UNIVERSIDADE FEDERAL DO ESTADO DO RIO DE JANEIRO
CENTRO DE CIÊNCIAS EXATAS E TECNOLOGIA
PROGRAMA DE PÓS-GRADUAÇÃO EM INFORMÁTICA

# AVALIAÇÃO COGNITIVA COM UM JOGO COMPUTACIONAL UTILIZANDO TÉCNICAS INTELIGENTES

Fábio Eduardo Gabriel dos Santos

**Orientadoras**

Prof.ª D.ra Leila Cristina Vasconcelos de Andrade

Prof.ª D.ra Kate Cerqueira Revoredo

RIO DE JANEIRO, RJ - BRASIL

SETEMBRO DE 2011

AVALIAÇÃO COGNITIVA COM UM JOGO COMPUTACIONAL UTILIZANDO TÉCNICAS INTELIGENTES

Fábio Eduardo Gabriel dos Santos

DISSERTAÇÃO APRESENTADA COMO REQUISITO PARCIAL PARA OBTENÇÃO DO TÍTULO DE MESTRE PELO PROGRAMA DE PÓS-GRADUAÇÃO EM INFORMÁTICA DA UNIVERSIDADE FEDERAL DO ESTADO DO RIO DE JANEIRO (UNIRIO). APROVADO PELA COMISSÃO EXAMINADORA ABAIXO ASSINADA.

Aprovado por:

Leila Cristina Vasconcelos de Andrade, D. Sc.

Universidade Federal do Estado do Rio de Janeiro / PPGI

Kate Cerqueira Revoredo, D. Sc.

Universidade Federal do Estado do Rio de Janeiro / PPGI

Carlo Emmanoel Tolla de Oliveira, Ph. D.

Universidade Federal do Rio de Janeiro / NCE

Flávia Maria Santoro, D. Sc.

Universidade Federal do Estado do Rio de Janeiro / PPGI

RIO DE JANEIRO, RJ – BRASIL

SETEMBRO DE 2011

Santos, Fábio Eduardo Gabriel dos.
S237 Avaliação cognitiva com um jogo computacional utilizando técnicas inteligentes / Fábio Eduardo Gabriel dos Santos, 2011.
xvi, 124f.

Orientador: Leila Cristina Vasconcelos de Andrade.
Coorientador: Kate Cerqueira Revoredo.
Dissertação (Mestrado em Informática) – Universidade Federal do Estado do Rio de Janeiro, Rio de Janeiro, 2011.

1. Jogo computacional. 2. Jogo do Supermercado. 3. Transtorno do Déficit de Atenção e Hiperatividade - Diagnóstico. 4. Meta-aprendizagem. I. Andrade, Leila Cristina Vasconcelos de. II. Revoredo, Kate Cerqueira. III. Universidade Federal do Estado do Rio de Janeiro (2003-). Centro de Ciências Exatas e Tecnologia Curso de Mestrado Informática. III. Título.

CDD – 006

## Agradecimentos

Inicialmente, à minha orientadora, Leila Andrade, por me oferecer a oportunidade de realizar este mestrado. Agradeço por sua presteza, paciência, suporte e incentivo. Foi uma honra ser orientado por você.

Ao Dr. Paulo Mattos, por ter acompanhado este trabalho como especialista do domínio, fornecendo conhecimentos e dando os devidos esclarecimentos sobre os pontos nebulosos. Agradeço por seu suporte, interesse e, sobre tudo, por sua humildade em conduzir todos os assuntos. A sua participação, definitivamente, enriquece este trabalho.

À Paula Bastos, pelo suporte colaborativo que me foi oferecido. Agradeço pelo seu incentivo, companheirismo, paciência e acima de tudo pelo trabalho que você realizou nesta dissertação comigo. A sua colaboração foi essencial para a realização desta pesquisa.

À minha coorientadora Kate Revoredo, por ter me conduzido aos caminhos corretos dentro da área de Aprendizagem de Máquina. Agradeço pelos conhecimentos fornecidos, pelas ideias, sugestões e pelo incentivo. Sua participação foi muito importante para o enfoque deste trabalho.

Ao amigo Herli Menezes, pelo incentivo incondicional que me foi oferecido durante a minha caminhada no mestrado. – Agradeço por seus esclarecimentos, presteza, pelos conhecimentos divididos e, acima de tudo, por sua amizade. Sem a sua presença, o mestrado teria sido um desafio solitário.

Aos professores Ângelo Ciarlini e Flávia Santoro, que me forneceram sugestões nos seminários, essenciais para o delineamento deste trabalho. Agradeço pelas palavras de incentivo e pelos conhecimentos transmitidos.

Aos professores Mariano Pimentel, Sean Siqueira, Renata Araújo, Fernanda Baião, além dos professores citados anteriormente, que foram os protagonistas da minha formação de base dentro do mestrado. Sou muito grato pelos valiosos ensinamentos.

Aos meus pais, Josias Eduardo dos Santos e Mariléia Gabriel dos Santos, que foram responsáveis não apenas pelos meus 3 nomes próprios, mas também pelo meu caráter, que tem sido uma peça fundamental em todas as minhas vitórias.

À minha irmã, Dra. Angélica Santos, pela energia positiva desde o processo seletivo,

pela presença marcante em todos os momentos, pelo incentivo que me foi oferecido, pelos conselhos e pela força transmitida.

À minha “amiga++” Dra. Elena Yangicher que, mesmo não falando português, contribuiu muito para os assuntos pertinentes à língua inglesa. – Agradeço por ter me auxiliado na dicção, pela formação da minha terminologia da área, pelo incentivo e pelo carinho.

SANTOS, Fábio Eduardo Gabriel dos. **Avaliação Cognitiva com um Jogo Computacional Utilizando Técnicas Inteligentes**. UNIRIO, 2011. 140 páginas. Dissertação de Mestrado. Departamento de Informática Aplicada, UNIRIO.

# RESUMO

Neste trabalho, propõe-se um método alternativo para o auxílio ao diagnóstico do Transtorno do Déficit de Atenção e Hiperatividade (TDAH) através de um jogo computacional, intitulado Jogo do Supermercado, e de técnicas de Aprendizagem de Máquina. **Objetivo:** Identificar um conjunto de modelos preditivos eficientes que possam localizar padrões nos dados do jogo, que estejam relacionados ao diagnóstico do TDAH. **Método:** Duas bases de treino foram submetidas a $48$ algoritmos de aprendizagem de acordo com $4$ estratégias de Meta-Aprendizagem. Estas bases foram obtidas de $2$ subgrupos populacionais diferentes, e consistem de dados de indivíduos que jogaram o Jogo do Supermercado e foram avaliados de acordo com o TDAH. Uma das $4$ estratégias de Meta-Aprendizagem – a Seleção Dinâmica de Vieses Especializada (SDVE) – foi elaborada no escopo deste trabalho. Duas métricas de desempenho foram adotadas para a avaliação dos modelos: Área Abaixo da Curva ROC (AUC) e Erro Absoluto Relativo (EAR). Os desempenhos mínimos considerados para comprovação da eficiência dos modelos foram $0{,}70$ em AUC e $0{,}50$ em EAR. **Resultados:** A estratégia SDVE foi a única que conseguiu identificar modelos eficientes. Foram obtidos $145$ modelos eficientes para a base Bas, com desempenhos entre $0{,}70$ e $0{,}76$ em AUC, e $92$ modelos eficientes para a base Mat, com desempenhos entre $0{,}70$ e $0{,}80$ em AUC. **Conclusão:** Pode-se comprovar a existência de um conjunto de modelos preditivos eficientes que relacionam os dados do Jogo do Supermercado ao diagnóstico do TDAH, porém, subpopulações diferentes possuem modelos diferentes.

**Palavras-chave:** TDAH, Jogo, Diagnóstico, Meta-Aprendizagem.

# ABSTRACT

In this work, it is proposed an alternative method to aid in Attention Deficit Hyperactivity Disorder (ADHD) diagnosis through a computer game, named Supermarket Game, and Machine Learning techniques. **Objective:** To identify a set of efficient predictive models that can look for patterns in the game's data that can be related to the ADHD diagnosis. **Method:** Two training bases underwent $48$ learning algorithms according to $4$ Meta-Learning strategies. These bases were drawn from $2$ different population subsets, and consist of data of individuals that played the Supermarket Game and were assessed according to the ADHD. One of the $4$ Meta-Learning strategies – the Specialized Dynamic Bias Selection (SDBS) – was developed in this work. Two performance metrics were used to assess the models: Area Under the ROC Curve (AUC) and Relative Absolute Error (REA). The minimal performances considered to prove the models efficiency were $0{,}70$ in AUC and $0{,}50$ in EAR. **Results:** Only the SDBS strategy could identify efficient models. It was obtained $145$ efficient models to Bas training base, with performance between $0{,}70$ and $0{,}76$ in AUC, and $92$ efficient models to Mat training base, with performance between $0{,}70$ and $0{,}80$ in AUC. **Conclusion:** One can prove that there is a set of efficient predictive models that relate the Supermarket Game's data to the ADHD diagnosis, although different sub-populations have different models.

**Keywords:** ADHD, Game, Diagnosis, Meta-Learning

# Lista de Figuras

# Lista de Tabelas

# Lista de Símbolos

| | |
|---|---|
| **ASRS** | *Adult Self-Report Scale* |
| **AUC** | Área Abaixo da Curva ROC |
| **BAGGING** | *Bootstrap Aggregating* |
| **CSV** | *Comma-separated values* |
| **DFI** | Delineamento Fatorial Incompleto |
| **DSM-IV** | *Diagnostic and Statistical Manual of Mental Disorder – Fourth Version* |
| **EAR** | Erro Absoluto Relativo |
| **FP** | Falso Positivo |
| **FN** | Falso Negativo |
| **GNU** | *General Public License* |
| **IC** | Intervalo de Confiança |
| **IGT** | *Iowa Gambling Task* |
| **KDD** | *Knowledge-Discovery in Databases* |
| **MAD** | Meta-árvore de Decisão |
| **MBE** | Medicina Baseada em Evidências |
| **ROC** | *Receiver Operating Characteristic* |
| **SVM** | *Support Vector Machine* |
| **SDVE** | Seleção Dinâmica de Vieses Especializada |
| **SNAP-IV** | *Swanson-Nolan-Pelham questionnaire – Fourth Version* |
| **TE** | Tempo de Execução |
| **TDAH** | Transtorno do Déficit de Atenção e Hiperatividade |

| | |
|---|---|
| **TPC** | Tabela de Probabilidade Condicional |
| **VA** | Viés de Avaliação |
| **VD** | Viés de Decisão |
| **VI** | Variável Independente |
| **VN** | Verdadeiro Negativo |
| **VP** | Verdadeiro Positivo |
| **VR** | Viés de Representação |

# Sumário

# Capítulo 1

# Introdução

Neste capítulo, apresentam-se o contexto, o problema, a hipótese, os objetivos e a organização da dissertação. Inicialmente, discute-se a importância dos testes para o processo de diagnóstico médico e como a Aprendizagem de Máquina pode contribuir para essa área. Em seguida, apresentam-se o problema e a hipótese que estão sendo investigados. Logo após, discutem-se os objetivos e apresenta-se a estrutura de organização da dissertação.

## 1.1 Contexto

Testes exercem um importante papel no processo de diagnóstico médico, pois contribuem significativamente na redução dos esforços para se obter conclusões (EPSTEIN *et al.*, 1986). O principal objetivo de um teste no processo de diagnóstico – quer seja psicológico, laboratorial, radiológico, etc – é a redução da incerteza. O grau de redução dependerá das características do teste e do contexto clínico que está sendo considerado (WALLACH, 2007). A Medicina moderna vem desenvolvendo métodos cada vez mais eficientes para a realização de um diagnóstico. Tais métodos normalmente consistem na aplicação de testes confiáveis e na correta interpretação dos resultados produzidos por esses testes. Contudo, alguns testes podem possuir conceitos extremamente complexos, cujos resultados não são de fácil interpretação (ZHOU *et al.*, 2002).

A construção de modelos preditivos é a abordagem mais utilizada para a interpretação

de testes que produzem resultados complexos. A modelagem preditiva na Medicina pode ser definida como o processo de aplicar dados de pacientes já diagnosticados para prospectivamente identificar indivíduos em risco no futuro (WEINER, 2005). Assim, ao invés de se tentar entender quais são as bases fundamentais que fazem com que um indivíduo obtenha um determinado resultado no teste, busca-se identificar padrões comuns a todos os indivíduos que receberam o mesmo resultado. Um aspecto desafiador dessa abordagem está relacionado a como identificar tais padrões. Nesse sentido, a Medicina tem aplicado conhecimentos de diferentes disciplinas e especialidades para a construção de modelos eficientes.

Uma disciplina que vem desempenhando um papel importante no processo de construção de modelos preditivos para a Medicina é a Aprendizagem de Máquina. Sua principal aplicação nesse contexto consiste na abstração das medidas dos parâmetros vitais de um paciente, através da manipulação de dados obtidos de testes – laboratoriais, genéticos, raio-X, etc – para a identificação de padrões nesses dados que estejam relacionados a um diagnóstico (SAMMUT e WEBB, 2011). Contudo, devido ao vasto número de técnicas e algoritmos de Aprendizagem de Máquina que normalmente são igualmente adequados a um determinado problema de diagnóstico, o problema de modelagem deixa de ser a busca por um modelo preditivo eficiente e passa a ser a identificação de uma estratégia de aprendizagem com bom desempenho.

Nesse sentido, técnicas de Meta-Aprendizagem podem conduzir o processo de aprendizagem através de metamodelos que sugerem planos estratégicos promissores para a identificação de um modelo preditivo adequado (VILALTA *et al.*, 2005; BRAZDIL *et al.*, 2009). Contudo, propostas automáticas para essa abordagem podem não produzir resultados satisfatórios se metaconhecimentos compatíveis com o domínio não estiverem disponíveis, e / ou se os dados para treino forem escassos. Uma alternativa para a construção de metamodelos eficientes em dadas circunstâncias é a elicitação de metaconhecimentos do domínio junto a um especialista para sugestão de vieses que possam reduzir o espaço de busca (NGUYEN, 2010). O objetivo dessa estratégia é utilizar os conhecimentos do especialista para a construção de um metamodelo que possa conduzir o processo de

aprendizagem, desconsiderando alternativas menos promissoras. Essa é a estratégia que está sendo sugerida nesta dissertação para solucionar o problema de pesquisa discutido na próxima seção.

## 1.2 Problema e Hipótese

O tema que está sendo discutido nesta dissertação é o auxílio ao diagnóstico do Transtorno do Déficit de Atenção e Hiperatividade (TDAH) através de um método alternativo. O TDAH é um transtorno psiquiátrico que possui um método subjetivo de diagnóstico, que exige perícia por parte do psiquiatra para a identificação de casos positivos. Esse diagnóstico normalmente baseia-se na aplicação de questionários, que são respondidos por informantes próximos ao paciente, e na aplicação de testes alternativos para a identificação de sintomas ou déficits relacionados – ou comorbidades. Os resultados obtidos nos questionários e nos testes são geralmente os únicos dados utilizados pelo psiquiatra para se estabelecer um diagnóstico, já que os indivíduos portadores do transtorno normalmente não manifestam seus sintomas durante a consulta médica (SIMITH *et al.*, 2007). Assim, por não haver um teste objetivo para a avaliação dos pacientes, o processo de diagnóstico tradicional do TDAH depende tanto da perícia do especialista em interpretar os resultados dos questionários e dos testes alternativos, quanto do nível de comprometimento dos informantes.

O **problema** que está sendo investigado nesta dissertação é o fato de não existir ainda um teste padronizado para o diagnóstico do TDAH. Embora alguns testes neuropsicológicos sejam utilizados durante o processo, eles não foram desenvolvidos especificamente para a identificação de casos do TDAH, e, dessa forma, precisam ser interpretados, assim como os questionários.

A hipótese de pesquisa que está sendo investigada está apoiada em $2$ pilares principais. O primeiro refere-se à utilização de um jogo computacional, intitulado Jogo do Supermercado, que foi desenvolvido inicialmente para provar que jogos fazem captura cognitiva. A motivação para essa abordagem deve-se ao fato de que este jogo já obteve bons resultados preditivos para o TDAH em uma pequena amostra de $10$ indivíduos adultos, através de um

modelo produzido pelo algoritmo Naive Bayes (ANDRADE, 2009). Como este estudo considerou apenas um algoritmo de aprendizagem e uma pequena amostra de treino – já que seu objetivo não era criar um teste para o TDAH – um modelo preditivo seguro ainda não pode ser definido.

O segundo pilar considera a utilização de conhecimentos obtidos através de um especialista do domínio para a elaboração de um plano de aprendizagem. A motivação para essa abordagem deve-se ao fato de que, no escopo deste trabalho, um especialista consagrado do domínio está disponível para a elicitação de conhecimentos. Além disso, como esse domínio não dispõe de uma grande quantidade de dados para processamento, algumas estratégias de Meta-Aprendizagem não poderiam, a princípio, ser consideradas – como, por exemplo, *Learn to Learn*.

A **hipótese** que está sendo aqui levantada é a de que se uma amostra suficientemente grande[1] – de indivíduos que jogaram o Jogo do Supermercado e foram avaliados em relação ao TDAH – for submetida a um metamodelo construído através de conhecimentos elicitados do domínio, então um conjunto de modelos preditivos eficientes poderá ser identificado para a elaboração de um teste.

## 1.3 Objetivos da Dissertação

O trabalho apresentado nesta dissertação visa, primeiramente, identificar um conjunto de modelos preditivos eficientes, capazes de localizar padrões nos dados do Jogo do Supermercado que estejam relacionados ao diagnóstico do TDAH. Modelos identificados nesse contexto são essenciais para a construção de um teste padronizado.

Para que as análises pudessem ser corretamente organizadas, nesta dissertação, a avaliação dos modelos está sendo conduzida na forma de um experimento com delineamento fatorial. Nessa abordagem, ao invés de se manipular apenas uma Variável Independente (VI), várias VIs (ou fatores) são manipuladas ao mesmo tempo. Cada diferente arranjo de fatores é um modelo preditivo avaliado no experimento.

[1] Nesse caso, uma amostra suficientemente grande deve possuir uma quantidade de instâncias igual ou superior a dimensionalidade das características do domínio (número de atributos relevantes).

Outra abordagem que está sendo considerada são os subdomínios amostrais. Como o TDAH manifesta-se de forma diferente em adultos e crianças, com o objetivo de verificar a capacidade preditiva do jogo nesses 2 subdomínios, duas bases de treino estão sendo utilizadas no experimento. Como essas são bases que pertencem a subdomínios diferentes, elas estão sendo, a princípio, processadas de forma independente. Assim, dois conjuntos de modelos estão sendo sugeridos, sendo um para cada subdomínio.

Outra questão que está sendo verificada refere-se à utilização de outras estratégias de Meta-Aprendizagem. Como este trabalho propõe a utilização de um metamodelo construído a partir de conhecimentos do domínio, considerando-se a disponibilidade de outras abordagens igualmente aplicáveis ao problema, verificou-se que uma análise mais completa deveria também considerar outras abordagens. Assim, outro aspecto que também está sendo avaliado nesta dissertação é a comparação do desempenho obtido pela abordagem sugerida, com outras 3 estratégias de Meta-Aprendizagem. O objetivo aqui é comparar o desempenho obtido pelo metamodelo sugerido com 2 estratégias de Meta-Aprendizagem automáticas e uma estratégia ingênua de seleção aleatória de algoritmos.

## 1.4 Organização da Dissertação

Além deste capítulo introdutório, esta dissertação está estruturada de acordo com o descrito a seguir.

O Capítulo 2 apresenta conceitos importantes para a compreensão do trabalho aqui apresentado. Descreve-se o TDAH com suas características e métodos de diagnóstico, o Jogo do Supermercado, além de conceitos e técnicas de Mineração de Dados utilizados no trabalho.

O Capítulo 3 apresenta a proposta de Meta-Aprendizagem que foi sugerida para solucionar o problema de pesquisa. Nesse capítulo, faz-se a descrição das propostas, discute-se a importância dos metaconhecimentos, apresenta-se a técnica utilizada para definição da estrutura de busca, discute-se como é realizada a execução de um metamodelo através dessa abordagem e apresenta-se um metamodelo para o problema de pesquisa aqui discutido.

O Capítulo 4 faz uma descrição do experimento. Estão sendo discutidos a caracterização do experimento, as amostras utilizadas, as ferramentas utilizadas, os mecanismos de base aplicados, as estratégias de Meta-Aprendizagem consideradas, as métricas de desempenho utilizadas, os métodos de validação, a estrutura que foi construída para aplicação da abordagem SDVE e o método utilizado para sua aplicação.

O Capítulo 5 apresenta os resultados obtidos. Discutem-se os resultados obtidos através das estratégias Seleção Aleatória de Técnicas de Aprendizagem, Meta-Aprendizagem Através de Mecanismos de Base, Mapeamento em Metanível e SDVE. Além disso, é realizada uma reavaliação de algumas estratégias, considerando-se outros argumentos hipotéticos. Ao final, discutem-se os resultados obtidos no experimento.

O Capítulo 6 é dedicado à apresentação das considerações finais, à analise de trabalhos relacionados, às contribuições do trabalho apresentado nesta dissertação e à sugestão de trabalhos futuros.

# Capítulo 2

# Fundamentação Teórica

Neste capítulo são discutidos alguns conceitos importantes utilizados no desenvolvimento do trabalho apresentado nesta dissertação. Inicialmente, é feita uma breve descrição do Transtorno do Déficit de Atenção e Hiperatividade (TDAH) e do estado da arte dos métodos utilizados para o seu diagnóstico. Em seguida, apresenta-se o Jogo do Supermercado, descrevendo-se suas principais características. Por fim, faz-se uma revisão na área de Mineração de Dados.

## 2.1 O Transtorno do Déficit de Atenção e Hiperatividade

O Transtorno do Déficit de Atenção e Hiperatividade (TDAH) é um transtorno psiquiátrico que, dependendo de sua gravidade, pode trazer sérios comprometimentos na vida acadêmica, profissional e familiar de indivíduos adultos e de crianças (SIMITH *et al.*, 2007; SHIMITZ *et al.*, 2002; BROOK e GEVA, 2001). De acordo com BARKLEY (1997), o TDAH está associado a prejuízos causados na vida do indivíduo, como baixo rendimento acadêmico, retenção em grades, suspensões e expulsões escolares, falta de relacionamento familiar, ansiedade, depressão, agressão, problemas de conduta, delinquência e experimentação prematura de substâncias. Além disso, no caso específico de indivíduos adultos, o TDAH pode estar relacionado a acidentes de trânsito e altas velocidades, dificuldades no relacionamento social adulto, casamento e emprego. Estudos mostram que o TDAH pode persistir na fase adulta em 60% a 70% dos casos observados

em crianças (SHIMITZ *et al.*, 2002; KESSLER *et al.*, 2005).

As características do TDAH são definidas no Manual Diagnóstico e Estatístico de Transtornos Mentais (DSM-IV, do inglês *Diagnostic and Statistical Manual of Mental Disorder – Fourth Version*), de acordo com 5 critérios:

- **Critério A:** Seis ou mais sintomas podem ser observados no indivíduo, em um dos grupos de sintomas listados na Tabela 2.1. Esses 2 grupos de sintomas – que compreendem sintomas de desatenção e sintomas de hiperatividade / impulsividade – definem 3 subtipos do transtorno: (1) Predominantemente desatento; (2) Predominantemente hiperativo / impulsivo; (3) Combinado.

- **Critério B:** Alguns sintomas de desatenção ou hiperatividade / impulsividade que hoje causam os prejuízos já estavam presentes antes dos 7 anos de idade.

- **Critério C:** Alguns prejuízos causados pelos sintomas estão presentes em pelo menos 2 cenários (Ex.: na escola e em casa).

- **Critério D:** Deve haver evidências claras de prejuízos clinicamente significantes na vida social, acadêmica e no lazer.

- **Critério E:** Os sintomas não devem ocorrer exclusivamente durante o curso de outros transtornos psicológicos como, por exemplo, esquizofrenia, e não devem ser melhor explicados por outro transtorno mental como, por exemplo, transtorno de humor, transtorno de ansiedade, transtorno de personalidade.

**Tabela 2.1** Sintomas de desatenção e sintomas de hiperatividade / impulsividade, de acordo com o critério A do DSM-IV para o TDAH.

| **Sintomas Desatenção** | **Sintomas Hiperatividade / Impulsividade** |
|---|---|
| – Frequentemente deixa de prestar atenção a detalhes ou comete erros por descuido em atividades escolares, de trabalho ou outras. | – Frequentemente agita as mãos ou os pés ou se remexe na cadeira. |
| – Com frequência tem dificuldades para manter a atenção em tarefas ou atividades lúdicas | – Frequentemente abandona sua cadeira em sala de aula ou outras situações nas quais se espera que permaneça sentado. |

Continua na próxima página...

**Tabela 2.1** – Continuação

| Sintomas Desatenção | Sintomas Hiperatividade / Impulsividade |
|---|---|
| – Com frequência parece não escutar quando lhe dirigem a palavra. | – Frequentemente corre ou escala em demasia, em situações nas quais isso é inapropriado (em adolescentes e adultos, pode estar limitado a sensações subjetivas de inquietação) |
| – Com frequência não segue instruções e não termina seus deveres escolares, tarefas domésticas ou deveres profissionais (não devido a comportamento de oposição ou incapacidade de compreender instruções) | – Com frequência tem dificuldade para brincar ou se envolver silenciosamente em atividades de lazer. |
| – Com frequência tem dificuldade para organizar tarefas e atividades. | – Frequentemente fala em demasia. |
| – Com frequência evita, antipatiza ou reluta a envolver-se em tarefas que exijam esforço mental constante (como tarefas escolares ou deveres de casa). | – Frequentemente dá respostas precipitadas antes de as perguntas terem sido completadas. |
| – Com frequência perde coisas necessárias para tarefas ou atividades (por ex., brinquedos, tarefas escolares, lápis, livros ou outros materiais). | – Com frequência tem dificuldade para aguardar sua vez em jogos ou situações em grupo. |
| – É facilmente distraído por estímulos alheios à tarefa. | – Frequentemente interrompe ou se mete em assuntos de outros (por ex., intromete-se em conversas ou brincadeiras). |
| – Com frequência apresenta esquecimento em atividades diárias. | – Frequentemente está se movimentando, como se estivesse sendo conduzido por um motor. |

O diagnóstico do TDAH é clínico. Em crianças e adolescentes, esse diagnóstico é baseado fundamentalmente na observação dos sintomas atuais. Já em adultos, o diagnóstico é normalmente realizado através da combinação entre os atuais sintomas observados e o histórico médico de comportamentos do indivíduo. A descrição de sintomas, baseada nos critérios do DSM-IV, é o método mais aceito para o diagnóstico do TDAH (BARKLEY, 1997; KUPFER, 2000). Como é descrito no critério C do DSM-IV, os demais critérios devem considerar a observação de sintomas em 2 ou mais contextos ou circunstâncias, tais como escola e casa. Uma forma de se obter um relatório desses sintomas é a aplicação de questionários padronizados. Esses questionários podem ser preenchidos por professores, pais, ou até mesmo pelo próprio indivíduo. Um exemplo de questionário padronizado é o Swanson-Nolan-Pelham-IV (SNAP-IV), baseado nos sintomas descritos no DSM-IV para o diagnóstico de crianças e adolescentes, através da avaliação e quantificação de sin-

tomas (SWANSON *et al.*, 2001). Outro questionário padronizado amplamente utilizado é o Adult Self-Report Scale (ASRS), que também é baseado no DSM-IV, porém adaptado para uso em adultos (KESSLER *et al.*, 2005). Como os pacientes normalmente não apresentam os sintomas do transtorno durante a consulta médica, o diagnóstico final dependerá da confiança nos relatórios e da perícia do psiquiatra em interpretar e avaliar os relatórios e o histórico do paciente.

Apesar de os relatórios serem amplamente utilizados no processo de diagnóstico, algumas questões sobre essa técnica devem ser observadas. Alguns estudos mostram que o grau correspondência entre os relatórios dos pais e dos professores é modesto, coincidindo em apenas metade dos itens (DE NIJS *et al.*, 2004), embora seja importante a utilização de vários informantes quando se diagnostica esse transtorno (MITSIS *et al.*, 2000). Pais normalmente relatam mais sintomas de hiperatividade / impulsividade em casa, e professores descrevem mais sintomas de desatenção na escola (SERRA-PINHEIRO *et al.*, 2008). Contudo, segundo SIMITH *et al.* (2007), esse desacordo entre as fontes pode estar refletindo alguma diferença real no comportamento dos indivíduos nesses diferentes contextos, que provavelmente ocorre em função de diferentes situações e demandas. Assim, a avaliação do TDAH através de relatórios talvez reflita diferentes julgamentos de diferentes pessoas, o que pode ser um problema porque não há bases científicas, até este momento, que comprovem a possibilidade de se validar um relatório através do outro. Esses relatórios são apenas informações fornecidas sobre o comportamento do indivíduo, em um determinado contexto, a partir do ponto de vista particular de um informante.

Nesse contexto, o clínico pode ser auxiliado no processo de diagnóstico através de testes neuropsicológicos a fim de se modelar o perfil cognitivo do paciente. Estudos mostram que o TDAH é associado a vários déficits neuropsicológicos[1] (FRAZIER *et al.*, 2004). O teste de atenção visual TAVIS-III, por exemplo, foi desenvolvido com o objetivo de avaliar crianças e adolescentes com idades entre 6 e 7 anos (MATTOS e DUCHESNE, 1997). Esse teste avalia vários níveis de atenção visual como sensibilidade, mudança de conceito e sustentação visual, produzindo vários escores de acordo com a tarefa execu-

[1]Muito embora pacientes pertencentes exclusivamente ao subtipo predominantemente hiperativo / impulsivo parecem não apresentar significativamente tais déficits (SHIMITZ *et al.*, 2002).

tada. Estudos mostram que esse teste pode contribuir para o diagnóstico do TDAH (COUTINHO *et al.*, 2007). Outro teste conhecido como *Iowa Gambling Task* (IGT) é também amplamente utilizado em diagnósticos do TDAH. Esse teste consiste de $4$ pilhas de cartas, em que a cada momento o indivíduo é solicitado a pegar uma carta de uma das pilhas, e dependendo da carta, pode-se ganhar ou perder dinheiro virtual. Em algumas pilhas o prêmio é grande, mas o jogador perde mais do que ganha. Em outras pilhas o prêmio é pequeno, mas o jogador ganha mais do que perde. Estudos com esse teste mostraram que um participante sadio, após ter selecionado em torno de $40$ cartas, pode identificar as boas pilhas. Por outro lado, participantes com disfunção executiva mantêm-se insistindo nas pilhas ruins (BECHARA *et al.*, 1997, 1994).

Apesar de os testes apresentados anteriormente poderem contribuir para o diagnóstico do TDAH, vale ressaltar que eles são utilizados em caráter sugestivo, pois não foram desenvolvidos especificamente para esse objetivo. Nesse contexto, o TDAH não possui ainda um teste padronizado.

## 2.2 O Jogo do Supermercado

O Jogo do Supermercado foi desenvolvido, inicialmente, para provar que jogos podem fazer captura cognitiva ANDRADE (2009), e foi inspirado em um teste neuropsicológico – chamado *Zoo Map Test* – aplicado na avaliação da disfunção executiva WILSON *et al.* (1997).

O jogo é basicamente um labirinto que deve ser atravessado enquanto o jogador adquire itens de uma lista de compras (ver Figura 2.1). Em sua interface, há um mapa do supermercado, uma lista de compras no lado direito, o escore de pontos para cada tarefa realizada, e o tempo gasto. O jogador personaliza um cliente do supermercado – através de um avatar – que deve ser controlado através das setas do teclado.

O jogo possui $18$ fases divididas em $2$ modos. O modo $1$ possui $10$ fases nas quais o avatar deve adquirir todos os itens mostrados na lista de compras no menor tempo possível, sem passar no mesmo lugar mais de uma vez. O modo $2$ possui $8$ fases nas quais o avatar também deve adquirir todos os itens mostrados na lista de compras no


**Figura 2.1** O Jogo do Supermercado.

menor tempo possível, sem passar pelo mesmo lugar mais de uma vez porém, dessa vez os itens devem ser adquiridos na mesma ordem da lista de compras. Em ambos os modos, a cada estágio a lista de compras é acrescida de 1 item. Cada modo inicia-se com 1 item na lista de compras. O modo 1 avalia a capacidade de planejamento do jogador e o modo 2 avalia a capacidade de execução do jogador.

As regras que devem ser seguidas em todos os estágios são: O avatar deve iniciar as compras na entrada de clientes e finalizar no caixa, localizado na saída do supermercado. O caminho enfatizado em azul pode ser atravessado quantas vezes forem necessárias, enquanto todos os demais caminhos só podem ser atravessados uma vez. Todos os itens presentes na lista devem ser adquiridos. Um ponto é ganho para cada item da lista de compras que é adquirido. Iniciar e finalizar nos locais corretos também soma um ponto positivo cada. Se o jogador atravessar novamente uma área já atravessada, comete-se uma falta, penalizada com a perda de 1 ponto. Considerando o pior cenário, pode-se obter uma pontuação negativa.

## 2.3 Mineração de Dados

Nesta seção, discutem-se os princípios da Mineração de dados, os fundamentos da aprendizagem supervisionada, métodos para validação de resultados, o efeito *overfitting*, métricas de desempenho, intervalos de confiança, técnicas de discretização, a delimitação do espaço de hipóteses através de vieses, conceitos básicos de Meta-Aprendizagem e de descoberta de conhecimento como um todo.

### 2.3.1 Princípios da Mineração de Dados

Bancos de dados podem ocultar informações potencialmente úteis que não serão descobertas facilmente através dos métodos tradicionais de análise de dados ou de linguagens de consulta. A Mineração de Dados é uma ciência multidisciplinar que tem por objetivo investigar técnicas para extração automática e conveniente de padrões que representam um conhecimento implícito armazenado em um conjunto de dados (HAN e KAMBER, 2006).

O princípio fundamental da tarefa de minerar dados está baseado na identificação de estruturas ou padrões através de exemplos. Um padrão identificado nos dados é sustentado por uma hipótese indutiva que pode ser expressa na forma de modelos produzidos por uma função. Como as hipóteses indutivas pertencem a um universo abstrato, apenas aquelas que podem ser expressas por modelos através de funções podem ser identificadas e, dessa forma, buscar por padrões nos dados é na verdade identificar os modelos eficientes que podem ser expressos por um conjunto disponível de funções adequadas aos dados. Em Mineração de Dados, funções normalmente são definidas através de algoritmos. Apesar de os termos "**hipótese**" e "**modelo**" serem normalmente utilizados como sinônimos, essa distinção está sendo explicitamente adotada neste trabalho.

O cerne da Mineração de Dados é a disciplina Aprendizagem de Máquina. Essa é uma disciplina derivada da Inteligência Artificial, interessada em construir sistemas computacionais capazes de se aprimorar através de experiências (MITCHELL, 1997). Essa disciplina fornece subsídios técnicos para a Mineração de Dados através de algoritmos especializados em analisar "dados brutos" e inferir qualquer estrutura que os sustente.

Algoritmos de Aprendizagem de Máquina aplicados a Mineração de Dados devem ser robustos o suficiente para lidar com dados imperfeitos, extraindo regularidades que muitas vezes são inexatas, porém úteis (WITTEN e FRANK, 2005).

A Mineração de Dados é normalmente aplicada a dois objetivos primários (FAYYAD *et al.*, 1996):

- **Predição:** Estimar acontecimentos futuros através da análise de dados do passado, como por exemplo, previsão do tempo, análise de risco de crédito e diagnóstico médico;

- **Descrição:** Compreender padrões encontrados nos dados e explicar comportamentos até então desconhecidos – como, por exemplo, a identificação de uma combinação de produtos normalmente adquiridos por um perfil específico de cliente.

Apesar de os limites entre predição e descrição não serem exatos (já que alguns modelos preditivos podem ser também descritivos e vice-versa) essa distinção é útil para o entendimento da tarefa de mineração como um todo.

A utilização de uma técnica de aprendizagem adequada é essencial para o sucesso da Mineração de Dados. Contudo a identificação dessa técnica pode exigir paciência e consumir tempo. Uma informação oculta em uma base de dados pode assumir diferentes formas que serão melhor identificadas através de técnicas específicas. Porém, a técnica mais adequada para cada situação é raramente óbvia. Normalmente não se sabe que tipos de padrões nos dados serão de interesse e, assim, a utilização de diferentes técnicas de mineração pode auxiliar na identificação de um modelo eficiente. É importante que um sistema para mineração possa lidar com diversos tipos de padrões e estruturas a fim de acomodar diferentes expectativas e aplicações do usuário (HAN e KAMBER, 2006). Por ser um processo tedioso, muitas vezes até mesmo o gerenciamento da sequência de tarefas de mineração deve ser realizado de forma automática, já que um possível modelo que se busca nos dados normalmente surge não de uma fórmula complexa, mas da aplicação contínua e paciente de um algoritmo simples, quase que em um processo de força bruta (ALPAYDIN, 2010).

A Mineração de Dados pode ser realizada de diferentes formas, através de diferentes técnicas. A escolha das técnicas que serão aplicadas dependerá da natureza dos dados e dos objetivos que se deseja alcançar. Essas técnicas definirão o tipo de padrão que se espera encontrar nos dados, como eles poderão ser minerados e como os resultados serão apresentados. Assim, "definir técnicas" é uma tarefa que não se restringe apenas à determinação dos algoritmos adequados para a mineração. Outros aspectos também deverão ser especificados, se possível através de um plano de ação. Um plano (ou projeto) de mineração deve contemplar:

1. A escolha dos algoritmos que processarão os dados;
2. A definição dos métodos que avaliarão o desempenho dos modelos;
3. A escolha das métricas que apresentarão os resultados;
4. A determinação do nível de confiança dos resultados produzidos.

A Mineração de Dados é uma ciência multidisciplinar que abrange áreas como Bancos de Dados, Aprendizagem de Máquina, Estatística, Recuperação de Informação, Inteligência Artificial, Computação de Alta Performance e Visualização (HAN e KAMBER, 2006). A Aprendizagem de Máquina e a Estatística são as áreas de maior influência dentro da Mineração de Dados e, de acordo com essas áreas, os algoritmos para mineração podem ser organizados em uma taxonomia baseada nos resultados desejados, da seguinte forma:

- Aprendizagem Supervisionada
  - Algoritmos de Classificação
  - Algoritmos de Regressão
- Aprendizagem Não Supervisionada
  - Algoritmos de Agrupamento

A aprendizagem supervisionada foi a estratégia adotada nesta dissertação e, dessa forma, está sendo discutida com mais detalhes nas próximas subseções. Informações sobre a aprendizagem não supervisionada podem ser vistas em ALPAYDIN (2010).

### 2.3.2 Aprendizagem Supervisionada Aplicada à Mineração de Dados

A aprendizagem supervisionada, quando aplicada a tarefas de Mineração de Dados, corresponde a aprender a como classificar os elementos de um conjunto, dado outro conjunto de elementos corretamente classificados. Aqui, o conceito de classificação está sendo estendido também à regressão, cujo problema corresponde à aproximação de um valor contínuo.

De acordo com VILALTA e DRISSI (2002), o processo de aprendizagem em si é realizado da seguinte forma: Um algoritmo $L$ é inicialmente treinado em um conjunto pré-classificado de exemplos $T\{(X_i,c_i)\}_i$. Cada objeto $X$ é caracterizado por um vetor de atributos, $X = (x_1,x_2,\ldots,x_n)$, podendo cada $x_k$ assumir diferentes valores. $X_i$ é rotulado com uma classe $c_i$ de acordo com uma função central $F$, $F(X_i) = c_i$. No caso de classificação, cada $c_i$ assume um valor de um conjunto pré-fixado de valores de categorias e, no caso de regressão, $c_i$ assume um valor contínuo dentro de uma extensão de valores numéricos. Um conjunto $T$ é composto de exemplos, que obedecem a uma distribuição de probabilidade conjunta fixa e desconhecida $\Phi$ no espaço de possíveis vetores de atributos versus classes, $X \times c$, que também é conhecido como espaço de entrada-saída. O objetivo da classificação é produzir uma hipótese $h$ que melhor aproxime $F$ no espaço de entrada-saída de acordo com $\Phi$.

O processo de aprendizagem inicia-se quando um algoritmo $L$ recebe como entrada um conjunto de treino $T$, e realiza buscas no espaço de hipóteses $H_L$ até encontrar uma hipótese $h$, $h \in H_L$, que aproxime a verdadeira função central $F$. A hipótese identificada $h$ pode, então, ser utilizada através de um modelo para prever valores de classe desconhecidos de novos exemplos.

A seguir, apresentam-se algumas das técnicas mais utilizadas por algoritmos que adotam essa abordagem.

### Regras

A aprendizagem através de Regras é um método muito popular para a construção de modelos preditivos. Esse método tem como vantagem poder ser facilmente utilizado por humanos na construção e interpretação de pequenos modelos. Na Tabela 2.2 é apresentado um conjunto de exemplos, onde cada uma das $10$ instâncias é descrita através de $3$ atributos (Sexo, País, Idade) e classificada através do atributo "Comprador", que pode assumir os valores "Sim" ou "Não".

**Tabela 2.2** Clientes potenciais compradores de um determinado livro.

| Sexo | País | Idade | **Comprador** |
|---|---|---|---|
| Masculino | Argentina | 25 | Sim |
| Masculino | Brasil | 21 | Sim |
| Feminino | Argentina | 23 | Sim |
| Feminino | Brasil | 34 | Sim |
| Feminino | Argentina | 30 | Não |
| Masculino | Bolívia | 21 | Não |
| Masculino | Bolívia | 20 | Não |
| Feminino | Bolívia | 18 | Não |
| Masculino | Bolívia | 34 | Não |
| Masculino | Argentina | 55 | Não |

Utilizando o conjunto de exemplos fictício apresentado na Tabela 2.2, pode-se construir um modelo para predição de possíveis compradores e não compradores do livro oferecido, através do seguinte conjunto de regras:

`SE País=Bolívia ENTÃO Comprar=Não`

`SE País=Brasil ENTÃO Comprar=Sim`

`SE País=Argentina E Idade` $\leqslant 25$ `ENTÃO Comprar=Sim`

`SE País=Argentina E Idade` $> 25$ `ENTÃO Comprar=Não`

O modelo apresentado no exemplo acima foi construído através de Regras de Classificação *SE-ENTÃO*. O antecedente, que encontra-se entre o *SE* e o *ENTÃO*, descreve uma série de testes (ou pré-condições) que deverão ser avaliados na regra. O consequente, que se encontra após o *ENTÃO*, pode apresentar como saída uma classe, um conjunto de classes, uma distribuição de probabilidade sobre um valor numérico, ou ainda uma função

que pode utilizar as características (ou atributos) apresentados no antecedente. Normalmente, os testes são realizados na pré-condição de uma regra através de uma conjunção simples, utilizando conectivos lógicos *AND* (E), e todo o conjunto de regras é visto como uma disjunção em que cada regra individual é separada por um conectivo lógico *OR* (OU). As pré-condições podem também utilizar outras expressões lógicas, além de conjunções WITTEN e FRANK (2005).

Regras também podem ser utilizadas para representar Associações (independentes de Classe) identificadas em um domínio. Ao invés de apresentar no consequente apenas o atributo Classe, Regras de Associação buscam relações entre outros atributos que possam ser observadas na base de treino (HAN e KAMBER, 2006). Por exemplo:

`SE Sexo=Masculino E País=Bolívia ENTÃO Idade<25 E Comprar=Não`

Normalmente, buscam-se regras de associação que cubram corretamente o maior número possível de casos em uma base de treino e que sejam acuradas o suficiente – ou seja, regras que possuam um alto percentual de acerto em suas predições. No exemplo anterior, a Cobertura seria a quantidade de clientes do sexo masculino, bolivianos, com menos de $25$ anos de idade que não foram compradores do livro (há $2$ casos na Tabela 2.2). A Acurácia é a proporção de casos de clientes realmente compradores do livro com idade menor que $25$ anos ($2$ casos), dentre todos os casos em que o cliente é do sexo masculino e boliviano ($3$ casos, resultando, então, em uma acurácia$= 2/3 = 0{,}67$).

Apesar de ser um método a princípio "sedutor" de aprendizagem, devido à sua facilidade de interpretação, modelos construídos através de Regras podem conter classificações equivocadas difíceis de serem identificadas. Uma nova regra pode ser facilmente adicionada a um modelo. Contudo essa facilidade pode ocasionar sérios conflitos lógicos se não forem consideradas as regras já existentes e o modo como essas regras são executadas. Modelos construídos através de regras podem ser interpretados como: (1) um Conjunto de Regras Independentes, em que cada regra é autossuficiente e pode ser tomada ao acaso; (2) uma Lista Ordenada de Decisão, em que as regras devem ser verificadas em uma ordem pré-estabelecida. Regras de um Conjunto de Regras Independentes devem ser capazes de lidar com qualquer subconjunto de exemplos do domínio. Regras de uma

lista ordenada de decisão são preparadas para lidar com casos específicos de exemplos e, dessa forma, uma regra tomada individualmente, fora do contexto, deverá estar incorreta. Conjuntos de Regras Independentes, apesar de proporcionarem maior liberdade, podem conduzir a problemas intratáveis como, por exemplo, 2 duas regras que produzem diferentes conclusões para uma mesma instância. Assim, dependendo da complexidade do problema de mineração, outros métodos como Árvores de Decisão, Redes Bayesianas, Redes Neurais, Aprendizagem Baseada em Instâncias e Máquinas de Vetores de Suporte poderão ser mais adequados.

### Árvores de Decisão

Uma Árvore de Decisão é uma estrutura lógica de fluxo, onde cada nó representa um teste no valor de um atributo e cada ramo representa um possível resultado para o teste (HAN e KAMBER, 2006). Os resultados da classificação são observados nas folhas que representam valores de classe ou distribuições de classe. Essa estrutura é construída recursivamente pelo algoritmo de Árvore de Decisão, que considera primeiramente os nós (atributos) mais puros na raiz da árvore. Dessa forma, os atributos cujos ramos possuem o maior número possível de instâncias de uma só classe são considerados primeiro. WITTEN e FRANK (2005) mostram que uma função para cálculo de entropia pode ser utilizada para a escolha dos melhores atributos que serão associados aos ramos da árvore na construção do modelo. Para cada ramo expandido – ou investigado – na construção da árvore, uma quantidade reduzida de instâncias será utilizada e apenas os atributos relevantes para essas instâncias são considerados. Essa abordagem, conhecida como "dividir para conquistar", possibilita que Árvores de Decisão produzam estruturas reduzidas que necessitam apenas de uma pequena quantidade de atributos para a construção do modelo. Uma Árvore de Decisão pode facilmente ser convertida em Regras de Classificação. No entanto, o caminho inverso é bem mais complexo, já que Árvores de Decisão não podem expressar facilmente a disjunção existente entre as diferentes regras de um Conjunto de Regras ou de uma Lista de Regras. A Figura 2.2 apresenta um modelo para os dados da Tabela 2.2 para o problema de classificação dos possíveis compradores do livro, similar ao modelo construído através de regras apresentado na seção anterior, mas dessa vez é

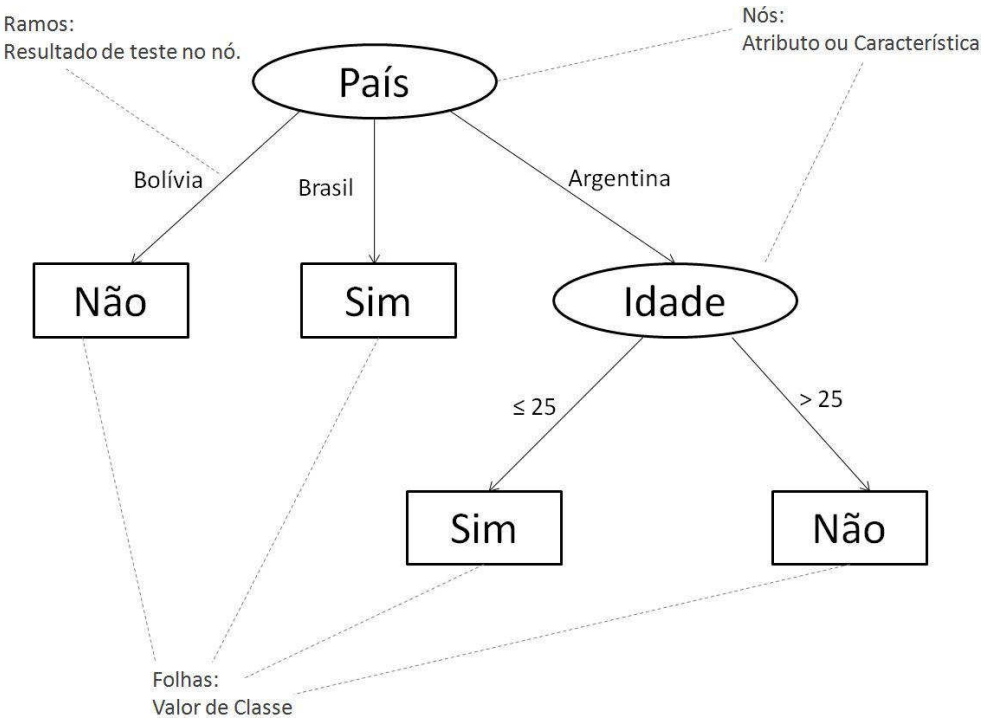



**Figura 2.2** Árvore de Decisão para o problema de classificação dos dados dos possíveis compradores do livro.

apresentada uma Árvore de Decisão.

### Máquinas de Vetores de Suporte

Máquina de Vetores de Suporte ou *Support Vector Machine* (SVM) é um método de classificação linear, binário e não probabilístico para aprendizagem supervisionada, proposto por CORTES e VAPNIK (1995), que consiste na identificação de uma dependência, seja ela mapeamento ou função, para classificação de dados linearmente separáveis. Para isso, um algoritmo SVM inicialmente representa cada instância de uma base de treino como sendo um ponto de dado em um espaço que possui $n$ dimensões – que podem ser atributos ou características. O objetivo do algoritmo é identificar um hiperplano ótimo que separe o espaço $n$-dimensional em dois grupos de pontos de dado – um grupo para cada classe –, de forma que os pontos mais próximos do hiperplano estejam também o mais afastado possível desse hiperplano. Os pontos de dado mais próximos da área de separação definem a posição do hiperplano e são chamados de vetores de suporte. O hiperplano identificado pode ser utilizado como modelo preditivo para classificação de

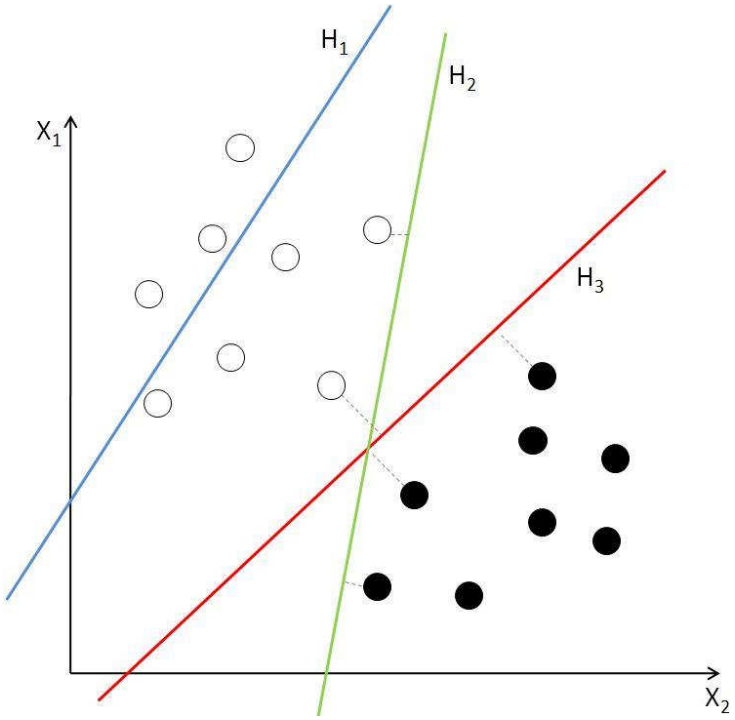



**Figura 2.3** O hiperplano $H_1$ (azul) não separa as duas classes corretamente. O hiperplano $H_2$ (verde) separa, mas com uma pequena margem entre o hiperplano e as instâncias mais próximas. O hiperplano $H_3$ (vermelho) separa as duas classes com a margem máxima.

novos exemplos. Um novo exemplo é mapeado e classificado de acordo com o lado do hiperplano que cair. Assim, uma SVM busca por uma superfície de decisão que esteja afastada ao máximo de qualquer ponto de dado de ambos os lados. A distância entre a superfície de decisão – ou hiperplano – e o ponto de dado mais próximo define a margem do classificador.

SVMs podem ser também aplicadas a problemas que necessitam utilizar mais do que 2 classes, reduzindo-se o problema multiclasse simples a vários problemas de classificação binária (DUAN e KEERTHI, 2005). Além disso, dados não linearmente separáveis também podem ser manipulados utilizando-se esse tipo de classificador, através da teoria de Cover da separabilidade de padrões (HOU *et al.*, 2005). SVMs podem ser utilizadas como técnica de aprendizagem supervisionada em problemas de classificação ou regressão.

### Redes Bayesianas

Uma Rede Bayesiana é um grafo direcionado acíclico, onde cada nó $X_i$ possui uma distribuição de probabilidade condicional $P(X_i|Pais(X_i))$ que quantifica os efeitos dos

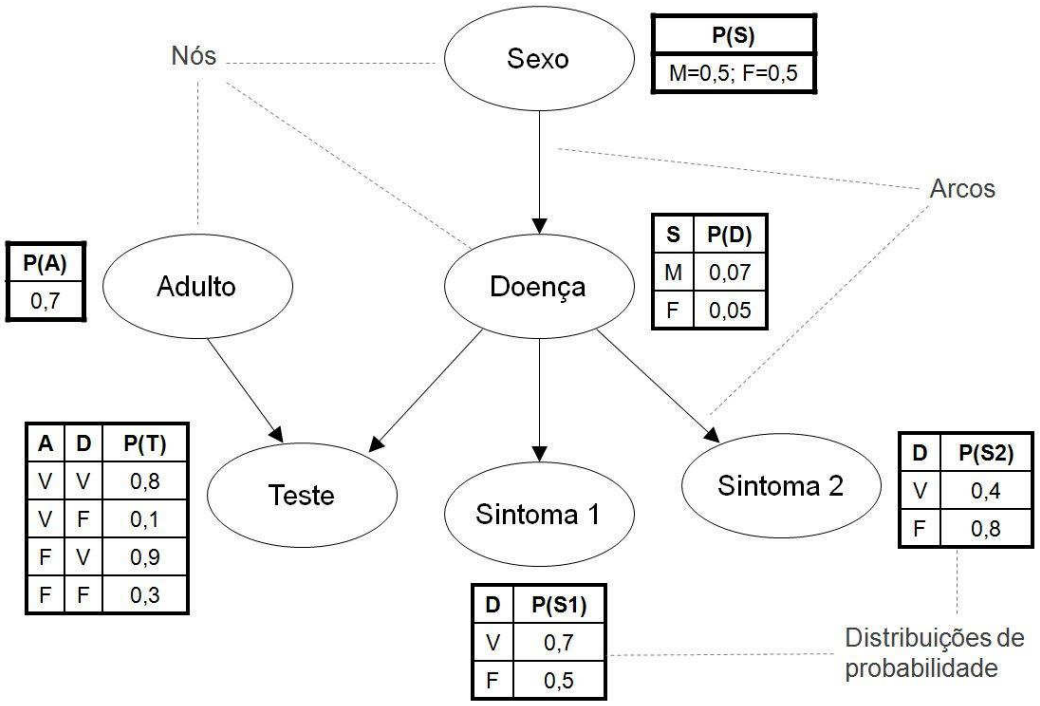



**Figura 2.4** Exemplo de uma Rede Bayesiana aplicada ao diagnóstico de uma doença fictícia que pode ser avaliada através da observação de 2 sintomas e de um teste. Neste exemplo, o sexo do indivíduo (gênero) influencia a probabilidade de ocorrência dessa doença, que por sua vez, influencia o sintoma 1 e o sintoma 2. O teste realizado no indivíduo é influenciado pela doença e pelo fato de o indivíduo ser adulto ou não.

nós pais sobre os nós filhos (RUSSELL e NORVIG, 2009). Um conjunto de variáveis aleatórias – que podem ser discretas ou contínuas – compõe os nós da rede que são conectados aos pares por um conjunto arcos. Um arco apontando do nó $X$ para o nó $Y$ indica que $X$ é pai de $Y$. Uma Rede Bayesiana é uma representação concisa e eficiente da distribuição de probabilidade conjunta de um domínio. Na Figura 2.4 é apresentado um exemplo de uma Rede Bayesiana, que, nesse caso, por possuir uma topologia livre, também pode ser chamada de Rede Bayesiana de Crença.

A distribuição de probabilidades de cada nó pode ser representada através de Tabelas de Probabilidade Condicional (TPC) quando esse nó se referir a uma variável aleatória discreta. No exemplo apresentado na Figura 2.4 todos os nós são discretos e, dessa forma, cada um possui uma TCP relacionada. Cada linha da TCP contém um caso condicional para cada valor do nó, que representa uma possível combinação de valores dos nós pais. Os valores de cada linha devem somar 1, pois as entradas representam um conjunto exaus-

tivo de casos para a variável. Para variáveis booleanas, uma vez definida a probabilidade do caso verdadeiro através de $(p)$, a probabilidade do caso falso é geralmente omitida, pois o mesmo pode ser obtido por complementação através de (1-p). Assim, na TPC do nó Doença, a probabilidade $0{,}07$ está sendo atribuída aos casos positivos da doença observados nos indivíduos do sexo masculino. Os casos negativos da doença observados nos indivíduos também do sexo masculino, apesar de omitidos, podem ser subentendidos – por complementação – através de $1 - 0{,}07 = 0{,}93$. A distribuição de probabilidade de variáveis aleatórias contínuas pode ser representada através de funções de densidade de probabilidade padrão (como distribuição Guassiana) ou através da discretização dos seus intervalos de valor. Um princípio similar pode ser aplicado a Redes Bayesianas Híbridas que acomodam variáveis discretas e contínuas (RUSSELL e NORVIG, 2009).

A semântica numérica de uma Rede Bayesiana parte do princípio de que cada nó é condicionalmente independente dos seus predecessores, dados seus pais. Esse princípio diminui bastante a complexidade da representação da probabilidade conjunta de um domínio, já que uma consulta pode ser realizada através do produto das probabilidades condicionais da rede, não necessitando, assim, de uma tabela de distribuição de probabilidade conjunta completa. Essa semântica pode também ser observada na Equação (2.1):

$$P(x_1, ..., x_n) = \prod_{i=1}^{n} P(x_i | pais(X_i)) \tag{2.1}$$

A topologia da rede especifica os relacionamentos de independência condicional considerados em um domínio. Essa estrutura, juntamente com as distribuições condicionais, pode ser definida manualmente por um especialista ou por um algoritmo de aprendizagem supervisionada, desde que uma quantidade suficientemente grande de exemplos esteja disponível para treino. O processo de classificação – ou consulta ao valor de um nó discreto – de um novo exemplo de dado em uma Rede Bayesiana pode ser realizado através de inferência exata, utilizando técnicas como Inferência por Enumeração, ou através de inferência aproximada, utilizando técnicas como Algoritmos de Monte Carlo (RUSSELL e NORVIG, 2009). Assim, utilizando-se o modelo apresentado na Figura 2.4 pode-se calcular, por exemplo, a probabilidade de o indivíduo estar doente, sabendo-se apenas

que os sintomas 1 e 2 foram observados, através de Inferência por Enumeração da seguinte forma:

1. Inicialmente, utiliza-se a seguinte equação para consulta:

   $P(X|e) = \alpha P(X,e) = \alpha \Sigma_Y P(X,e,y)$.

   Onde:

   X = Variável de consulta que no domínio está sendo representada pela variável aleatória Doença;

   e = Variáveis que representam a evidência observada. No problema estão sendo representadas pelas variáveis aleatórias Sintoma 1 e Sintoma 2;

   Y = Variáveis ocultas, não especificadas, mas existentes no domínio. Estão sendo representadas no problema pelas variáveis aleatórias Sexo, Adulto e Teste.

2. Em seguida, substituem-se as variáveis na equação:

   $P(D|s2,s1) = \alpha P(D,s2,s1) = \alpha \Sigma_s \Sigma_a \Sigma_t P(D,s2,s1,s,a,t)$

3. Aplica-se, logo após, a semântica numérica das Redes Bayesianas (de acordo com a Equação (2.1)):

   $P(d|s2,s1) = \alpha \Sigma_s \Sigma_a \Sigma_t P(d|s)P(s2|d)P(s1|d)P(s)P(a)P(t|d,a)$

4. Simplifica-se o problema

   $= \alpha P(s2|d)P(s1|d)(\Sigma_s P(s)P(d|s))(\Sigma_a P(a)\Sigma_t P(t|d,a))$

5. Para Doença=Verdadeiro, temos:

   $= \alpha 0.4 * 0.7 * (0.5 * 0.07 + 0.5 * 0.05) * [(0.7 * (0.8 + 0.2)) + (0.3 * (0.9 + 0.1)]$

   $= \alpha 0.0168$

6. Para Doença=Falso, temos:

   $= \alpha 0.8 * 0.5 * (0.5 * 0.93 + 0.5 * 0.95) * [(0.7 * (0.1 + 0.9)) + (0.3 * (0.3 + 0.7))]$

   $= \alpha 0.376$

7. Normalizando, temos:

   $= \alpha \langle 0.0168,0.376 \rangle = \langle 0.04,0.96 \rangle$

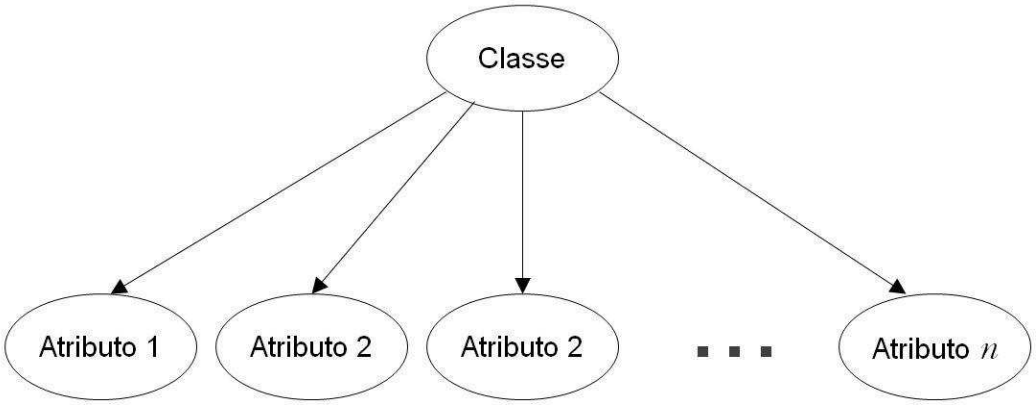



**Figura 2.5** Topologia de uma Rede Bayesiana estruturada como Classificador Naive Bayes. Todos os nós atributo da rede são independentes entre si, dado um nó classe.

Assim, de acordo com essa consulta, existe $4\%$ de probabilidade de um indivíduo estar doente, dado que o Sintoma 1 e o Sintoma 2 foram observados.

### Classificador Naive Bayes

Um tipo específico de Rede Bayesiana amplamente investigado em problemas de classificação que necessitam de acurácia e velocidade na identificação de modelos é conhecido como Naive. Por ser aplicado especificamente a problemas de classificação, esse tipo de Rede Bayesiana é geralmente conhecido como Classificador Bayesiano ou Classificador Naive Bayes. Esse classificador simples é considerado ingênuo – ou naive – por assumir que o efeito do valor de um atributo em uma dada classe é independente do efeito dos valores dos demais atributos – ao contrário de uma Rede Bayesiana de Crença, que permite representações de independência entre subconjuntos de atributos. Graças a esse pressuposto ingênuo, as tarefas de computação são bruscamente simplificadas nesse tipo de Rede Bayesiana. A Figura 2.5 mostra a topologia de uma Rede Bayesiana estruturada como Classificador Naive Bayes.

O teorema de Bayes, apresentado na Equação (2.2), pode ser aplicado diretamente na construção de um Classificador Naive Bayes. Esse teorema propõe um método de se calcular a probabilidade a posteriori de uma hipótese $H$ dada uma observação $X$ – ou seja, $P(H|X)$ – através de $P(H)$, $P(X|H)$ e $P(X)$, já que essas probabilidades podem ser estimadas a partir da base de treino.

$$P(H|X) = \frac{P(X|H)P(H)}{P(X)} \tag{2.2}$$

Utilizando-se os dados da Tabela 2.2 pode-se construir, por exemplo, um modelo preditivo para classificação de um novo exemplo – ou instância –
$X = (Comprador = Sim; Sexo = Masculino; País = Argentina; Idade = 23)$,
através do Classificador Naive Bayes. O objetivo do modelo é identificar qual valor de classe possui maior probabilidade de ocorrência para esse novo exemplo. Isso pode ser realizado através dos seguintes passos:

1. Inicialmente, calcula-se a probabilidade a priori $P(C_i)$ para cada valor de classe, onde $C_1 = Sim$ e $C_2 = Não$. Isso pode ser facilmente realizado através da base de treino:
   $P(Comprador = Sim) = 4/10 = 0{,}4$
   $P(Comprador = Não) = 6/10 = 0{,}6$

2. Em seguida, calcula-se a probabilidade condicional $P(X|C_i)$ de cada um dos atributos de valor discreto da base de treino, onde $C_1 = Sim$ e $C_2 = Não$. Isso também pode ser realizado através da base de treino:
   $P(Sexo = Masculino|Comprar = Sim) = 2/4 = 0{,}5$
   $P(Sexo = Masculino|Comprar = Não)) = 4/6 = 0{,}67$
   $P(País = Argentina|Comprar = Sim) = 2/4 = 0{,}5$
   $P(País = Argentina|Comprar = Não)) = 2/6 = 0{,}33$

3. Calcula-se, então, a probabilidade condicional dos atributos de valor contínuo. Na base de treino, idade é o único atributo desse tipo. Geralmente, um algoritmo Naive Bayes assume que um atributo de valor contínuo tem uma distribuição Gaussiana – ou Normal – com média $\mu$ e desvio padrão $\sigma$. A probabilidade condicional do atributo será, então, calculada a partir da densidade da curva normal. Isso pode ser facilmente realizado através da seguinte equação:

$$g(x,\mu,\sigma) = \frac{1}{\sqrt{2\pi\sigma}} e^{-\frac{(x-\mu)^2}{2\sigma^2}} \tag{2.3}$$

A probabilidade condicional poderá, então, ser calculada através de:

$$P(x_k|C_i) = g(x_k,\mu_{C_i},\sigma_{C_i}) \tag{2.4}$$

Observando a base de treino, para $Comprador = Sim$ tem-se para o atributo $Idade$ $\mu = 25{,}75$ e $\sigma = 5{,}74$. Substituindo na função, obtém-se:

$P(Idade = 23|Comprador = Sim) = g(23; 25{,}75; 5{,}74) = 0{,}15$

Para $Comprador = Não$ tem-se para o atributo $Idade$ $\mu = 29{,}67$ e $\sigma = 14{,}54$. Substituindo na função, obtém-se:

$P(Idade = 23|Comprador = Não) = g(23; 29{,}67; 14{,}54) = 0{,}09$

4. Utilizando-se as probabilidades calculadas anteriormente, obtém-se para $Comprador = Sim$:

$$\begin{aligned} P(X|Comprador = Sim) &= P(Comprador = Sim) \times \\ &\quad P(Sexo = Masculino|Comprar = Sim) \times \\ &\quad P(País = Argentina|Comprar = Sim) \times \\ &\quad P(Idade = 23|Comprador = Sim) \\ &= 0{,}4 \times 0{,}5 \times 0{,}5 \times 0{,}15 = 0{,}015 \end{aligned}$$

e para $Comprador = Não$:

$$\begin{aligned} P(X|Comprador = Não) &= P(Comprador = Não) \times \\ &\quad P(Sexo = Masculino|Comprar = Não) \times \\ &\quad P(País = Argentina|Comprar = Não) \times \\ &\quad P(Idade = 23|Comprador = Não) \\ &= 0{,}6 \times 0{,}67 \times 0{,}33 \times 0{,}09 = 0{,}012 \end{aligned}$$

Normalizando-se os resultados obtidos $\langle 0{,}15; 0{,}12\rangle$, tem-se $\langle 0{,}56; 0{,}44\rangle$ para $Comprador = Sim$ e $Comprador = Não$, respectivamente. Assim, o algoritmo Naive Bayes classificaria o novo exemplo $X$ como $Comprador = Sim$, pois este é o valor de classe que atingiu a probabilidade máxima no modelo $(0{,}56)$.

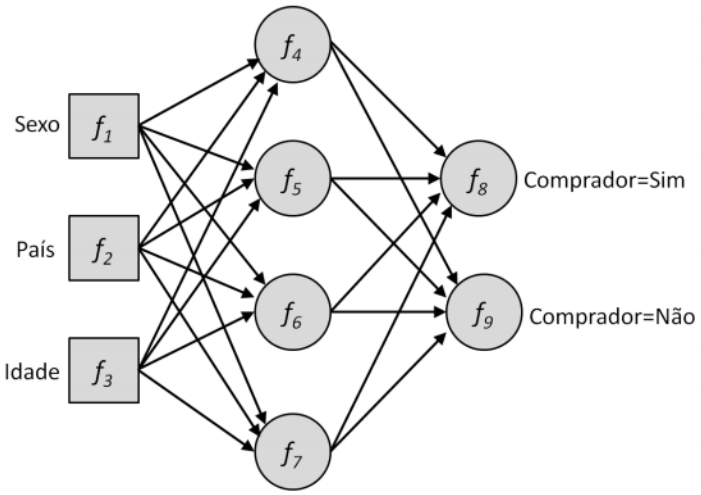



**Figura 2.6** Uma Rede Neural simples.

### Redes Neurais

Segundo HAN e KAMBER (2006), em uma descrição rudimentar, uma rede Rede Neural é um conjunto de unidades com entradas e saídas conectadas, similares a neurônios, em que cada uma dessas conexões possui um peso associado. O processo de aprendizagem consiste, então, em, dada uma base de treino, ajustar os pesos das conexões de forma que as instâncias dessa base de treino sejam corretamente classificadas. Assim como em outros métodos de aprendizagem, um modelo construído através da calibragem dos pesos das conexões pode ser utilizado no processo de classificação ou, no caso de funções, aproximação de novos exemplos. A Figura 2.6 mostra um exemplo simples de uma Rede Neural.

Redes Neurais surgiram da tentativa de se encontrar, em sistemas biológicos, representações matemáticas úteis para o processamento de informação (BISHOP, 2006). A ideia central parte do princípio de que um cérebro biológico e um computador possuem modelos de processamento bem diferentes. Enquanto um computador normalmente possui apenas um processador rápido – com tempo de comutação girando em torno de $10^{-10}$ segundos) –, um cérebro humano pode possuir aproximadamente $(10^{11})$ unidades de processamento simples e relativamente lentas – com tempo de comutação em torno de $10^{-3}$ segundos –, que estão ligadas a uma média de $(10^4)$ outros neurônios (MITCHELL, 1997). Redes Neurais artificiais são baseadas na crença de que o que torna o cérebro humano

diferente e capaz de solucionar problemas que os computadores ainda não conseguem resolver é a vasta conectividade entre os neurônios, além do fato de eles operarem em paralelo (ALPAYDIN, 2010).

A pesquisa em Redes Neurais artificiais busca simular o comportamento de um cérebro biológico na resolução de problemas difíceis. Várias propostas têm sido apresentadas nessa área, mas, segundo RUSSELL e NORVIG (2009), existem duas categorias principais de estruturas de Redes Neurais artificiais: Redes Neurais sem alimentação – ou *feed-forward* – e Redes Neurais com alimentação – ou recorrentes –. Uma Rede Neural *feed-forward* é a representação de uma função para seus valores de entrada, sem nenhum estado interno além dos pesos das conexões. Uma Rede Neural Recorrente retorna suas saídas de volta para suas entradas, formando um sistema dinâmico que pode atingir um estado estável, exibir oscilações ou mesmo um comportamento caótico inesperado. Esse tipo de estrutura proporciona a simulação de memórias de curto prazo, sendo, dessa forma, mais adequada para modelagem de um cérebro biológico, embora bem mais difícil de ser compreendida. Segundo BISHOP (2006), Redes Neurais *feed-forward* têm demonstrado maior aplicação prática, em particular, um tipo específico conhecido como Perceptron Multicamadas.

Redes Neurais têm sido algumas vezes criticadas pela sua difícil interpretabilidade, ao contrário das Regras de Classificação. Além disso, o longo tempo de treino necessário para construção de um modelo inviabiliza sua aplicação em certos domínios. Contudo, essa técnica de aprendizagem tem como vantagens:

- Ser altamente tolerante a ruídos nos dados;
- Ser capaz de classificar padrões para os quais não foi treinada;
- Produzir bons resultados, mesmo quando não se sabe sobre a existência de relacionamentos entre atributos e classes;
- Ser adequada para a produção de modelos que manipulam valores contínuos de entrada e saída;

- Apesar de o processo de aprendizagem ser muito demorado, uma vez que o modelo esteja pronto, sua aplicação em tarefas de classificação ou regressão de novos casos é muito ágil, produzindo respostas satisfatórias, mesmo para sistemas de tempo real (MITCHELL, 1997).

**Aprendizagem Baseada em Instâncias**

A Aprendizagem Baseada em Instâncias é um método utilizado para classificação ou regressão de novos exemplos de um domínio através da observação de exemplos com características similares já avaliados existentes em uma base de treino. O algoritmo utiliza a própria base de treino como modelo, tentando identificar um exemplo próximo do novo exemplo que se deseja classificar. A classe – ou valor – do exemplo mais próximo é, então, atribuída a este novo exemplo. Assim, sempre que um novo exemplo estiver sendo classificado – ou aproximado por regressão –, as instâncias já avaliadas da base de treino, cujas características mais se assemelham a este novo exemplo, são utilizadas como palpite. Esse método é também conhecido como "preguiçoso", pois nenhuma tarefa de aprendizagem é executada até que um novo exemplo deva ser classificado.

A semelhança entre um novo exemplo – ou instância – e um exemplo já avaliado de uma base de treino pode ser medida através da proximidade de suas características – ou atributos. Quando os exemplos possuem apenas um atributo, essa medida pode ser facilmente calculada através da diferença entre os valores dos atributos. Quando os exemplos possuem vários atributos, esse cálculo pode ser realizado através de Distância Euclidiana, desde que os atributos considerados estejam normalizados e sejam de igual importância para o domínio. Quando o atributo comparado for nominal, a distância entre valores idênticos pode ser considerada como $0$ e a distância entre valores diferentes como $1$ (WITTEN e FRANK, 2005). Na Aprendizagem Baseada em Instâncias, a classificação de um novo exemplo através da identificação de um único exemplo da base de treino com menor Distância Euclidiana é chamado de método do Vizinho-mais-próximo (*nearest-neighbor*). Às vezes, é desejável se avaliar um novo exemplo não através de apenas um vizinho próximo, mas através de uma vizinhança inteira. Contudo, a definição da distância dessa vizinhança pode ser muito abstrata, já que uma vizinhança curta pode não conter nenhum

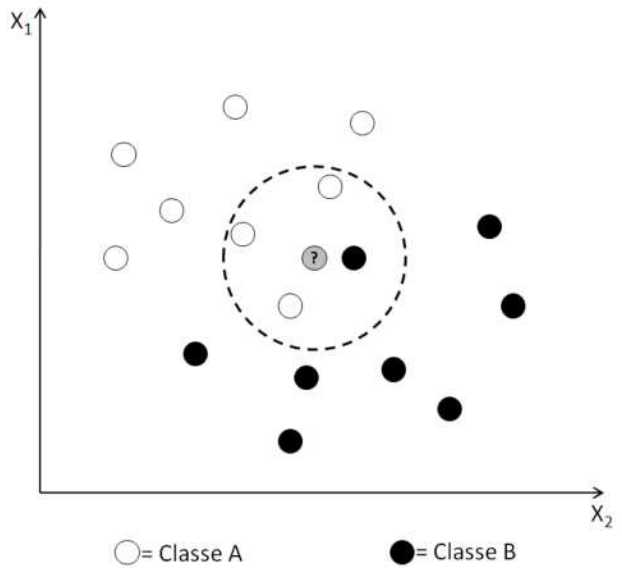



**Figura 2.7** Classificação Baseada em Instâncias em um espaço bidimensional, considerando 2 atributos.

exemplo – ou vizinho próximo –, e uma vizinhança extremamente grande pode conter todos os exemplos da base de treino. Uma solução normalmente utilizada para contornar esse problema é a definição de uma vizinhança que seja grande o suficiente para comportar apenas $k$ vizinhos. A classe majoritária – ou a média ponderada, para o caso de regressão – dos $k$ vizinhos mais próximos é então atribuída ao novo exemplo. Esse método é conhecido como método dos $k$-Vizinho-mais-próximos (*k-nearest-neighbor*)(RUSSELL e NORVIG, 2009).

A escolha da quantidade de vizinhos que serão considerados no processo de aprendizagem pode mudar drasticamente os resultados da classificação. Na Figura 2.7 é apresentado , em um espaço bidimensional, um novo exemplo sendo classificado de acordo com seu(s) vizinho(s) mais próximo(s). Considerando-se o algoritmo 1-Vizinho-mais-próximo, o novo exemplo estaria sendo classificado como pertencente à Classe $B$. Contudo, se o algoritmo $k$-Vizinhos-mais-próximos for considerado, este mesmo novo exemplo seria classificado como pertencente à Classe $A$.

A Aprendizagem Baseada em Instâncias apresenta como desvantagem o fato de não apresentar uma estrutura explícita de aprendizagem, como um modelo de classificação ou uma função de regressão, devendo o algoritmo recorrer à base de treino a cada nova

avaliação de um novo exemplo. Além disso, bases de treino excessivamente grandes podem tornar o processo de aprendizagem lento e dessa forma, algumas estratégias devem ser adotadas para decidir quais exemplos deverão ser guardados para avaliações futuras. Contudo, a Aprendizagem Baseada em Instâncias tem como grande vantagem o fato de que, a cada nova instância incorporada à base de treino, um novo conhecimento é adquirido, fazendo com que o classificador se torne cada vez mais robusto, sem a necessidade de criação de novos modelos para o domínio.

### 2.3.3 Métodos de Validação

A validação tem por objetivo verificar a capacidade de generalização de um modelo produzido por um algoritmo de aprendizagem. Existem várias alternativas para a execução dessa tarefa, que podem ser consideradas de acordo com a quantidade de dados e com os recursos computacionais disponíveis.

Normalmente, verificar o desempenho de um modelo através dos mesmos dados que foram utilizados no treinamento do algoritmo que o originou não é uma boa estratégia de validação. Isso porque, certamente, esse modelo apresentará desempenho preditivo otimista nos dados de treino, independente da sua capacidade preditiva em outros dados da mesma população. O desempenho de um modelo é melhor avaliado através de uma amostra de teste, composta por instâncias cujas classes são conhecidas e não foram utilizadas para o treinamento do algoritmo. O desempenho obtido pelo modelo no próprio conjunto de dados que o originou é conhecido como erro de resubstituição. Apesar de esse valor possuir algumas aplicações práticas, não é adequado para a verificação da capacidade de generalização de um modelo.

Quando uma vasta quantidade de dados está disponível no domínio, um modelo pode ser construído a partir de uma amostra suficientemente grande e testado através de outra amostra suficientemente grande. Contudo, em alguns problemas de aprendizagem em que os dados são escassos – como, por exemplo, domínios em que os dados de treino são baseados na perícia de um especialista humano –, os modelos devem ser construídos e testados através da mesma amostra. Nesse caso, devem ser considerados métodos de validação, capazes de lidar com uma quantidade reduzida de dados, sem que se deixe de

produzir boas generalizações para o domínio. Algumas dessas estratégias estão sendo apresentadas a seguir.

- ***Holdout***: Esse método considera reservar uma parte dos dados para treino e outra parte dos dados para teste. Normalmente, $1/3$ da amostra é reservado para teste, e o restante para o treinamento do algoritmo (ver Figura 2.8(a)).

- **Validação Cruzada**: Esse método utiliza um processo iterativo onde a amostra é dividida em um número pré-estabelecido de *folds* – ou partes –, considerando a cada iteração uma parte para teste e as demais partes para treino do algoritmo. Ao final do processo, todas as partes devem ter sido utilizadas uma vez como teste (ver Figura 2.8(b)).

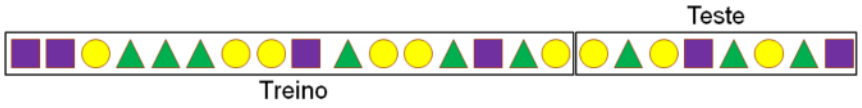
- ***Leave-one-out***: Esse método estabelece uma validação cruzada, em que o número de *folds* é igual ao número de instâncias da amostra. A cada iteração, uma instância é utilizada para teste e todas as demais instâncias são utilizadas para treino do algoritmo. Ao final do processo, todas as instâncias devem ter sido utilizadas uma vez como teste (ver Figura 2.8(c)).

- ***Bootstrap Aggregating* (BAGGING)**: Esse método faz uso de um número pré-estabelecido de amostras bootstrap[2], ao invés da amostra original, para a validação de um modelo. Normalmente, este método é aplicado juntamente com outros métodos de validação como *holdout* e validação cruzada. Se o modelo final considerar também amostras bootstrap em sua composição, o BAGGING atuará como uma estratégia de Meta-Aprendizagem (ver Figura 2.8(e)). Outras informações sobre BAGGING podem ser vistas em BREIMAN (1996).

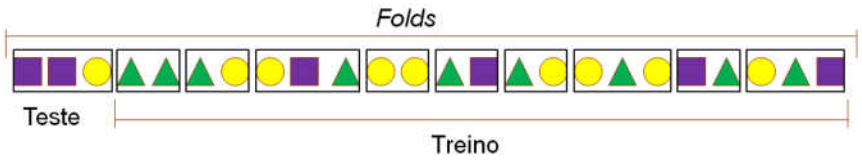
Os métodos de validação apresentados acima também podem considerar as seguintes estratégias para garantir a legitimidade dos modelos:

- **Estratificação**: Essa estratégia impõe que as partes utilizadas para treino e para

[2]Amostras bootstrap são amostras com reposição obtidas de uma amostra original (EFRON e TIBSHIRANI, 1993; CHERNICK, 2008).

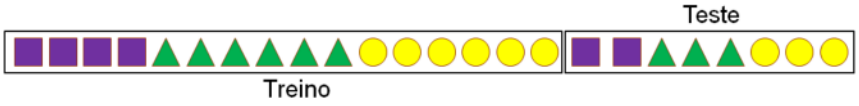



(a) Método de validação *holdout*. Os dados são divididos em 2 partes distintas, destinadas ao treino dos algoritmos e ao teste dos modelos.

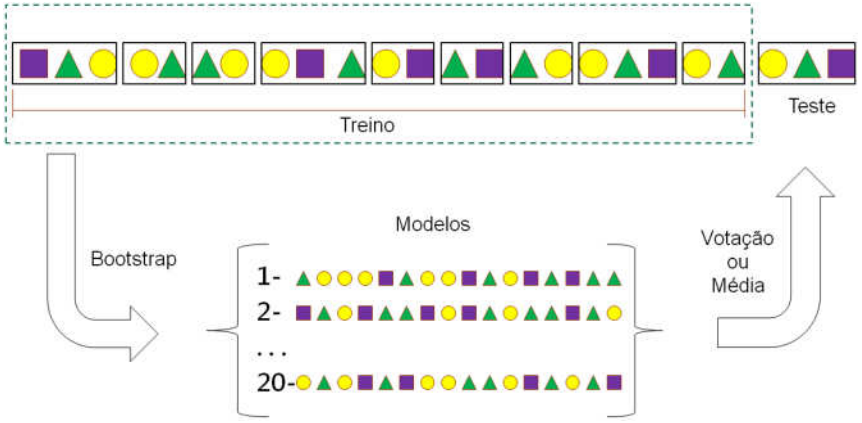



(b) Validação Cruzada. Os dados são divididos em *folds* e, durante um processo iterativo, cada *fold* é utilizado para testar o modelo produzido pelos demais *folds*.


(c) Método de validação *leave-one-out*. Em um processo iterativo, cada instância é utilizada para testar um modelo produzido por todas as demais instâncias.




(d) Estratégia de estratificação. Os dados de treino e teste mantêm a mesma distribuição da amostra original.




(e) Bootstrap Aggregating (BAGGING). O processo de aprendizagem dá-se através de amostras bootstrap da amostra original. A predição é realizada por votação ou, no caso de regressão, através da média dos submodelos.

**Figura 2.8** Estratégias de validação. Nos exemplos, cada instância é representada por uma figura geométrica (quadrado, círculo, triângulo). Diferentes figuras geométricas representam diferentes classes.

teste possuam instâncias representativas para cada classe na mesma proporção observada na amostra original – sempre que possível (ver Figura 2.8(d)).

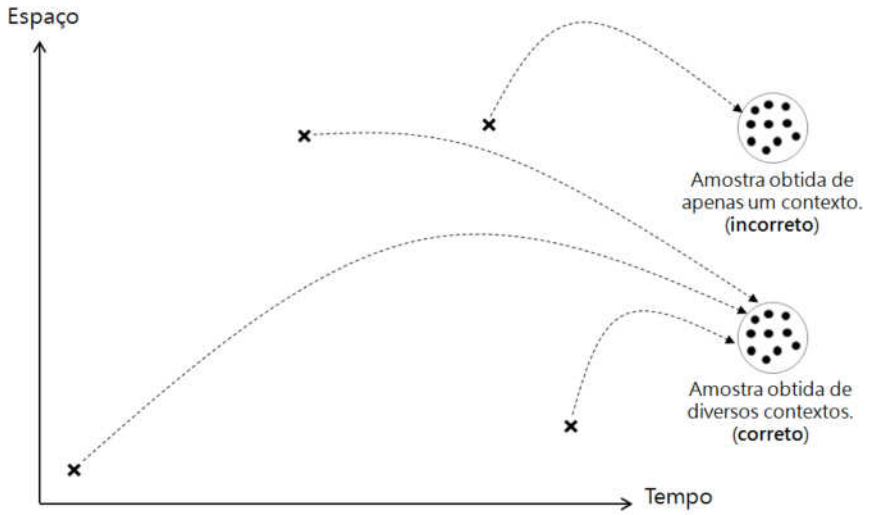
- **Repetição**: Essa estratégia impõe que o processo de aprendizagem básico seja repetido um número pré-estabelecido de vezes, utilizando diferentes arranjos aleatórios para a amostra. O desempenho final é calculado a partir da média dos desempenhos obtidos em cada arranjo aleatório.

Um procedimento amplamente utilizado em Mineração de Dados para a correta validação de modelos é a validação cruzada com $10$ folds estratificados, repetida em $10$ arranjos aleatórios. Outras informações sobre métodos de validação podem ser vistas em WITTEN e FRANK (2005).

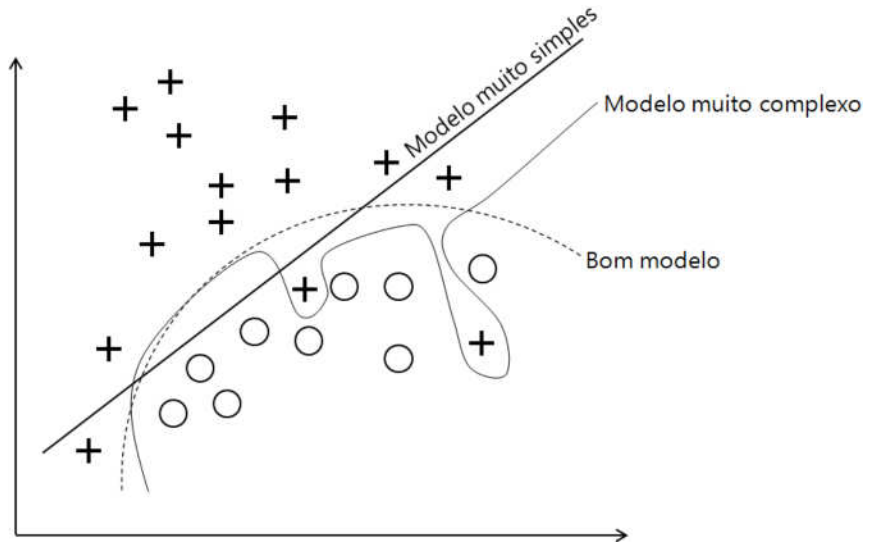
### 2.3.4 *Overfitting*

Um problema que nem sempre é solucionado por métodos de validação é a possibilidade de a amostra de treino considerada possuir características próprias que não são normalmente observadas na população. Nesse caso, os modelos poderão estar refletindo comportamentos específicos da amostra, e sua aplicação na população certamente não produzirá bons desempenhos. Esse fenômeno é conhecido como efeito *overfitting*, e uma das formas de evitá-lo é assegurar que as amostras utilizadas sejam realmente representativas da população. Amostras muito pequenas em relação à dimensionalidade – ou número de atributos – do domínio ou obtidas em contextos muito específicos devem ser evitadas (ver Figura 2.9(a)). Contudo, mesmo considerando uma amostra representativa, há casos em que modelos extremamente específicos também podem sofrer o efeito *overfitting* (ver Figura 2.9(b)). Nesses casos, os modelos costumam seguir todas as flutuações da amostra de treino, ao invés de produzirem generalizações (ver Figura 2.9).

O *overfitting* pode ser detectado testando-se o modelo em outras amostras da mesma população, e pode ser contornado aumentando-se o tamanho da amostra ou simplificando-se os modelos – através da diminuição do número de atributos, poda de Árvores de Decisão, utilização de um número menor de unidades na camada oculta de um perceptron, etc. Outras informações sobre o efeito *overfitting* podem ser vistas em TUFFÉRY (2011).




(a) Amostras que consideram diversos contextos do domínio evitam o *overfitting*, ao contrário de amostras que contemplam apenas um contexto.




(b) Modelos relativamente simples e genéricos evitam o *overfitting*, ao contrário de modelos extremamente precisos nos dados amostrais. Fonte: TUFFÉRY (2011).

**Figura 2.9** Tipos de *overfitting*.

### 2.3.5 Métricas de Desempenho

O desempenho preditivo obtido por um mecanismo de aprendizagem é expresso através de métricas. Dependendo da natureza da métrica e da estratégia de aprendizagem aplicada, diferentes informações podem ser demonstradas, sendo a interpretação destas informações essencial para a identificação de modelos eficientes. Algumas métricas apre-

sentam a quantidade de acertos, e outras preocupam-se em demonstrar a quantidade de erros. Contudo essas duas filosofias nem sempre são tão complementares quanto parecem. Nesta seção, estão sendo discutidas métricas que avaliam tanto a taxa de acertos quanto a taxa de erros para estratégias de classificação e regressão. Está sendo dado um enfoque especial às métricas utilizadas em Medicina.

**Classificação**

A métrica mais simples para a representação do desempenho em problemas de classificação é a acurácia. Esta métrica informa a porcentagem de instâncias na base de teste que foram corretamente classificadas pelo modelo, independente da classe considerada, ou seja:

$$Acurácia = \frac{Instâncias\ Corretamente\ Classificadas}{Todas\ as\ Instâncias} \tag{2.5}$$

Um recurso normalmente utilizado para complementar a acurácia – avaliando o quanto um modelo é capaz de reconhecer instâncias de classes específicas – é a matriz confusão. Uma matriz confusão é um tipo específico de tabela de contingência que "cruza" os valores de classe previstos por um modelo com as classes reais observadas na base de teste. A diagonal principal da matriz representa o número de instâncias corretamente previsto para cada classe, e a soma de seus valores deve, idealmente, aproximar-se do número total de instâncias testadas (ver Figura 2.10). Contudo, deve-se observar que uma matriz confusão apresenta os resultados obtidos através de apenas uma tarefa básica de aprendizagem. Se os modelos estiverem sendo validados através de repetições aleatórias, outra estratégia deverá ser utilizada para a fusão das diversas matrizes ou para a produção de uma métrica única.

Uma abordagem normalmente aplicada em problemas de classificação é a dicotomia. Nessa abordagem, cada valor de classe é calculado – e ponderado – separadamente em relação a todas as demais classes, de forma que o problema de classificação, como um todo, se torne um teste dicotômico. Em suma, trata-se de considerar apenas 2 classes em um problema de classificação. Quando essa abordagem passa a ser considerada, outras mé-

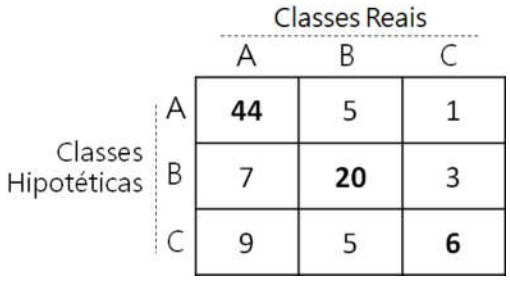



**Figura 2.10** Exemplo de uma matriz confusão.

tricas – e conceitos – mais eficientes podem ser utilizados. A abordagem dicotômica faz uso de uma tabela de contingência – ou matrix confusão – que considera apenas $2$ classes (ver Figura 2.11(a)). A partir dessa tabela, podem ser obtidos $4$ valores essenciais para a avaliação dos modelos, equivalentes às seguintes classes: Verdadeiro Positivo (VP), Verdadeiro Negativo (VN), Falso Negativo (FN), Falso Positivo (FP). Considerando-se um modelo aplicado a uma amostra em que todas as instâncias estejam devidamente classificadas, esta tabela de contingência equivale ao diagrama apresentado na Figura 2.11(b).

A partir dos $4$ valores produzidos na avaliação de um teste dicotômico, diversas métricas básicas podem ser estabelecidas. A Tabela 2.3 apresenta algumas das métricas mais utilizadas para essa abordagem.

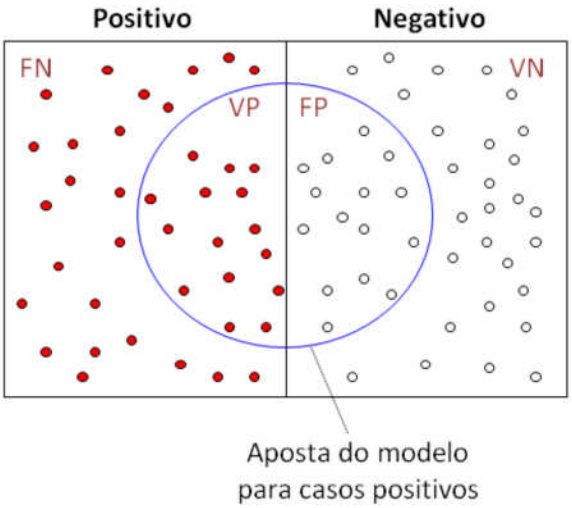
Apesar de as métricas apresentadas anteriormente serem igualmente aplicáveis na avaliação de testes dicotômicos, algumas delas fazem uso da prevalência[3] para determinar a importância da probabilidade de cada classe. Essa é uma suposição arbitrária e, na maioria dos casos, inválida. Mesmo considerando amostras estratificadas da população, métricas que fazem uso da prevalência para avaliar a capacidade preditiva de um modelo podem gerar interpretações extremamente equivocadas. Por exemplo, dada uma base de teste, composta por $100$ indivíduos previamente classificados segundo uma doença, em que $1$ indivíduo é doente – ou caso positivo – e $99$ indivíduos são sadios – ou controles–, um modelo que sempre produza "negativo" como resultado, quer o indivíduo seja controle ou não, obterá $99\%$ de acurácia nesta amostra, mesmo não sendo capaz de identificar nenhum caso positivo da doença.

Além disso, há casos em que, dependendo do tipo de estudo realizado, o uso da pre-

[3] A prevalência é a proporção de casos existentes para uma determinada classe numa determinada população e num determinado momento temporal (PORTA, 2008).

| | | Classes Reais | |
|---|---|---|---|
| | | **Positivo** | **Negativo** |
| Classes Hipotéticas | **Positivo** | Verdadeiro Positivo (VP) | Falso Positivo (FP) |
| | **Negativo** | Falso Negativo (FN) | Verdadeiro Negativo (VN) |

(a) Tabela de contingência para um teste dicotômico.




(b) Representação da avaliação de um teste dicotômico. O círculo interno representa um modelo aplicado à amostra. Todas as instâncias internas ao círculo foram consideradas como casos positivos, e todas as instâncias externas, como casos negativos. Pode-se perceber que o modelo acertou algumas instâncias e errou outras.

**Figura 2.11** Representações de um teste dicotômico.

valência é proibitivo. Por exemplo, um estudo de caso controle em epidemiologia normalmente faz uso de amostras adquiridas de forma não aleatória e dessa forma, uma provável prevalência artificial inserida na amostra influenciará nos resultados produzidos por métricas que utilizarem tal prevalência. De uma forma geral, já há muito tempo que a prevalência é vista como um aspecto perigoso de ser considerado, pois ela dá poderes ao investigador de predeterminar métricas através da manipulação dos elementos da amostra em estudo (OTA, 1980).

Para contornar os problemas normalmente relacionados à prevalência na avaliação de um modelo dicotômico, apenas métricas que utilizam valores para uma classe devem ser

**Tabela 2.3** Métricas de desempenho para predições categóricas (classificação), considerando testes dicotômicos.

| Métrica | Fórmula | Sensível à Prevalência |
|---|---|---|
| Acurácia (ou microacurácia) | $\frac{VP+VN}{VP+VN+FP+FN}$ | Sim |
| Precisão (Valor Preditivo Positivo) | $\frac{VP}{VP+FP}$ | Sim |
| Valor Preditivo Negativo | $\frac{VN}{VN+FN}$ | Sim |
| Sensibilidade (ou revocação) | $\frac{VP}{VP+FN}$ | Não |
| Especificidade | $\frac{VN}{VN+FP}$ | Não |
| F-measure | $\frac{2 \times Precisão \times Revocação}{Precisão + Revocação}$ | Sim |
| Macroacurácia | $\frac{Sensibilidade + Especificidade}{2}$ | Não |

consideradas, e nesse sentido, a sensibilidade e a especificidade são as mais utilizadas. Como foi dito anteriormente, um teste dicotômico avalia uma classe – considerada como caso positivo – em relação a todas as demais – conjunto de classes considerado como caso negativo. A sensibilidade é a proporção de casos positivos classificados corretamente, dentre todos os casos realmente positivos. A especificidade é a proporção de casos negativos classificados corretamente, dentre todos os casos realmente negativos. Um modelo que apresente ao mesmo tempo maior sensibilidade e maior especificidade do que outro, pode ser considerado mais eficiente, se nenhum aspecto relacionado a custos for considerado. Contudo, essa abordagem nem sempre é possível, já que testes com altas sensibilidades normalmente apresentam especificidades relativamente reduzidas, e vice-versa. De uma forma geral, é impossível deixar de avaliar as implicações de $2$ tipos de erros produzidos, quando essas métricas estão sendo consideradas: $(1)$ Se as consequências de um falso positivo são maiores – por exemplo, determinar uma terapia perigosa para um indivíduo que, na realidade, não está doente –, especificidades altas são desejá-

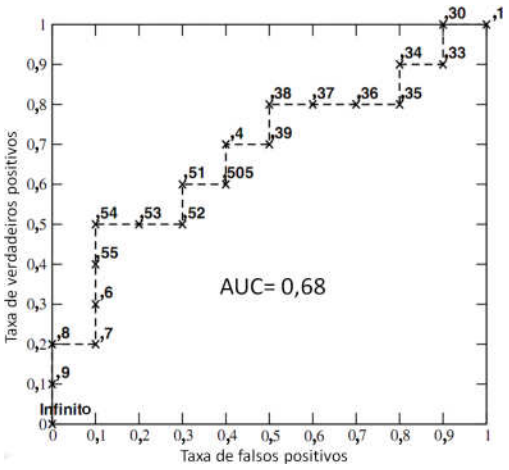
veis; (2) Se as implicações de um falso negativo são temíveis – por exemplo, deixar de identificar um individuo doente que após algum tempo, atinge um estágio irremediável –, altas sensibilidades são preferíveis (OTA, 1980).

Pode-se perceber que, apesar de a sensibilidade e de a especificidade não serem afetadas pela prevalência, sua interpretação conjunta não é intuitiva e, dessa forma, um método seguro para a produção de uma métrica sumarizada é desejável. Dois conceitos que fazem uso destas métricas para a produção de interpretações objetivas, são: discriminação e calibragem. A discriminação refere-se à capacidade dos modelos em distinguir corretamente entre 2 possíveis valores de classe. Um modelo com alto poder discriminativo é capaz de produzir um *ranking* que apresente maiores probabilidades para casos positivos do que para casos negativos. Umas das métricas de discriminação mais utilizadas é a Área Abaixo da Curva ROC (AUC)(ver Figura 2.12). A Calibragem de um modelo mostra o quão aproximado as probabilidades previstas coincidem numericamente com os resultados reais, independente de um *ranking*. Uma estatística amplamente utilizada para avaliação da calibragem de um modelo é o teste *Goodness-of-fit* de Hosmer-Lemeshow (HOSMER e LEMESHOW, 2000). Contudo, métricas que avaliam a quantidade de erros probabilísticos produzidos por um modelo (discutidas na próxima seção) também podem ser utilizadas para esse propósito, além da macroacurácia. Deve-se observar, que uma boa discriminação é sempre preferível a uma boa calibragem: Um modelo com boa capacidade discriminativa pode sempre ser recalibrado, mas o *ranking* ordenado das probabilidades não pode ser modificado para o aprimoramento da discriminação (BALAKRISHNAN e RAO, 2004).

Apesar de toda a discussão anterior, a prevalência populacional pode ser utilizada para determinar o valor preditivo de um modelo – ou seja, o quanto o resultado de um modelo é capaz de modificar as crenças atuais, dada a quantidade de casos existentes na população. Por exemplo, em Medicina, a prevalência populacional, quando conhecida, é útil para a determinação da probabilidade do pós-teste – ou seja, medir o grau de incerteza de um diagnóstico após execução de um teste diagnóstico, através de uma abordagem bayesiana. Nessa abordagem, a verossimilhança – positiva ou negativa – do primeiro teste é multi-

| Inst# | Classe | Prob. | Inst# | Classe | Prob. |
|---|---|---|---|---|---|
| 1 | p | ,9 | 11 | p | ,4 |
| 2 | p | ,8 | 12 | n | ,39 |
| 3 | n | ,7 | 13 | p | ,38 |
| 4 | p | ,6 | 14 | n | ,37 |
| 5 | p | ,55 | 15 | n | ,36 |
| 6 | p | ,54 | 16 | n | ,35 |
| 7 | n | ,53 | 17 | p | ,34 |
| 8 | n | ,52 | 18 | n | ,33 |
| 9 | p | ,51 | 19 | p | ,30 |
| 10 | n | ,505 | 20 | n | ,1 |

(a) Vinte instâncias, com suas respectivas classes reais, ordenadas pelo escore de probabilidade que foi atribuído por um modelo.




(b) A curva é criada através dos dados observados na tabela. Inicialmente, assume-se que todas as instâncias pertencem a classe "p", que está sendo avaliada. Para cada acerto do modelo, sobe-se um intervalo no gráfico. Para cada erro, anda-se um intervalo para a direita. A área abaixo da curva formada é a métrica AUC.

**Figura 2.12** Exemplo de uma AUC, adaptado de FAWCETT (2006).

plicada pela prevalência populacional, produzindo um pós-teste. Já a verossimilhança do segundo teste é multiplicada pelo pós-teste do primeiro teste, e assim sucessivamente, até que o último teste seja considerado. O pós-teste final deve ser, então, avaliado através de uma interpretação padrão para essa sequência de testes. Normalmente, um diagnóstico é considerado confiável se a probabilidade de seu pós-teste excede $0{,}90$ (KUKAR, 2001; MARCHEVSKY e WICK, 2011).

### Regressão

Quando valores numéricos devem ser preditos ao invés de classes, o desempenho dos modelos deve ser apresentado através de métricas que demonstrem o quão aproximado – ou afastado – seus palpites foram da realidade. Isso porque, ao contrário da classificação que deve predizer uma das categorias de um conjunto finito de opções, em regressão, existem infinitos valores igualmente adequados a um dado problema dentro de um espaço contínuo. Normalmente, essas métricas são baseadas no erro médio – absoluto ou relativo – obtido por um modelo, quando aplicado a uma base de teste. O objetivo aqui é

demonstrar o afastamento global entre os valores previstos pelo modelo e os valores reais para as classes numéricas na base de teste. A Tabela 2.4 apresenta um conjunto sumarizado de métricas para tarefas de regressão, sugerido por WITTEN e FRANK (2005).

**Tabela 2.4** Métricas de desempenho para predições numéricas (regressão). $p$ são valores preditos, $a$ são os valores reais, $n$ representa o número total de instâncias.

| Métrica | Fórmula |
|---|---|
| Erro Médio Quadrado | $\frac{(p_1 - a_1)^2 + \cdots + (p_n - a_n)^2}{n}$ |
| Raiz do Erro Médio Quadrado | $\sqrt{\frac{(p_1 - a_1)^2 + \cdots + (p_n - a_n)^2}{n}}$ |
| Erro Médio Absoluto | $\frac{\lvert p_1 - a_1\rvert + \cdots + \lvert p_n - a_n\rvert}{n}$ |
| Erro Quadrado Relativo | $\frac{(p_1 - a_1)^2 + \cdots + (p_n - a_n)^2}{(a_1 - \bar{a}_1)^2 + \cdots + (a_n - \bar{a}_n)^2}$, onde $\bar{a} = \frac{1}{n}\sum_i a_i$ |
| Raiz do Erro Quadrado Relativo | $\sqrt{\frac{(p_1 - a_1)^2 + \cdots + (p_n - a_n)^2}{(a_1 - \bar{a}_1)^2 + \cdots + (a_n - \bar{a}_n)^2}}$ |
| Erro Absoluto Relativo (EAR) | $\frac{\lvert p_1 - a_1\rvert + \cdots + \lvert p_n - a_n\rvert}{\lvert a_1 - \bar{a}_1\rvert + \cdots + \lvert a_n - \bar{a}_n\rvert}$ |

Essas métricas também podem ser utilizadas em problemas de classificação, para avaliação da calibragem, verificando assim, o quanto as predições sugeridas para cada classe se aproximam da realidade. Caso seja necessário produzir uma métrica única, problemas com a prevalência também podem ser contornados, considerando-se a macromédia – ou média não ponderada – dos erros obtidos por todas as classes. Outras informações sobre a utilização de métricas de erro em problemas de classificação podem ser vistas em FERRI *et al.* (2009)

### 2.3.6 Intervalos de Confiança

A utilização de modelos preditivos baseia-se na ideia de que através de observações feitas em uma amostra, podem-se inferir padrões sobre todos os indivíduos - ou elementos – de uma população. Se a amostra utilizada não for representativa para a população,

os modelos identificados não serão úteis. Contudo, mesmo considerando uma amostra representativa, obtida de um estudo corretamente elaborado, um modelo identificado poderá conter apenas uma ideia da população, em virtude de um provável viés amostral. Resultados obtidos de uma pequena amostra estão sujeitos a incertezas estatísticas, fortemente relacionadas ao tamanho da amostra.

Para contornar este problema, é importante que seja incorporada alguma medida de imprecisão aos resultados expressos pelas métricas. Neste sentido, apresentar os resultados obtidos diretamente na escala original da métrica, juntamente com Intervalos de Confiança (IC) – a fim de se demonstrar as imprecisões amostrais – possui grandes vantagens sobre os valores $P$ usuais, que são dicotomizados em "significante" e "não significante" (ALTMAN *et al.*, 2002).

Um IC tem como principal objetivo indicar a imprecisão de uma estimativa obtida através de uma amostra. Esta estimativa pode indicar um parâmetro estatístico, uma métrica de desempenho, ou qualquer outra característica da população. Dependendo da métrica considerada, o IC pode ser calculado de diversas formas. Contudo, neste trabalho esta sendo adotado apenas o método proporcional para modelos dicotômicos (WILSON, 1927; NEWCOMBE, 1998; ALTMAN *et al.*, 2002).

O IC proporcional pode ser calculado da seguinte forma: Sendo $r$, o número de indivíduos – ou elementos – com uma determinada característica, em uma população de tamanho $n$, pode-se determinar a proporção estimada de tais indivíduos com tal característica através de $p = r/n$. A proporção de indivíduos – ou elementos – que não possuem a característica pode também ser estimada através de $q = 1 - p$. $z$ corresponde a pontuação padrão mono-caudal – ou seja, a quantidade de desvios padrão em uma direção a partir de uma média. Este valor é utilizado para definir o nível de significância $\alpha$ – ou o grau de confiança $1-\alpha$ – desejado, e pode ser obtido através de uma tabela de distribuição normal mono-caudal. Por exemplo, para $95\%$ de confiança $z_{1-\alpha/2} = 1{,}96$. Através dos valores $r$,$n$,$p$,$q$ e $z$, pode-se calcular as quantidades

$$A = 2r + z^2; \quad B = z\sqrt{z^2 + 4rq}; \quad C = 2(n + z^2);$$

e então, o intervalo de confiança

$$\text{de} \quad (A - B)/C \quad \text{até} \quad (A + B)/C.$$

Deve-se observar ainda que ICs cobrem apenas efeitos de variação da amostra, não sendo capazes de controlar vieses causados por erros na elaboração, condução, ou análise do estudo. Outras informações sobre ICs podem ser vistas em (ALTMAN *et al.*, 2002).

### 2.3.7 Discretização

A discretização é a tarefa de dividir o intervalo numérico de um atributo contínuo em categorias – ou seja, transformar um atributo contínuo em um atributo categórico. Essa é uma prática normalmente considerada em Aprendizagem de Máquina, já que um atributo corretamente discretizado costuma apresentar desempenho preditivo superior ao mesmo atributo no formato numérico. Além disso, algumas implementações para Árvore de Decisão e Regras trabalham de forma muito mais lenta quando atributos numéricos estão presentes. Basicamente, a discretização pode ser realizada através das seguintes formas:

- **Discretização Não Supervisionada**: Em que os atributos numéricos são particionados sem o conhecimento das classes das instâncias, sendo agrupados de acordo com algum critério de separação. Dois critérios normalmente considerados são: $(1)$ Divisão da extensão numérica em um número pré-determinado de intervalos do mesmo tamanho; $(2)$ Divisão da extensão numérica em intervalos com a mesma frequência de exemplos – ou seja, intervalos que possuam o mesmo número instâncias da base de treino.

- **Discretização Supervisionada**: Em que os atributos numéricos são particionados considerando as classes das instâncias. Uma das técnicas mais utilizadas é a discretização baseada em entropia, em que a extensão numérica é dividida – ou biparticionada – recursivamente, considerando para cada possível ponto de divisão a partição com maior ganho de informação.

- **Força Bruta**: Nessa estratégia, procura-se pela melhor forma de particionar um atributo em $k$ intervalos. Essa estratégia é exponencial em $k$, e, dessa forma, in-

vestigar todas as possíveis alternativas torna-se inviável. Regras ou heurísticas para a delimitação do espaço de busca devem ser definidas. Essa estratégia pode ser aplicada em conjunto tanto com a discretização não supervisionada quanto com a supervisionada.

Outras informações relacionadas a estratégias de discretização podem ser vistas em (WITTEN e FRANK, 2005).

### 2.3.8 Vieses

Segundo MITCHELL (1980), um viés é todo fator que influencia a escolha de uma hipótese indutiva, em detrimento a outra, sendo tal preferência baseada em evidências extras que não dependem dos dados. GORDON e DESJARDINS (1995), estendem essa definição incluindo qualquer fator – inclusive os dados – que influencie a seleção de uma hipótese indutiva. De uma forma geral, vieses criam restrições no espaço de hipóteses que devem ser consideradas no processo de aprendizagem. Vieses podem ser declarativos – ou de representação – ou procedurais (BRAZDIL *et al.*, 2009). Vieses declarativos descrevem o escopo do espaço de hipóteses, podendo definir, por exemplo, características – ou atributos – relevantes, tipos permitidos de características, ou mesmo uma linguagem de representação, como, por exemplo, restrição à utilização de regras com expressões na forma normal disjuntiva. Vieses procedurais definem restrições de ordem para as hipóteses indutivas, como, por exemplo, preferir primeiro hipóteses simples ou específicas. Considerando essa abordagem, pode-se dizer que um algoritmo de aprendizagem é guiado não apenas pelos dados da base de treino, mas também pelas delimitações impostas pelos vieses no espaço de hipóteses (BRISCOE e CAELLI, 1996).

De acordo com UTGOFF (1986), vieses declarativos podem ser classificados através de várias características. Dentre as mais consideradas, estão força e correção. Um viés declarativo forte implica em um espaço de hipóteses pequeno (viés 1 na Figura 2.13). Já um viés declarativo fraco implica em um espaço de hipóteses grande, com muitas alternativas (viés 2 na Figura 2.13). Um viés declarativo é considerado correto se ele define um espaço de hipóteses que inclui o conceito ou conhecimento desejado – caso

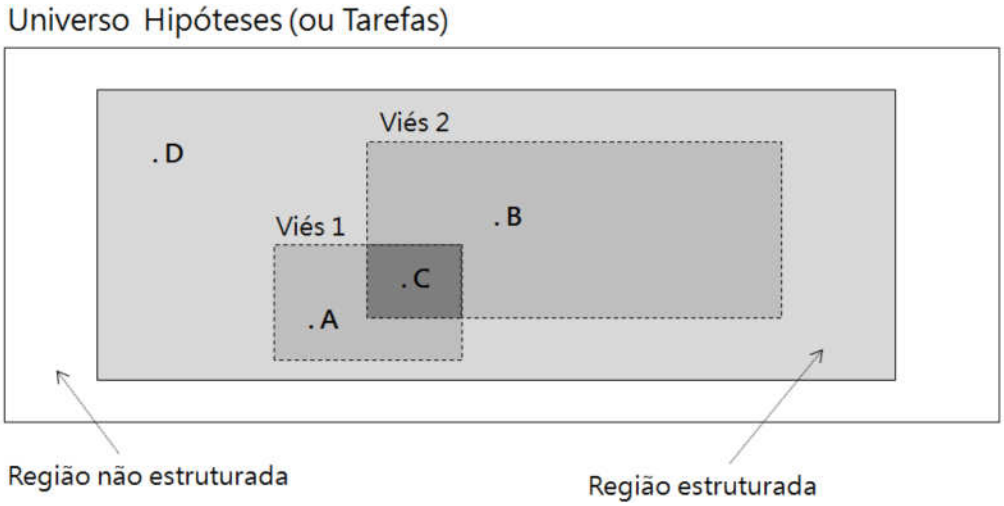



**Figura 2.13** Universo de hipóteses delimitado através de vieses. O viés 1 é relativamente fraco em relação ao viés 2. A hipótese A pertence ao escopo do viés 1. A hipótese B pertence ao escopo do viés 2. A hipótese C está contida na interseção entre o viés 1 e o viés 2. A hipótese D está fora do escopo dos vieses A e B.

contrário, será um viés incorreto.

Vieses são aplicados a um espaço maior, conhecido como universo de hipóteses. Esse universo pode ser dividido em 2 regiões: região aleatória e região estruturada (ver Figura 2.13). A região estruturada contém grupos de hipóteses que obedecem a algum padrão. Essa região está contida no universo de hipóteses. Toda área entre a região estruturada e os limites do universo de hipóteses representa a área não estruturada, que compreende hipóteses que não obedecem a padrões. Essa definição é uma adaptação da proposta feita por VILALTA e DRISSI (2002), que considera um espaço de tarefas de aprendizagem ao invés de um espaço de hipóteses.

### 2.3.9 Meta-Aprendizagem

A Meta-Aprendizagem é uma disciplina que estuda como o processo de aprendizagem pode ser aprimorado através de experiências (VILALTA e DRISSI, 2002). A ideia central está na utilização de todo tipo de conhecimento produzido sobre um domínio para aplicação de Vieses. Esse conhecimento - que na verdade é um metaconhecimento - pode ser obtido através de um especialista do domínio ou através do histórico de execução de um processo de aprendizagem que não necessariamente ocorreu no mesmo domínio. Os metaconhecimentos constituem a base de formação dos vieses, que, por sua vez, são tra-

duzidos em estratégias para restringir o espaço de modelos. Assim, pode-se dizer que a Meta-Aprendizagem investiga as relações entre os domínios e as estratégias de aprendizagem, buscando compreender em que condições determinadas estratégias são mais apropriadas (VILALTA *et al.*, 2005).

A Meta-Aprendizagem é baseada no seguinte princípio: Um mecanismo de base considera apenas um número limitado de hipóteses relativo ao viés fixo, que é incorporado à sua definição. Esse tipo de mecanismo de aprendizagem nunca poderá ser desenvolvido para lidar eficientemente com todas as alternativas do universo de hipóteses, enquanto seu viés permanecer fixo (VILALTA e DRISSI, 2002). Um mecanismo de alto nível que seja capaz de sugerir mecanismos de base adequados a um determinado problema de aprendizagem, terá melhor desempenho do que um simples mecanismo de base, já que, neste caso, o processo de aprendizagem passa a considerar não apenas a identificação de uma hipótese em uma representação, mas também a escolha da representação – que nesse caso é um mecanismo de base. Esse princípio pode ser estendido para a sugestão de um plano de aprendizagem que considera não apenas mecanismos de base, mas também qualquer outra estratégia envolvida no processo.

Uma das aplicações mais exploradas da Meta-Aprendizagem é a recomendação de algoritmos e técnicas. Nesse sentido, a Meta-Aprendizagem busca reduzir conscientemente o espaço de modelos para facilitar a busca por alternativas eficientes. O objetivo da Meta-Aprendizagem, nesse contexto, é o de reduzir o número de alternativas a serem testadas em um dado problema, com um mínimo de perda na qualidade dos resultados obtidos, se comparados a um resultado ótimo (BRAZDIL *et al.*, 2009). Como um modelo ótimo está contido na região estruturada de vieses (ver Seção 2.3.8), a melhor estratégia de busca por um modelo eficiente, nesse sentido, é a aplicação de vieses adequados. Esta é a base para a recomendação de algoritmos e técnicas através de Meta-Aprendizagem.

A Meta-Aprendizagem pode ser aplicada através de diferentes métodos, que consideram não apenas a estratégia de recomendação. Esses métodos, normalmente, seguem uma estrutura padrão que pode ser generalizada através da arquitetura básica apresentada na Figura 2.14. Embora alguns métodos não utilizem exatamente essa arquitetura,

ela serve de referência para a compreensão dos princípios e técnicas utilizados em Meta-Aprendizagem. Basicamente, um sistema de Meta-Aprendizagem pode ser subdividido em 2 módulos: Aquisição de Conhecimento e Recomendação. O módulo Aquisição de Conhecimento é responsável por obter dados sobre os domínios auxiliares similares, transformando esses dados em metadados e armazenando-os em um repositório de metaconhecimentos. Já o módulo Recomendação utiliza os dados previamente armazenados no repositório de metaconhecimentos para sugerir estratégias de aprendizagem. Além disso, todos os metadados obtidos com os experimentos realizados através das técnicas sugeridas são também armazenados no repositório de metaconhecimentos durante o processo de Meta-Aprendizagem – criando, assim, uma espécie de metabase de treino, que se aprimora com o passar do tempo. Domínios que não possuem metaconhecimentos claros são obrigados a definir estratégias de aprendizagem aleatoriamente. Contudo, com a aquisição dos metadados obtidos através dos experimentos realizados – mesmo que aleatoriamente –, aumenta-se também a perícia do sistema em decidir qual estratégia é a mais adequada.

Segundo VILALTA *et al.* (2005) os métodos mais utilizados de Meta-Aprendizagem podem ser classificados da seguinte forma:

- **Mapeamento:** O Mapeamento é uma estratégia de Meta-Aprendizagem que busca construir mecanismos que mapeiam bases de dados – ou aplicações – em modelos preditivos. Os critérios de avaliação normalmente utilizados são acurácia, complexidade de espaço e tempo necessário para execução. Os seguintes métodos aplicam essa estratégia:
  - **Construção Manual de Metarregras:** Constitui a definição manual dos Vieses que serão considerados no sistema através de metarregras. Normalmente esse método utiliza os conhecimentos adquiridos de um especialista - através de elucidação -, que são traduzidos em estratégias de aprendizagem. Esse método tem como desvantagem poder falhar em identificar muitas metarregras importantes.
  - **Aprendizagem em Metanível:** Esse método busca construir um repositório

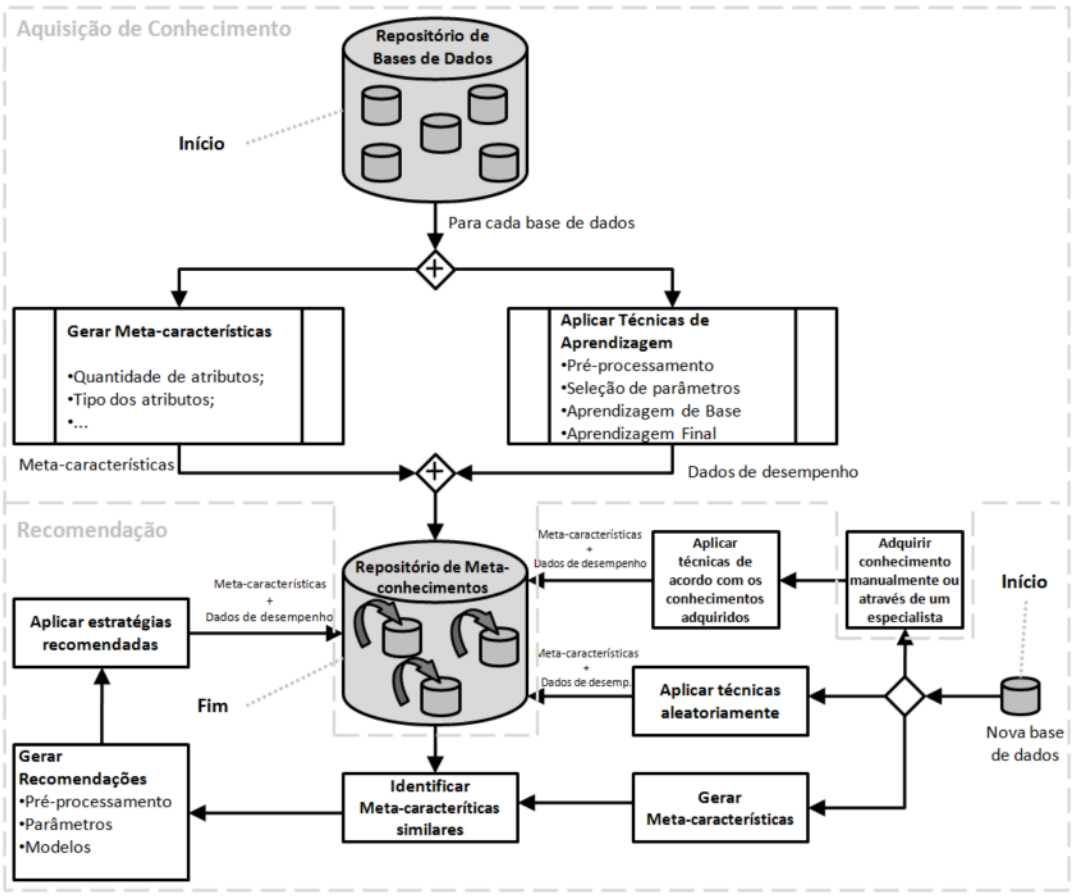



**Figura 2.14** Arquitetura básica de um sistema de Meta-Aprendizagem.

de metaconhecimentos, onde cada elemento tem a forma:

( vetor de metacaracterísticas , melhor modelo )

Um novo elemento pode, então, ser classificado através da identificação de um elemento similar no repositório de metaconhecimentos através de suas metacaracterísticas. Cada elemento do repositório normalmente possui uma lista de Modelos que obtiveram desempenho satisfatório no passado. O melhor modelo do elemento de maior similaridade no repositório é, então, sugerido ao novo elemento. O repositório de metaconhecimentos pode ser composto por elementos obtidos através de metarregras construídas manualmente ou por outras abordagens automáticas.

- **k-Vizinhos-mais-próximo:** Similar à mesma técnica aplicada na aprendizagem de base, esse método busca identificar um número $k$ predefinido de vizinhos mais próximos através das metacaracterísticas – utilizando a mesma abordagem da Aprendizagem de Metanível. O modelo com maior desempenho médio entre os $k$ elementos vizinhos é, então, recomendado para o novo elemento.
- **Ranking:** Essa abordagem busca mapear um banco de dados não em apenas um modelo, mas em um ranking de diferentes modelos. Essa estratégia, além de dar maior flexibilidade de escolha, facilita a identificação de um modelo eficiente já que a recomendação não se limita a apenas uma alternativa. Essa abordagem pode ser incorporada a todos os métodos discutidos anteriormente.

- **Meta-Aprendizagem através de Mecanismos de Base:** Essa estratégia é fundamentada na utilização explícita das informações de desempenho obtidas através de algoritmos de aprendizagem de base. Os seguintes métodos aplicam essa estratégia:
  - **Empilhamento:** Esse é um método que trabalha em uma arquitetura de camadas. Cada camada representa um algoritmo de base. Os algoritmos são treinados um após o outro, até o final da pilha de algoritmos. O primeiro algoritmo é treinado através das características – ou atributos– originais da base

de treino. O segundo algoritmo considera, além das características originais, a predição do primeiro algoritmo como sendo uma característica adicional. Já o terceiro algoritmo considera, além das características originais, a predição do segundo algoritmo, e assim sucessivamente. O último algoritmo a ser treinado realiza a predição final.

  - ***Boosting*:** Esse método é baseado na ideia de gerar um conjunto de modelos de base através de variações da base de treino. Cada variação é uma amostra com reposição – similar às amostras bootstrap –, mas geradas sob um conceito de distribuição ponderada. Essa distribuição é modificada a cada nova variação, dando-se mais peso aos exemplos – ou instâncias – incorretamente classificados na distribuição anterior.
  - **Meta-Aprendizagem através de *Landmarkers*:** *Landmarkers* são mecanismos de aprendizagem de base, extremamente simples, peritos em atuar em áreas específicas. A Meta-Aprendizagem através de *Landmarkers* busca criar um mecanismo de aprendizagem mais avançado através da combinação desses mecanismos simples.
  - **Meta-árvores-de-decisão:** Esse método busca combinar vários modelos através de uma Árvore de Decisão. Cada nó interno da árvore é um conjunto de modelos que produz uma metacaracterística medindo a capacidade de predição da distribuição de probabilidade para uma variável classe de um dado exemplo. Cada ramo representa um Viés que define um subconjunto menor de modelos. Os nós folha correspondem a um único modelo preditivo. Quando um novo exemplo – neste caso, uma base de dados – deve ser classificado, a Meta-árvore-de-decisão recomenda o modelo que parece ser mais adequado na tarefa de predição de sua variável de classe.

- **Seleção Dinâmica de Vieses:** Esta é uma estratégia de Meta-Aprendizagem que se baseia na modificação dinamicamente da Região de Perícia no espaço de hipóteses, através da aplicação de Vieses promissores. O objetivo é buscar por uma melhor cobertura da hipótese desejada, considerando metaconhecimentos que pos-

sam conduzir a vieses corretos – que contenham a hipótese procurada – e fortes – que possuam um tamanho reduzido. O desempenho dos Mecanismos de Base aplicados durante o processo é utilizado como uma métrica simples para guiar aos melhores caminhos a serem seguidos. Essa é uma abordagem diferente da apresentada na arquitetura da Figura 2.14, já que o processo de aprendizagem não ocorre em 2 etapas – considerando a Aquisição de Conhecimento e a Recomendação –, mas sim em uma simples etapa. Essa estratégia pode ser aplicada manualmente através da observação do comportamento dos dados no domínio, ou automaticamente através de aplicação de um algoritmo de Meta-Aprendizagem. A Seleção de Atributos Relevantes é uma das técnicas mais utilizados da Seleção Dinâmica de Vieses.

### 2.3.10 Descoberta de Conhecimento em Bases de Dados

A Mineração de Dados faz parte de um processo maior conhecido como Descoberta de Conhecimento em Bases de Dados (KDD, do inglês *Knowledge-Discovery in Databases*). O KDD é sucintamente descrito na literatura como uma disciplina interessada em investigar técnicas para extração não-trivial de informações implícitas, inicialmente desconhecidas e potencialmente úteis de um conjunto de dados (FRAWLEY *et al.*, 1992; SIETSMA *et al.*, 2002; SCHERMER, 2007). A Mineração de Dados é o cerne do processo de KDD, sendo considerada a etapa mais importante. Apesar de essas duas abordagens possuírem definições muito parecidas, a Mineração de Dados compõe apenas uma fase do processo. Segundo MAIMON e ROKACH (2005), o KDD, como um todo, é composto por uma sequência de atividades que, juntamente à Mineração de Dados, podem ser descritas através de 9 etapas distintas:

1. **Reconhecimento do domínio:** Fase inicial e preparatória,em que os objetivos são definidos. As ferramentas que serão utilizadas durante o processo – técnicas de transformação, algoritmos de aprendizagem e métodos de representação dos resultados – também são definidas nesta etapa.

2. **Captação dos dados:** Seleção e criação de um conjunto de dados no qual será reali-

zada a Mineração de Dados. Esta pode ser uma etapa muito custosa, principalmente se os dados forem provenientes de fontes distintas;

3. **Pré-processamento e limpeza:** Essa etapa inclui o preenchimento manual ou automático de valores inexistentes, e a exclusão de valores ruidosos e discrepantes. Algoritmos de aprendizagem poderão ser utilizados nesta etapa para prever valores inexistentes através da observação de outros valores do conjunto de dados aplicados ao mesmo contexto.

4. **Transformação de dados:** Neste estágio, os dados são aprimorados para serem utilizados pelos algoritmos. Uma redução dimensional da base pode ser aplicada nesta fase, através de seleção de atributos ou registros relevantes. Os dados também podem ser transformados através de técnicas de discretização – transformação de atributos numéricos em categóricos –, ou de técnicas de normalização – transformação de valores astronômicos de um atributo para uma escala padrão, por exemplo, 0 a 1.

5. **Escolha das técnicas de Mineração de Dados adequadas:** Nesta fase, deve-se decidir que tipo de mineração de dados será utilizado. As abordagens mais aplicadas são: classificação, regressão e clusterização. A escolha de uma delas depende do objetivo do processo. Os objetivos mais buscados em Mineração de Dados estão relacionados a predição e descrição. Normalmente, a construção de modelos preditivos está relacionada a técnicas de Mineração de Dados supervisionada, que incluem aí classificação e regressão. Por outro lado, quando se pretende entender um domínio, através da descrição de informações ocultas em seus dados, técnicas de Mineração de Dados não supervisionada são geralmente aplicadas através de clusterização. Alguns aspectos pertinentes à visualização dos dados também são avaliados nesta etapa, quando se deseja descrever os dados de um domínio.

6. **Escolha dos Algoritmos de Mineração de Dados:** Neste estágio, são selecionados os algoritmos que realizarão a busca por padrões nos dados. Poderá ser realizada uma investigação experimental ou mesmo literária, em que o foco será entender

o comportamento de determinados algoritmos na resolução de problemas particulares. O objetivo é tentar identificar os algoritmos de Mineração de Dados mais adequados a realização da tarefa.

7. **Aplicação dos algoritmos de Mineração de Dados:** Nesta etapa, os algoritmos selecionados serão aplicados. Normalmente, um algoritmo de Mineração de Dados possui um conjunto de parâmetros que devem ser "regulados" a cada iteração. Assim, o algoritmo deverá ser aplicado várias vezes até que um resultado satisfatório seja alcançado, com uma combinação de parâmetros ideais.

8. **Avaliação:** Nesta etapa, os padrões obtidos são avaliados em relação aos objetivos definidos no primeiro estágio. Como esta fase do processo está focada na compreensão e utilidade dos modelos obtidos, o reprocessamento de alguns dos passos anteriores pode ser considerado. O conhecimento que foi descoberto sobre os dados é também documentado para uso futuro nesta fase.

9. **Utilização do conhecimento obtido:** Neste estágio, o novo conhecimento adquirido é incorporado a outro sistema e seus efeitos são avaliados. O sucesso desta última etapa determina a efetividade de todo o processo de KDD. O maior desafio deste estágio é adequar às condições produzidas em laboratório para um sistema especialista de uso prático. Muitas vezes, podemos estar lidando com um domínio dinâmico, em que uma simples amostra utilizada para extração do conhecimento poderá estar refletindo apenas um momento estático da população. Em outras situações, o domínio pode sofrer alterações estruturais, com a inclusão de novos atributos, ou conceituais, em que novos valores para alguns atributos passam a ser considerados. Em ambos os casos, o sistema deve ser capaz de lidar com situações que não foram consideradas no momento da aprendizagem.

O processo de KDD é cíclico e, dessa forma, as informações obtidas ao final de um processo poderão servir de base para o início de um novo processo. Além disso, etapas realizadas de forma não satisfatória poderão ser retornadas e reavaliadas durante o ciclo (MAIMON e ROKACH, 2005).

Embora a Mineração de Dados seja apenas uma etapa dentro do processo de KDD, esses termos costumam ser interpretados como sinônimos, sendo “Mineração de Dados” o termo mais utilizado popularmente (SCHERMER, 2007; HAN e KAMBER, 2006; WITTEN e FRANK, 2005). Mineração de Dados é um termo cunhado no mundo dos negócios que se refere à aplicação de algoritmos de Aprendizagem de Máquina a grandes volumes de dados, enquanto KDD é um termo utilizado pela Ciência da Computação e está relacionado a um sentido mais amplo do processo (ALPAYDIN, 2010).

# Capítulo 3

# Proposta de Meta-Aprendizagem

Neste capítulo, está sendo apresentada a estratégia de Meta-Aprendizagem que foi utilizada para auxiliar no desenvolvimento do trabalho aqui apresentado. Inicialmente, é feita uma descrição da proposta e da arquitetura sugerida. Em seguida, apresentam-se alguns trabalhos relacionados. Logo após, discutem-se a construção do espaço de metabusca e a execução de um metamodelo. Por fim, propõe-se um metamodelo para o problema de pesquisa aqui discutido através dessa abordagem.

## 3.1 Descrição da Proposta

Considerando um problema de aprendizagem, assume-se a seguinte abstração: Dado um domínio, e todas as suas possíveis formas de representação[1], existe um universo composto por hipóteses, a princípio desconhecidas, que tentam mapear cada instância deste domínio em um conceito – classe ou valor numérico. No centro desse universo, existe um conceito central que é capaz de classificar corretamente qualquer instância do domínio (ver Figura 3.1 (a)). As hipóteses podem ser interpretadas como funções que tentam se aproximar do conceito central – ou função-verdade – através de diferentes teorias. Hipóteses mais próximas do conceito central possuem melhor desempenho preditivo do que hipóteses mais afastadas. Algumas destas hipóteses podem ser expressas através de algo-

[1]Representações, neste contexto, são as diferentes formas de se descrever um domínio através de um mesmo conjunto de dados. Pode-se considerar, por exemplo, a discretização de alguns atributos, a utilização de apenas alguns valores de classe, a adequação a estratégias de aprendizagem (Classificação / Regressão), a seleção de atributos relevantes, etc.

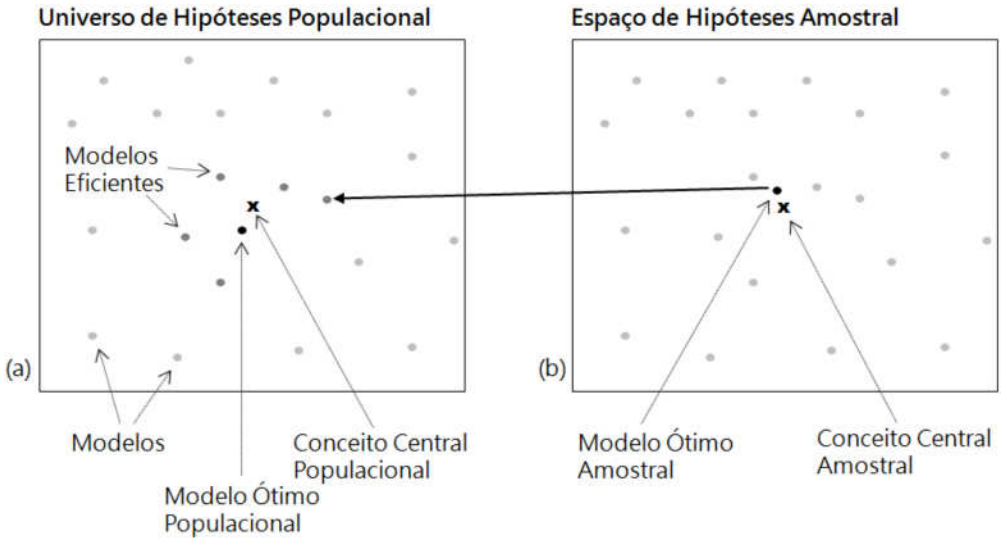



**Figura 3.1** Representação de um universo de hipóteses populacional e de um espaço de hipóteses amostral.

ritmos de aprendizagem, na forma de modelos. Hipóteses que não podem ser expressas através de um algoritmo são invisíveis nesse universo. Ainda nessa abstração, modelos que expressam hipóteses próximas ao conceito central populacional são considerados modelos eficientes, dos quais o mais próximo é considerado o modelo ótimo – dados os algoritmos e as representações disponíveis.

Da mesma forma, considerando uma tarefa de aprendizagem específica, que contemple apenas uma amostra e uma representação do domínio, pode-se abstrair um espaço de hipóteses amostral similar (ver Figura 3.1 (b)). Contudo, devido às possíveis diferenças existentes entre os dados de uma população e os dados de uma amostra, o modelo ótimo amostral, provavelmente, não será o mesmo considerado no universo de hipóteses populacional. Isso ocorre porque este modelo "ótimo" está sujeito às distorções dos dados da amostra – ou seja, está sendo induzido por um viés amostral. Dessa forma, em condições normais, um modelo considerado ótimo no espaço de hipóteses amostral não estará tão próximo do conceito central se ele for considerado no universo de hipóteses populacional. Esse fenômeno – que é conhecido como efeito *overfitting* – pode ser contornado através da aquisição de outras amostras da mesma população para que os vieses amostrais sejam anulados[2]. Em Meta-Aprendizagem, essa estratégia é conhecida

[2]Dependendo do domínio, é necessário que as amostras alternativas sejam obtidas em diferentes contextos de espaço e tempo. A simples divisão de uma grande amostra em subamostras deve não ser considerada

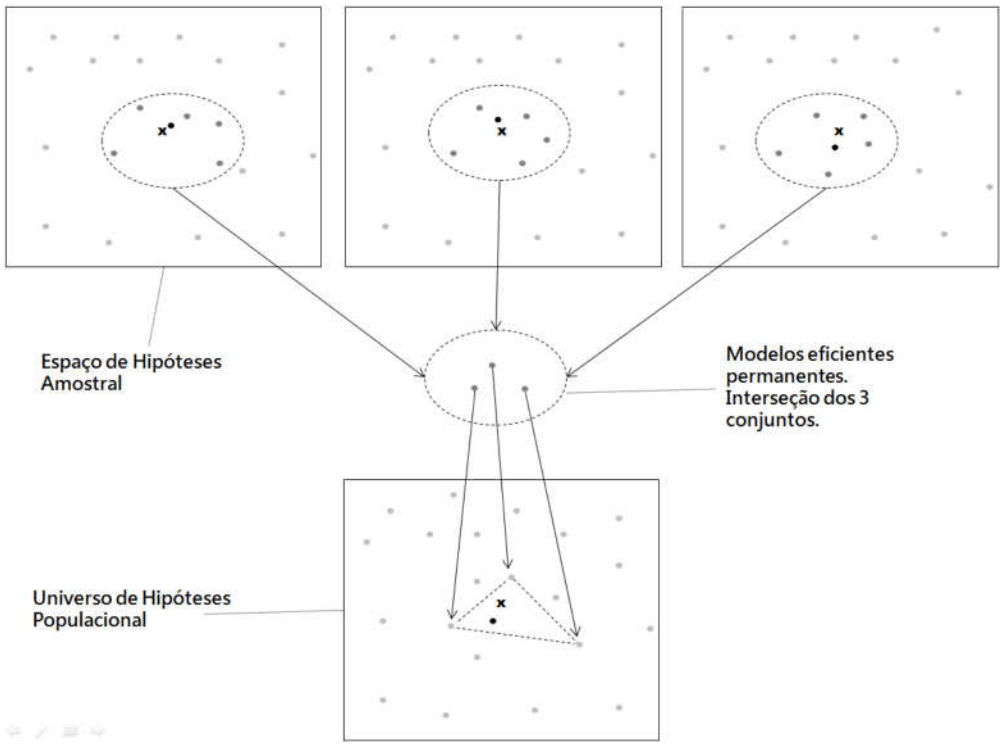



**Figura 3.2** Estratégia de Meta-Aprendizagem *Learn-to-Learn*.

como *Learn to Learn* e seu objetivo é, a cada nova tarefa de aprendizagem, utilizar apenas o conjunto de modelos eficientes que se manteve permanente nas tentativas de aprendizagem anteriores (ver Figura 3.2). No entanto, se apenas uma amostra da população estiver disponível, essa estratégia não poderá ser considerada, pois não haverá conhecimentos anteriores para serem utilizados.

A proposta aqui apresentada sugere que conhecimentos **sobre** o domínio – ou metaconhecimentos – também podem ser utilizados para a aproximação do conceito central populacional, além dos dados da amostra. Esses metaconhecimentos podem ser utilizados para a sugestão de vieses que estabelecem representações mais eficientes do domínio (ver Figura 3.3). Argumenta-se que, mesmo considerando apenas uma amostra, algumas representações do domínio podem possuir desempenho preditivo melhor do que outras. Argumenta-se ainda que se uma nova representação – sugerida pelo domínio e não pela amostra – possui melhor desempenho do que uma representação original, então existe grande possibilidade de os modelos estarem sendo conduzidos ao conceito central popu-

em algumas ocasiões (ALTMAN *et al.*, 2002).

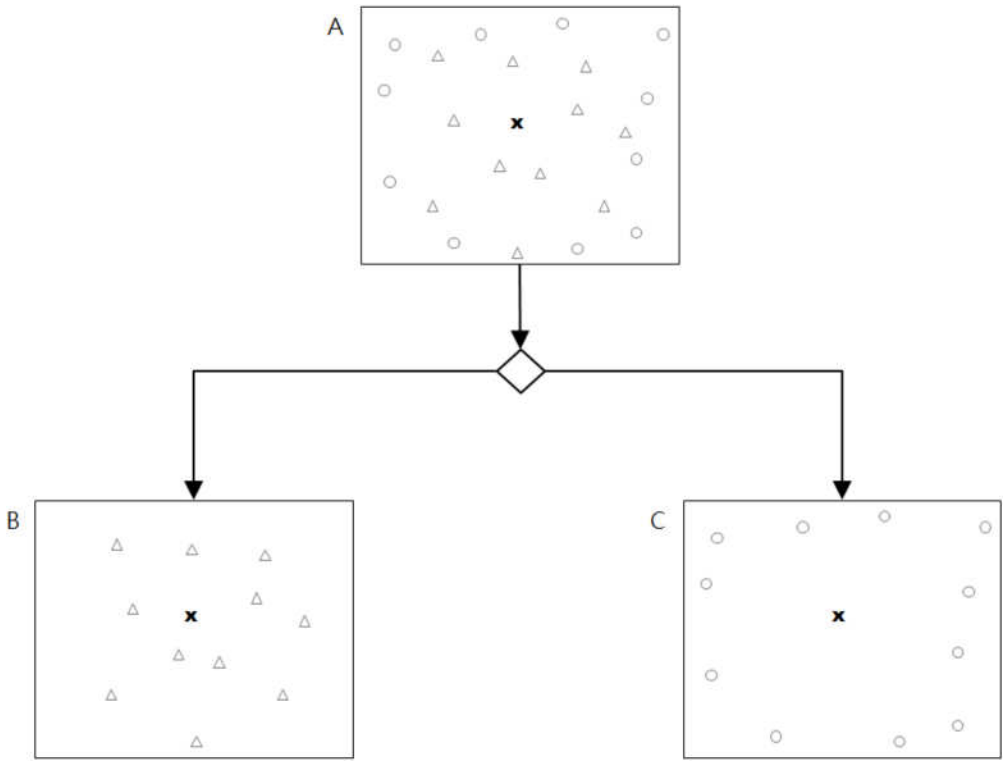



**Figura 3.3** Fundamentação da proposta. No espaço de hipóteses amostral "A" existem 2 representações: triângulos e círculos. O espaço de hipóteses "B" produz modelos preditivos mais eficientes – ou mais próximos do conceito central amostral – do que o espaço de hipóteses "C", pois considera apenas a representação "triângulos", que é mais eficiente do que a representação "círculos".

lacional através dessa representação. Um aspecto que ajuda a validar essa abordagem é o fato de que o desempenho das representações é definido não por apenas um modelo, mas sim pelo desempenho coletivo[3] do conjunto de modelos considerado, diminuindo, assim, as chances de um efeito *overfitting*. Deve-se ainda observar ,que, nessa proposta, uma representação pode estar conduzindo o processo de busca a um modelo eficiente que não é exatamente o modelo ótimo populacional, já que este último pode não estar sendo contemplado nesta representação. Contudo, esse também é um aspecto interessante, já que, se for considerada uma amostra com informações escassas – por exemplo, poucas ou nenhuma instâncias para uma determinada classe –, a busca poderá estar sendo induzida ao melhor modelo que pode ser obtido com os dados que se têm em mãos. O sucesso dessa abordagem depende de que a única amostra disponível para treino tenha uma re-

[3]O desempenho coletivo, neste caso, pode ser verificado através da média dos desempenhos obtidos por todos os modelos. Em uma estratégia mais sofisticada, pode-se considerar, ainda, o desempenho médio obtido pela metade mais eficiente dos modelos.

presentatividade mínima satisfatória para o domínio – ou seja, possua um conjunto de modelos eficientes que também sejam eficientes na população. Pode-se dizer que essa proposta através de 3 dimensões – amostra(s), representações e algoritmos – busca alcançar 2 objetivos: (1) Identificar modelos, em representações alternativas, que sejam mais eficientes do que os que seriam obtidos através da representação original; (2) Induzir o processo de busca ao conceito central populacional, mesmo considerando uma quantidade reduzida de instâncias para treino. Essa proposta está sendo chamada Seleção Dinâmica de Vieses Especializada (SDVE).

## 3.2 Arquitetura da Proposta

Para que essa proposta seja aplicada, é necessário que, inicialmente, metaconhecimentos do domínio sejam adquiridos. Nesse contexto, 2 estratégias podem ser empregadas: (1) Investigação direta, que compreende a busca direta por oportunidades de conhecimentos do domínio que possam ser aplicados no aprimoramento da aprendizagem; (2) Elicitação, que corresponde à tarefa de adquirir, junto a um especialista do domínio, metaconhecimentos especializados que não seriam facilmente observados nos dados ou não seriam considerados por mecanismos de aprendizagem.

Os metaconhecimentos fornecidos pelo especialista ou observados nos dados podem sugerir vieses que delimitam o escopo de diferentes formas. Para facilitar a referência aos diferentes tipos de vieses, estão sendo introduzidos os seguintes conceitos:

- **Viés de Representação (VR):** É um delimitador guiado por um metaconhecimento que cria um novo espaço de hipóteses dentro do espaço de hipóteses atual. Em relação ao contexto, estão sendo consideradas as seguintes subclasses de VRs:
  - **Genérico:** É o VR maior que define o escopo atual e contém os demais vieses;
  - **Interno:** São VRs contidos dentro de um VR genérico. VRs internos também podem ser genéricos em relação aos VRs contidos em seu escopo;
  - **Externo:** São VRs externos ao escopo do VR genérico atual;

  - **Primário:** São VRs que devem ser aplicados ao escopo inicial antes de se considerar qualquer estratégia de aprendizagem.

  Esses vieses estão sendo representados por um retângulo pontilhado nos diagramas.

- **Viés de Decisão (VD):** É um processo decisório para identificação do melhor VR de um conjunto de VRs. Normalmente, os VRs que compõem o conjunto são baseados em possíveis alternativas para um único metaconhecimento. Esses vieses estão sendo representados por um losango nos diagramas.

- **Viés de Avaliação (VA):** É um conjunto de VRs obrigatórios que normalmente se refere a possíveis alternativas para um único metaconhecimento. Cada VA possui sua própria lista de VRs. Em um processo de aprendizagem, todos os VAs devem ser considerados ao mesmo tempo, através da combinação exaustiva de cada um de seus elementos, para a formação de um novo conjunto de VRs. Esses vieses estão sendo representados por um retângulo pequeno contendo uma exclamação nos diagramas.

Considerando-se os princípios apresentados anteriormente, a Meta-Aprendizagem através da proposta SDVE pode ser vista como um processo composto por duas dimensões: $(1)$ Dimensão-macro: Que mapeia um espaço de busca no universo de hipóteses; $(2)$ Dimensão-micro: Que define uma região específica do VR que deverá ser explorada exaustivamente. A Figura 3.4 mostra como essas dimensões se relacionam. VDs – produzidos a partir das várias alternativas para um mesmo metaconhecimento – definem VRs alternativos, que dessa forma, guiam a execução do processo. O desempenho médio obtido em cada dimensão-micro é calculado através de VAs. Esse desempenho médio é atribuído ao VR relacionado, sendo utilizado pelo VD para decidir qual caminho seguir. Assumindo-se que cada dimensão-micro será exaustivamente explorada através dos VAs – considerando os recursos computacionais e algoritmos disponíveis –, a busca no espaço de hipóteses reduz-se a um problema de identificação das regiões mais promissoras na dimensão-macro. Um aspecto desafiador desta abordagem está relacionado a

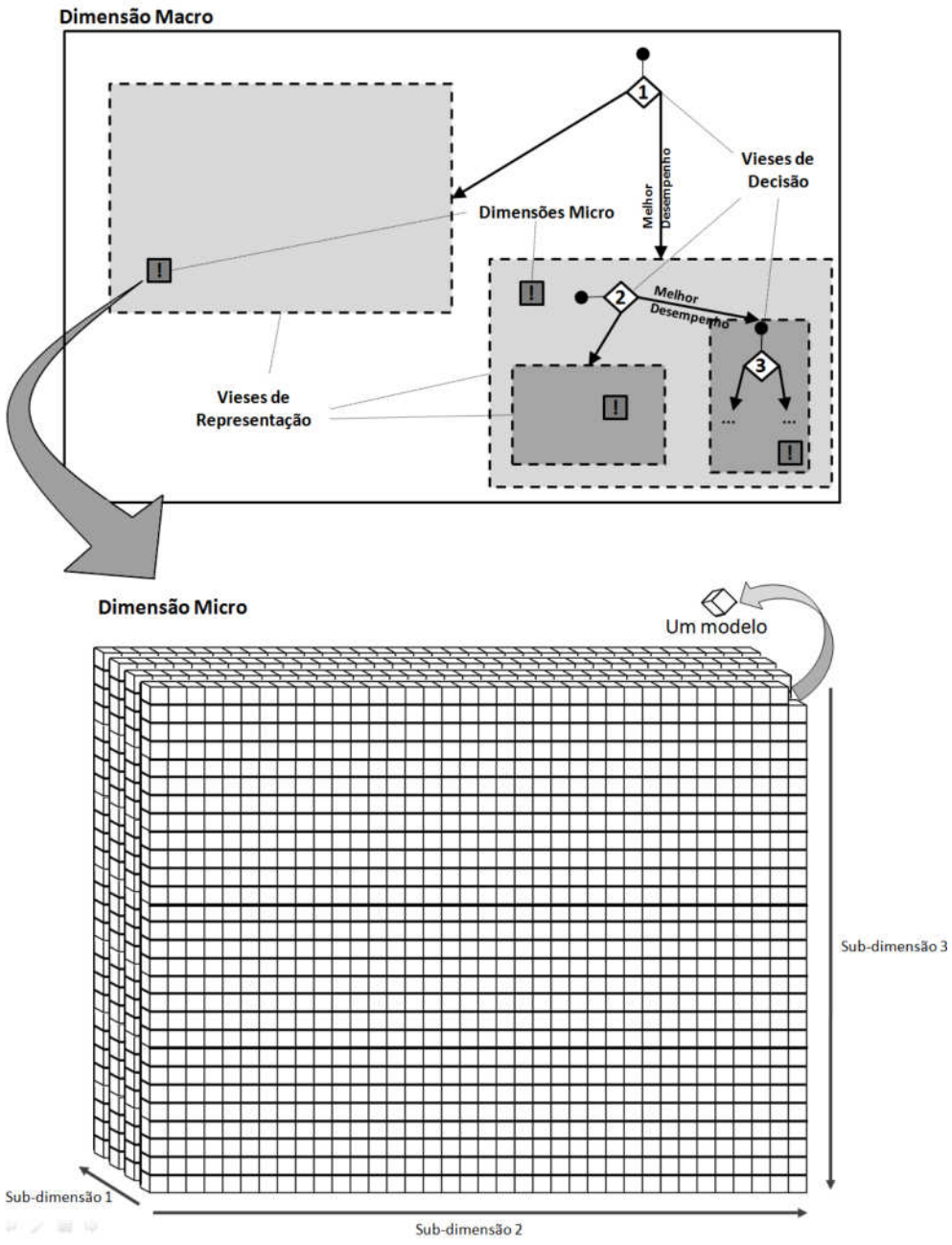



**Figura 3.4** Dimensões de busca utilizadas na abordagem Seleção Dinâmica de Vieses Especializada. A dimensão-macro explora um espaço de busca através de decisões. Já a dimensão-miscro explora um espaço de busca exaustivamente. Juntas, as 2 dimensões fazem buscas especializadas no universo de modelos.

como construir o plano de aprendizagem que irá orientar os VDs. Uma adaptação feita à estratégia de Meta-Aprendizagem Seleção Dinâmica de Vieses está sendo considerada.

O produto final dessa abordagem é um espaço de busca, composto por vieses especializados – ou fundamentados – que, ao ser explorado – através de seleção dinâmica –, é capaz de sugerir um conjunto de modelos promissores. Esse espaço de busca pode ser definido sem a necessidade de se processar nenhum algoritmo de aprendizagem, bastando apenas que os metaconhecimentos – fornecidos pelo especialista ou observados nos dados – estejam disponíveis.

## 3.3 Trabalhos Relacionados

Em GORDON e DESJARDINS (1995), é proposto um *framework* que trata a tarefa de identificar uma hipótese eficiente como um processo de busca no espaço de vieses e "metavieses". Apesar de essa abordagem considerar a aprendizagem como um processo de múltiplas camadas – como é feito na abordagem SDVE –, a arquitetura do espaço de busca proposto considera apenas VRs para delimitação do espaço e aquilo que está sendo chamado de "vieses procedurais", que são, na verdade, os algoritmos adequados.

Em TODOROVSKI e DŽEROSKI (2003), é proposta uma abordagem conhecida como Meta-árvore de Decisão (MAD) para identificação dos mecanismos de base adequados a uma dada Base de Treino. A proposta aqui apresentada diferencia-se dessa abordagem nos seguintes aspectos: $(1)$ O objetivo de uma MAD é o de identificar um mecanismo de base adequado, enquanto uma SDVE tem o objetivo de identificar um conjunto de modelos preditivos eficientes, considerando não apenas mecanismos de base, mas também outros aspectos pertinentes ao processo de aprendizagem, como pré-processamento, metaconhecimentos e até mesmo outras estratégias automáticas de Meta-Aprendizagem; $(2)$ A estrutura de uma MAD é construída automaticamente, utilizando-se o mesmo paradigma de construção de uma Árvore de Decisão, enquanto a estrutura de uma SDVE é construída manualmente, através de interpretações do domínio;$(3)$ Uma MAD utiliza em seus nós conjuntos de mecanismos de base que são avaliados na construção árvore através do mesmo cálculo de ganho de informação utilizado em Árvores de Decisão, enquanto a

SDVE utiliza em seus nós VDs que são avaliados na construção da estrutura, através de um método próprio sugerido para se calcular o ganho de informação.

Em (NGUYEN, 2010), é apresentado um *framework* para facilitar a elicitação de conhecimentos de um domínio para a elaboração de sistemas de aprendizagem. A proposta é baseada em um esquema de aprendizagem conceitual hierárquico, para tratar conhecimentos imprecisos e complexos que normalmente são manipulados neste tipo de tarefa. Contudo, essa abordagem trata a hierarquia dos conhecimentos de um modo diferente daquele definido através do conceito de vieses, que é utilizado na proposta SDVE.

Em BAXTER (2011), é investigado um modelo para seleção automática de vieses. Essa abordagem é baseada na ideia de que o conceito central está contido em um ambiente de tarefas de aprendizagem. Assim, através desse ambiente, um mecanismo de aprendizagem pode obter amostras de diversas tarefas, para poder buscar por um espaço de hipóteses que contenha uma boa solução para muitos dos problemas do ambiente. Advoga-se que um espaço de hipóteses (ambiente) que produza bons desempenhos para um número suficientemente grande de tarefas de aprendizagem produzirá também bons resultados para uma nova tarefa no mesmo ambiente. Contudo, uma certa quantidade de tarefas de aprendizagem deve estar disponível no mesmo ambiente para a aplicação dessa abordagem. O domínio aqui apresentado não dispõe nem de novas bases de treino para produção de novas tarefas de aprendizagem, nem de domínios similares de onde pudessem ser extraídos metaconhecimentos úteis, como foi demonstrado através da estratégia Mapeamento em Metanível.

## 3.4 Definição do Espaço de Busca de uma SDVE

Como foi discutido anteriormente, para a elaboração do espaço de busca de uma SDVE, inicialmente, deve-se obter metaconhecimentos do domínio para a sugestão de vieses. Esses metaconhecimentos podem ser obtidos de especialistas do domínio através de reuniões explanatórias ou através de técnicas de elicitação como as discutidas em AYYUB (2001) e GAAG *et al.* (2002). Outros metaconhecimentos importantes - como estruturas, tipos de dados e distribuições estatísticas - também podem ser obtidos pelo

minerador através da simples observação do comportamento dos dados. De uma forma geral, é importante que os metaconhecimentos considerados sejam realmente consistentes com o domínio, pois eles serão utilizados para conduzir a busca no espaço de hipóteses.

Para cada metaconhecimento obtido, vieses equivalentes são sugeridos, formando um conjunto de metadados, em que cada elemento é um par metaconhecimento / viés. Elementos desse conjunto que sugerem VRs primários devem ser retirados e aplicados imediatamente aos dados para redução do espaço de busca. Esses VRs normalmente são sustentados por fundamentos consagrados, altamente consistentes com o domínio, que devem ser considerados antes de qualquer tentativa de aprendizagem. Um exemplo de um VR primário seria a eliminação de atributos que claramente não são informativos para a aprendizagem. Alguns desses VRs primários podem ser observados pelo minerador, e outros, que são próprios do domínio, devem ser inferidos através dos metaconhecimentos fornecidos pelo especialista.

Logo após a retirada – e aplicação – dos VRs primários, deve-se selecionar os elementos do conjunto de metadados cujos vieses serão utilizados como VAs no processo de aprendizagem. Vieses utilizados como VAs normalmente são caracterizados por teorias que devem ser investigativas juntamente com qualquer outra teoria do domínio – por exemplo, os algoritmos de aprendizagem. Esses vieses devem ser selecionados com cautela já que são avaliados de forma conjunta – ou seja, os VRs alternativos de todos os VAs considerados serão combinados exaustivamente e todas as combinações devem ser avaliadas a cada etapa do processo de aprendizagem. O critério para seleção de VAs dependerá dos objetivos da aprendizagem, e aqui, mais uma vez, o especialista poderá dar sua contribuição – que neste caso, também pode ser considerada como um metaconhecimento.

O passo seguinte na definição do espaço de busca para uma SDVE consiste na definição dos VDs. Após a seleção dos VAs, eles são também retirados do conjunto de metadados, e os vieses restantes são utilizados como VDs. Esse conjunto de vieses remanescentes deve ser reorganizado em uma lista de aplicação, de forma que os vieses mais fracos sejam considerados primeiro. O objetivo aqui é fazer com que os VDs que apontam

para VRs mais genéricos – ou fracos – sejam considerados antes dos VDs que apontam para VRs mais específicos – ou fortes –. Para facilitar essa interpretação, a capacidade média de generalização de todos os VRs alternativos está sendo atribuída ao VD avaliado. Como essa é uma tarefa baseada em suposições subjetivas do minerador e do especialista, é importante que haja uma interpretação que mostre o quão genérico é o VD. Em alguns casos, a escolha do VD mais genérico pode ser difícil, já que alguns VDs alternativos podem parecer estar no mesmo nível de generalização. Para contornar esse problema, está sendo proposta uma medida de peso, baseada na consistência dos metaconhecimentos, que pode ser utilizada para auxiliar na escolha de VDs que pareçam estar no mesmo nível de generalização. A Tabela 3.1 sugere uma escala para a interpretação destes pesos.

**Tabela 3.1** Escala para interpretação da consistência dos metaconhecimentos.

| Peso | Consistência do Metaconhecimento |
|---|---|
| 0,5 | **Fraca**.<br>– Fenômenos nunca observados.<br>– Especulações sobre o domínio. |
| 1 | **Média**.<br>– Fenômenos que podem ser observados ocasionalmente.<br>– Suposições sobre o domínio com alguma fundamentação. |
| 2 | **Alta**.<br>– O fenômeno pode ser claramente observado no domínio. |

VDs devem ser avaliados através de uma métrica de desempenho pré-determinada. Dependendo dos objetivos da aprendizagem, essa métrica pode ser baseada em critérios como: (1) Capacidade preditiva média dos modelos; (2) Complexidade de tempo; (3) Capacidade de adaptação dos mecanismos de base. Cada VD pode utilizar uma métrica diferente, independente da opção feita para os demais VDs. Além disso, uma lista de métricas também pode ser aplicada a um VD. Contudo, apenas um valor de desempenho – que será utilizado para análise comparativa durante a busca no espaço de hipóteses – deve ser produzido.

A estrutura final de uma SDVE é definida da seguinte forma: Todo o escopo de busca encontra-se dentro da área de interseção considerada pelos VRs primários. Nessa área, encontram-se os VDs que apontam para seus respectivos VRs alternativos que serão ava-

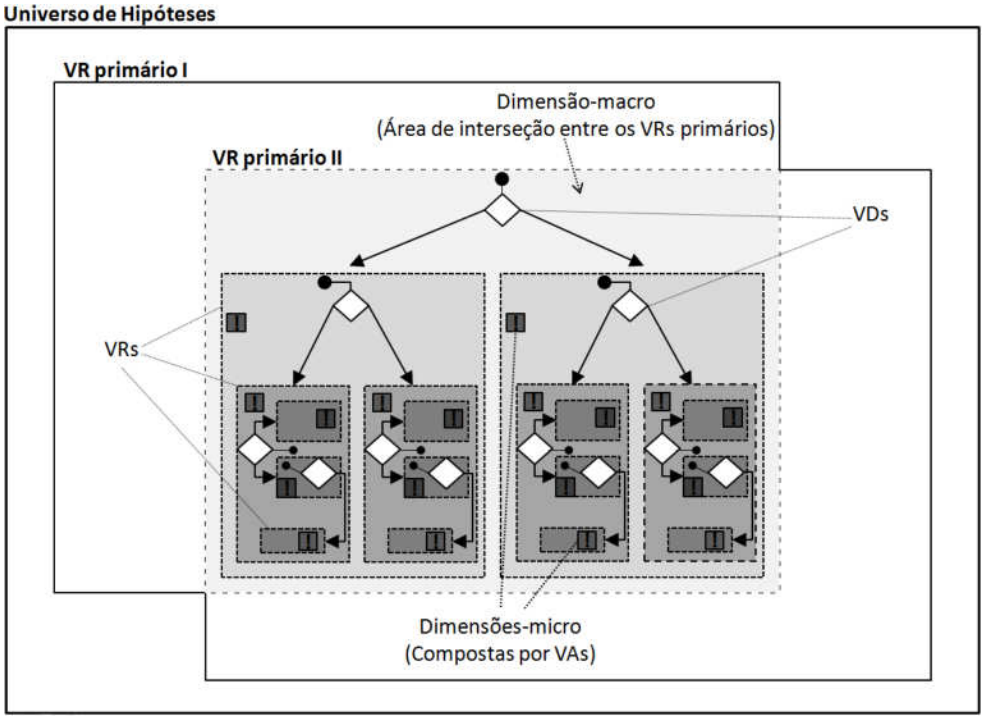


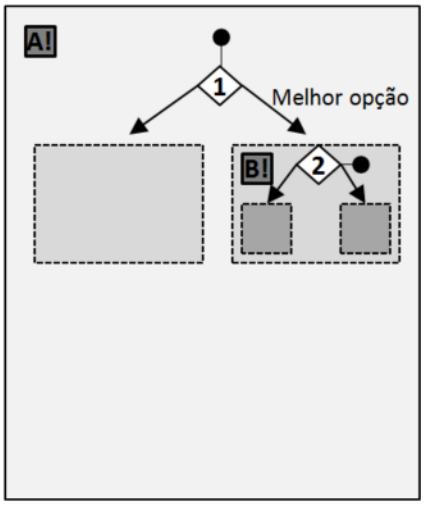

**Figura 3.5** Estrutura de busca de uma SDVE.

liados em tempo de execução. Todo VD está dentro de um VR e aponta para outros VRs, que podem ser internos ou externos. Em tempo de execução, o desempenho do melhor VR alternativo será comparado ao VR genérico atual. Um VD pode apontar para apenas um VR. Um VR pode conter vários VDs que são avaliados em sequência – no diagrama essa sequência é definida por setas tracejadas. VRs que devem ser avaliados contêm uma dimensão-micro composta por VAs. A Figura 3.5 mostra um espaço de busca para uma SDVE pronto para ser avaliado.

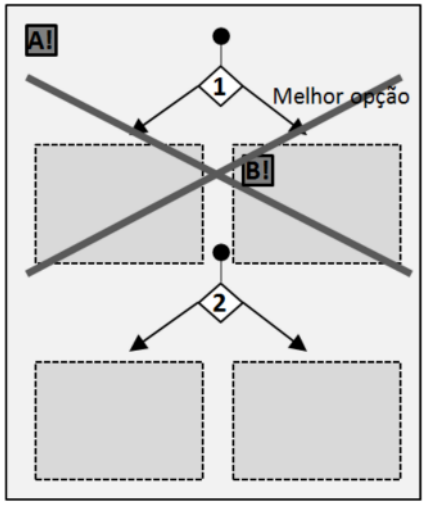
## 3.5 Execução da Busca no Espaço de Hipóteses através de SDVE

Após a elaboração do escopo da SDVE, o processo de busca / aprendizagem pode ser iniciado. A estratégia que está sendo utilizada para a realização da busca no espaço de hipóteses é a Seleção Dinâmica de Vieses. Essa estratégia determina que os caminhos a serem percorridos no espaço de busca devem ser guiados por métricas de desempenho pré-determinadas.

O processo de aprendizagem é executado da seguinte forma: O VD pertencente ao




(a) Um exemplo considerando que o melhor VR alternativo possui desempenho médio na dimensão-micro superior ao desempenho médio da dimensão-micro do VR genérico atual. O próximo VD é aplicado dentro do VR alternativo.




(b) Um exemplo considerando que o melhor VR alternativo **não** possui desempenho médio na dimensão-micro superior ao desempenho médio da dimensão-micro do VR genérico atual. O próximo VD é aplicado no VR genérico atual.

**Figura 3.6** Sequências alternativas de processo na proposta SDVE.

VR mais genérico no universo de hipóteses é chamado, e este avalia as dimensões-micro dos VRs alternativos. O VR alternativo cuja dimensão-micro apresentou melhor desempenho médio é selecionado para continuar o processo (ver Figura 3.6(a)). Se o maior desempenho médio obtido dentre as dimensões-micro para VRs alternativos for inferior ao desempenho médio da dimensão-micro para o VR genérico – no qual está contido o VD –, o próximo VD é aplicado no VR genérico atual sem considerar o melhor VR alternativo do VD anterior, ou seja, considerando todo espaço de busca do VR genérico[4] (ver Figura 3.6(b)). Da mesma forma, VRs que não são aplicáveis ao VR genérico devem ser desconsiderados. Se todos os VRs de um VD forem desconsiderados, o VD também será desconsiderado. O processo continua até que o último VD seja chamado. Ao final, esse processo apresenta como resultado: (1) A sequência de VRs considerada na busca; (2) Um conjunto contendo todos os modelos avaliados nas dimensões-micro com seus respectivos desempenhos. Para identificação de modelos eficientes, esse conjunto pode ser facilmente analisado através de *rankings*.

[4]Esta regra não é válida para o primeiro VD do modelo, pois ele não possui um VA. Nesse caso, o melhor VR alternativo será o escolhido.

## 3.6 Construção de um Metamodelo para o Problema de Pesquisa

Nesta seção, está sendo demonstrado como foi construído o metamodelo alternativo utilizado para auxiliar na resolução do problema de pesquisa aqui apresentado. Inicialmente, são relacionados os metaconhecimentos elicitados, juntamente aos vieses correspondentes que foram sugeridos. Em seguida, discute-se como os Vieses foram ponderados. Ao final, apresenta-se o metamodelo, que na verdade é um espaço de busca especializado, definido através da abordagem SDVE.

### 3.6.1 Metaconhecimentos Adquiridos

Para que a abordagem SDVE fosse aplicada no domínio aqui apresentado, inicialmente, metaconhecimentos foram elicitados junto ao especialista, mesmo antes de os dados terem sido fornecidos. Nessa etapa, o especialista que, neste caso, é um psiquiatra, forneceu vários dados sobre as características do TDAH, incluindo as particularidades do transtorno, informações epidemiológicas, características dos dados que são produzidos no diagnóstico, além de especulações sobre as possíveis relações entre o TDAH e o Jogo do Supermercado. Após a aquisição dos dados, outras informações sobre suas características puderam ser observadas, e esses novos metaconhecimentos foram também considerados juntamente aos metaconhecimentos fornecidos pelo especialista. Nas subseções a seguir, estão sendo apresentados todos os metaconhecimentos que foram considerados na construção do metamodelo alternativo – através da abordagem SDVE – utilizado neste trabalho.

**Três pontos de corte podem ser considerados:** $4$**,** $5$ **ou** $6$ **sintomas (Especialista)**

Normalmente, o diagnóstico do TDAH é realizado através do número de sintomas de desatenção e hiperatividade/impulsividade que são observados no indivíduo. De acordo com o DSM IV, indivíduos que possuem $6$ ou mais sintomas de desatenção e/ou hiperatividade/impulsividade são classificados como TDAH. Contudo, segundo o especialista, existe a possibilidade de indivíduos que apresentaram uma quantidade menor de sintomas serem também portadores do transtorno – ou casos subsindrômicos. Para ele, como os

dados de diagnóstico do TDAH são expressos na forma de quantidade de sintomas observados, seria importante que um mecanismo preditivo fosse avaliado considerando outros pontos de corte além do sugerido pelo DSM IV. Os seguintes pontos de corte foram, então, propostos:

- $6$ ou mais sintomas, como é sugerido pelo DSM IV;
- $4$ ou mais sintomas e $5$ ou mais sintomas, como foi sugerido pelo especialista.

**Os indivíduos podem ser classificados através de $4$ ou $2$ classes (Especialista)**

De acordo com o DSM IV, o TDAH é diagnosticado através de $4$ classes: Desatento, Hiperativo/Impulsivo, Combinado, e Não-TDAH. Contudo, segundo o especialista, um modelo capaz de dividir um grupo de indivíduos em apenas $2$ classes – TDAH e Não-TDAH – é quase tão importante quanto um modelo capaz de identificar as $4$ classes – ou subtipos – do diagnóstico. Essa é uma informação importante, já que, com um número menor de classes, aumenta-se o desempenho do mecanismo de aprendizagem.

**Alguns atributos do Jogo do Supermercado parecem ser redundantes (Especialista)**

De acordo com o especialista, como as $18$ fases do jogo consistem em tarefas similares, deve-se verificar a possibilidade de diminuição deste número de fases, considerando apenas aquelas que apresentarem desempenho preditivo mais relevante. Como cada fase é representada na base de treino através de um par de atributos ponto / tempo, a identificação de fases relevantes pode ser entendida como uma seleção de atributos relevantes, que também é uma técnica de Meta-aprendizagem baseada na estratégia de Seleção Dinâmica de Vieses.

Foi considerada, também, a existência de $4$ grupos distintos de atributos: (1) Atributos de Pontos; (2) Atributos de Tempo; (3) Atributos das Primeiras Fases do Jogo; (4) Atributos das Últimas Fases do Jogo.

Essas são informações importantes, já que, em condições normais, diminuindo-se o número de atributos, aumenta-se o desempenho do mecanismo de aprendizagem.

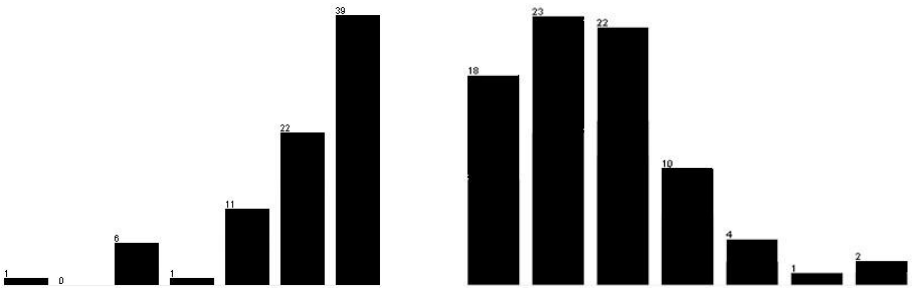
**Os dados são aplicáveis a estratégias de Classificação ou Regressão**

As bases de treino que foram fornecidas consistem de dados de indivíduos que jogaram o Jogo do Supermercado e foram avaliados de acordo com o número de sintomas de desatenção e hiperatividade observados. Essa é uma configuração que certamente sugeriria a aplicação de técnicas de regressão para aproximação do número de sintomas. Contudo, o domínio do problema aqui apresentado possui um forte viés para classificação que também deve ser ponderado. Apesar de os indivíduos na base de treino fornecida terem sido avaliados através do número de sintomas, o diagnóstico do TDAH é categórico – ou seja, o número de sintomas observados deve ser traduzido através de uma função padronizada – como, por exemplo, a sugerida pelo DSM IV – para produção de um resultado categórico final. Assim, 2 estratégias de aprendizagem poderão ser consideradas: Classificação e Regressão.

**Os atributos numéricos do Jogo do Supermercado apresentam distribuições assimétricas**

A maioria dos algoritmos de aprendizagem para Classificação ou Regressão que utilizam atributos numéricos assume que esses atributos possuem distribuição estatística normal(HAN e KAMBER, 2006; WITTEN e FRANK, 2005). Observando os atributos de pontos e tempo produzidos pelo Jogo do Supermercado, pode-se perceber que eles possuem distribuições consideravelmente assimétricas. Atributos de pontos e atributos de tempo apresentam diferentes tipos de assimetria. Muitos dos jogadores parecem alcançar altos escores no jogo, mas nem todos – fazendo com que a distribuição para esse atributo seja assimétrica à esquerda (ver Figura 3.7(a)). Por outro lado, apesar do tempo gasto em cada fase ser relativamente baixo para muitos jogadores, alguns gastam mais tempo do que o normal, o que faz desta uma distribuição assimétrica à direita (ver Figura 3.7(b)).

Como esse comportamento pode ser observado em todos os atributos de pontos e tempo produzidos pelo jogo, acredita-se que os resultados obtidos não serão muito satisfatórios, se esses atributos forem considerados no formato – ou tipo – numérico.




(a) Atributos de pontos ganhos. (b) Atributos de tempo gasto.

**Figura 3.7** Exemplos de distribuições normalmente observadas nas amostras. As colunas representam frequências de instâncias para intervalos iguais de valores.

**Uma quantidade reduzida de exemplos está disponível para treino.**

As amostras obtidas para treino dos modelos apresentam uma quantidade de exemplos relativamente pequena, se for considerada a quantidade de atributos delas. Nesse caso, mesmo considerando amostras de treino representativas da população, a variância[5] pode ser muito alta – o que pode produzir métricas de desempenho imprecisas e intervalos de confiança muito largos.

**Foram disponibilizadas** $2$ **bases de treino originadas do mesmo domínio**

Duas bases de dados, compostas de diferentes exemplos do mesmo domínio, foram fornecidas para treino dos mecanismos de aprendizagem. Uma das bases consiste de uma amostra de crianças / adolescentes que jogaram o Jogo do Supermercado, e outra de uma amostra de adultos que também jogaram o jogo (ver Seção 4.5.1). A princípio, na visão de um minerador, essas bases poderiam ser somadas para aprimorar o desempenho mecanismos de aprendizagem. Contudo, o especialista do domínio não encorajou a adoção dessa abordagem, já que as bases pertencem a subdomínios diferentes, em que crianças / adolescentes e adultos apresentam comportamentos distintos em relação ao TDAH. Como esse é um metaconhecimento que não tem influência direta na escolha dos vieses, decidiu-se aplicá-lo apenas ao VR que apresentar melhor desempenho no experimento.

[5] A variância, nesse caso, mede as variações dos modelos produzidos pelo mecanismo de aprendizagem, que podem ser observadas de uma amostra de treino para outra.

### 3.6.2 Definição do Espaço de Busca

Para que o espaço de busca seja definido, é necessário que primeiro sejam sugeridos vieses para os metaconhecimentos elicitados e observados. A Tabela 3.2 apresenta um conjunto de metadados, composto por pares metaconhecimento / viés, que representam os metaconhecimentos discutidos na Seção 3.6.1, juntamente aos seus respectivos vieses sugeridos.

**Tabela 3.2** Metaconhecimentos e vieses sugeridos.

| Id | Metaconhecimento | Viés Sugerido |
|---|---|---|
| 1 | Uma quantidade reduzida de exemplos está disponível para treino. | Aplicar a técnica de reamostragem *Bootstrap Aggregating* (BAGGING) e comparar seu desempenho ao desempenho obtido utilizando-se a base de treino original. |
| 2 | Os indivíduos podem ser classificados através de $4$ ou de $2$ classes. | Escolher o conjunto de classes que apresenta melhor desempenho. |
| 3 | Os atributos numéricos do Jogo do Supermercado apresentam distribuições assimétricas. | Transformar os atributos numéricos em categóricos através de técnicas de discretização. Comparar o desempenho entre atributos numéricos e atributos categóricos. |
| 4 | Três pontos de corte podem ser considerados: $4$, $5$ ou $6$ sintomas. | Investigar todas essas possibilidades. |
| 5 | Os dados são aplicáveis a diversos mecanismos de aprendizagem disponíveis. | Investigar o maior número possível de mecanismos de aprendizagem. |
| 6 | Alguns atributos do Jogo do Supermercado parecem ser redundantes. | Verificar o desempenho isolado de: atributos de pontos, atributos de tempo, atributos das primeiras fases, atributos das últimas fases. Aplicar uma técnica de seleção de atributos relevantes. Comparar o desempenho de todas essas estratégias. |
| 7 | Os dados são aplicáveis a estratégias de Classificação e Regressão. | Escolher a estratégia mais eficiente. |

Inicialmente, deve-se decidir quais dos elementos desse conjunto serão considerados vieses primários. O viés de número $3$ no conjunto de metadados parece ser um forte candidato, já que a assimetria das distribuições dos atributos numéricos pode facilmente ser observada. Contudo, a título de verificação, optou-se por manter os modelos que

**Tabela 3.3** Ordem dos VDs considerada para a elaboração do metamodelo.

| Id | Metaconhecimento | Viés Sugerido | Interpretação das Generalizações |
|---|---|---|---|
| 7 | Os dados são aplicáveis a estratégias de Classificação e Regressão. | Escolher a estratégia mais eficiente. | A decisão por um dos grupos de estratégias definirá a aplicação de outros metaconhecimentos e, dessa forma, esse deve ser um dos primeiros VDs a ser considerado. **Consistência Alta** (2)**:** Os dados são realmente aplicáveis a essas 2 estratégias. |
| 2 | Os indivíduos podem ser classificados através de 4 ou de 2 classes. | Escolher o conjunto de classes que apresenta melhor desempenho. | A definição das classes que serão utilizadas pelos mecanismos de aprendizagem deve ser realizada antes das estratégias que dependem desse metaconhecimento. **Consistência alta** (2)**:** Segundo o especialista, essa abordagem é altamente pertinente ao domínio. |
| 1 | Uma quantidade reduzida de exemplos está disponível para treino. | Aplicar a técnica de reamostragem. *Bootstrap Aggregating* (BAGGING) e comparar seu desempenho ao desempenho obtido utilizando-se a base de treino original. | Esse VD verificará se o pequeno número de exemplos na amostra aumenta a variância dos modelos. Esta estratégia deve ser considerada antes de VDs baseados em refinamentos ou especulações. **Consistência Média** (1)**:** A amostra é relativamente pequena. |
| 6 | Alguns atributos do Jogo do Supermercado parecem ser redundantes. | Verificar o desempenho isolado de: Atributos de pontos, atributos de tempo, atributos das primeiras fases, atributos das últimas fases. Aplicar uma técnica de seleção de atributos relevantes. Comparar o desempenho de todas essas estratégias. | Este VD aponta para VRs fortes, e desta forma deve ser considerado por último. Contudo, deve ser considerado antes de VDs ainda mais fortes. **Consistência Baixa** (0,5)**:** São apenas especulações. |

consideram atributos numéricos no metamodelo, para que seus desempenhos possam ser comparados aos desempenhos dos modelos que consideram atributos discretos. O viés de número 2 é outro candidato a viés primário, já que, assumindo-se que as bases de treino são pequenas e que a classificação binária dos indivíduos – em TDAH e Não-TDAH – não diminuiria a importância do modelo, considerar previamente apenas modelos binário seria, nesse contexto, a melhor decisão a ser tomada. Contudo, mais uma vez, optou-se por avaliar o desempenho de ambas as estratégias e decidir qual seria a melhor. Ao final, decidiu-se não aplicar nenhum viés primário ao espaço de hipóteses do metamodelo.

Em seguida, foram selecionados os VAs. Como foi definido na Seção 3.1, o conjunto

de mecanismos de aprendizagem que será aplicado aos dados é um VA obrigatório. Assim, o viés de número 5 no conjunto de metadados está sendo considerado como VA. O viés de número 3, que poderia ter sido utilizado como viés primário, será também utilizado como VA por possuir muitas opções de intervalos de discretização alternativos que devem ser avaliadas. O viés de número 4 no conjunto de metadados foi aplicado no metamodelo como VA por sugestão do especialista, já que esta é uma hipótese de caráter investigativo que deveria ser avaliada em todas as etapas da aprendizagem.

Após a seleção dos VAs, os vieses remanescentes do conjunto de metadados foram definidos como VDs. Como foi sugerido na Seção 3.4, esses VDs estão sendo ordenados com base na estimativa de suas generalizações e, em caso de indecisão, com base na consistência de seus metaconhecimentos. A Tabela 3.3 mostra como os VDs foram ordenados através desses critérios.

Para ordenar os VDs, o peso das consistências não precisou ser utilizado, já que a interpretação das generalizações foi suficiente para realizar essa tarefa. Após a definição da ordem de aplicação dos vieses, o metamodelo – ou espaço de busca especializado – pode facilmente ser definido. A Figura 3.8 apresenta o modelo final, proposto através da abordagem SDVE, para identificação de indivíduos de acordo com o TDAH, utilizando os dados produzidos pelo Jogo do Supermercado.

No metamodelo apresentado na Figura 3.8, o VD 7 é o primeiro a ser chamado. Esse VD compara 2 VRs: "Classificação" e "Regressão". Como os VRs "2 Classes" e "4 Classes" não são compatíveis com o VR "Regressão", o VD 2 é utilizado apenas no VR "Classificação". Como o VR "4 Classes" representa a definição original das classes do TDAH, está sendo utilizado em comparação ao VR "Regressão". Interno ao VR "4 Classes" e ao VR "Regressão" está sendo considerado um VR específico, "Validação Cruzada", prevendo que futuramente os resultados obtidos serão comparados ao VR "BAGGING". Do lado do VR "Classificação", a comparação entre o VR "4 classes" e o VR "2 classes" é realizada no VD 2, que faz uso dos resultados já obtidos anteriormente – quando se comparou o VR "Classificação" com o VR "Regressão". Novamente, essa comparação é realizada dentro do VR "Validação Cruzada", pelo mesmo motivo discutido

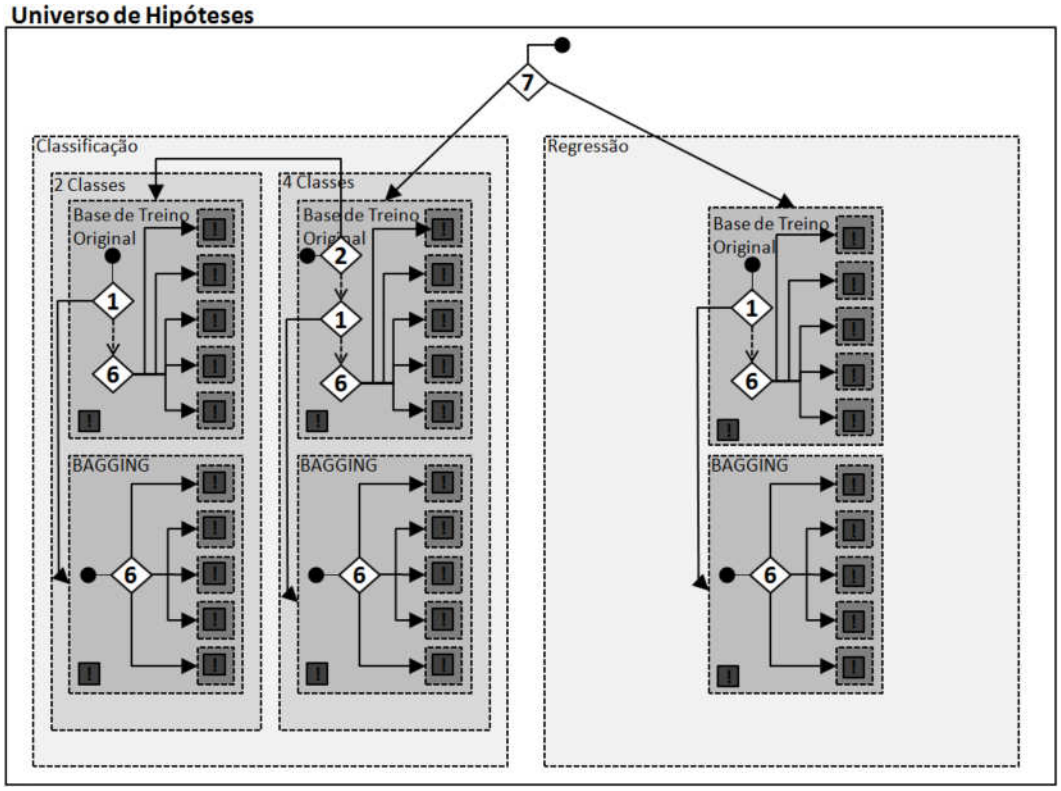



**Figura 3.8** Metamodelo final, utilizando a estratégia SDVE, para busca por um conjunto de modelos eficientes no problema de classificação de indivíduos de acordo com o TDAH, através dos dados produzidos pelo Jogo do Supermercado. A numeração dos VDs equivale à identificação dos mesmos na lista de prioridades.

anteriormente. Após decidir entre o VR “4 classes” e o VR “2 classes”, o processo caminha para o VD 1. A partir daqui, essa mesma sequência de processos é válida também para o VR “Regressão”. O VD 1 verifica se o VR “BAGGING”, que é externo, apresenta desempenho melhor do que o VR genérico atual, “Validação Cruzada”. Por fim, o VD 6 é chamado e testa 5 VRs alternativos.

# Capítulo 4

# Experimento

Neste capítulo, descreve-se o experimento que foi realizado para a comprovação da hipótese de pesquisa. Inicialmente, apresentam-se a caracterização, o enfoque, o objetivo e a motivação do experimento. Em seguida, apresentam-se os dados utilizados, as técnicas consideradas e as ferramentas desenvolvidas. Por fim, discute-se como foi realizado o desenvolvimento do experimento.

## 4.1 Objetivos do Experimento

O objetivo principal do experimento é a comprovação da existência de um conjunto de modelos preditivos eficientes que sejam capazes de relacionar os dados do Jogo do Supermercado ao diagnóstico do TDAH. Para tanto, está sendo aplicada uma estratégia de Meta-Aprendizagem que considera metaconhecimentos do domínio para a orientação da busca no espaço de modelos. Essa estratégia, que está sendo chamada SDVE, foi desenvolvida no escopo deste trabalho com o objetivo de contornar as dificuldades impostas pelo domínio considerado. Contudo, para que seja comprovada a eficiência dessa estratégia, outras 3 abordagens de Meta-Aprendizagem igualmente aplicáveis também estão sendo contempladas. Assim, além do objetivo principal, outro aspecto que também está sendo avaliado neste experimento é o desempenho da abordagem SDVE em identificar modelos preditivos eficientes, em um cenário em que os dados para treino são escassos.

## 4.2 Enfoque do Experimento

Como foi discutido na Seção 2.1, o TDAH manifesta-se de diferentes formas em adultos e crianças. Normalmente, os sintomas observados nos portadores do transtorno em fase adulta são sutilmente diferentes dos sintomas observados em crianças e adolescentes. Assim, um conjunto de modelos eficientes na tarefa de identificar adultos TDAH pode não ser adequado para a identificação de crianças TDAH.

Considerando o aspecto distinto dessas duas populações, e com o objetivo de se aumentar a relevância dos resultados obtidos, 2 focos independentes estão sendo contemplados ao mesmo tempo neste experimento:

- Análise da capacidade preditiva do Jogo do Supermercado em identificar portadores do TDAH em uma população de crianças / adolescentes, estudantes de escolas públicas;
- Análise da capacidade preditiva do Jogo do Supermercado em identificar portadores do TDAH em uma população de adultos universitários;

## 4.3 Caracterização do Experimento

Foram obtidas 2 amostras de dados, oriundas de 2 estudos de caso-controle, que descrevem os comportamentos de indivíduos dentro do Jogo do Supermercado, e a avaliação desses indivíduos de acordo com o TDAH. Cada amostra compreende um grupo diferente de indivíduos: (1) Crianças / Adolescentes, entre 10 e 17 anos, estudantes de escolas públicas; (2) Adultos, entre 21 e 27 anos, universitários. Essas amostras foram processadas através de algoritmos de aprendizagem, que consideraram os dados produzidos pelo jogo como atributos de entrada para a construção de uma generalização – ou modelo – para predição da situação dos indivíduos em relação ao TDAH. O experimento seguiu o padrão de um delineamento fatorial, com 7 fatores ou variáveis independentes, buscando a melhor relação de causalidade entre o jogo e o TDAH. Os fatores considerados foram: (1) Algoritmo de aprendizagem; (2) Técnica de discretização; (3) Método de validação; (4) Configuração de classes (2 classes / 4 classes); (5) Ponto de corte (4,5 ou

6 sintomas); (6) Configuração de atributos considerados; (7) Estratégia de aprendizagem – Classificação / Regressão. A escolha dos fatores foi orientada por 4 estratégias de Meta-Aprendizagem.

## 4.4 Terminologia e Delineamento do Experimento

De acordo com a terminologia experimental, os seguintes elementos foram considerados:

- **Fatores ou Variáveis Independentes (VI):** São os elementos ajustados no experimento. Estão sendo representados pelos 7 fatores apresentados anteriormente;
- **Tratamentos:** São os possíveis valores para os fatores. Estão sendo representados pelos possíveis valores que cada um dos 7 fatores poderá assumir;
- **Interações:** São as combinações de fatores. Estão sendo representadas pelos modelos produzidos a partir da combinação dos fatores considerados;
- **Variáveis Dependentes:** São os elementos utilizados para avaliar os tratamentos realizados. Compreendem as métricas de desempenho discutidas na Seção 4.10.

O experimento foi elaborado nos moldes de um Delineamento Fatorial Incompleto (DFI), sendo essa uma estratégia em que cada tratamento aplicado a uma VI depende dos tratamentos aplicados às demais VIs consideradas (BLACK, 2010). Essa estratégia foi considerada devido ao fato de que cada **modelo preditivo** (interação) é afetado pelas **alterações** (tratamentos) feitas em cada um dos 7 **conjuntos de vieses** (fatores ou VIs), sendo os resultados produzidos observados nas **métricas de desempenho** (variáveis dependentes). O caráter incompleto do delineamento deve-se ao fato de que não é possível avaliar todas as interações.

Para facilitar a legibilidade deste trabalho, no restante do texto, os termos experimentais – entre parênteses, no parágrafo anterior – serão evitados, em favor dos termos próprios das áreas de Mineração de Dados e KDD – em negrito no parágrafo anterior.

## 4.5 Amostras

As amostras utilizadas foram obtidas de 2 estudos de caso-controle, em que cada um considerou uma população diferente de indivíduos. Esses estudos foram realizados fora do escopo do trabalho aqui apresentado e, dessa forma, com o objetivo de concentrar o foco apenas no experimento, assume-se que esses dados foram corretamente produzidos. A descrição dessas amostras está sendo apresentada nas subseções a seguir.

### 4.5.1 Amostra de Crianças e Adolescentes (Base de Treino Bas)

A primeira amostra foi obtida de 3 escolas públicas do Estado do Rio de Janeiro, através de um estudo de caso-controle conduzido entre julho de 2009 e agosto de 2010. Sua composição contempla 98 escolares, que são crianças e adolescentes com idades entre 10 e 17 anos.

Essa amostra foi obtida da seguinte forma: 141 indivíduos – crianças e adolescentes – escolhidos aleatoriamente nas 3 escolas foram avaliados por seus professores através do questionário SNAP-IV (ver Seção 2.1), sob a supervisão externa de um psiquiatra. Desse total, 40 indivíduos que apresentaram 6 ou mais sintomas em desatenção ou hiperatividade / impulsividade[1] foram classificados no estudo de caso-controle como Casos – ou seja, indivíduos que apresentam o transtorno. Da mesma forma, 58 indivíduos que apresentaram simultaneamente menos de 6 sintomas em desatenção e hiperatividade / impulsividade foram classificados no estudo de caso-controle como Controles – ou seja, indivíduos que não apresentam o transtorno. Esses indivíduos foram inicialmente submetidos a uma sessão de jogos, com o objetivo de avaliar suas habilidades com computadores[2]. Como todos demonstraram aptidão com ambientes computacionais, foram, então, submetidos a outra sessão de jogos, dessa vez com o Jogo do Supermercado. Os dados produzidos pelo jogo, e os dados obtidos do questionário SNAP-IV preenchido pelos professores foram, então, utilizados para a composição dessa amostra. Essa amostra está

[1] Foram considerados sintomas com presença significante, avaliados como muito ou demais no SNAP-IV (ver Seção 2.1).

[2] Essa questão surgiu já que inabilidades com ambientes computacionais poderiam ser interpretadas como déficits neuropsicológicos pelos mecanismos de aprendizagem, produzindo resultados incorretos.

sendo identificada no experimento como base de treino Bas.

Para a realização deste estudo, todas as permissões necessárias foram obtidas dos órgãos competentes, incluindo comitê de ética e Secretaria de Educação Estadual. Outros detalhes sobre este estudo podem ser vistos em SANTOS *et al.* (2011b).

### 4.5.2 Amostra de Adultos (Base de Treino Mat)

A segunda amostra foi obtida de uma universidade pública, através de um estudo de caso-controle conduzido entre julho de 2009 e julho de 2010. Sua composição contempla 50 indivíduos adultos, todos estudantes de medicina, com idades entre 21 e 27 anos.

Essa amostra foi obtida da seguinte forma: 300 indivíduos universitários, escolhidos aleatoriamente, foram submetidos a uma auto-avaliação através do questionário ASRS (ver Seção 2.1). Deste grupo, 17 indivíduos classificados como positivos pelo ASRS foram também diagnosticados como casos positivos do TDAH por uma equipe de psiquiatras. Esses indivíduos foram classificados no estudo de caso-controle como Casos. No mesmo grupo, 33 indivíduos que não apresentaram quantidade suficiente de sintomas no relatório ASRS e foram também diagnosticados pela equipe de psiquiatras como casos negativos do TDAH foram classificados no estudo de caso-controle como Controles. Todos os 50 indivíduos foram submetidos a uma sessão com o Jogo do Supermercado. Os dados produzidos pelo jogo e os dados obtidos do diagnóstico dos psiquiatras foram, então, utilizados para composição dessa amostra.

Essa amostra está sendo identificada no experimento como base de treino Mat.

## 4.6 Descrição dos Dados

Os dados fornecidos pelas 2 amostras descritas na Seção 4.5 apresentam os seguintes atributos:

1. Identificação: A identificação do indivíduo;
2. Idade: A idade do indivíduo;
3. Sexo: O gênero do indivíduo;

4. Pontos: 18 atributos de pontos – um para cada fase;

5. Tempo: 18 atributos de tempo – um para cada fase;

6. Número de Sintomas de Desatenção: Número de sintomas de desatenção constatados;

7. Número de Sintomas de Hiperatividade/Impulsividade: Número de sintomas de hiperatividade/impulsividade constatados.

Cada um dos 18 atributos de pontos e tempo refere-se a uma diferente fase do jogo. Os atributos de tempo foram convertidos para valores inteiros, de forma que foram expressos em segundos. Os atributos Número de Sintomas de Desatenção e Número de Sintomas de Hiperatividade/Impulsividade foram fornecidos pelo avaliador – que, no caso da base de treino Bas, é um professor e, no caso da base de treino Mat, é um psiquiatra. Os atributos Idade e Sexo foram fornecidos pelo supervisor da sessão de jogos. O atributo Identificação não foi considerado para o treinamento dos modelos e, dessa forma, estão sendo utilizados apenas 40 atributos, sendo 38 atributos preditores e 2 atributos classe numéricos – Número de Sintomas de Desatenção e Número de Sintomas de Hiperatividade/Impulsividade.

## 4.7 Ferramentas de Software Utilizadas

Os algoritmos para os mecanismos de base utilizados foram obtidos de pacotes de implementações de uso livre. Inicialmente, 3 alternativas foram consideradas: Weka[3], Orange[4] e RapidMiner[5].

- **Weka:** Essa é uma coleção de algoritmos de Aprendizagem de Máquina para tarefas de Mineração de Dados. Além dos algoritmos, é disponibilizado também um *framework* para execução de teste e experimentos simples. O Weka é um software livre distribuído sob licença GNU (General Public License) (HALL *et al.*, 2009).

[3]http://www.cs.waikato.ac.nz/ml/weka/
[4]http://orange.biolab.si/
[5]http://rapid-i.com/

- **Orange:** Essa é uma coleção de algoritmos de Aprendizagem de Máquina, implementados na linguagem C++, e gerenciados por *scripts* na linguagem Python. Além disso, é disponibilizado um *framework* para elaboração de experimentos simples. O Orange também é um software livre distribuído sob licença GNU (DEMŠAR *et al.*, 2004).

- **RapidMiner:** Esse é um sistema integrado de mineração de dados com foco em *workflows*. O RapidMiner possui uma versão livre distribuída sob licença GNU (RAPIDMINER, 2009).

Após algumas investigações, e com o objetivo de tornar a implementação do experimento uniforme, optou-se por utilizar apenas um pacote de algoritmos. A alternativa escolhida foi o Weka, por ter demonstrado agilidade na elaboração de hipóteses experimentais. O RapidMiner também foi utilizado, porém, apenas para a elaboração da estratégia de Mapeamento em Metanível – discutida na Seção 4.11. O plano de aprendizagem gerado pelo RapidMiner foi processado no Weka, com o objetivo de utilizar o mesmo método de validação considerada pelas demais estratégias do experimento.

Além do pacote de algoritmos, a ferramenta BrOffice Calc[6] foi utilizada nas etapas de limpeza e pós-processamento dos dados, e a ferramenta NetBeans[7] foi utilizada no desenvolvimento de aplicativos auxiliares.

## 4.8 Mecanismos de Aprendizagem de Base

Os mecanismos de aprendizagem de base utilizados compreendem aqueles disponíveis no Weka, que foram adequados aos dados das bases de treino e apresentaram um consumo de tempo máximo aceitável para realização do experimento, considerando os recursos computacionais disponíveis. O critério de tempo utilizado para escolha dos algoritmos foi a capacidade de se produzir um modelo preditivo em menos de $60$ segundos para as $2$ bases de treino. A Tabela 4.1 apresenta todas as $48$ implementações para mecanismos de aprendizagem de base que foram utilizados no experimento. Os algoritmos

[6] http://www.broffice.org/
[7] http://netbeans.org/

que não apresentam autoria são implementações próprias do Weka.

**Tabela 4.1** Algoritmos de aprendizagem de base utilizados no experimento – Implementações do Weka.

| # | Identificação | E* | A** | Descrição |
|---|---|---|---|---|
| **Algoritmos Bayesianos** | | | | |
| 1 | NaiveBayes | C | CN | Implementa um algoritmo Naive Bayes. |
| 2 | AODE | C | C | Implementa um algoritmo que considera a média de classificadores alternativos Naive Bayes simples, cuja abordagem de independência não é tão forte quanto em um Naive Bayes real (WEBB *et al.*, 2005). |
| 3 | AODEsr | C | C | Implementa uma versão aprimorada do AODE, que identifica e elimina especializações entre atributos. (ZHENG e WEBB, 2006). |
| 4 | BayesNet | C | CN | Implementa um algoritmo para construção de uma Rede Bayesiana utilizando vários algoritmos e métricas de qualidade. |
| 5 | HNB | C | C | Implementa um algoritmo para um classificador *Hidden Naive Bayes.* (ZHANG *et al.*, 2005). |
| 6 | NaiveBayesSimple | C | C | Implementa um algoritmo para um algoritmo Naive Bayes simples. |
| 7 | NaiveBayesUpdateable | C | CN | Implementa uma versão atualizável do Naive Bayes. (JOHN e LANGLEY, 1995). |
| 8 | WAODE | C | C | Implementa um algoritmo chamado Estimador de Uma-Dependência com Média Ponderada (ou *Weightily Averaged One-Dependence Estimators*) (JIANG e ZHANG, 2006). |
| **Funções** | | | | |
| 9 | RBFNetwork | CR | CN | Implementa uma rede radial gaussiana. |
| 10 | SMO | CR | CN | Implementa uma algoritmo de otimização mínima sequencial para treinamento de uma Máquina de Vetores de Suporte. (PLATT, 1999). |
| 11 | Logistic | C | N | Implementa um algoritmo de regressão logística multinominal (LE CESSIE e VAN HOUWELINGEN, 1992). |
| 12 | SimpleLogistic | C | N | Implementa um algoritmo para construção de modelos de regressão logística (LANDWEHR *et al.*, 2005). |
| 13 | GaussianProcesses | R | CN | Implementa um algoritmo de regressão sem ajuste de hiperparâmetros (MACKAY, 1998). |
| 14 | LeastMedSq | R | CN | Implementa um algoritmo de regressão linear através da mediana quadrada mínima utilizando os algoritmos existentes no Weka para regressão (ROUSSEEUW *et al.*, 1987). |
| 15 | LinearRegression | R | CN | Implementa um algoritmo de regressão linear. |
| 16 | MultiLayerPerceptron | R | CN | Implementa um algoritmo *backpropagation*. |
| 17 | SMOreg | R | CN | Implementa um algoritmo de Máquina de Vetores de Suporte para regressão (SHEVADE *et al.*, 2000). |

Continua na próxima página...

**Tabela 4.1** – Continuação

| # | Identificação | E* | A** | Descrição |
|---|---|---|---|---|
| 18 | IsotonicRegression | R | N | Implementa um algoritmo para aprendizagem de um modelo isotônico. |
| 19 | PLSClassifier | R | N | Implementa um algoritmo de regressão através de filtros. |
| 20 | SimpleLinearRegression | R | N | Implementa um algoritmo para regressão linear simples. |
| | **Baseados em Instâncias** | | | |
| 21 | IB1 | C | CN | Implementa um classificador baseado nos vizinhos mais próximos. Utiliza uma distância euclidiana normalizada para encontrar a instância mais próxima à instância testada (AHA *et al.*, 1991). |
| 22 | IBK | CR | CN | Implementa um algoritmo baseado nos k-Vizinhos-Mais-Próximos (AHA *et al.*, 1991). |
| 23 | KStar | CR | CN | Implementa um algoritmo baseado nos k-Vizinhos-Mais-Próximos (AHA *et al.*, 1991). |
| 24 | LBR | C | C | Implementa um algoritmo Naive Bayes preguiçoso (ZHENG e WEBB, 2000). |
| 25 | LWL | CR | CN | Implementa um algoritmo de aprendizagem através de ponderação local (FRANK *et al.*, 2003). |
| | **Miscelânea** | | | |
| 26 | HyperPipes | CR | CN | Implementa um algoritmo de classificação através de *HyperPipes*. |
| 27 | VFI | C | CN | Implementa um algoritmo de aprendizagem que funciona através da votação de intervalos de características (DEMIRÖZ e GÜVENIR, 1997). |
| | **Regras** | | | |
| 28 | ConjuntiveRule | CR | CN | Implementa um algoritmo simples para aprendizagem de regras conjuntivas. |
| 29 | DecisionTable | CR | CN | Implementa um algoritmo de Tabela de Decisão (KOHAVI, 1995). |
| 30 | JRip | C | CN | Implementa o algoritmo de regra proposicional *Repeated Incremental Pruning to Produce Error Reduction* (RIPPER) (COHEN, 1995). |
| 31 | NNGE | C | CN | Implementa um algoritmo similar aos algoritmos de Vizinhos-Mais-Próximos, utilizando exemplares generalizados não aninhados (ROY, 2002). |
| 32 | OneR | C | CN | Implementa um algoritmo de classificação através de apenas 1 regra. (HOLTE, 1993). |
| 33 | PART | C | CN | Implementa um algoritmo que constrói uma Árvore de Decisão parcial em cada iteração e utiliza a melhor folha na regra (FRANK *et al.*, 1998). |

Continua na próxima página...

**Tabela 4.1** – Continuação

| # | Identificação | E* | A** | Descrição |
|---|---|---|---|---|
| 34 | Prism | C | C | Implementa um algoritmo para classificação que considera apenas atributos categóricos (CENDROWSKA, 1987). |
| 35 | Ridor | C | CN | Implementa um algoritmo baseado em uma técnica de aprendizagem conhecida como Ripple-Down (KANG *et al.*, 1995). |
| 36 | ZeroR | C | CN | Implementa um algoritmo que faz predições sem utilizar regras, considerando apenas a média (para classes numéricas) ou a moda (para classes nominais). |
| 37 | M5R | R | CN | Implementa um algoritmo que generaliza listas de decisão para problemas de regressão, utilizando a estratégia dividir-para-conquistar (HOLMES *et al.*, 1999). |
| **Árvores de Decisão** | | | | |
| 38 | BFTree | C | CN | Implementa um algoritmo de Árvore de Decisão baseado na técnica primeiro-melhor (ou *best-first*) (SHI, 2007) |
| 39 | DecisionStump | CR | CN | Implementa um algoritmo para construção de modelos baseados em Árvores de Decisão. |
| 40 | FT | C | CN | Implementa um algoritmo para Árvores de Decisão "funcionais", que são árvores que podem possuir funções de regressão em seus nós ou em suas folhas. (GAMA, 2004). |
| 41 | Id3 | C | C | Implementa um algoritmo de Árvores de Decisão não podada (QUINLAN, 1986). |
| 42 | J48 | C | CN | Implementa o algoritmo C4.5 para Árvores de Decisão (QUINLAN, 1993). |
| 43 | J48graf | C | CN | Implementa um algoritmo C4.5 para Árvores de Decisão enxertadas (WEBB, 1999). |
| 44 | LADTree | C | CN | Implementa um algoritmo que gera uma Árvore de Decisão alternativa multi-classe utilizando a estratégia *LogitBoost* (HOLMES *et al.*, 2002). |
| 45 | RandomForest | C | CN | Implementa um algoritmo para construção de floresta de Árvores de Decisão aleatórias (BREIMAN, 2001). |
| 46 | RandomTree | C | CN | Implementa um algoritmo que constrói uma Árvore de Decisão que considera k atributos aleatórios em cada nó. |
| 47 | REPTree | CR | CN | Implementa um algoritmo de Árvore de Decisão rápida. |
| 48 | M5P | R | CN | Implementa o algoritmo M5P para Árvores de Decisão (QUINLAN, 1992). |

* Estratégia de aprendizagem: **C** = Classificação; **R** = Regressão.

** Tipo de atributo aplicável: **C** = Categórico; **N** = Numérico.

## 4.9 Estratégias de Discretização

Para transformar os atributos de intervalos contínuos em atributos categóricos, foram utilizadas as seguintes estratégias de discretização:

- **Discretização em Intervalos de Igual Tamanho:** Nessa estratégia, os atributos numéricos foram discretizados em intervalos de igual tamanho, através da implementação *discretize* do Weka;
- **Discretização em Intervalos com a Mesma Frequências de Instâncias:** Nessa estratégia, os atributos numéricos foram discretizados em intervalos que possuem aproximadamente o mesmo número de instâncias. Também foi utilizada a implementação *discretize* do Weka;
- **Discretização Otimizada em Intervalos de Igual Tamanho:** Nessa estratégia, os atributos numéricos foram discretizados em intervalos de igual tamanho, porém, o número de intervalos foi otimizado, ou seja, utilizou-se de 2 ao número máximo de intervalos permitidos. Também foi utilizada a implementação *discretize* do Weka;
- **Discretização Proporcional em Intervalos com a Mesma Frequências de Instâncias:** Nessa estratégia, os atributos numéricos foram discretizados em intervalos com o mesmo número de instâncias equivalentes, sendo o número de intervalos igual à raiz quadrada do número total de instâncias – ou valores não faltosos. O algoritmo utilizado foi elaborado por YANG e WEBB (2001).

Excetuando-se a proposta de intervalos proporcionais, foram utilizadas 9 configurações para cada estratégia, que consideraram de 2 a 10 intervalos categóricos.

## 4.10 Métricas

As seguintes métricas estão sendo consideradas no experimento:

- **Área abaixo da Curva ROC (AUC):** Essa é uma métrica de discriminação que está sendo utilizada para avaliar modelos em problemas de classificação. Busca-se

a maximização de seus valores.

- **Erro Absoluto Relativo (EAR):** Essa é uma métrica de calibragem que está sendo utilizada para avaliar modelos em problemas de regressão. Busca-se a minimização de seus valores.

- **Tempo de Execução (TE):** Além das métricas AUC e EAR a proposta SDVE considerou também o tempo de execução dos VRs. VRs que levaram mais de $24$ horas de processamento foram considerados, mas seus caminhos não foram mais expandidos. O objetivo da adoção dessa métrica foi possibilitar a realização do experimento no tempo disponível.

Como, de acordo com algumas pesquisas, a métrica AUC começa a apresentar resultados relevantes a partir de $0{,}70$ (ZHU *et al.*, 2010; FISCHER *et al.*, 2003), esse ponto de corte está sendo adotado para confirmação da eficiência dos modelos.

A métrica EAR não possui uma interpretação consensual para definição de um limite comum a todos os domínios. Contudo, após alguns debates, decidiu-se considerar $0{,}50$ em EAR como ponto de corte, já que nesse limite os modelos começam a cometer menos da metade dos erros de um preditor simples – que, nesse caso, é a média.

## 4.11 Estratégias de Meta-Aprendizagem

Para que o espaço de modelos pudesse ser explorado, foram aplicadas $4$ estratégias de Meta-Aprendizagem. Uma dessas estratégias, a proposta SDVE (ver Capítulo 3, pág. 73), foi elaborada no escopo deste trabalho , e contempla a hipótese de pesquisa. As demais estratégias estão sendo utilizadas como ponto de referência para análise do desempenho da proposta SDVE. Essas $4$ estratégias estão sendo discutidas nas subseções a seguir.

### 4.11.1 Seleção Aleatória

Essa é uma estratégia ingênua que aplica diferentes técnicas aleatoriamente com o objetivo de identificar um modelo eficiente. Normalmente, são aplicados mecanismos de base escolhidos ao acaso sem nenhuma justificativa de aplicação. Apenas domínios com

um grande número de modelos eficientes conseguem produzir bons resultados – o que faz dessa uma estratégia inadequada para domínios em que modelos eficientes são raros. Conceitualmente, essa não seria exatamente uma estratégia de Meta-Aprendizagem – já que nenhum metaconhecimento é utilizado. Contudo, com o objetivo de comparar diferentes abordagens, essa estratégia foi aplicada. Foram considerados 3 algoritmos de classificação e 2 algoritmos capazes de lidar com classificação e regressão. As seguintes implementações apresentadas na Tabela 4.1 foram utilizadas: NaiveBayes, SMO, IBK, ConjuntiveRule, J48. As classes utilizadas na estratégia de classificação foram definidas a partir da interpretação sugerida pelo DSM-IV – ou seja, 4 rótulos (TDAH-H, TDAH-C, TDAH-I, Não-TDAH) produzidos a partir da observação de 6 ou mais sintomas em desatenção e / ou hiperatividade / impulsividade (ver Seção 2.1). A estratégia de regressão foi aplicada em 2 avaliações: (1) Aproximação do número de sintomas de desatenção; e (2) Aproximação do número de sintomas de hiperatividade / impulsividade. Os atributos numéricos não foram discretizados nessa estratégia. Os resultados obtidos foram avaliados através das métricas EAR e AUC.

### 4.11.2 Meta-Aprendizagem através de Mecanismos de Base

Como foi discutido na Seção 2.3.9, essa é uma estratégia fundamentada na combinação de algoritmos de base. Para essa estratégia, foram consideradas as seguintes técnicas:

- **Empilhamento:** É uma técnica que utiliza uma série de algoritmos que se auxiliam através da transmissão de suas predições. Para essa técnica, foi utilizada a implementação do Weka StackingC (SEEWALD, 2002).

- **Votação:** É uma técnica baseada na votação de um conjunto de algoritmos para inferir uma predição. Para essa técnica, foi utilizada a implementação do Weka Vote (KUNCHEVA, 2004)

- **Boosting:** É uma técnica baseada na revisão iterativa. Para essa técnica, foi utilizada a implementação do Weka ADABoostM1 (FREUND e SCHAPIRE, 1996).

Como essas são técnicas normalmente aplicadas à tarefa de classificação, algoritmos de regressão foram desconsiderados. Todos os demais algoritmos da Tabela 4.1, capazes de lidar com classificação, foram utilizados. A técnica *Boosting* foi aplicada utilizando apenas a implementação DecisionStump. Da mesma forma que na Seleção Aleatória, a classificação considerou $4$ valores de classe, obtidos a partir do corte em $6$ sintomas, além dos atributos numéricos em seu formato natural. Essa estratégia foi avaliada através da métrica AUC.

### 4.11.3 Mapeamento em Metanível

Como foi discutido na Seção 2.3.9, essa é uma estratégia que busca armazenar e adquirir metaconhecimentos de repositórios de metadados para realização da Meta-Aprendizagem. Para a implementação dessa estratégia no experimento, foi utilizada a ferramenta PaREn (SHAFAIT *et al.*, 2010), que pode ser acoplada como *plugin* no software RapidMiner para construção automática de um plano de aprendizagem utilizando metadados de outras bases. Foram utilizados metadados de $90$ bases de dados do repositório UCI [8]. Devido a limitações da ferramenta PaREn, essa estratégia foi aplicada apenas a problemas de classificação, utilizando-se a métrica AUC para avaliação dos resultados.

### 4.11.4 Proposta SDVE

Foi considerada também a aplicação da abordagem SDVE, introduzida no Capítulo 3. Ao contrário do Mapeamento em Metanível, essa estratégia avalia metaconhecimentos explicitamente observados ou sugeridos para o domínio. Foram avaliados $7$ metaconhecimentos, que podem ser vistos como fatores de um experimento baseado em Delineamento Fatorial. Esses metaconhecimentos foram ou observados diretamente no domínio, ou obtidos do especialista que acompanhou este trabalho. Essa estratégia foi avaliada através das métricas EAR, AUC e TE.

[8] http://archive.ics.uci.edu/ml/index.html

## 4.12 Métodos de Validação dos Resultados

Foram utilizados 2 métodos para validação dos modelos identificados:

- **Validação Cruzada:** Esse foi o método utilizado na maior parte do tempo para validação dos modelos. Para cada modelo, foram produzidos 10 arranjos aleatórios da base de treino; e para cada arranjo foi realizada uma validação cruzada com 10 *folds*, totalizando 100 submodelos testados. O desempenho médio dos 100 submodelos testados foi, então, atribuído ao modelo final, que foi produzido a partir de todos os dados da base treino.

- **BAGGING:** Esse método foi utilizado dentro da proposta SDVE com o objetivo de diminuir variância dos modelos. O BAGGING foi aninhado dentro de um processo de validação cruzada – como explicado anteriormente – e considerou 20 amostras *bootstrap* para cada um dos 100 submodelos testados. O desempenho de cada um dos 100 submodelos foi calculado então através da votação dos 20 subsubmodelos. Da mesma forma que na validação cruzada, o desempenho médio dos 100 submodelos testados foi, então, atribuído ao modelo final. Como a proposta BAGGING muda não apenas o método de validação, mas o modelo em si, o objetivo dessa abordagem foi comparar o desempenho entre um modelo construído a partir da base original e um modelo elaborado a partir de amostras *bootstrap* da mesma base original.

Todos os modelos utilizaram os mesmos arranjos aleatórios e a mesma sequência de 10 *folds* para a validação cruzada. A validação através de BAGGING também considerou as mesmas 20 amostras bootstrap para cada um dos 100 submodelos dos modelos testados.

## 4.13 Estrutura do Mecanismo Utilizado para Aplicação da proposta SDVE

A proposta SDVE foi implementada a partir de uma série de estruturas de processamento. A estrutura do espaço de busca foi implementada através de um script, que ao ser executado, busca dinamicamente os VRs dos VDs através dos resultados produzidos por

suas dimensões-micro.

As dimensões-micro foram definidas em XML e foram interpretadas através do módulo de experimentação do Weka. Três dimensões-micro padrão foram criadas, cada uma considerando um conjunto de algoritmos adequados a um tipo de tarefa de aprendizagem: (1) Regressão com atributos numéricos, (2) classificação com atributos numéricos, (3) classificação com atributos categóricos.

Foi desenvolvido, ainda, um módulo de discretização próprio – na linguagem Java –, que é executado antes de qualquer tarefa de aprendizagem que considere atributos categóricos. Os resultados do processamento são armazenados em arquivos CSV (*Comma-separated values*) e são interpretados por planilhas do BrOffice Calc, que fazem o refinamento dos dados e produzem informações sintetizadas sobre o experimento.

## 4.14 Método

Nesta seção, estão sendo discutidas as 4 etapas que foram executadas para a realização do experimento.

### 4.14.1 Etapa de Aquisição de Dados e Metaconhecimentos

Nesta etapa, foi realizada uma pesquisa para identificação de bases de treino do domínio que apresentassem dados de qualidade. Foram selecionadas 2 bases de treino, as quais estão sendo discutidas na Seção 4.5.

Nesta etapa também foram elicitados metaconhecimentos do domínio através do especialista que participou deste trabalho. Esses metaconhecimentos foram utilizados para a proposta SDVE, apresentada no Capítulo 3.

### 4.14.2 Etapa de Análise dos Dados

Nesta etapa, os dados da base Bas e da base Mat foram analisados e pré-processados manualmente através do software BrOffice Calc. Inicialmente, foram desconsideradas as instâncias com valores faltosos ou discrepantes para algum atributo. Os atributos de tempo gasto nas fases do jogo foram convertidos para o tipo inteiro, sendo esses expressos em

segundos. Para cada base de treino, foram criados 6 atributos classe, calculados a partir dos atributos "Quantidade de Sintomas de Desatenção" e "Quantidade de Sintomas de Hiperatividade/Impulsividade" da seguinte forma:

- Considerando 4 valores de classe (TDAH-H, TDAH-C, TDAH-I, Não-TDAH) e 6 ou mais sintomas de desatenção e / ou hiperatividade / impulsividade;
- Considerando 4 valores de classe (TDAH-H, TDAH-C, TDAH-I, Não-TDAH) e 5 ou mais sintomas de desatenção e / ou hiperatividade / impulsividade;
- Considerando 4 valores de classe (TDAH-H, TDAH-C, TDAH-I, Não-TDAH) e 4 ou mais sintomas de desatenção e / ou hiperatividade / impulsividade;
- Considerando 2 valores de classe (TDAH, Não-TDAH) e 6 ou mais sintomas de desatenção e / ou hiperatividade / impulsividade;
- Considerando 2 valores de classe (TDAH, Não-TDAH) e 5 ou mais sintomas de desatenção e / ou hiperatividade / impulsividade;
- Considerando 2 valores de classe (TDAH, Não-TDAH) e 4 ou mais sintomas de desatenção e / ou hiperatividade / impulsividade.

Para as tarefas de classificação, apenas um dos atributos classe foi utilizado por vez. As planilhas foram convertidas para o formato CSV, para que pudessem ser utilizadas pelos algoritmos implementados no Weka.

### 4.14.3 Etapa de Processamento dos Dados

Nesta etapa, os dados foram efetivamente processados, através das 4 estratégias de Meta-Aprendizagem apresentadas na Seção 4.11. Essas estratégias foram aplicadas na seguinte ordem:

1. Seleção Aleatória;
2. Meta-Aprendizagem através de Mecanismos de Base;
3. Mapeamento em Metanível;

4. Proposta SDVE.

O módulo *Experimenter* do Weka foi utilizado para o processamento de algoritmos nas estratégias: Seleção Aleatória; Meta-Aprendizagem através de Mecanismos de Base; e Mapeamento em Metanível. Já a proposta SDVE, apesar de ter considerado também os algoritmos implementados no Weka, gerenciou seus processos externamente, através de um script de orquestração desenvolvido na própria proposta.

### 4.14.4 Etapa de Pós-Processamento e Análise dos Resultados

Nesta etapa, os resultados produzidos através das estratégias de Meta-Aprendizagem foram pós-processados para a produção de gráficos e *rankings* que pudessem facilitar a visualização das conclusões. A ferramenta utilizada nessa etapa foi o BrOffice Calc. Foram produzidos resumos para estratégias aplicadas, algoritmos e técnicas consideradas. Foram, também, calculadas outras métricas não consideradas pelos mecanismos de aprendizagem durante o processo – mas que foram necessárias para as conclusões –, além de intervalos de confiança para todas as métricas. Foram produzidos rankings para modelos, algoritmos e estratégias de discretização mais eficientes.

# Capítulo 5

# Resultados Obtidos

Este capítulo apresenta os resultados obtidos a partir do experimento que foi realizado. Estão sendo apresentados os resultados produzidos para as $4$ estratégias de Meta-Aprendizagem, além de uma etapa de reavaliação de algumas propostas. Ao final deste capítulo, estão sendo discutidos os desempenhos obtidos em cada estratégia.

## 5.1 Seleção Aleatória de Técnicas de Aprendizagem

Os resultados apresentados nesta seção foram obtidos a partir da estratégia de Seleção Aleatória de Técnicas de Aprendizagem. Como foi discutido na Seção 4.11.1 (pag. 106), foram utilizados $5$ algoritmos de aprendizagem de base aplicados a classificação e regressão. Os algoritmos de classificação consideraram $4$ valores de classe, obtidos a partir do corte em $6$ sintomas. Ambas as estratégias de aprendizagem aplicaram os atributos numéricos em seu formato natural. A Tabela 5.1(a) apresenta os resultados obtidos através da métrica EAR e a Tabela 5.1(b) apresenta os resultados obtidos através da métrica AUC.

Essa estratégia de Meta-Aprendizagem não produziu resultados satisfatórios. Nenhum dos modelos conseguiu atingir os pontos de corte definidos para comprovação da eficiência. O melhor desempenho em EAR ($0{,}74$) foi obtido utilizando-se a implementação IBK para a base de treino Mat. A mesma implementação também apresentou o melhor desempenho em AUC ($0{,}57$).

**Tabela 5.1** Desempenhos obtidos através da estratégia de Seleção Aleatória de Técnicas de Aprendizagem, considerando 4 classes nas estratégias de classificação.

(a) Desempenho em EAR.

| Base | Estr. de Aprendizagem | Atr. Num. | NaiveBayes | SMO | IBK | Conjuntive | J48 |
|---|---|---|---|---|---|---|---|
| **Bas** | Classificação (4 Classes) | Não Discr. | 0,87 | 1,06 | 0,86 | 0,94 | 0,99 |
| **Bas** | Regressão (S. Desatenção) | Não Discr. | – | – | 1,12 | 1,03 | – |
| **Bas** | Regressão (S. Hiper./Imp.) | Não Discr. | – | – | 1,03 | 0,96 | – |
| **Mat** | Classificação (4 Classes) | Não Discr. | 1,15 | 1,16 | 0,74 | 0,96 | 1,03 |
| **Mat** | Regressão (S. Desatenção) | Não Discr. | – | – | 1,29 | 1,09 | – |
| **Mat** | Regressão (S. Hiper./Imp.) | Não Discr. | – | – | 1,17 | 1,01 | – |

(b) Desempenho em AUC.

| Base | Estr. de Aprendizagem | Atr. Num. | NaiveBayes | SMO | IBK | Conjuntive | J48 |
|---|---|---|---|---|---|---|---|
| **Bas** | Classificação (4 Classes) | Não Discr. | 0,54 | 0,50 | 0,54 | 0,55 | 0,52 |
| **Bas** | Regressão (S. Desatenção) | Não Discr. | – | – | – | – | – |
| **Bas** | Regressão (S. Hiper./Imp.) | Não Discr. | – | – | – | – | – |
| **Mat** | Classificação (4 Classes) | Não Discr. | 0,40 | 0,43 | 0,57 | 0,46 | 0,53 |
| **Mat** | Regressão (S. Desatenção) | Não Discr. | – | – | – | – | – |
| **Mat** | Regressão (S. Hiper./Imp.) | Não Discr. | – | – | – | – | – |

## 5.2 Meta-Aprendizagem Através de Mecanismos de Base

Os resultados apresentados nesta seção foram obtidos a partir da estratégia Meta-Aprendizagem Através de Mecanismos de Base. Como foi discutido na Seção 4.11.2 (pág. 107), essa estratégia utilizou 3 técnicas de Meta-Aprendizagem aplicadas à classificação, que consideraram 4 valores de classe, obtidos a partir do corte em 6 sintomas. Além disso, os atributos numéricos foram aplicados em seu formato natural. A Tabela 5.2(a) apresenta os resultados obtidos através da métrica EAR e a Tabela 5.2(b) apresenta os resultados obtidos através da métrica AUC.

Essa estratégia também não apresentou resultados satisfatórios. Da mesma forma que a Seleção Aleatória de Técnicas de Aprendizagem, nenhum dos modelos conseguiu atingir os pontos de corte definidos para comprovação da eficiência. O melhor desempenho em EAR (0,95) foi obtido utilizando-se a implementação Vote para a base de treino Bas. O melhor desempenho em AUC (0,59) foi obtido utilizando-se a implementação Boosting, também para a base de treino Bas.

**Tabela 5.2** Desempenhos obtidos com a estratégia Meta-Aprendizagem Através de Mecanismos de Base, considerando 4 classes.

(a) Desempenho em EAR.

| Base | Estr. de Aprendizagem | Atr. Num. | StackingC | Vote | Boosting |
|---|---|---|---|---|---|
| **Bas** | Classificação (4 Classes) | Não Discr. | 0,98 | 0,95 | 1,1 |
| **Mat** | Classificação (4 Classes) | Não Discr. | 0,96 | 0,99 | 1,02 |

(b) Desempenho em AUC.

| Base | Estr. de Aprendizagem | Atr. Num. | StackingC | Vote | Boosting |
|---|---|---|---|---|---|
| **Bas** | Classificação (4 Classes) | Não Discr. | 0,55 | 0,57 | 0,59 |
| **Mat** | Classificação (4 Classes) | Não Discr. | 0,52 | 0,42 | 0,37 |

## 5.3 Mapeamento em Metanível

Nesta seção, estão sendo apresentados os resultados obtidos a partir da estratégia de Mapeamento em Metanível. A Tabela 5.3 apresenta, de forma conjunta, os resultados obtidos através das métricas EAR e AUC para as bases Bas e Mat.

**Tabela 5.3** Desempenhos obtidos através das métricas EAR e AUC, utilizando a estratégia Mapeamento em Metanível, considerando 4 classes.

| | | | Base Bas | | Base Mat | |
|---|---|---|---|---|---|---|
| **Algoritmos (Impl.)** | **Estr. de Aprendizagem** | **Atr. Num.** | **EAR** | **AUC** | **EAR** | **AUC** |
| **MultiLayerPerceptron** | Classificação (4 Classes) | Não Discr. | 0,86 | 0,57 | 1,12 | 0,37 |
| **RandomForest** | Classificação (4 Classes) | Não Discr. | 1,00 | 0,49 | 0,98 | 0,52 |
| **IBK** | Classificação (4 Classes) | Não Discr. | 0,86 | 0,53 | 0,73 | 0,57 |
| **NaiveBayes** | Classificação (4 Classes) | Não Discr. | 0,87 | 0,54 | 1,15 | 0,40 |
| **OneR** | Classificação (4 Classes) | Não Discr. | 0,98 | 0,49 | 0,75 | 0,47 |

Em um primeiro momento, o módulo de busca do algoritmo PaRen foi aplicado às bases de treino, gerando um *ranking* de algoritmos promissores através das experiências adquiridas nas 90 bases da UCI consideradas. Em seguida, os parâmetros dos algoritmos selecionados foram aprimorados, e um plano de aprendizagem foi gerado considerando cada um dos algoritmos. Ao final, produziu-se exatamente o mesmo plano de aprendizagem para ambas as bases. Basicamente, os planos sugeriram uma técnica de verificação de valores faltosos, a normalização dos atributos numéricos e, finalmente, a aplicação dos algoritmos sugeridos. Os planos sugeridos pelo PaRen foram aplicados no Weka, consi-

derando a mesma abordagem de validação aplicada às estratégias anteriores. Apesar de essa abordagem ter consumido $5$ horas de processamento os resultados não foram satisfatórios. Os melhores resultados em EAR e AUC para a base Bas foram, respectivamente, $0{,}86$ e $0{,}57$, obtidos na implementação MultiLayerPerceptron. Já para a base Mat, os melhores resultados para as mesmas métricas foram, respectivamente, $0{,}73$ e $0{,}57$, obtidos na implementação IBK.

## 5.4 SDVE

Nesta seção, estão sendo apresentados os resultados obtidos a partir da estratégia SDVE. Para essa estratégia, inicialmente, foi realizada a busca no espaço de hipóteses descrito na Figura 3.8 da página 93. O critério utilizado para seleção dos melhores caminhos (VRs) foi o desempenho médio obtido pela meta mais eficiente dos modelos produzidos. Assim, para a métrica EAR, considerou-se a média da metade dos modelos que obtiveram menor valor em cada VR. Da mesma forma, para a métrica AUC, considerou-se a média da metade dos modelos que obtiveram maior valor em cada VR. Além disso, como foi discutido na Seção 4.10 (pag. 105), VRs que levaram mais do que $24$ horas para serem processados foram descontinuados. A Figura 5.1 mostra qual foi a rota percorrida pelo metamodelo através dos VRs no espaço de busca.

As setas pretas mostram quais foram os caminhos escolhidos durante a busca. As duas bases de treino consideradas seguiram rotas similares, divergindo apenas no último VD, onde a base Bas produziu melhores resultados considerando os atributos de pontos do jogo, e a base Mat produziu melhores resultados considerando os atributos de tempo do jogo. O melhor resultado foi obtido na base Mat, considerando apenas os atributos das últimas fases. Contudo, esse modelo está sendo avaliado com cautela, já que o desempenho médio deste VR não foi o melhor encontrado no VD.

Ao final do processo de busca – que, ao todo, durou $122$ horas – foram produzidos $53.040$ diferentes modelos, que foram ranqueados através de seus valores obtidos na métrica AUC. Como a métrica EAR foi utilizada apenas para comparar as estratégias de classificação e regressão – e esta última obteve desempenho médio inferior a um preditor

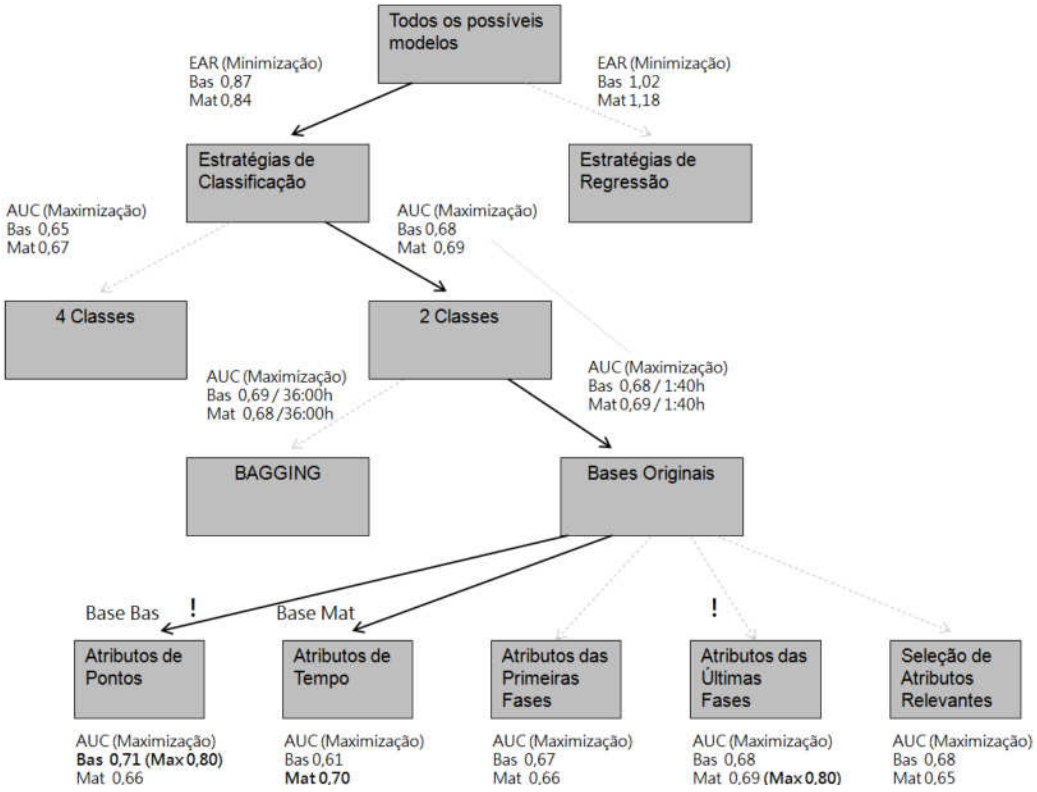



**Figura 5.1** Rota percorrida pelo metamodelo proposto através de SDVE. O critério utilizado para seleção dos melhores caminhos (VRs) foi o desempenho médio obtido pela meta mais eficiente dos modelos produzidos.

simples em ambas as bases –, com o objetivo de facilitar a legibilidade dos resultados, essa métrica foi desconsiderada. Ao final, puderam ser contabilizados 145 modelos eficientes sugeridos para a base de treino Bas e 92 modelos eficientes sugeridos para a base de treino Mat. Todos esses modelos conseguiram atingir o ponto de corte definido para comprovação da eficiência através da métrica AUC. As Tabelas 5.4(a) e 5.4(b) apresentam, respectivamente, os 10 melhores modelos obtidos para a base Bas e para a base Mat.

Para a base de treino Bas, em todos os 10 melhores modelos foi considerada a classificação através de 2 valores de classe, e apenas os atributos de pontos do jogo. Quase todos os modelos utilizaram classes que consideraram o ponto de corte em 6 sintomas de desatenção e hiperatividade / impulsividade. Duas diferentes técnicas de discretização apresentaram bons resultados. Porém, ambas dividiram o espaço contínuo dos atributos numéricos em 3 intervalos discretos. O melhor resultado obtido foi 0,76 em AUC, através da implementação LBR.

**Tabela 5.4** Melhores resultados obtidos através da proposta SDVE.

(a) **Base de treino Bas.**

| **AUC** (95% **IC)** | **Macroacurácia** (95% **IC)** | **Discretização** | **Algo** | **Filtro** | **Atributos do jogo** | **Nº de Classes** | **Corte Sintomas** |
|---|---|---|---|---|---|---|---|
| 0,76 *(0,67–0,83)* | 0,66 *(0,56–0,75)* | DscEqIntOp3B | [14] | Nenhum | Pontos | 2 | 6 |
| 0,75 *(0,66–0,83)* | 0,68 *(0,58–0,76)* | DscEqInt3B | [17] | Nenhum | Pontos | 2 | 5 |
| 0,75 *(0,66–0,83)* | 0,67 *(0,57–0,76)* | DscEqInt3B | [1] | Nenhum | Pontos | 2 | 6 |
| 0,75 *(0,66–0,83)* | 0,64 *(0,54–0,73)* | DscEqIntOp3B | [1] | Nenhum | Pontos | 2 | 6 |
| 0,75 *(0,66–0,83)* | 0,67 *(0,57–0,76)* | DscEqInt3B | [14] | Nenhum | Pontos | 2 | 6 |
| 0,75 *(0,66–0,83)* | 0,68 *(0,58–0,76)* | DscEqInt3B | [17] | Nenhum | Pontos | 2 | 6 |
| 0,75 *(0,66–0,83)* | 0,68 *(0,58–0,76)* | DscEqIntOp3B | [17] | Nenhum | Pontos | 2 | 6 |
| 0,75 *(0,66–0,83)* | 0,67 *(0,57–0,76)* | DscEqInt3B | [6] | Nenhum | Pontos | 2 | 6 |
| 0,75 *(0,66–0,83)* | 0,64 *(0,54–0,73)* | DscEqIntOp3B | [6] | Nenhum | Pontos | 2 | 6 |
| 0,75 *(0,66–0,83)* | 0,67 *(0,57–0,76)* | DscEqInt3B | [7] | Nenhum | Pontos | 2 | 6 |

**Implementações de algoritmos:**
3 x [17] VFI
2 x [1] NaiveBayes
2 x [6] NaiveBayesSimple
2 x [14] LBR
1 x [7] NaiveBayesUpdateable

**Técnicas de Discretização:**
4 x DscEqInt3B: Discretização em 3 intervalos de igual tamanho.
6 x DscEqIntOp3B: Discretização em até 3 intervalos de igual tamanho.

(b) **Base de treino Mat.**

| **AUC** (95% **IC)** | **Macroacurácia** (95% **IC)** | **Discretização** | **Algo** | **Filtro** | **Atributos do jogo** | **Nº de Classes** | **Corte Sintomas** |
|---|---|---|---|---|---|---|---|
| 0,80 *(0,67–0,89)* | 0,71 *(0,57–0,82)* | DscEqFrq4B | [4] | Nenhum | Últimos | 2 | 5 |
| 0,78 *(0,65–0,87)* | 0,73 *(0,59–0,83)* | DscEqInt4B | [33] | Nenhum | Todos | 2 | 5 |
| 0,78 *(0,65–0,87)* | 0,69 *(0,55–0,80)* | DscEqInt8B | [13] | Nenhum | Tempo | 2 | 5 |
| 0,77 *(0,64–0,86)* | 0,70 *(0,56–0,81)* | DscEqInt4B | [33] | Nenhum | Tempo | 2 | 5 |
| 0,77 *(0,64–0,86)* | 0,69 *(0,55–0,80)* | DscEqInt8B | [5] | Nenhum | Tempo | 2 | 5 |
| 0,77 *(0,64–0,86)* | 0,68 *(0,54–0,79)* | DscEqInt8B | [8] | Nenhum | Tempo | 2 | 5 |
| 0,76 *(0,63–0,86)* | 0,71 *(0,57–0,82)* | DscEqInt4B | [24] | BAGGING | Todos | 2 | 5 |
| 0,76 *(0,63–0,86)* | 0,65 *(0,51–0,77)* | DscEqInt8B | [3] | Nenhum | Tempo | 2 | 5 |
| 0,75 *(0,62–0,85)* | 0,70 *(0,56–0,81)* | DscEqInt4B | [24] | BAGGING | Todos | 2 | 6 |
| 0,75 *(0,62–0,85)* | 0,72 *(0,58–0,83)* | DscEqInt8B | [12] | Nenhum | Tempo | 2 | 5 |

**Implementações de algoritmos:**
2 x [33] LADTree
2 x [24] Prism
1 x [4] BayesNet
1 x [13] KStar
1 x [5] HNB
1 x [8] WAODE
1 x [3] AODEsr
1 x [12] IBK

**Técnicas de Discretização:**
5 x DscEqInt8B: Discretização em 8 intervalos de igual tamanho.
4 x DscEqInt4B: Discretização em 4 intervalos de igual tamanho.
1 x DscEqFrq4B: Discretização em 4 intervalos de igual frequência.

Para a base de treino Mat, em todos os 10 melhores modelos também foi considerada a classificação através de 2 valores. Contudo 3 tipos de conjuntos de atributos apresentaram bons resultados: últimos atributos das fases, atributos de tempo e todos os atributos. Desses 3 grupos, os atributos de tempo foram os mais presentes. Diferentemente dos melhores modelos sugeridos para a base Bas, quase todos os modelos para a base Mat utilizaram classes que consideraram o ponto de corte em 5 sintomas de desatenção e hiperatividade / impulsividade. Três diferentes técnicas de discretização apresentaram bons resultados, considerando 4 ou 8 intervalos discretos. O melhor resultado obtido foi 0,80 em AUC, através da implementação BayesNet.

## 5.5 Outras Avaliações

Nesta seção, estão sendo apresentados resultados referentes a 2 avaliações que inicialmente não faziam parte do escopo do experimento. Como essas são hipóteses com fortes argumentações que surgiram durante o desenvolvimento do trabalho, decidiu-se incorporar essas abordagens ao conjunto de estratégias. Está sendo feita uma reavaliação das estratégias Mapeamento em Metanível e Meta-Aprendizagem Através de Mecanismos de Base, considerando 2 classes ao invés de 4. Estão sendo apresentados, também, os resultados obtidos a partir de uma segunda avaliação com a proposta SDVE, mas, dessa vez, considerando a união das 2 bases de treino.

### 5.5.1 4 Classes versus 2 Classes

Após a aplicação das estratégias de Meta-Aprendizagem que foram inicialmente consideradas, observou-se que um argumento imposto durante as avaliações poderia estar influenciando nos resultados. Esse argumento refere-se à utilização de 4 ou 2 classes na tarefa de classificação. Como essa seria uma abordagem certamente aplicada por um minerador com um mínimo de experiência, resolveu-se reavaliar as estratégias Meta-Aprendizagem Através de Mecanismos de Base e Mapeamento em Metanível, considerando dessa vez apenas 2 classes. As Tabelas 5.5 (a e b) e 5.6 apresentam, respectivamente, os resultados obtidos para essas duas técnicas, considerando essa nova abordagem.

**Tabela 5.5** Desempenhos obtidos com a estratégia Meta-Aprendizagem Através de Mecanismos de Base, considerando 2 classes.

(a) Desempenho em EAR.

| Base | Estr. de Aprendizagem | Atr. Num. | StackingC | Vote | Boosting |
|---|---|---|---|---|---|
| **Bas** | Classificação (4 Classes) | Não Discr. | 0,88 | 0,89 | 0,89 |
| **Mat** | Classificação (4 Classes) | Não Discr. | 0,92 | 1,07 | 1,10 |

(b) Desempenho em AUC.

| Base | Estr. de Aprendizagem | Atr. Num. | StackingC | Vote | Boosting |
|---|---|---|---|---|---|
| **Bas** | Classificação (4 Classes) | Não Discr. | 0,67 | 0,62 | 0,57 |
| **Mat** | Classificação (4 Classes) | Não Discr. | 0,61 | 0,32 | 0,33 |

**Tabela 5.6** Desempenhos obtidos através das métricas EAR e AUC, utilizando a estratégia Mapeamento em Metanível, considerando 2 classes.

| | | | Base Bas | | Base Mat | |
|---|---|---|---|---|---|---|
| **Algoritmos (Impl.)** | **Estr. de Aprendizagem** | **Atr. Num.** | **EAR** | **AUC** | **EAR** | **AUC** |
| **MultiLayerPerceptron** | Classificação (4 Classes) | Não Discr. | 0,77 | 0,68 | 1,14 | 0,32 |
| **RandomForest** | Classificação (4 Classes) | Não Discr. | 0,92 | 0,58 | 1,06 | 0,37 |
| **IBK** | Classificação (4 Classes) | Não Discr. | 0,74 | 0,61 | 1,00 | 0,47 |
| **NaiveBayes** | Classificação (4 Classes) | Não Discr. | 0,67 | 0,67 | 1,46 | 0,29 |
| **OneR** | Classificação (4 Classes) | Não Discr. | 0,88 | 0,52 | 1,19 | 0,39 |

Para a estratégia Meta-Aprendizagem Através de Mecanismos de Base, quando apenas $2$ classes foram consideradas, a base de treino Bas apresentou as seguintes diferenças de desempenho, em relação à aprendizagem com $4$ classes: StackingC: $+10\%$ de desempenho em EAR e $+21\%$ de desempenho em AUC; Vote: $+7\%$ de desempenho em EAR e $+9\%$ de desempenho em AUC; Boosting: $+19\%$ de desempenho em EAR e $-4\%$ de desempenho em AUC. Já a base de treino Mat apresentou as seguintes diferenças relativas: StackingC: $+4\%$ de desempenho em EAR e $+19\%$ de desempenho em AUC; Vote: $-8\%$ de desempenho em EAR e $-26\%$ de desempenho em AUC; Boosting: $-8\%$ de desempenho em EAR e $-12\%$ de desempenho em AUC.

Para a estratégia Mapeamento em Metanínvel, quando apenas $2$ classes foram consideradas, a base de treino Bas apresentou uma melhora significante chegando a atingir $0{,}57$ de desempenho em AUC. Por outro lado, a base treino Mat produziu resultados piores em todos os algoritmos. Os mesmos algoritmos sugeridos anteriormente através desta estra-

tégia considerando $4$ classes foram também sugeridos na abordagem com $2$ classes para ambas as bases de treino.

De uma forma geral, mesmo considerando essa nova abordagem, nenhum dos modelos produzidos conseguiu atingir os pontos de corte definidos para comprovação da eficiência nas $2$ estratégias.

### 5.5.2 União das Bases

Uma outra hipótese que surgiu durante o desenvolvimento do experimento foi a possibilidade de se unir as bases para se obter um conjunto mais representativo do domínio. Aqui, observaram-se duas teorias conflitantes. De um lado, uma teoria amplamente aceita em Mineração de Dados afirma que bases de treino maiores são mais representativas e, dessa forma, tendem a produzir modelos de melhor desempenho. Do outro lado, segundo o especialista, crianças / adolescentes e adultos manifestam comportamentos diferentes em relação ao TDAH e, desta forma, considerar as duas bases de treino como sendo apenas uma poderia confundir um algoritmo de aprendizagem. Com o objetivo de avaliar essas possibilidades, resolveu-se aplicar novamente a proposta SDVE em uma nova base de treino composta pela união das bases de treinio Bas e Mat. Essa nova base está sendo chamada de base de treino BasMat.

A rota percorrida pelo metamodelo utilizando a base BasMat foi similar às percorridas pelas bases Bas e Mat (ver Figura 5.1), divergindo apenas no final, onde nenhuma das propostas de seleção de atributos obteve desempenho médio superior à utilização de todos os atributos ao mesmo tempo. Ao final do processo de busca – que também durou $122$ horas –, foram produzidos $26.520$ diferentes modelos, que foram ranqueados através de seus valores obtidos na métrica AUC. A Tabela 5.7 apresenta os resultados obtidos.

Considerando a união das $2$ bases de treino, o melhor resultado produzido em AUC foi $0{,}69$. Figuram entre os $10$ melhores modelos $9$ algoritmos de aprendizagem e $6$ estratégias de discretização. Foram também considerados os filtros BAGGING e Seleção Automática de Atributos Relevantes. O conjunto de atributos mais representativo considerou todos os atributos. Nenhum dos modelos produzidos conseguiu atingir o ponto de corte em AUC $(0{,}70)$, definido para comprovação da eficiência.

**Tabela 5.7** Melhores resultados obtidos na proposta SDVE utilizando-se a base de treino BasMat.

| AUC (95% IC) | Macroacurácia (95% IC) | Discretização | Algo | Filtro | Atributos do jogo | Nº de Classes | Corte Sintomas |
|---|---|---|---|---|---|---|---|
| 0,69 *(0,61–0,76)* | 0,61 *(0,53–0,68)* | DscEqInt3B | [15] | Nenhum | Tempo | 2 | 5 |
| 0,67 *(0,59–0,74)* | 0,63 *(0,55–0,70)* | DscEqInt4B | [19] | SelecAtrib | Todos | 2 | 5 |
| 0,67 *(0,59–0,74)* | 0,61 *(0,53–0,68)* | DscEqInt3B | [15] | BAGGING | Todos | 2 | 6 |
| 0,67 *(0,59–0,74)* | 0,65 *(0,57–0,72)* | DscEqIntOp3B | [33] | Nenhum | Todos | 2 | 4 |
| 0,66 *(0,58–0,73)* | 0,61 *(0,53–0,68)* | DscEqIntOp5B | [1] | Nenhum | Últimos | 2 | 5 |
| 0,66 *(0,58–0,73)* | 0,66 *(0,58–0,73)* | DscEqFrq8B | [31] | SelecAtrib | Todos | 2 | 5 |
| 0,66 *(0,58–0,73)* | 0,66 *(0,58–0,73)* | DscEqFrq8B | [32] | Nenhum | Tempo | 2 | 5 |
| 0,66 *(0,58–0,73)* | 0,61 *(0,53–0,68)* | DscEqInt4B | [6] | Nenhum | Pontos | 2 | 5 |
| 0,66 *(0,58–0,73)* | 0,61 *(0,53–0,68)* | DscEqIntOp7B | [7] | BAGGING | Todos | 2 | 5 |
| 0,65 *(0,57–0,72)* | 0,65 *(0,57–0,72)* | DscEqInt3B | [10] | Nenhum | Todos | 2 | 4 |

**Implementações de algoritmos:**
2 x [15] LWL
1 x [19] DecisionTable
1 x [33] LADTree
1 x [1] NaiveBayes
1 x [31] J48
1 x [32] J48Graft
1 x [6] NaiveBayesSimple
1 x [7] NaiveBayesUpdateable
1 x [10] SMO

**Técnicas de Discretização:**
3 x DscEqInt3B: Discretização em 8 intervalos de igual tamanho.
2 x DscEqInt4B: Discretização em 4 intervalos de igual tamanho.
2 x DscEqFrq8B: Discretização em 4 intervalos de igual frequência.
1 x DscEqIntOp7B: Discretização em até 7 intervalos de igual tamanho.
1 x DscEqIntOp3B: Discretização em até 3 intervalos de igual tamanho.
1 x DscEqIntOp5B: Discretização em até 5 intervalos de igual tamanho.

## 5.6 Análise dos Resultados

Como já era esperado, a Seleção Aleatória de Técnicas de Aprendizagem não é uma estratégia ideal para se lidar com domínios que possuem conceitos centrais difíceis de serem interpretados. Os resultados obtidos, além de ruins, foram tão aleatórios quanto à escolha dos algoritmos aplicados.

A estratégia Meta-Aprendizagem Através de Mecanismos de Base, que normalmente apresenta bons resultados em outros domínios, pode não ter sido bem sucedida neste contexto por não haver ainda um conjunto de algoritmos promissores que pudessem ser sugeridos – principalmente para as estratégias de votação e Boosting. Contudo, deve-se observar que bons resultados produzidos através dessa estratégia têm boa chance de ser

*overfintting* já que essa abordagem não considera nenhum conhecimento extra além dos algoritmos e da amostra.

A estratégia de Mapeamento em Metanível certamente produziria bons resultados se um problema similar fosse identificado nas bases da UCI, através de metacaracterísticas. Contudo, como foram consideradas apenas $90$ bases de treino – e seus respectivos problemas de aprendizagem –, talvez seja difícil a identificação de um caso similar, já que esse é um número muito pequeno de bases alternativas para ser aplicado a esse tipo de abordagem.

A proposta de Meta-Aprendizagem SDVE apresentou bons resultados para ambas as bases, apesar de haver chances de *overfitting* devido ao tamanho das amostras. Contudo, há duas motivações que dão credibilidade aos modelos produzidos através da proposta SDVE: $(1)$ As estratégias aplicadas foram sugeridas a partir de metaconhecimentos do domínio e não da amostra e, dessa forma, existe grande possibilidade de os mecanismos de aprendizagem estarem sendo conduzidos a um conceito central populacional, e não a um conceito central amostral. $(2)$ Analisando os $10$ melhores modelos produzidos para cada base de treino, pode-se perceber uma certa similaridade em suas metacaracterísticas (discretização, atributos utilizados, ponto de corte no número de sintomas), levando a crer que essas hipóteses estejam próximas a um conceito central. Considerando essas $2$ motivações, pode-se concluir que, mesmo que o modelo ótimo não esteja presente no *ranking* apresentado, existe grande possibilidade de essas alternativas estarem próximas.

A reavaliação da proposta SDVE utilizando a base de treino BasMat produziu resultados ruins pois as duas bases utilizadas na união são realmente oriundas de subdomínios diferentes, como havia sugerido o especialista. Em condições normais – ou seja, bases de treino obtidas de um mesmo domínio –, essa abordagem produziria desempenhos melhores. Contudo, um fato interessante que pode ser observado é a possibilidade de se construir um modelo híbrido a partir da união dos modelos – e não das bases – que foram produzidos para cada uma das bases de treino. Nessa abordagem, o atributo Idade poderia ser utilizado para selecionar um submodelo interno a um modelo maior que agregaria as $2$ hipóteses ao mesmo tempo.

Finalmente, com os resultados apresentados pela proposta SDVE, pode-se concluir que existe uma relação de causalidade entre o TDAH e o Jogo do Supermercado. Contudo, essa relação é sustentada por hipóteses específicas, sensíveis a diferentes grupos populacionais.

# Capítulo 6

# Conclusão

Através dos resultados produzidos pela proposta SDVE, pode-se concluir que existe um conjunto de modelos preditivos eficientes, capazes de relacionar os dados do Jogo do Supermercado ao diagnóstico do TDAH. Como foi discutido na Seção 5.6, essa conclusão baseia-se tanto no fato de que os mecanismos de aprendizagem utilizados foram conduzidos por conhecimentos do próprio domínio, quanto na similaridade dos $10$ melhores modelos identificados para cada população. Além disso, o fato de o objetivo de pesquisa ter sido alcançado através de $2$ subpopulações diferentes reforça ainda mais a hipótese de pesquisa discutida na introdução. Mesmo considerando a escassez de dados, esses são fortes argumentos que ajudam a comprovar a existência de uma relação de causalidade entre o TDAH e o Jogo do Supermercado, e afastam a possibilidade de um efeito *overfintting*. Contudo, para que os modelos produzidos neste trabalho possam ser utilizados em um teste diagnóstico para o TDAH, é importante que sejam revalidados em outras amostras das mesmas subpopulações consideradas para a correta certificação de suas funções preditivas.

Concluiu-se, também, que indivíduos de diferentes grupos populacionais possuem diferentes tipos de comportamentos no jogo, pois cada subpopulação produziu um conjunto de modelos preditivos diferente. Um teste diagnóstico que venha a ser definido através desta abordagem deve considerar essa questão.

Pode-se concluir, ainda, que qualquer tentativa de descoberta de conhecimento nesse

domínio deve considerar, também, heurísticas específicas – já que, como foi observado, parte dos conhecimentos necessários para se identificar bons modelos não está presente nos dados, mas sim em metadados que devem ser investigados ou elicitados através de um especialista.

# Capítulo 7

# Considerações Finais

Neste capítulo, apresentam-se as principais contribuições do trabalho apresentado nesta dissertação e recomendam-se alguns trabalhos que podem / devem ser desenvolvidos no futuro.

## 7.1 Contribuições

O trabalho aqui apresentado reforça a hipótese de que a utilização de um jogo computacional pode ser uma abordagem interessante para a elaboração de um teste diagnóstico. A ideia básica desta abordagem consiste em utilizar um mecanismo lúdico para quantificar e qualificar conceitos pré-estabelecidos na mente indivíduo. Essa é uma abordagem que deve ser investigada não apenas no contexto da psiquiatria, mas também em outras áreas em que a captura cognitiva possa apresentar bons resultados.

No contexto da Meta-Aprendizagem, outro aspeto importante, também demonstrado nesta dissertação, é o fato de que o conhecimento específico de um domínio, quando disponível, constitui uma excelente fonte de informação para identificação de modelos preditivos eficientes. Essa é uma abordagem que deve ser considerada em domínios cujos conceitos centrais sejam extremamente complexos, ou em situações em que os dados para treino são escassos.

Outras contribuições específicas que podem ser destacadas são:

- A análise e avaliação de estratégias de Meta-Aprendizagem igualmente adequadas

ao problema, através de um experimento;

- A definição de um método para Meta-Aprendizagem que considera a utilização de conhecimentos elicitados do domínio;
- Um conjunto de modelos preditivos eficientes sugeridos para a avaliação de crianças / adolescentes em relação ao TDAH através de um jogo computacional;
- Um conjunto de modelos preditivos eficientes sugeridos para a avaliação de adultos em relação ao TDAH através de um jogo computacional;

Além disso, alguns resultados obtidos neste trabalho podem ser apreciados em artigos publicados que apresentam e discutem algumas das abordagens propostas nesta dissertação. São eles:

- ANDRADE (2009): É apresentada a proposta inicial, que serviu de base para a utilização do Jogo do Supermercado no trabalho aqui apresentado.
- SANTOS *et al.* (2011b): Analisa uma série de testes preditivos iniciais, que foram realizados na base de treino Bas. As métricas utilizadas neste trabalho foram sensibilidade, especificidade, valor preditivo positivo (precisão) e valor preditivo negativo.
- SANTOS *et al.* (2011a): Apresenta os resultados que foram obtidos através da base de treino Mat. As métricas utilizas neste trabalho foram macroácurácia e AUC.

## 7.2 Trabalhos Futuros

Nesta seção, estão sendo discutidos pontos que ainda devem / podem ser investigados, tendo como base os conhecimentos produzidos nesta dissertação.

Em relação aos modelos produzidos pela proposta SDVE, os mesmos devem ser validados e / ou refinados através de outras bases de treino, para cada subdomínio. Embora os resultados produzidos tenham sido satisfatórios, ainda podem estar sofrendo de um viés amostral. Esse problema poderia ser corrigido se outras amostras fossem submetidas

ao mesmo espaço de metabusca, de forma que os modelos que apresentassem maior eficiência em todas as amostras pudessem ser considerados como eficientes populacionais. Outra abordagem interessante seria a aplicação de uma espécie de validação cruzada entre os melhores modelos de cada amostra – considerando sempre o mesmo subdomínio.

Uma outra proposta que tange aos modelos produzidos na proposta SDVE – porém, sem considerar outras bases de treino – seria a aplicação das estratégias de votação e *boosting* utilizando os modelos mais eficientes produzidos. Ao contrário da mesma abordagem aplicada na estratégia de Meta-Aprendizagem Através de Mecanismos de Base, ao invés de algoritmos de aprendizagem, seriam utilizados os melhores modelos do *ranking* para produção de um metamodelo.

O mecanismo de inferência para o jogo é também um trabalho futuro essencial que deve ser considerado. Esse mecanismo deve ser elaborado após a validação dos modelos produzidos, a fim de se garantir um bom desempenho no teste.

A proposta SDVE, que apresentou bons resultados neste domínio, pode ser aplicada a outros contextos que, da mesma forma, necessitem de conhecimentos específicos do domínio para definição de modelos preditivos. Além disso, essa proposta pode ainda ser aprimorada através de novas abordagens para a representação dos vieses e de outras estratégias para a definição da ordem de aplicação dos mesmos.

Ainda no contexto da proposta SDVE, a construção de um *framework* genérico poderia facilitar a aplicação dessa abordagem em outros domínios. Esse *framework* poderia contar com estratégias eficientes de elicitação – através de tabelas, questionários e até mesmo jogos – para a correta aquisição de conhecimentos, a partir de um especialista. Esses conhecimentos poderiam subjetivamente sugerir possíveis vieses, que seriam a base para a construção do espaço de busca. Além disso, a execução da busca – ou seleção dinâmica de vieses – poderia estar sendo orquestrada por esse *framework*, já que, essa foi a tarefa mais trabalhosa na aplicação da proposta SDVE neste trabalho.

# Referências Bibliográficas


AHA, D., KIBLER, D., ALBERT, M., 1991, "Instance-based learning algorithms", *Machine learning*, v. 6, n. 1, pp. 37–66.

ALPAYDIN, E., 2010, *Introduction to machine learning*. The MIT Press. ISBN: 978-0-262-01243-0.

ALTMAN, D., MACHIN, D., BRYANT, T., et al., 2002, *Statistics with confidence*. BMJ books Bristol.

ANDRADE, L. C. V., 2009, "Avaliação Cognitiva Utilizando Técnicas Inteligentes e um Jogo Computacional". In: *XX Simpósio Brasileiro de Informática na Educação*, Florianópolis, SC.

AYYUB, B., 2001, *Elicitation of expert opinions for uncertainty and risks*. CRC.

BALAKRISHNAN, N., RAO, C., 2004, *Advances in Survival Analysis,(Handbook of Statistics, vol. 23)*. North-Holland, Amsterdam.

BARKLEY, R., 1997, "Behavioral inhibition, sustained attention, and executive functions: Constructing a unifying theory of ADHD." *Psychological bulletin*, v. 121, n. 1, pp. 65. ISSN: 1939-1455.

BAXTER, J., 2011, "A model of inductive bias learning", *Arxiv preprint arXiv:1106.0245*.

BECHARA, A., DAMASIO, A., DAMASIO, H., et al., 1994, "Insensitivity to future consequences following damage to human prefrontal cortex* 1", *Cognition*, v. 50, n. 1-3, pp. 7–15. ISSN: 0010-0277.

BECHARA, A., DAMASIO, H., TRANEL, D., et al., 1997, "Deciding advantageously before knowing the advantageous strategy", *Science*, v. 275, n. 5304, pp. 1293.

BISHOP, C., 2006, *Pattern recognition and machine learning*, v. 4. Springer New York.

BLACK, K., 2010, "Business statistics: contemporary decision making, international student version (paperback)", .

BRAZDIL, P., GIRAUD-CARRIER, C., SOARES, C., et al., 2009, *Metalearning: Applications to data mining*. Springer-Verlag New York Inc.

BREIMAN, L., 1996, "Bagging predictors", *Machine learning*, v. 24, n. 2, pp. 123–140.

BREIMAN, L., 2001, "Random forests", *Machine learning*, v. 45, n. 1, pp. 5–32.

BRISCOE, G., CAELLI, T., 1996, *A Compendium of Machine Learning: Symbolic machine learning*. Ablex series in artificial intelligence. Ablex Pub. Corp. ISBN: 9781567501797. Disponível em: <http://books.google.com/books?id=xJwrEBxgyQcC>.

BROOK, U., GEVA, D., 2001, "Knowledge and attitudes of high school pupils towards peers' attention deficit and learning disabilities", *Patient Education and Counseling*, v. 43, n. 1, pp. 31–36. ISSN: 0738-3991.

CENDROWSKA, J., 1987, "PRISM: An algorithm for inducing modular rules", *International Journal of Man-Machine Studies*, v. 27, n. 4, pp. 349–370.

CHERNICK, M., 2008, *Bootstrap methods: A guide for practitioners and researchers*, v. 619. Wiley-Interscience.

COHEN, W., 1995, "Fast effective rule induction". In: *MACHINE LEARNING-INTERNATIONAL WORKSHOP THEN CONFERENCE-*, pp. 115–123. MORGAN KAUFMANN PUBLISHERS, INC.

CORTES, C., VAPNIK, V., 1995, "Support-vector networks", *Machine learning*, v. 20, n. 3, pp. 273–297.

COUTINHO, G., MATTOS, P., ARAÚJO, C., 2007, "Desempenho neuropsicológico de tipos de transtorno do déficit de atenção e hiperatividade (TDAH) em tarefas de atenção visual", *J Bras Psiquiatr*, v. 56, n. 1, pp. 13–6.

DE NIJS, P., FERDINAND, R., DE BRUIN, E., et al., 2004, "Attention-deficit/hyperactivity disorder (ADHD): parents' judgment about school, teachers' judgment about home", *European child & adolescent psychiatry*, v. 13, n. 5, pp. 315–320.

DEMIRÖZ, G., GÜVENIR, H., 1997, "Classification by voting feature intervals", *Machine Learning: ECML-97*, pp. 85–92.

DEMŠAR, J., ZUPAN, B., LEBAN, G., et al., 2004, "Orange: From experimental machine learning to interactive data mining", *Knowledge discovery in databases: PKDD 2004*, pp. 537–539.

DUAN, K., KEERTHI, S., 2005, "Which is the best multiclass SVM method? An empirical study", *Multiple Classifier Systems*, pp. 278–285.

EFRON, B., TIBSHIRANI, R., 1993, *An introduction to the bootstrap*, v. 57. Chapman & Hall/CRC.

EPSTEIN, A., BEGG, C., MCNEIL, B., 1986, "The use of ambulatory testing in prepaid and fee-for-service group practices", *New England Journal of Medicine*, v. 314, n. 17, pp. 1089–1094.

FAWCETT, T., 2006, "An introduction to ROC analysis", *Pattern recognition letters*, v. 27, n. 8, pp. 861–874. ISSN: 0167-8655.

FAYYAD, U., PIATETSKY-SHAPIRO, G., SMYTH, P., 1996, "From data mining to knowledge discovery in databases", *AI magazine*, v. 17, n. 3, pp. 37.

FERRI, C., HERNÁNDEZ-ORALLO, J., MODROIU, R., 2009, "An experimental comparison of performance measures for classification", *Pattern Recognition Letters*, v. 30, n. 1, pp. 27–38. ISSN: 0167-8655.

FISCHER, J., BACHMANN, L., JAESCHKE, R., 2003, "A readers' guide to the interpretation of diagnostic test properties: clinical example of sepsis", *Intensive care medicine*, v. 29, n. 7, pp. 1043–1051.

FRANK, E., WITTEN, I., OF WAIKATO. DEPT. OF COMPUTER SCIENCE, U., 1998, "Generating accurate rule sets without global optimization", Citeseer.

FRANK, E., HALL, M., PFAHRINGER, B., 2003, "Locally weighted naive Bayes". In: *Proceedings of the Conference on Uncertainty in Artificial Intelligence*, v. 256. Citeseer.

FRAWLEY, W., PIATETSKY-SHAPIRO, G., MATHEUS, C., 1992, "Knowledge discovery in databases: An overview", *Ai Magazine*, v. 13, n. 3, pp. 57. ISSN: 0738-4602.

FRAZIER, T., DEMAREE, H., YOUNGSTROM, E., 2004, "Meta-Analysis of Intellectual and Neuropsychological Test Performance in Attention-Deficit/Hyperactivity Disorder* 1", *Neuropsychology*, v. 18, n. 3, pp. 543–555. ISSN: 0894-4105.

FREUND, Y., SCHAPIRE, R., 1996, "Experiments with a new boosting algorithm". In: *MACHINE LEARNING-INTERNATIONAL WORKSHOP THEN CONFERENCE-*, pp. 148–156. Citeseer.

GAAG, L. V. D., RENOOIJ, S., WITTEMAN, C., et al., 2002, "Probabilities for a probabilistic network: A case-study in oesophageal carcinoma", *Artificial Intelligence in Medicine*, v. 25, pp. 123–148.

GAMA, J., 2004, "Functional trees", *Machine Learning*, v. 55, n. 3, pp. 219–250.

GORDON, D., DESJARDINS, M., 1995, "Evaluation and selection of biases in machine learning", *Machine Learning*, v. 20, n. 1, pp. 5–22.

HALL, M., FRANK, E., HOLMES, G., et al., 2009, "The WEKA data mining software: An update", *ACM SIGKDD Explorations Newsletter*, v. 11, n. 1, pp. 10–18. ISSN: 1931-0145.

HAN, J., KAMBER, M., 2006, *Data mining: concepts and techniques*. Morgan Kaufmann. ISBN: 1558609016.

HOLMES, G., HALL, M., PRANK, E., 1999, "Generating rule sets from model trees", *Advanced Topics in Artificial Intelligence*, pp. 1–12.

HOLMES, G., PFAHRINGER, B., KIRKBY, R., et al., 2002, "Multiclass alternating decision trees", *Machine Learning: ECML 2002*, pp. 105–122.

HOLTE, R., 1993, "Very simple classification rules perform well on most commonly used datasets", *Machine learning*, v. 11, n. 1, pp. 63–90.

HOSMER, D., LEMESHOW, S., 2000, *Applied logistic regression*, v. 354. Wiley-Interscience.

HOU, Z., OTHERS, 2005, "Texture defect detection using support vector machines with adaptive gabor wavelet features". Publishe by the IEEE Computer Society.

JIANG, L., ZHANG, H., 2006, "Weightily averaged one-dependence estimators", *PRICAI 2006: Trends in Artificial Intelligence*, pp. 970–974.

JOHN, G., LANGLEY, P., 1995, "Estimating continuous distributions in Bayesian classifiers". In: *Proceedings of the eleventh conference on uncertainty in artificial intelligence*, v. 1, pp. 338–345. Citeseer.

KANG, B., COMPTON, P., PRESTON, P., 1995, "Multiple classification ripple down rules: Evaluation and possibilities". In: *The 9th Knowledge Acquisition for Knowledge Based Systems Workshop*. Citeseer.

KESSLER, R., ADLER, L., BARKLEY, R., et al., 2005, "Patterns and predictors of attention-deficit/hyperactivity disorder persistence into adulthood: results from the na-

tional comorbidity survey replication", *Biological Psychiatry*, v. 57, n. 11, pp. 1442–1451. ISSN: 0006-3223.

KOHAVI, R., 1995, "The power of decision tables", *Machine Learning: ECML-95*, pp. 174–189.

KUKAR, M., 2001, "Making reliable diagnoses with machine learning: A case study", *Artificial Intelligence in Medicine*, pp. 88–98.

KUNCHEVA, L., 2004, *Combining pattern classifiers: methods and algorithms*. Wiley-Interscience.

KUPFER, D., 2000, "The Consensus Development Panel. National Institutes of Health Consensus Development Conference statement: diagnosis and treatment of Attention-Deficit/Hyperactivity Disorder (ADHD)", *J Am Acad Child Adolesc Psychiatry*, v. 39, pp. 182–93.

LANDWEHR, N., HALL, M., FRANK, E., 2005, "Logistic model trees", *Machine Learning*, v. 59, n. 1, pp. 161–205.

LE CESSIE, S., VAN HOUWELINGEN, J., 1992, "Ridge estimators in logistic regression", *Applied Statistics*, pp. 191–201.

MACKAY, D., 1998, "Introduction to Gaussian processes", *NATO ASI Series F Computer and Systems Sciences*, v. 168, pp. 133–166.

MAIMON, O., ROKACH, L., 2005, *Data mining and knowledge discovery handbook*. Springer-Verlag New York Inc. ISBN: 0387244352.

MARCHEVSKY, A., WICK, M., 2011, *Evidence Based Pathology and Laboratory Medicine*. Springer.

MATTOS, P., DUCHESNE, M., 1997, "Normalização de um teste computadorizado de atenção visual", *Arq. Neuropsiquiatria*, v. 55, pp. 62–69.

MITCHELL, T., 1980, *The need for biases in learning generalizations*. Dep. of Computer Science, Laboratory for Computer Science Research, Rutgers Univ.

MITCHELL, T. M., 1997, *Machine Learning*. McGraw-Hill Science/Engineering/Math.

MITSIS, E., MCKAY, K., SCHULZ, K., et al., 2000, "Parent–teacher concordance for DSM-IV attention-deficit/hyperactivity disorder in a clinic-referred sample." *Journal of the American Academy of Child & Adolescent Psychiatry*. ISSN: 1527-5418.

NEWCOMBE, R., 1998, "Interval estimation for the difference between independent proportions: comparison of eleven methods", *Statistics in Medicine*, v. 17, n. 8, pp. 873–890.

NGUYEN, T., 2010, "Layered approximation approach to knowledge elicitation in machine learning". In: *Rough Sets and Current Trends in Computing*, pp. 446–455. Springer.

OTA, 1980, *The implications of cost-effectiveness analysis of medical technology*. N. 3-4. DIANE Publishing.

PLATT, J., 1999, "Fast training of support vector machines using sequential minimal optimization". In: *Advances in Kernel Methods*, pp. 185–208. MIT press. ISBN: 0262194163.

PORTA, M., 2008, *A dictionary of epidemiology*. Oxford University Press, USA.

QUINLAN, J., 1986, "Induction of decision trees", *Machine learning*, v. 1, n. 1, pp. 81–106.

QUINLAN, J., 1992, "Learning with continuous classes". In: *5th Australian joint conference on artificial intelligence*, pp. 343–348. Citeseer.

QUINLAN, J., 1993, *C4. 5: programs for machine learning*. Morgan Kaufmann.

RAPIDMINER, R., 2009. "Open Source Data Mining". .

ROUSSEEUW, P., LEROY, A., WILEY, J., 1987, *Robust regression and outlier detection*, v. 3. Wiley Online Library.

ROY, S., 2002, "Nearest neighbor with generalization", *Unpublished, University of Canterbury, Christchurch, New Zealand*.

RUSSELL, S., NORVIG, P., 2009, *Artificial intelligence: a modern approach*. Prentice hall. ISBN: 0136042597.

SAMMUT, C., WEBB, G., 2011, *Encyclopedia of machine learning*. Springer-Verlag New York Inc.

SANTOS, F. E. G., BASTOS, A. P. Z., ANDRADE, L. C. V., et al., 2011a, "Assessment of ADHD in a Sample of Adults through a Computer Game". In: *International Conference on Applied Comuting*, Rio de Janeiro - RJ - Brasil, a.

SANTOS, F. E. G., BASTOS, A. P. Z., ANDRADE, L. C. V., et al., 2011b, "Assessment of ADHD through a Computer Game: An Experiment with a Sample of Students". In: *Third International Conference on Games and Virtual Worlds for Serious Applications*, pp. 104–111, Athens, Greece, b. IEEE Computer Society Order Number E4419, BMS Part Number CPF 1138G-CDR, ISBN 978-0-7695-4419-9.

SCHERMER, B., 2007, *Software agents, surveillance, and the right to privacy: a legislative framework for agent-enabled surveillance*. Amsterdam Univ Pr. ISBN: 9087280211.

SEEWALD, A., 2002, "How to make stacking better and faster while also taking care of an unknown weakness". In: *Proceedings of the Nineteenth International Conference on Machine Learning*, pp. 554–561. Morgan Kaufmann Publishers Inc.

SERRA-PINHEIRO, M., MATTOS, P., ANGÉLICA REGALLA, M., 2008, "Inattention, Hyperactivity, and Oppositional–Defiant Symptoms in Brazilian Adolescents", *Journal of Attention Disorders*, v. 12, n. 2, pp. 135. ISSN: 1087-0547.

SHAFAIT, F., REIF, M., KOFLER, C., et al., 2010, “Pattern Recognition Engineering”. In: *RapidMiner Community Meeting and Conference*, Online, 9.

SHEVADE, S., KEERTHI, S., BHATTACHARYYA, C., et al., 2000, “Improvements to the SMO algorithm for SVM regression”, *Neural Networks, IEEE Transactions on*, v. 11, n. 5, pp. 1188–1193.

SHI, H., 2007, “Best-first decision tree learning”. The University of Waikato.

SHIMITZ, M., CADORE, L., PACZKO, M., et al., 2002, “Neuropsychological performance in DSM-IV ADHD subtypes: an exploratory study with untreated adolescents”, *Can J Psychiatry*, v. 47, pp. 863–869.

SIETSMA, R., VERBEEK, J., VAN DEN HERIK, J., 2002, *Datamining en opsporing: toepassing van datamining ten behoeve van de opsporingstaak: strafprocesrecht versus recht op privacy*. Sdu Uitgevers. ISBN: 9012095034.

SIMITH, B. H., BARKLEY, R. A., SHAPIRO, C. J., 2007, *Attention deficit hyperactivity disorder*. 4th ed. New York, Guilford.

SWANSON, J., KRAEMER, H., HINSHAW, S., et al., 2001, “Clinical relevance of the primary findings of the MTA: success rates based on severity of ADHD and ODD symptoms at the end of treatment”, *Journal of the American Academy of Child & Adolescent Psychiatry*, v. 40, n. 2, pp. 168–179. ISSN: 0890-8567.

TODOROVSKI, L., DŽEROSKI, S., 2003, “Combining classifiers with meta decision trees”, *Machine Learning*, v. 50, n. 3, pp. 223–249.

TUFFÉRY, S., 2011, *Data Mining and Statistics for Decision Making*. Wiley.

UTGOFF, P., 1986, “Shift of bias for inductive concept learning”, *Machine learning: An artificial intelligence approach*, v. 2, pp. 107–148.

VILALTA, R., DRISSI, Y., 2002, “A perspective view and survey of meta-learning”, *Artificial Intelligence Review*, v. 18, n. 2, pp. 77–95.

VILALTA, R., GIRAUD-CARRIER, C., BRAZDIL, P., 2005, "Meta-learning", *Data Mining and Knowledge Discovery Handbook*, pp. 731–748.

WALLACH, J., 2007, *Interpretation of diagnostic tests*. Lippincott Williams & Wilkins.

WEBB, G., 1999, "Decision tree grafting from the all-tests-but-one partition". In: *INTERNATIONAL JOINT CONFERENCE ON ARTIFICIAL INTELLIGENCE*, v. 16, pp. 702–707. Citeseer.

WEBB, G., BOUGHTON, J., WANG, Z., 2005, "Not so naive bayes: Aggregating one-dependence estimators", *Machine Learning*, v. 58, n. 1, pp. 5–24.

WEINER, J., 2005, "The Johns Hopkins ACG Case-Mix System Reference Manual, Version 7.0", *Health Services Research & Development Center, The Johns Hopkins University Bloomberg School of Public Health. Baltimore, Maryland*.

WILSON, B., EVANS, J., ALDERMAN, N., et al., 1997, "Behavioural assessment of the dysexecutive syndrome", *Methodology of frontal and executive function*, pp. 239–250.

WILSON, E., 1927, "Probable inference, the law of succession, and statistical inference", *Journal of the American Statistical Association*, v. 22, n. 158, pp. 209–212.

WITTEN, I., FRANK, E., 2005, *Data Mining: Practical machine learning tools and techniques*. Morgan Kaufmann Pub. ISBN: 0120884070.

YANG, Y., WEBB, G., 2001, "Proportional k-interval discretization for naive-Bayes classifiers", *Machine Learning: ECML 2001*, pp. 564–575.

ZHANG, H., JIANG, L., SU, J., 2005, "Hidden naive bayes", *A A*, v. 1, n. 2, pp. 3.

ZHENG, F., WEBB, G., 2006, "Efficient lazy elimination for averaged one-dependence estimators", pp. 1113–1120.

ZHENG, Z., WEBB, G., 2000, "Lazy learning of Bayesian rules", *Machine Learning*, v. 41, n. 1, pp. 53–84.

ZHOU, X., OBUCHOWSKI, N., MCCLISH, D., 2002, *Statistical methods in diagnostic medicine*, v. 414. LibreDigital.

ZHU, W., ZENG, N., WANG, N., 2010, "Sensitivity, Specificity, Accuracy, Associated Confidence Interval and ROC Analysis with Practical SAS® Implementations", *NESUG*.