considerar estes fatores, são necessários estudos sobre quais são esses fatores (Borsboom *et al.*, 2011).

Além do SVM, testamos o *Random Forest*, que obteve sucesso em trabalhos anteriores. Tem a vantagem de não assumir nenhuma distribuição para os dados, é um algoritmo eficiente e o único parâmetro que precisamos ajustar é o número de árvores. Os testes foram realizados com o mesmo número de árvores para todos os casos, sem nenhum ajuste, e obteve sucesso em menos casos que o SVM. É possível que com um ajuste do número de árvores o algoritmo tenha melhor desempenho. Apesar de termos utilizado o *Greedy Search* para o ajuste dos parâmetros do SVM, o classificador ainda precisa ser mais explorado, como, por exemplo, testar outros tipos de funções para o Kernel. Ao utilizarmos apenas um transtorno como classe positiva os resultados de AUC foram melhores, sendo que SNPs em regiões regulatórias foram os que apresentaram o melhor desempenho para classificação, da mesma forma que ocorreu quando consideramos todas as doenças.

## 5.2 Conclusões

O nosso objetivo era usar técnicas de reconhecimento de padrões classificando as amostras da população em saudável ou doente para transtornos psiquiátricos a partir de um conjunto de SNPs como características dessas amostras. Este objetivo foi concluído assim como os seguintes objetivos específicos: Avaliar os achados realizados por técnicas de AM para seleção de características com as anotações biológica e funcional em um conjunto de SNPs pré selecionados para construir um classificador. definir critérios para a identificação dos SNPs no conjunto de dados selecionado; estudar técnicas de reconhecimento de padrões para a classificação das amostras; selecionar as características (SNPs) do conjunto amostral; integrar as características; escolher um modelo para a classificação; realizar os treinamento e os testes de classificação comparar os métodos de seleção de características utilizados e os métodos de classificação

Concluímos que para a nossa população usar critérios estatísticos e biológicos traz melhores resultados. Em acordo com a literatura a localização do SNP em regiões regulatórias do genoma parece ser a mais relevante. O método de FISHER e uso de SVM foram os que atingiram melhores AUC. Vale notar que usar endofenótipos mais específicos, assim como a informação de comorbidade, fatores de exposição ambiental, sexo dos indivíduos que não foi levado em consideração precisam ser melhor explorados. Da mesma forma buscar mais características relacionadas ao potencial regulador dos SNPs por exemplo, informação de eQTL também precisam ser melhor exploradas.

# Apêndice A

# Modelo Relacional

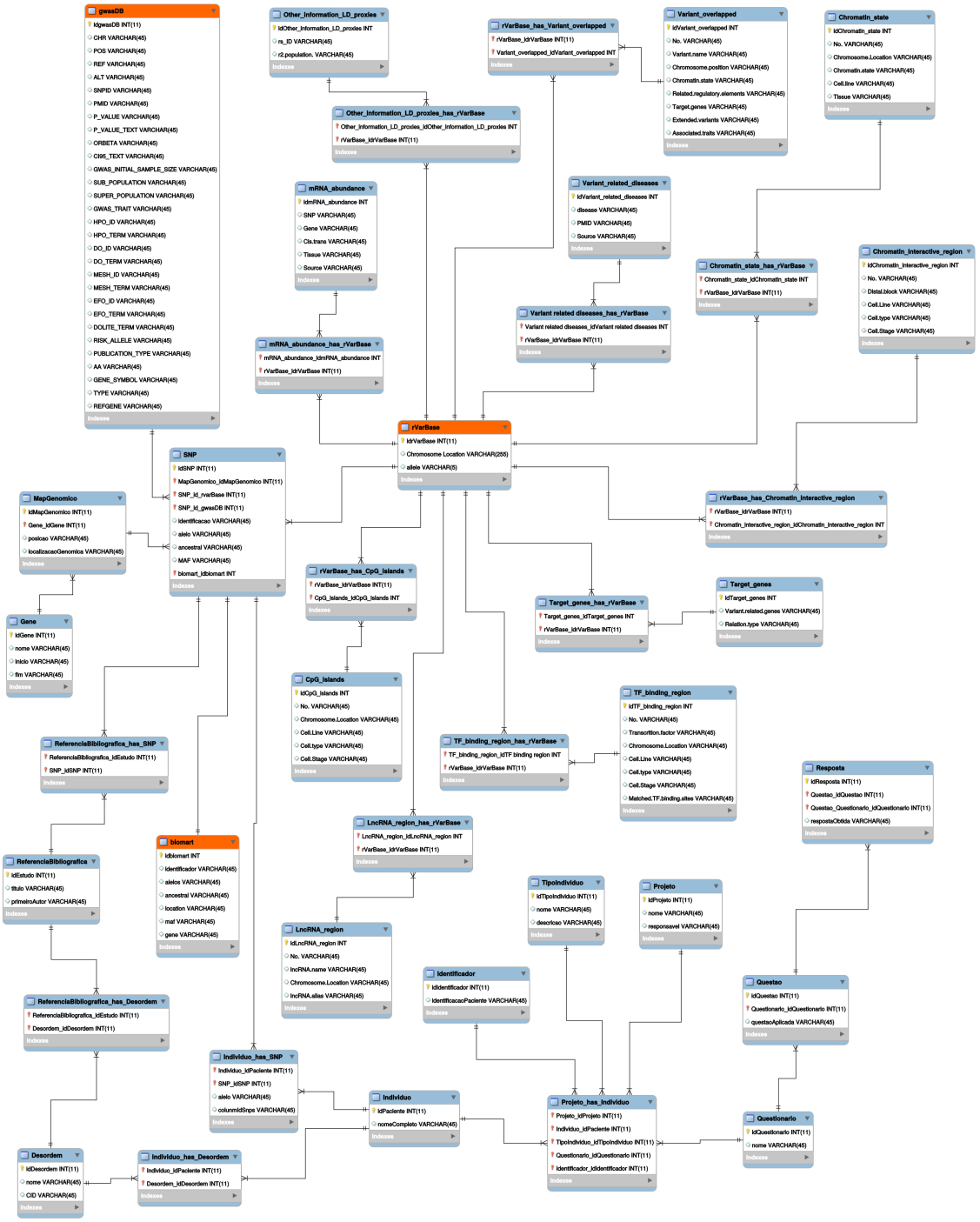


**Figura A.1:** *Modelo Relacional do Banco de Dados para Integração de informações biológicas*

# Referências Bibliográficas


**Aguiar-Pulido** *et al.***(2010)** Vanessa Aguiar-Pulido, José A Seoane, Juan R Rabuñal, Julián Dorado, Alejandro Pazos e Cristian R Munteanu. Machine learning techniques for single nucleotide polymorphism?disease classification models in schizophrenia. *Molecules*, 15(7):4875–4889. Citado na pág. 2

**Anyfantis** *et al.***(2007)** D Anyfantis, M Karagiannopoulos, S Kotsiantis e P Pintelas. Robustness of learning techniques in handling class noise in imbalanced datasets. Em *IFIP International Conference on Artificial Intelligence Applications and Innovations*, páginas 21–28. Springer. Citado na pág. 19

**Ardlie** *et al.***(2002)** Kristin G Ardlie, Leonid Kruglyak e Mark Seielstad. Patterns of linkage disequilibrium in the human genome. *Nature Reviews Genetics*, 3(4):299–309. Citado na pág. 7

**Arlot** *et al.***(2010)** Sylvain Arlot, Alain Celisse *et al.* A survey of cross-validation procedures for model selection. *Statistics surveys*, 4:40–79. Citado na pág. 12

**Baharudin** *et al.***(2010)** Baharum Baharudin, Lam Hong Lee e Khairullah Khan. A review of machine learning algorithms for text-documents classification. *Journal of advances in information technology*, 1(1):4–20. Citado na pág. 14

**Balding(2006)** David J Balding. A tutorial on statistical methods for population association studies. *Nature Reviews Genetics*, 7(10):781–791. Citado na pág. 7

**Bekkar e Alitouche(2013)** Mohamed Bekkar e Taklit Akrouf Alitouche. Imbalanced data learning approaches review. *International Journal of Data Mining & Knowledge Management Process*, 3 (4):15. Citado na pág. 19

**Ben-David** *et al.***(2011)** Shai Ben-David, N Srebro e R Urner. Universal learning vs. no free lunch results. Em *Philosophy and Machine Learning Workshop NIPS*. Citado na pág. 11

**Bengio** *et al.***(2013)** Yoshua Bengio, Aaron Courville e Pierre Vincent. Representation learning: A review and new perspectives. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 35(8):1798–1828. Citado na pág. 19

**Bickel** *et al.***(2006)** Peter J Bickel, Bo Li, Alexandre B Tsybakov, Sara A van de Geer, Bin Yu, Teófilo Valdés, Carlos Rivero, Jianqing Fan e Aad van der Vaart. Regularization in statistics. *Test*, 15(2):271–344. Citado na pág. 16

**Bishop(2007)** C Bishop. Pattern recognition and machine learning (information science and statistics), 1st edn. 2006. corr. 2nd printing edn. *Springer, New York*. Citado na pág. 10

**Borsboom** *et al.***(2011)** Denny Borsboom, Angélique OJ Cramer, Verena D Schmittmann, Sacha Epskamp e Lourens J Waldorp. The small world of psychopathology. *PloS one*, 6(11):e27407. Citado na pág. 1, 44

**Breiman(2001)** Leo Breiman. Random forests. *Machine learning*, 45(1):5–32. Citado na pág. 20

**Bulik** *et al.***(2000)** Cynthia M Bulik, Patrick F Sullivan, Tracey D Wade e Kenneth S Kendler. Twin studies of eating disorders: a review. *International Journal of Eating Disorders*, 27(1):1–20. Citado na pág. 7

**Burmeister** *et al.***(2008)** Margit Burmeister, Melvin G McInnis e Sebastian Zöllner. Psychiatric genetics: progress amid controversy. *Nature Reviews Genetics*, 9(7):527–540. Citado na pág. 2

**Bush e Moore(2012)** William S Bush e Jason H Moore. Genome-wide association studies. *PLoS Comput Biol*, 8(12):e1002822. Citado na pág. 8

**Castellanos e Tannock(2002)** F Xavier Castellanos e Rosemary Tannock. Neuroscience of attention-deficit/hyperactivity disorder: the search for endophenotypes. *Nature Reviews Neuroscience*, 3(8):617–628. Citado na pág. 8

**Categorical(1998)** A Categorical. Glossary of terms. *Machine Learning*, 30(2/3):271–274. Citado na pág. 10

**Chandrashekar e Sahin(2014)** Girish Chandrashekar e Ferat Sahin. A survey on feature selection methods. *Computers & Electrical Engineering*, 40(1):16–28. Citado na pág. ix, 18

**Chawla(2005)** Nitesh V Chawla. Data mining for imbalanced datasets: An overview. Em *Data mining and knowledge discovery handbook*, páginas 853–867. Springer. Citado na pág. 19

**Chen e Blostein(2007)** Nawei Chen e Dorothea Blostein. A survey of document image classification: problem statement, classifier architecture and performance evaluation. *International Journal on Document Analysis and Recognition*, 10(1):1–16. Citado na pág. 29

**Chen** *et al.***(2008)** Shyh-Huei Chen, Jielin Sun, Latchezar Dimitrov, Aubrey R Turner, Tamara S Adams, Deborah A Meyers, Bao-Li Chang, S Lilly Zheng, Henrik Grönberg, Jianfeng Xu *et al.* A support vector machine approach for detecting gene-gene interaction. *Genetic epidemiology*, 32(2):152–167. Citado na pág. 21

**Cichon** *et al.***(2009)** Sven Cichon, Nick Craddock, Mark Daly, Stephen V Faraone, Pablo V Gejman, John Kelsoe, Thomas Lehner, Douglas F Levinson, Audra Moran, Pamela Sklar e Patrick F Sullivan. Genomewide association studies: history, rationale, and prospects for psychiatric disorders. *Am. J. Psychiatry*, 166(5):540–56. ISSN 1535-7228. doi: 10.1176/appi.ajp.2008.08091354. URL http://www.ncbi.nlm.nih.gov/pubmed/19339359. Citado na pág. 7

**Committee** *et al.***(2009)** Psychiatric GWAS Consortium Coordinating Committee *et al.* Genomewide association studies: history, rationale, and prospects for psychiatric disorders. *The American journal of psychiatry*, 166(5):540–556. Citado na pág. 7, 8, 9

**Consortium** *et al.***(2010)** 1000 Genomes Project Consortium *et al.* A map of human genome variation from population-scale sequencing. *Nature*, 467(7319):1061–1073. Citado na pág. 2

**Cousin** *et al.***(2003)** E Cousin, E Genin, S Mace, S Ricard, C Chansac, M Del Zompo e JF Deleuze. Association studies in candidate genes: strategies to select snps to be tested. *Human heredity*, 56 (4):151–159. Citado na pág. 7

**Cover e Van Campenhout(1977)** Thomas M Cover e Jan M Van Campenhout. On the possible orderings in the measurement selection problem. *IEEE Transactions on Systems, Man, and Cybernetics*, 7(9):657–661. Citado na pág. 18

**Dash e Liu(1997)** Manoranjan Dash e Huan Liu. Feature selection for classification. *Intelligent data analysis*, 1(1-4):131–156. Citado na pág. 14

**de Andrade Oliveira** *et al.***()** Joanito de Andrade Oliveira, Luciano Vieira Dutra e Camilo Daleles Rennó. Seleção e extração de atributos para classificação de regiões. Citado na pág. 17

**de Bakker** *et al.***(2005)** Paul IW de Bakker, Roman Yelensky, Itsik Pe'er, Stacey B Gabriel, Mark J Daly e David Altshuler. Efficiency and power in genetic association studies. *Nature genetics*, 37 (11):1217–1223. Citado na pág. 8

**De La Vega** *et al.***(2005)** FRANCISCO M De La Vega, Hadar I Isaac e Charles R Scafe. A tool for selecting snps for association studies based on observed linkage disequilibrium patterns. Em *Pacific Symposium on Biocomputing. Pacific Symposium on Biocomputing*, páginas 487–498. Citado na pág. 3

**Di Deco** *et al.***(2013)** Javier Di Deco, Ana M Gonzalez, Julia Diaz, Virginia Mato, Daniel Garcia-Frank, Juan Alvarez-Linera, Ana Frank e Juan A Hernandez-Tamames. Machine learning and social network analysis applied to alzheimer?s disease biomarkers. *Current topics in medicinal chemistry*, 13(5):652–662. Citado na pág. 11

**Dietterich(2002)** Thomas G Dietterich. Machine learning for sequential data: A review. Em *Structural, syntactic, and statistical pattern recognition*, páginas 15–30. Springer. Citado na pág. 11

**Domingue** *et al.***(2014)** Benjamin W Domingue, Daniel W Belsky, Kathleen Mullan Harris, Andrew Smolen, Matthew B McQueen e Jason D Boardman. Polygenic risk predicts obesity in both white and black young adults. Citado na pág. 3

**Domschke(2013)** Katharina Domschke. Clinical and molecular genetics of psychotic depression. *Schizophrenia bulletin*, página sbt040. Citado na pág. 1

**Dua e Du(2016)** Sumeet Dua e Xian Du. *Data mining and machine learning in cybersecurity*. CRC press. Citado na pág. 12

**Duda** *et al.***(2012)** Richard O Duda, Peter E Hart e David G Stork. *Pattern classification*. John Wiley & Sons. Citado na pág. 10, 11

**Dudbridge(2013)** Frank Dudbridge. Power and predictive accuracy of polygenic risk scores. *PLoS Genet.*, 9(3):e1003348. ISSN 1553-7404. doi: 10.1371/journal.pgen.1003348. URL http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3605113&tool=pmcentrez&rendertype=abstract. Citado na pág. 2, 3, 9

**Efron e Tibshirani(1995)** Bradley Efron e Robert J Tibshirani. *Cross-validation and the bootstrap: Estimating the error rate of a prediction rule*. Division of Biostatistics, Stanford University. Citado na pág. 29

**Euesden** *et al.***(2014)** Jack Euesden, Cathryn M Lewis e Paul F O'Reilly. Prsice: Polygenic risk score software. *Bioinformatics*, página btu848. Citado na pág. 9

**Fanous e Kendler(2005)** AH Fanous e KS Kendler. Genetic heterogeneity, modifier genes, and quantitative phenotypes in psychiatric illness: searching for a framework. *Molecular psychiatry*, 10(1):6–13. Citado na pág. 1

**Fawcett(2006)** Tom Fawcett. An introduction to roc analysis. *Pattern recognition letters*, 27(8): 861–874. Citado na pág. 30

**Fornito e Bullmore(2012)** Alex Fornito e Edward T Bullmore. Connectomic intermediate phenotypes for psychiatric disorders. *Frontiers in psychiatry*, 3. Citado na pág. 1

**Frazer** *et al.***(2009)** Kelly A Frazer, Sarah S Murray, Nicholas J Schork e Eric J Topol. Human genetic variation and its contribution to complex traits. *Nature Reviews Genetics*, 10(4):241–251. Citado na pág. 2

**Fukunaga(2013)** Keinosuke Fukunaga. *Introduction to statistical pattern recognition*. Academic press. Citado na pág. 10

**Gaspar** *et al.***(2012)** Paulo Gaspar, Jaime Carbonell e José Luís Oliveira. On the parameter optimization of support vector machines for binary classification. *Journal of Integrative Bioinformatics (JIB)*, 9(3):33–43. Citado na pág. 21

**Gibson(2012)** Greg Gibson. Rare and common variants: twenty arguments. *Nature Reviews Genetics*, 13(2):135–145. Citado na pág. 8

**Glessner e Hakonarson(2009)** Joseph T Glessner e Hakon Hakonarson. Common variants in polygenic schizophrenia. *Genome Biol*, 10(9):236. Citado na pág. 9

**Goh** *et al.***(2007)** Kwang-Il Goh, Michael E Cusick, David Valle, Barton Childs, Marc Vidal e Albert-László Barabási. The human disease network. *Proceedings of the National Academy of Sciences*, 104(21):8685–8690. Citado na pág. 2

**Golberg e Cho(2004)** Michael A Golberg e Hokwon A Cho. *Introduction to regression analysis*. Wit Press. Citado na pág. 21

**Goldstein** *et al.***(2011)** Benjamin A Goldstein, Eric C Polley e Farren Briggs. Random forests for genetic association studies. *Statistical applications in genetics and molecular biology*, 10(1). Citado na pág. 20

**Gratten** *et al.***(2014)** Jacob Gratten, Naomi R Wray, Matthew C Keller e Peter M Visscher. Large-scale genomics unveils the genetic architecture of psychiatric disorders. *Nature neuroscience*, 17 (6):782–790. Citado na pág. 8

**Gu** *et al.***(2012)** Quanquan Gu, Zhenhui Li e Jiawei Han. Generalized fisher score for feature selection. *arXiv preprint arXiv:1202.3725*. Citado na pág. 16

**Guo** *et al.***(2015)** Liyuan Guo, Yang Du, Susu Qu e Jing Wang. rvarbase: an updated database for regulatory features of human variants. *Nucleic acids research*, página gkv1107. Citado na pág. 26

**Guo** *et al.***(2016)** Yiran Guo, Zhi Wei, Brendan J Keating e Hakon Hakonarson. Machine learning derived risk prediction of anorexia nervosa. *BMC medical genomics*, 9(1):4. Citado na pág. 22

**Guyon e Elisseeff(2003)** Isabelle Guyon e André Elisseeff. An introduction to variable and feature selection. *The Journal of Machine Learning Research*, 3:1157–1182. Citado na pág. 12, 14

**Hawkins(2004)** Douglas M Hawkins. The problem of overfitting. *Journal of chemical information and computer sciences*, 44(1):1–12. Citado na pág. 11

**He** *et al.***(2009)** Haibo He, Edwardo Garcia *et al.* Learning from imbalanced data. *Knowledge and Data Engineering, IEEE Transactions on*, 21(9):1263–1284. Citado na pág. 19

**Hira e Gillies(2015)** Zena M Hira e Duncan F Gillies. A review of feature selection and feature extraction methods applied on microarray data. *Advances in bioinformatics*, 2015. Citado na pág. 14

**Hirschhorn e Daly(2005)** Joel N Hirschhorn e Mark J Daly. Genome-wide association studies for common diseases and complex traits. *Nature Reviews Genetics*, 6(2):95–108. Citado na pág. 2, 28

**Hua** *et al.***(2005)** Jianping Hua, Zixiang Xiong, James Lowey, Edward Suh e Edward R Dougherty. Optimal number of features as a function of sample size for various classification rules. *Bioinformatics*, 21(8):1509–1515. Citado na pág. 18

**Jain e Zongker(1997)** Anil Jain e Douglas Zongker. Feature selection: Evaluation, application, and small sample performance. *IEEE transactions on pattern analysis and machine intelligence*, 19(2):153–158. Citado na pág. 17

**Jiang** *et al.***(2009)** Rui Jiang, Wanwan Tang, Xuebing Wu e Wenhui Fu. A random forest approach to the detection of epistatic interactions in case-control studies. *BMC bioinformatics*, 10(1):S65. Citado na pág. 20

**Jović** *et al.***(2015)** Alan Jović, Karla Brkić e Nikola Bogunović. A review of feature selection methods with applications. Em *Information and Communication Technology, Electronics and Microelectronics (MIPRO), 2015 38th International Convention on*, páginas 1200–1205. IEEE. Citado na pág. 15

**Kohavi** *et al.***(1995)** Ron Kohavi *et al.* A study of cross-validation and bootstrap for accuracy estimation and model selection. Em *Ijcai*, volume 14, páginas 1137–1145. Stanford, CA. Citado na pág. 29

**Koller e Sahami(1996)** Daphne Koller e Mehran Sahami. Toward optimal feature selection. Relatório técnico, Stanford InfoLab. Citado na pág. 14, 19

**Korkmaz(2011)** Baris Korkmaz. Theory of mind and neurodevelopmental disorders of childhood. *Pediatric research*, 69:101R–108R. Citado na pág. 1

**Korte e Farlow(2013)** Arthur Korte e Ashley Farlow. The advantages and limitations of trait analysis with gwas: a review. *Plant methods*, 9(1):29. Citado na pág. 8

**Kumar e Minz(2014)** Vipin Kumar e Sonajharia Minz. Feature selection. *SmartCR*, 4(3):211–229. Citado na pág. 14

**Li** *et al.***(2011)** Mulin Jun Li, Panwen Wang, Xiaorong Liu, Ee Lyn Lim, Zhangyong Wang, Meredith Yeager, Maria P Wong, Pak Chung Sham, Stephen J Chanock e Junwen Wang. Gwasdb: a database for human genetic variants identified by genome-wide association studies. *Nucleic acids research*, página gkr1182. Citado na pág. 26

**Li** *et al.***(2012)** Mulin Jun Li, Panwen Wang, Xiaorong Liu, Ee Lyn Lim, Zhangyong Wang, Meredith Yeager, Maria P Wong, Pak Chung Sham, Stephen J Chanock e Junwen Wang. Gwasdb: a database for human genetic variants identified by genome-wide association studies. *Nucleic acids research*, 40(D1):D1047–D1054. Citado na pág. 4

**Lima()** Leandro de Araujo Lima. *Uma abordagem integrativa usando dados de interação proteína-proteína e estudos genéticos para priorizar genes e funções biológicas em transtorno de déficit de atenção e hiperatividade*. Tese de Doutorado, Universidade de São Paulo. Citado na pág. 7

**Listgarten** *et al.***(2004)** Jennifer Listgarten, Sambasivarao Damaraju, Brett Poulin, Lillian Cook, Jennifer Dufour, Adrian Driga, John Mackey, David Wishart, Russ Greiner e Brent Zanke. Predictive models for breast cancer susceptibility from multiple single nucleotide polymorphisms. *Clinical cancer research*, 10(8):2725–2737. Citado na pág. 22

**Liu(2004)** Alexander Yun-chung Liu. *The effect of oversampling and undersampling on classifying imbalanced text datasets*. Tese de Doutorado, Citeseer. Citado na pág. 19

**Lopes(2011)** Fabrício Martins Lopes. *Redes complexas de expressão gênica: síntese, identificação, análise e aplicações*. Tese de Doutorado, Universidade de São Paulo. Citado na pág. 17

**Malhotra** *et al.***(2004)** Anil K Malhotra, Greer M Murphy Jr e James L Kennedy. Pharmacogenetics of psychotropic drug response. *American Journal of Psychiatry*, 161(5):780–796. Citado na pág. 2

**Manolio e Collins(2009)** Teri A Manolio e Francis S Collins. The hapmap and genome-wide association studies in diagnosis and therapy. *Annual review of medicine*, 60:443–456. Citado na pág. 8

**Manor e Segal(2013)** Ohad Manor e Eran Segal. Predicting disease risk using bootstrap ranking and classification algorithms. *PLoS Comput Biol*, 9(8):1003200. Citado na pág. 8, 9

**Marian(2012)** Ali J Marian. Molecular genetic studies of complex phenotypes. *Translational Research*, 159(2):64–79. Citado na pág. 1

**Maurano** *et al.***(2012)** Matthew T Maurano, Richard Humbert, Eric Rynes, Robert E Thurman, Eric Haugen, Hao Wang, Alex P Reynolds, Richard Sandstrom, Hongzhu Qu, Jennifer Brody *et al.* Systematic localization of common disease-associated variation in regulatory dna. *Science*, 337(6099):1190–1195. Citado na pág. 3

**McCarthy** *et al.***(2008)** Mark I McCarthy, Gonçalo R Abecasis, Lon R Cardon, David B Goldstein, Julian Little, John PA Ioannidis e Joel N Hirschhorn. Genome-wide association studies for complex traits: consensus, uncertainty and challenges. *Nature Reviews Genetics*, 9(5):356–369. Citado na pág. 2, 8

**Molina** *et al.***(2002)** Luis Carlos Molina, Lluís Belanche e Àngela Nebot. Feature selection algorithms: A survey and experimental evaluation. Em *Data Mining, 2002. ICDM 2003. Proceedings. 2002 IEEE International Conference on*, páginas 306–313. IEEE. Citado na pág. 15

**Monard e Baranauskas(2003)** Maria Carolina Monard e José Augusto Baranauskas. Conceitos sobre aprendizado de máquina. *Sistemas Inteligentes-Fundamentos e Aplicações*, 1(1). Citado na pág. xi, 11, 12, 13

**Murthy e Salzberg(1995)** Kolluru Venkata Sreerama Murthy e Steven L Salzberg. *On growing better decision trees from data*. Tese de Doutorado, Citeseer. Citado na pág. 18

**Naghibi** *et al.***(2015)** Tofigh Naghibi, Sarah Hoffmann e Beat Pfister. A semidefinite programming based search strategy for feature selection with mutual information measure. *IEEE transactions on pattern analysis and machine intelligence*, 37(8):1529–1541. Citado na pág. 15

**Natarajan** *et al.***(2013)** Nagarajan Natarajan, Inderjit S Dhillon, Pradeep K Ravikumar e Ambuj Tewari. Learning with noisy labels. Em *Advances in neural information processing systems*, páginas 1196–1204. Citado na pág. 44

**Nurnberger Jr** *et al.***(2016)** John I Nurnberger Jr, Wade Berrettini e Alexander B Niculescu III. Genetics of psychiatric disorders. Em *The medical basis of psychiatry*, páginas 553–600. Springer. Citado na pág. 8

**Owoeye** *et al.***(2013)** Olabisi Owoeye, Tara Kingston, Paul J Scully, Patrizia Baldwin, David Browne, Anthony Kinsella, Vincent Russell, Eadbhard O?Callaghan e John L Waddington. Epidemiological and clinical characterization following a first psychotic episode in major depressive disorder: comparisons with schizophrenia and bipolar i disorder in the cavan-monaghan first episode psychosis study (camfeps). *Schizophrenia bulletin*, 39(4):756–765. Citado na pág. 1

**Özgür** *et al.***(2008)** Arzucan Özgür, Thuy Vu, Güneş Erkan e Dragomir R Radev. Identifying gene-disease associations using centrality on a literature mined gene-interaction network. *Bioinformatics*, 24(13):i277–i285. Citado na pág. 21

**Patnala** *et al.***(2013)** Radhika Patnala, Judith Clements e Jyotsna Batra. Candidate gene association studies: a comprehensive guide to useful in silico tools. *BMC genetics*, 14(1):39. Citado na pág. 9

**Pereira** *et al.***(2008)** Carlos A de B Pereira, Julio Michael Stern, Sergio Wechsler *et al.* Can a significance test be genuinely bayesian? *Bayesian Analysis*, 3(1):79–100. Citado na pág. 16

**Pereira e Stern(2001)** Carlos AB Pereira e Julio M Stern. Fbst regularization and model selection. Em *Annals of the 7th International Conference on Information Systems Analysis and Synthesis (ISAS 2001), Orlando, Florida, FL, USA*, volume 7, páginas 60–65. Citado na pág. 16

**Phuong** *et al.***(2005)** Tu Minh Phuong, Zhen Lin e Russ B Altman. Choosing snps using feature selection. Em *Computational Systems Bioinformatics Conference, 2005. Proceedings. 2005 IEEE*, páginas 301–309. IEEE. Citado na pág. 7

**Pirooznia** *et al.***(2012)** Mehdi Pirooznia, Fayaz Seifuddin, Jennifer Judy, Pamela B Mahon, James B Potash, Peter P Zandi, Bipolar Genome Study (BiGS) Consortium *et al.* Data mining approaches for genome-wide association of mood disorders. *Psychiatric genetics*, 22(2):55. Citado na pág. 9

**Plomin e Kosslyn(2001)** Robert Plomin e Stephen M Kosslyn. Genes, brain and cognition. *Nature neuroscience*, 4(12):1153–1154. Citado na pág. 1

**Plomin** *et al.***(2001)** Robert Plomin, Kathryn Asbury e Judith Dunn. In review. *Can J Psychiatry*, 46:225–233. Citado na pág. 1

**Plomin** *et al.***(2009)** Robert Plomin, Claire MA Haworth e Oliver SP Davis. Common disorders are quantitative traits. *Nature Reviews Genetics*, 10(12):872–878. Citado na pág. 2, 9

**Pritchard e Przeworski(2001)** Jonathan K Pritchard e Molly Przeworski. Linkage disequilibrium in humans: models and data. *The American Journal of Human Genetics*, 69(1):1–14. Citado na pág. 8

**Purcell** *et al.***(2007)** Shaun Purcell, Benjamin Neale, Kathe Todd-Brown, Lori Thomas, Manuel AR Ferreira, David Bender, Julian Maller, Pamela Sklar, Paul IW De Bakker, Mark J Daly *et al.* Plink: a tool set for whole-genome association and population-based linkage analyses. *The American Journal of Human Genetics*, 81(3):559–575. Citado na pág. 21, 25

**Purcell** *et al.***(2009)** Shaun M Purcell, Naomi R Wray, Jennifer L Stone, Peter M Visscher, Michael C O'Donovan, Patrick F Sullivan e Pamela Sklar. Common polygenic variation contributes to risk of schizophrenia and bipolar disorder. *Nature*, 460(7256):748–52. ISSN 1476-4687. doi: 10.1038/nature08185. URL http://www.ncbi.nlm.nih.gov/pubmed/19571811. Citado na pág. 9

**Ramensky** *et al.***(2002)** Vasily Ramensky, Peer Bork e Shamil Sunyaev. Human non-synonymous snps: server and survey. *Nucleic acids research*, 30(17):3894–3900. Citado na pág. 2

**Saeys** *et al.***(2007)** Yvan Saeys, Iñaki Inza e Pedro Larrañaga. A review of feature selection techniques in bioinformatics. *bioinformatics*, 23(19):2507–2517. Citado na pág. 12

**Schwarz** *et al.***(2010)** Daniel F Schwarz, Inke R König e Andreas Ziegler. On safari to random jungle: a fast implementation of random forests for high-dimensional data. *Bioinformatics*, 26 (14):1752–1758. Citado na pág. 20

**Schwender e Ickstadt(2008)** Holger Schwender e Katja Ickstadt. Identification of snp interactions using logic regression. *Biostatistics*, 9(1):187–198. Citado na pág. 22

**Sebastiani(2002)** Fabrizio Sebastiani. Machine learning in automated text categorization. *ACM computing surveys (CSUR)*, 34(1):1–47. Citado na pág. 11

**Service** *et al.***(2012)** SK Service, KJH Verweij, J Lahti, E Congdon, J Ekelund, M Hintsanen, K Räikkönen, T Lehtimäki, M Kähönen, E Widen *et al.* A genome-wide meta-analysis of association studies of cloninger's temperament scales. *Translational psychiatry*, 2(5):e116. Citado na pág. 10

**Shastry(2009)** Barkur S Shastry. Snps: impact on gene function and phenotype. *Single Nucleotide Polymorphisms: Methods and Protocols*, páginas 3–22. Citado na pág. 2

**Shen *et al.*(2012)** Yuanyuan Shen, Zhe Liu e Jurg Ott. Support vector machines with l 1 penalty for detecting gene-gene interactions. *International journal of data mining and bioinformatics*, 6 (5):463–470. Citado na pág. 21

**Shih *et al.*(2004)** Regina A Shih, Pamela L Belmonte e Peter P Zandi. A review of the evidence from family, twin and adoption studies for a genetic contribution to adult psychiatric disorders. *International review of psychiatry*, 16(4):260–283. Citado na pág. 8

**Shirbani e Soltanian Zadeh(2013)** F Shirbani e H Soltanian Zadeh. Fast sffs-based algorithm for feature selection in biomedical datasets. *Amirkabir International Journal of Electrical & Electronics Engineering*, 45(2):43–56. Citado na pág. 17

**Silva(2011)** Sérgio Francisco da Silva. *Seleção de características por meio de algoritmos genéticos para aprimoramento de rankings e de modelos de classificação*. Tese de Doutorado, Universidade de São Paulo. Citado na pág. 15

**Simonoff *et al.*(2008)** Emily Simonoff, Andrew Pickles, Tony Charman, Susie Chandler, Tom Loucas e Gillian Baird. Psychiatric disorders in children with autism spectrum disorders: prevalence, comorbidity, and associated factors in a population-derived sample. *Journal of the American Academy of Child & Adolescent Psychiatry*, 47(8):921–929. Citado na pág. 1

**Skafidas *et al.*(2014)** Efstratios Skafidas, Renee Testa, Daniela Zantomio, Gursharan Chana, Ian P Everall e Christos Pantelis. Predicting the diagnosis of autism spectrum disorder using gene pathway analysis. *Molecular psychiatry*, 19(4):504–510. Citado na pág. 28

**Smedley *et al.*(2015)** Damian Smedley, Syed Haider, Steffen Durinck, Luca Pandini, Paolo Provero, James Allen, Olivier Arnaiz, Mohammad Hamza Awedh, Richard Baldock, Giulia Barbiera *et al.* The biomart community portal: an innovative alternative to large, centralized data repositories. *Nucleic acids research*, página gkv350. Citado na pág. 26

**Sobel *et al.*(2002)** Eric Sobel, Jeanette C Papp e Kenneth Lange. Detection and integration of genotyping errors in statistical genetics. *The American Journal of Human Genetics*, 70(2): 496–508. Citado na pág. 25

**Strachan e Read(2016)** Tom Strachan e Andrew Read. *Genética molecular humana*. Artmed Editora. Citado na pág. 8

**Stranger *et al.*(2011)** Barbara E Stranger, Eli A Stahl e Towfique Raj. Progress and promise of genome-wide association studies for human complex trait genetics. *Genetics*, 187(2):367–383. Citado na pág. 8

**Sullivan *et al.*(2000)** Patrick F Sullivan, Michael C Neale e Kenneth S Kendler. Genetic epidemiology of major depression: review and meta-analysis. *American Journal of Psychiatry*, 157(10): 1552–1562. Citado na pág. 7

**Sutha e Tamilselvi(2015)** K Sutha e J Jebamalar Tamilselvi. A review of feature selection algorithms for data mining techniques. *International Journal on Computer Science and Engineering*, 7(6):63. Citado na pág. 14

**Tang *et al.*(2009)** Yuchun Tang, Yan-Qing Zhang, Nitesh V Chawla e Sven Krasser. Svms modeling for highly imbalanced classification. *IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics)*, 39(1):281–288. Citado na pág. 19

**Tape((accessado em Dezembro de 2016))** Thomas G. Tape. The area under an roc curve. http://gim.unmc.edu/dxtests/roc3.htm, (accessado em Dezembro de 2016). Citado na pág. 30

**Tibshirani(2011)** Robert Tibshirani. Regression shrinkage and selection via the lasso: a retrospective. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 73(3):273–282. Citado na pág. 17

**Tibshirani** *et al.***(1997)** Robert Tibshirani *et al.* The lasso method for variable selection in the cox model. *Statistics in medicine*, 16(4):385–395. Citado na pág. 17

**Touw** *et al.***(2012)** Wouter G Touw, Jumamurat R Bayjanov, Lex Overmars, Lennart Backus, Jos Boekhorst, Michiel Wels e Sacha AFT van Hijum. Data mining in the life sciences with random forest: a walk in the park or lost in the jungle? *Briefings in bioinformatics*, página bbs034. Citado na pág. 20

**Tsymbal** *et al.***(2003)** Alexey Tsymbal, Pádraig Cunningham, Mykola Pechenizkiy e Seppo Puuronen. Search strategies for ensemble feature selection in medical diagnostics. Em *Computer-Based Medical Systems, 2003. Proceedings. 16th IEEE Symposium*, páginas 124–129. IEEE. Citado na pág. 15

**Turner** *et al.***(2011)** Stephen Turner, Loren L Armstrong, Yuki Bradford, Christopher S Carlson, Dana C Crawford, Andrew T Crenshaw, Mariza Andrade, Kimberly F Doheny, Jonathan L Haines, Geoffrey Hayes *et al.* Quality control procedures for genome-wide association studies. *Current protocols in human genetics*, páginas 1–19. Citado na pág. 25

**Uppu** *et al.***(2016)** Suneetha Uppu, Aneesh Krishna e Raj Gopalan. A review of machine learning and statistical approaches for detecting snp interactions in high-dimensional genomic data. *IEEE/ACM transactions on computational biology and bioinformatics*. Citado na pág. 19

**Vidaurre** *et al.***(2013)** Diego Vidaurre, Concha Bielza e Pedro Larrañaga. A survey of l1 regression. *International Statistical Review*, 81(3):361–387. Citado na pág. 17

**Visscher** *et al.***(2012)** Peter M Visscher, Matthew A Brown, Mark I McCarthy e Jian Yang. Five years of gwas discovery. *The American Journal of Human Genetics*, 90(1):7–24. Citado na pág. 7, 8

**Viswanathan** *et al.***(2011)** Viswa Viswanathan, Anup K Sen e Soumyakanti Chakraborty. Stochastic greedy algorithms. *International Journal on Advances in Software Volume 4, Number 1 & 2, 2011*. Citado na pág. 28

**Walder(2006)** Christian J Walder. Support vector machines for business applications. *Business Applications and Computational Intelligence*, página 267. Citado na pág. 11

**Wang** *et al.***(2016)** Lipo Wang, Yaoli Wang e Qing Chang. Feature selection methods for big data bioinformatics: A survey from the search perspective. *Methods*, 111:21–31. Citado na pág. 14, 15

**Wei** *et al.***(2009)** Zhi Wei, Kai Wang, Hui-Qi Qu, Haitao Zhang, Jonathan Bradfield, Cecilia Kim, Edward Frackleton, Cuiping Hou, Joseph T Glessner, Rosetta Chiavacci *et al.* From disease association to risk assessment: an optimistic view from genome-wide association studies on type 1 diabetes. *PLoS Genet*, 5(10):e1000678. Citado na pág. 21

**Wray** *et al.***(2008)** Naomi R Wray, Michael E Goddard e Peter M Visscher. Prediction of individual genetic risk of complex disease. *Current opinion in genetics & development*, 18(3):257–263. Citado na pág. 8

**Wray** *et al.***(2014)** Naomi R Wray, Sang Hong Lee, Divya Mehta, Anna AE Vinkhuyzen, Frank Dudbridge e Christel M Middeldorp. Research review: polygenic methods and their application to psychiatric traits. *Journal of Child Psychology and Psychiatry*, 55(10):1068–1087. Citado na pág. 9

**Wu** *et al.***(2010)** Michael C Wu, Peter Kraft, Michael P Epstein, Deanne M Taylor, Stephen J Chanock, David J Hunter e Xihong Lin. Powerful snp-set analysis for case-control genome-wide association studies. *The American Journal of Human Genetics*, 86(6):929–942. Citado na pág. 8

**Xu** *et al.***(2002)** Jianfeng Xu, Aubrey Turner, Joy Little, Eugene R Bleecker e Deborah A Meyers. Positive results in association studies are associated with departure from hardy-weinberg equilibrium: hint for genotyping error? *Human genetics*, 111(6):573–574. Citado na pág. 26

**Xuegong(2000)** Zhang Xuegong. Introduction to statistical learning theory and support vector machines. *Acta Automatica Sinica*, 26(1):32–42. Citado na pág. 20

**Zhou e Wang(2007)** Nina Zhou e Lipo Wang. Effective selection of informative snps and classification on the hapmap genotype data. *BMC bioinformatics*, 8(1):484. Citado na pág. 2

**Zhu** *et al.***(2005)** Xiaojin Zhu, John Lafferty e Ronald Rosenfeld. *Semi-supervised learning with graphs*. Carnegie Mellon University, language technologies institute, school of computer science. Citado na pág. 10

**Ziegler** *et al.***(2008)** Andreas Ziegler, Inke R König e John R Thompson. Biostatistical aspects of genome-wide association studies. *Biometrical Journal*, 50(1):8–28. Citado na pág. 19

**Zou e Hastie(2005)** Hui Zou e Trevor Hastie. Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 67(2):301–320. Citado na pág. 17