O Uso de Técnicas de Aprendizado de Máquina na Predição de Desempenho Acadêmico de Alunos em Cursos Superiores

Maitê Marques Caetano

Outubro / 2016

Dissertação de Mestrado em Ciência da Computação

O Uso de Técnicas de Aprendizado de Máquina na Predição de Desempenho Acadêmico de Alunos em Cursos Superiores

Esse documento corresponde à dissertação de mestrado apresentada à Banca Examinadora da Dissertação no curso de Mestrado em Ciência da Computação da Faculdade Campo Limpo Paulista.

Campo Limpo Paulista, 27 de outubro de 2016.

Maitê Marques Caetano

Profa. Dra. Maria do Carmo Nicoletti
Orientadora

Faculdade Campo Limpo Paulista

Programa de Mestrado em Ciência da Computação

"O Uso de Técnicas de Aprendizado de Máquina na Predição de Desempenho Acadêmico de Alunos em Cursos Superiores"

Maitê Marques Caetano

Dissertação de Mestrado apresentado ao Programa de Mestrado em Ciência da Computação da Faculdade Campo Limpo Paulista, como parte dos requisitos para a obtenção do título de Mestre em Ciência da Computação.

Membros da Banca:

Profa. Dra. Maria do Carmo Nicoletti
(Orientadora –FACCAMP)

Prof. Dr. Marcelo de Paiva Guimarães
(FACCAMP)

Prof. Dr. Estevam Rafael Hruschka Jr.
(UFSCar)

Campo Limpo Paulista, 24 de outubro de 2016.

FICHA CATALOGRÁFICA

Dados Internacionais de Catalogação na Publicação (CIP)

Câmara Brasileira do Livro, São Paulo, Brasil.

Caetano, Maitê Marques

O uso de técnicas de aprendizado de máquina na predição de desempenho acadêmico de alunos em cursos superiores / Sérgio Santos Silva Filho. Campo Limpo Paulista, SP: FACCAMP, 2016.

Orientador: Profª. Drª. Maria do Carmo Nicoletti
Dissertação (Programa de Mestrado em Ciência da Computação) – Faculdade Campo Limpo Paulista – FACCAMP.

1. Aprendizado de máquina. 2. Mineração de dados educacionais. 3. Desempenho acadêmico. I. Nicoletti, Maria do Carmo. II. Campo Limpo Paulista. III. Título.

CDD-370.2854

Agradecimentos

Agradeço primeiro a Deus pela oportunidade e habilidades concedidas para a realização deste projeto.

À minha querida família, os meus pais Sebastião de Souza Marques e Fernanda Pereira Marques, minha irmã Sophia Marques e o meu estimado esposo Gustavo Abreu Caetano, meus queridos sogros Ilson Tercio Caetano e Neila de Abreu Caetano e a minha cunhada Carolina Caetano, muito obrigada. Sem o apoio destes está conquista não seria possível. Obrigada por todos os bons conselhos, pela compreensão na ausência e pelo apoio nos momentos mais difíceis.

Ressalto o agradecimento à minha mãe pelas palavras de incentivo sempre a mim direcionadas em momentos cruciais. E ao Prof. Me. Ilson Tercio Caetano, meu chefe, meu sogro, pelo incentivo inicial e suporte oferecido durante todo o desenvolvimento desta pesquisa.

À minha amiga Roberta Carla Gomes da Silva que acompanhou todo o processo aconselhando nos momentos de dificuldade.

À professora e orientadora Dra. Maria do Carmo Nicoletti pela orientação e, principalmente, pela dedicação, ensinamentos e incentivo durante o período de realização desta dissertação.

À instituição Faccamp através do coordenador do programa de mestrado em Ciência da Computação Prof. Dr. Osvaldo Luiz de Oliveira pelo apoio e incentivo.

Aos professores Dr. Estevam R. Hruschka Jr. e Dr. Marcelo de Paiva Guimarães pela disposição e contribuições ao trabalho examinado.

A todos aqueles que de uma forma ou de outra auxiliaram para que a realização deste trabalho fosse possível.

Resumo. *Embora originalmente direcionadas a negócios, presentemente as técnicas de mineração de dados podem ser aplicadas a qualquer tipo de dado, tais como os relacionados à meteorologia, finanças, seguros e, particularmente, aqueles produzidos por instituições educacionais. Está dissertação apresenta um trabalho de pesquisa na área Mineração de Dados Educacionais, cujo principal objetivo é a predição do desempenho de estudantes universitários em ambiente convencional de ensino. Os experimentos descritos neste documento usam aprendizado automático, com três dos mais conhecidos algoritmos de classificação, Naïve Bayes, Nearest Neighbor (1Bk) e J48 (versão Weka do C4.5), com 4-validação cruzada. Os resultados mostram que é possível predizer o desempenho de estudantes com precisão em torno de 70% com informações iniciais, e em torno de 90% se usados o conjunto completo de atributos, ou o conjunto reduzido com os atributos mais impactantes na classe. Os resultados podem ser fornecidos a professores e gestores acadêmicos como informações úteis sobre o desempenho dos estudantes, a fim de que sirvam como base para a modificação, se necessária, da condução de disciplinas nos cursos de graduação. Também podem ser usados para identificar estudantes com desempenho a quem do esperado para aprovação, com o propósito de recuperar o desempenho destes estudantes, ainda no período regular da disciplina.*

Abstrac*t. Even though it was originally business oriented, nowadays data mining techniques can be applied to a variety of types of data such as those from meteorology, medicine, finance, insurance and, particularly for educational institutions. This present thesis is a research in the area of Educational Data Mining; its main objective is to predict the performance of university students in a conventional teaching environment. The experiments here described use automatic learning, with three of the best known classification algorithms, Naïve Bayes, Nearest Neighbor (1Bk) and J48 (Weka version of C4.5) with a 4-fold cross validation. The results show that it is possible to accurately predict in about 70% the students' performance with initial information, and around 90% if it is used the complete set of attributes, or in a reduced group with the most impressive attributes of its class. The results can be provided to professors and academic managers in the form of useful information from the students´ performance in order to serve as a basis for change if necessary of the subjects in the undergraduate courses. They can also be used to identify students with lower performance for approval in order to help them to increase their performance, even in the regular period of the subject.*

Sumário

Lista de Tabelas

Lista de Figuras

Lista de Algoritmos

Capítulo 1
Introdução

1.1 Contextualização & Motivação

A possibilidade de prever o desempenho de alunos em disciplinas de cursos universitários, dentro de um intervalo de acurácia pré-estabelecido, tem papel relevante em ambientes educacionais uma vez que pode subsidiar, de maneira decisiva, a condução do processo ensino-aprendizagem. Antever, com alguma margem de tempo, o desempenho de um aluno pode, entre outros, motivar o docente responsável pela disciplina a usar e aplicar recursos pedagógicos que permitam, por exemplo, a recuperação daqueles alunos que se enquadram em um determinado perfil de desempenho, ainda durante o tempo regular da disciplina, sem a necessidade de tempo extra de recuperação, normalmente requerido.

De uma maneira simplista, uma possibilidade de antever o desempenho escolar é considerar dados relativos a desempenhos escolares anteriores e, com base nesses dados, predizer um desempenho futuro. Obviamente existe um número bem grande de variáveis envolvidas em tal processo e, portanto, dependendo da qualidade dos dados disponíveis, a previsão pode ter (ou não) uma acurácia satisfatória e ser usada, se conveniente, para corrigir eventuais problemas. Algoritmos e técnicas da área de Mineração de Dados (MD) podem subsidiar o processo de predição, desde que existam dados representativos e confiáveis (i.e., sem ruídos, sem valores ausentes, etc.) e em volume suficiente sobre o processo que representam.

A MD pode ser caracterizada, de maneira abrangente e no que tange à pesquisa, como a área voltada ao estudo e investigação de formalismos matemáticos e técnicas computacionais, com vistas ao desenvolvimento de algoritmos para a exploração do volume (cada vez maior) de dados, relativos a um número crescente de domínios de conhecimento. Por meio do uso de algoritmos de MD pretende-se a extração de informação implícita que seja potencialmente útil (tal como regras de associação entre variáveis), a partir de dados relativos a um determinado domínio de conhecimento. Muitas

das técnicas caracterizadas como de MD são algoritmos da área de Aprendizado de Máquina (AM) que foram customizados para lidar com grandes volumes de dados e muitas vezes, algoritmos de MD e AM são entendidos como tendo a mesma funcionalidade. A área intitulada Mineração de Dados Educacionais (MDE) diz respeito ao uso de técnicas de MD em domínios que abrangem dados educacionais.

Como apontado em [García *et al.* 2011] e [Baker & Yacef 2010], a MDE emergiu nos últimos anos como uma área de pesquisa para pesquisadores das mais diversas áreas (e.g., Computação, Educação, Psicologia, Psicométrica, Estatística, Sistemas Tutores Inteligentes, E-learning, etc.), quando da análise de grandes volumes de dados com o objetivo de resolver questões voltadas às áreas educacionais associadas às respectivas áreas de pesquisa individuais. O aumento no número de repositórios com dados acadêmicos de alunos e, também, a rápida obtenção (embora restrita) de tais dados por meio de *queries* a banco de dados, tem viabilizado, de certa forma, o seu uso em um contexto acadêmico, para extração de informações que podem reverter em replanejamento de disciplinas, redirecionamento de práticas pedagógicas, revisão de ementas, etc... Processos utilizados em MDE podem ser vistos como conversores daqueles dados acadêmicos brutos que foram obtidos por sistemas educacionais tradicionais (sejam eles automatizados ou não), em informação útil ao sistema educacional como um todo, que pode ser utilizada por desenvolvedores de software, professores, pesquisadores educacionais, etc.

1.2 Organização do Documento

Esta pesquisa tem por objetivo principal a investigação das potencialidades no uso de determinadas técnicas de MD e de AM, com o objetivo de prever o desempenho de alunos em uma disciplina de um curso universitário de Bacharelado em Computação de um Centro Universitário situado na grande São Paulo.

Além desse capítulo inicial, que essencialmente investe na contextualização e motivação para a pesquisa realizada, o documento está organizado em mais sete capítulos, cujos conteúdos são brevemente informados a seguir.

O Capítulo 2 traz um relato das pesquisas que se mostraram potencialmente relevantes para subsidiar parte do trabalho de pesquisa realizado e apresentado nesta

dissertação. Grande parte dos trabalhos, foram identificados durante a atividade de revisão bibliográfica e as demais referências, durante a finalização do projeto.

O Capítulo 3 apresenta o ambiente universitário provedor dos dados para a pesquisa realizada. No capítulo são descritos o ambiente educacional alvo, as particularidades do curso superior pretendido, bem como a disciplina escolhida para o estudo. O capítulo apresenta, também, uma descrição detalhada dos dados existentes relativos ao curso e disciplina. Tais dados estão armazenados no banco de dados da instituição e são relativos à caracterização do aluno bem como à descrição de seu histórico escolar.

O Capítulo 4 investe na revisão de alguns dos algoritmos usados na pesquisa descrita, discutindo os respectivos pseudocódigos, as entradas requeridas e saídas geradas e, eventualmente, parâmetros esperados.

O Capítulo 5 detalha a metodologia adotada na realização dos experimentos relacionados ao uso de algoritmos de AM na predição de desempenho acadêmico. O capítulo também descreve o tratamento de valores ausentes de atributos e a formatação dos arquivos usados nos experimentos, além de uma breve apresentação de alguns métodos para seleção de atributos, disponíveis no ambiente Weka.

O Capítulo 6 aborda os experimentos relativos ao processo de seleção de atributos relevantes dos dados considerados. Esta seleção fez uso de duas técnicas disponíveis no ambiente Weka: a que seleciona subconjuntos de atributos relevantes e aquela que ranqueia atributos individuais por ordem de relevância em detrimento da classe.

O Capítulo 7 descreve os experimentos de AM utilizando os sete conjuntos de dados educacionais coletados e três dos algoritmos de ML disponíveis no ambiente Weka, a saber, o Naïve Bayes, o J48 e o 1-Nearest Neighbor (1Bk).

O Capítulo 8 apresenta as conclusões relativas ao trabalho realizado, resume as principais atividades desenvolvidas na pesquisa, comenta os resultados obtidos nos experimentos, bem como sugere pontos a serem investigados como continuidade a pesquisa realizada descrita nesta dissertação.

Capítulo 2
Revisão Bibliográfica

2.1 Sobre a Predição de Desempenho Acadêmico

Como comentado no Capítulo 1, a predição, com uma certa antecipação, de desempenho acadêmico de estudantes em disciplinas de cursos superiores pode contribuir, de diferentes maneiras, nos processos de aprendizagem dos próprios alunos envolvidos, na revisão do cronograma planejado, associado ao conteúdo programático da disciplina e nos processos relacionados à administração escolar de tais cursos.

Dentre os vários trabalhos acadêmicos evidenciados durante a fase de levantamento bibliográfico, norteado pela pesquisa pretendida, alguns foram bastante relevantes para subsidiar o trabalho de pesquisa realizado e descrito nesta dissertação. Outros, entretanto, apesar de não terem influenciado diretamente o desenvolvimento da pesquisa, foram selecionados para compor a revisão apresentada neste capítulo em virtude de sua potencialidade e sua caracterização diferenciada, como é o caso no uso de sistemas de recomendação e acoplamento de classificadores, como tratado na Seção 2.4 deste capítulo.

2.2 Predição de Desempenho Acadêmico

No que segue o relato do trabalho realizado por Kotsiantis e colaboradores, descrito em [Kotsiantis *et al.* 2003], é abordado de maneira a extrair e compor as informações mais relevantes que subsidiaram a pesquisa realizada, descrita neste documento.

2.2.1 Considerações Iniciais

A referência [Kotsiantis *et al.* 2003] descreve uma pesquisa que teve por principal objetivo a investigação da eficiência de técnicas de aprendizado de máquina, na predição da performance de alunos, em um ambiente universitário de educação a distância. Tal pesquisa realizou um conjunto de experimentos envolvendo cinco algoritmos de MDE,

que foram treinados usando dados obtidos junto ao curso de Informática, da Hellenic Open University. Nos experimentos foram usados os algoritmos listados na Tabela 2.1.

Tabela 2.1 Algoritmos de MDE utilizados nos experimentos realizados e descritos em [Kotsiantis *et al.* 2003].

ALGORITMO	SIGLA	Comentários sobre a representação do conceito utilizada
(1) C4.5	AD	C4.5 [Quinlan 1993] representa o conceito aprendido como uma árvore de decisão.
(2) Naïve Bayes	NB	Conceito aprendido representado como uma rede Bayesiana [Domingos & Pazzani 1997].
(3) 3-NN	NN	A versão 3-NN do *Nearest Neighbor* [Cover & Hart 1967] representa o conceito com base nas instâncias armazenadas.
(4) RIPPER	RD	O algoritmo RIPPER [Cohen 1995] induz a representação do conceito como um conjunto de regras de decisão.
(5) WINNOW	PE	O WINNOW induz a representação do conceito com base no Perceptron [Rosenblatt 1958]

2.2.2 Sobre os Dados Utilizados e Resultados Obtidos

Os dados utilizados nos experimentos foram aqueles relativos à um módulo específico, *Introdução à Informática* (INF10), da grade curricular do curso de Informática. Durante o ano acadêmico os alunos inscritos em INF10 devem entregar 4 trabalhos escritos, opcionalmente participar de 4 encontros presenciais com seus respectivos tutores e então, prestar um exame final após um período de 11 meses. É mandatório que o aluno entregue pelo menos 3 dos 4 trabalhos escritos. A nota final associada aos trabalhos escritos entregues por um aluno deve ser maior ou igual a 20, para que tal aluno se qualifique para realizar o exame final.

Cada instância (ou registro) de dado é descrita por um conjunto de atributos, que foi considerado dividido em dois subconjuntos disjuntos:

(1) o dos atributos associados a informações gerais (identificado como o conjunto IG) do respectivo aluno e

(2) aquele dos atributos associados ao desempenho do respectivo aluno no módulo INF10 (identificado como o conjunto D_INF).

A Tabela 2.2 apresenta uma descrição dos dois grupos de atributos. Para cada atributo considerado são listados, na coluna correspondente à direita, os seus possíveis valores.

Tabela 2.2 Conjuntos de atributos (IG & D_INF) que descrevem cada instância de dado e respectivos valores que cada um dos atributos pode assumir.

CONJUNTO DE ATRIBUTOS ASSOCIADOS A INFORMAÇÕES GERAIS (IG)	
NOME DO ATRIBUTO	VALORES POSSÍVEIS
(1) Sexo	masculino, feminino
(2) Idade	idade<32, idade>32
(3) Estado civil	solteiro, casado, divorciado, viúvo
(4) Número de filhos	0, 1, 2, mais
(5) Ocupação	não, parcial, integral
(6) Alfabetizado computacionalmente	não, sim
(7) Trabalho associado com computação	não, júnior, sênior

CONJUNTO DE ATRIBUTOS ASSOCIADOS A DESEMPENHO (D_INF)		
NOME DO ATRIBUTO	SIGLA	VALORES POSSÍVEIS
(8) 1º. encontro presencial	EP1	ausente, presente
(9) 1º. trabalho escrito	TE1	nota<3, 3≤nota≤6, nota>6
(10) 2º. encontro presencial	EP2	ausente, presente
(11) 2º. trabalho escrito	TE2	nota<3, 3≤nota≤6, nota>6
(12) 3º. encontro presencial	EP3	ausente, presente
(13) 3º. trabalho escrito	TE3	nota<3, 3≤nota≤6, nota>6
(14) 4º. encontro presencial	EP4	ausente, presente
(15) 4º. trabalho escrito	TE4	nota<3, 3≤nota≤6, nota>6

Dois experimentos foram conduzidos com os cinco algoritmos de MDE listados na Tabela 2.1. O primeiro experimento usou como conjunto de treinamento 354 instâncias (*i.e.*, registros de alunos) e o segundo experimento usou 28 instâncias, associadas ao módulo INF10. O uso de apenas 28 instâncias como conjunto de treinamento para o segundo experimento, foi justificado com base no número aproximado de alunos inscritos em uma turma regular do módulo INF10, que dificilmente ultrapassa 30 estudantes.

Os resultados do primeiro experimento alcançaram até 62% de precisão nas previsões iniciais (quando o número de atributos foi sendo variado) e acima de 82% quando todos os atributos foram considerados. Já os resultados obtidos com o segundo experimento, evidenciaram que 28 instâncias não foram suficientes para a indução de uma expressão do conceito que fosse representativa de todo o conjunto de dados. Depois de realizados testes com diferentes números de instâncias nos conjuntos de treinamentos, observou-se que são necessárias pelo menos 70 instâncias para uma melhor predição.

Os valores dos atributos de D_INF foram fornecidos pelos respectivos tutores. Quando um trabalho escrito não é entregue por um determinado aluno, o correspondente valor em seu registro é zero. As notas atribuídas aos trabalhos escritos foram discretizadas em cinco grupos:

(1) *não*: não houve entrega do respectivo trabalho;

(2) *falho*: nota menor que 5,0;

(3) *bom*: nota $\in$ [5,0 6,5) ;

(4) *muito bom*: nota $\in$ [6,5 8,5) e

(5) *excelente*: nota $\in$ [8,5 10,0].

O resultado do exame final (considerado como a classe do desempenho acadêmico à qual o estudante em questão pertence) tem dois valores possíveis:

(1) *reprovado*: estudantes com desempenho fraco e

(2) *aprovado:* estudantes que completaram o módulo tendo pelo menos nota 5 no teste final.

É importante mencionar que os autores não deixam claro qual a influência, das notas obtidas nos trabalhos escritos no valor do resultado final.

O primeiro objetivo do trabalho realizado foi inferir (usando um algoritmo de MDE) o resultado do exame final do aluno (antes de sua realização), com base nos valores dos 15 atributos que refletem sua participação no módulo INF10. Os resultados evidenciaram que, com a ajuda de tais algoritmos, tutores de cursos a distância estão aptos a anteciparem, quase que desde o início do módulo (no caso INF10) e com base apenas nos dados do currículo do aluno, quais deles irão ser aprovados (e quais não), antes do exame final ser realizado.

Como objetivo secundário os experimentos procuraram, também, selecionar os atributos que se mostraram efetivamente relevantes para a categorização de um aluno no grupo reprovado ou no grupo aprovado, ou seja, antecipar seu desempenho no exame final, com base no seu desempenho durante os 11 meses cursando a INF10.

A seleção de atributos relevantes pode contribuir substancialmente na redução de espaço de armazenamento de dados, bem como promover uma maior rapidez na indução dos conceitos que representam cada uma das classes (a de aprovados e a de reprovados).

Dentre os 15 atributos iniciais, os resultados mostraram que os atributos (1) sexo, (2) idade, (3) estado civil, (4) número de filhos e (6) ocupação não contribuem para um aumento na precisão dos conceitos induzidos. Também, os resultados dos experimentos correlacionam fortemente o desempenho do aluno à existência de educação prévia em Informática ou, então, à experiência profissional em áreas associadas à computação.

2.2.3 Uma Breve Descrição da Metodologia Empregada nos Experimentos

A metodologia utilizada para a realização dos experimentos descritos na Seção 2.2.2 abordou cada um dos experimentos divididos em duas fases distintas

(1) fase de treinamento e

(2) fase de classificação.

Na fase (1) os cinco algoritmos foram treinados com dados acadêmicos da disciplina de INF10 do ano 2000-1. Esta fase foi subdividida em 9 passos, cada um deles referente a um determinado conjunto de atributos utilizado para descrever as instâncias de treinamento, como mostra a Tabela 2.3. O significado de cada sigla está explicado na parte inferior da Tabela 2.2. O sufixo numérico acrescentado a cada sigla de algoritmo de MDE (na coluna CLASSIFICADORES da Tabela 2.3) representa o índice do passo; por exemplo, AD1 representa a árvore de decisão gerada no Passo 1.

Tabela 2.3 Atributos usados como conjunto de treinamento em cada um dos passos (P).

CONJUNTO DE TREINAMENTO		
P	ATRIBUTOS	CLASSIFICADORES
1º	IG + AD	AD1, NB1, NN1, RD1, PE1
2º	IG + EP1 + AD	AD2, NB2, NN2, RD2, PE2
3º	IG + EP1 + TE1 + AD	AD3, NB3, NN3, RD3, PE3
4º	IG + EP1 + TE1 + EP2 + AD	AD4, NB4, NN4, RD4, PE4
5º	IG + EP1 + TE1 + EP2 + TE2 + AD	AD5, NB5, NN5, RD5, PE5
6º	IG + EP1 + TE1 + EP2 + TE2 + EP3 + AD	AD6, NB6, NN6, RD6, PE6
7º	IG + EP1 + TE1 + EP2 + TE2 + EP3 + TE3 + AD	AD7, NB7, NN7, RD7, PE7
8º	IG + EP1 + TE1 + EP2 + TE2 + EP3 + TE3 + EP4 + AD	AD8, NB8, NN8, RD8, PE8
9º	IG + EP1 + TE1 + EP2 + TE2 + EP3 + TE3 + EP4 + TE4 + AD	AD9, NB9, NN9, RD9, PE9

Como mostra a Tabela 2.3, o 1º Passo consistiu em usar cada um dos cinco algoritmos listados na Tabela 2.1 para gerar classificadores – neste passo o conjunto de treinamento foi descrito apenas pelos 7 atributos referentes às informações gerais (IG)

(parte superior da Tabela 2.2) mais a avaliação na disciplina (AD) (*i.e.*, a classe associada à instância). No 2º Passo as instâncias do conjunto de treinamento foram descritas pelos 7 atributos de IG, mais o resultado obtido no encontro presencial 1 (EP1) mais a AD. O 3º Passo utilizou os mesmos atributos usados no 2º. Passo para descrever as instâncias, mais as informações do teste escrito 1 (TE1), e assim sucessivamente até que no último passo, o 9º. Passo, todos os atributos apresentados na Tabela 2.2 foram utilizados na descrição do conjunto de treinamento.

Para a fase (2), de classificação, foram utilizados 10 grupos de dados do ano 2001-2. Tais grupos de dados foram coletados por 10 tutores, 1 grupo de dado por tutor. A fase de classificação teve também 9 passos. No 1º Passo, classificadores induzidos usando instâncias descritas apenas pelos 7 atributos de IG (e a classe) foram utilizados para predizer a classe (aprovado/reprovado) de cada aluno. Este processo foi repetido por 10 vezes (cada uma associada a um dos grupos de dados fornecidos pelos tutores) e a média da acurácia das previsões (IG_M) foi calculada para cada algoritmo e está na Tabela 2.4.

Tabela 2.4 Precisão dos algoritmos em cada passo testado.

	Naïve Bayes	3-NN	RIPPER	C4.5	WINNOW
IG_M	62,59%	63,13%	62,97%	61,40%	54,77%
EP1_M	61,95%	62,67%	63,17%	61,00%	54,54%
TE1_M	66,59%	63,48%	65,27%	64,02%	62,61%
EP2_M	72,43%	67,00%	71,75%	73,57%	62,78%
TE2_M	77,18%	73,80%	78,55%	78,73%	69,85%
EP3_M	78,43%	77,75%	78,92%	77,59%	70,31%
TE3_M	81,13%	79,82%	79,06%	77,53%	76,51%
EP4_M	81,17%	80,38%	79,99%	77,39%	76,37%
TE4_M	83,00%	83,03%	81,34%	78,09%	77,88%
MÉDIA	73,83%	72,34%	73,44%	72,15%	67,29%

O 2º Passo abordou os classificadores induzidos usando as instâncias descritas pelos atributos relativos à IG e o valor de EP1. A avaliação dos classificadores em cada um dos 10 conjuntos coletados por tutores foi conduzida e a precisão média de classificação para cada algoritmo, está indicada na linha "EP1_M" da Tabela 2.4. Os passos restantes também foram executados 10 vezes utilizando o mesmo procedimento descrito anteriormente. A precisão média da predição está representada nas linhas "TE1_M", "EP2_M", "TE2_M", "EP3_M", "TE3_M", "EP4_M" e "TE4_M", da Tabela 2.4, para cada algoritmo.

Após realizar a classificação dos estudantes com os algoritmos citados, os autores trabalharam na classificação destes algoritmos baseado no critério "precisão da predição". Para realizar tal comparação foi utilizado o teste estatístico T-Student. Este teste calcula as diferenças entre cada algoritmo e considera essa diferença estatisticamente significativa quando p>0,001.

O resultado da comparação, como mostrado na Tabela 2.5, classificou as 9 etapas do processo em 3 valores (x/y/z), que representam as diferenças significativas entre o algoritmo da linha em relação ao algoritmo da coluna. O valor de x representa a quantidade de passos em que houve perdas do algoritmo em comparação ao outro, o valor de y representa a quantidade de passos em que não houve diferença significativa e o valor de z representa a quantidade de passos em que houve ganho. Como pode ser visto na tabela, o algoritmo Naïve Bayes 'ganhou' em 3 passos quando comparado ao algoritmo 3-NN, em 6 passos não houve diferença significativa relevante. Já na comparação entre o Naïve Bayes e o RIPPER, não foi detectada diferença significativa em qualquer dos passos, o que significa que o desempenho de ambos é similar.

Tabela 2.5 Comparação entre 3-NN, RIPPER, C4.5 e WINNOW.

Algoritmo	3-NN	RIPPER	C4.5	WINNOW
Naïve Bayes	0/6/3	0/9/0	0/5/4	0/0/9
3-NN		2/7/0	2/4/3	0/1/8
RIPPER			1/5/3	0/1/8
C4.5				0/4/5

Os resultados da comparação evidenciaram que o Naïve Bayes e o RIPPER possuem os melhores desempenhos dentre os 5 classificadores avaliados. Para a escolha de apenas um classificador, foi levado em conta o menor tempo computacional obtido bem como a possibilidade de trabalhar com dados ausentes, o que não é possível com o classificador RIPPER. Dessa forma, os autores concluíram que o Naïve Bayes é o mais apropriado para ser usado na construção de uma ferramenta para apoio ao tutor escolar, com as características daquela definida no trabalho feito.

2.2.4 Relato de uma Extensão do Trabalho Realizado

Um trabalho subsequente realizado por Kotsiantis e colaboradores e descrito em [Kotsiantis *et al.* 2004] pode ser considerado uma extensão do trabalho descrito em [Kotsiantis *et al.* 2003]; nele foram repetidos os dois experimentos abordados nas seções

2.2.1-2.2.3, com algumas pequenas variações, particularmente aquela relativa aos algoritmos de aprendizado utilizados, como pode ser evidenciado na Tabela 2.6.

Tabela 2.6 Algoritmos utilizados nos experimentos descritos em [Kotsiantis *et al.* 2004].

ALGORITMO	SIGLA	Comentários sobre a representação do conceito
(1) C4.5	AD	C4.5 [Quinlan 1993] representa o conceito aprendido como uma árvore de decisão.
(2) Naïve Bayes	NB	Conceito aprendido representado como uma rede Bayesiana [Domingos and Pazzani 1997].
(3) 3-NN	NN	A versão 3-NN do *Nearest Neighbor* [Cover & Hart 1967] representa o conceito com base nas instâncias armazenadas.
(4) BACKPROPAGATION	BP	O *Backpropagation* [Rumelhart *et al.*1986] induz a representação do conceito como uma rede neural.
(5) MLE	ML	O MLE (*Estimativa de Probabilidade Máxima*) é um método estatístico de regressão logística [Long 1997].
(6) SMO	SM	O *Sequential Minimal Optimization* é um algoritmo SVM [Platt 1998].

Os seis algoritmos foram treinados e testados utilizando os mesmos conjuntos de dados utilizados nos experimentos descritos em [Kotsiantis *et al.* 2003] e revisado nas seções anteriores deste documento. Os atributos que descrevem as instâncias de treinamento utilizadas pelos autores são os mesmos 15 atributos listados na Tabela 2.2 da Seção 2.2.2, com o acréscimo do atributo (16) nomeado de Teste Final, com duas possíveis opções de valores: ‘aprovação’ e ‘reprovação’.

Na extensão do trabalho os autores avaliaram o desempenho dos algoritmos não apenas pelo critério ‘precisão na classificação’ dos estudantes, mas, também, pela ‘sensibilidade’ e ‘especificidade’, critérios subsidiados por métodos estatísticos. A análise dos resultados do primeiro experimento foram similares e evidenciaram, levando em consideração os critérios de precisão e sensibilidade, que o algoritmo Naïve Bayes obteve os melhores desempenhos; já com relação ao critério especificidade, o algoritmo SMO foi o mais bem sucedido.

A análise dos resultados do segundo experimento, que foi realizado com um conjunto de treinamento de 28 instâncias, mostrou que o algoritmo Naïve Bayes obteve

o melhor desempenho em todos os critérios (precisão, sensibilidade e especificidade), para ambos os testes estatísticos utilizados (ANOVA e Post-hoc).

As comparações realizadas entre os algoritmos evidenciaram empiricamente que o Naïve Bayes é o mais apropriado para ser utilizado em uma ferramenta de suporte a tutores. Obteve o melhor desempenho nos três critérios do segundo experimento e em dois critérios do primeiro. Além disso, tem como vantagem o pouco tempo (comparativamente) envolvido nas fases de treinamento e teste.

2.3 Previsão de Desempenho Discente em Ambiente Virtual de Aprendizagem (AVA)

Esta seção aborda o trabalho realizado por Gottardo e colaboradores [Gottardo *et al.* 2012b], com foco em previsão de desempenho de alunos em cursos a distância, realizados via Ambiente Virtual de Aprendizagem (AVA); o AVA em questão foi implementado via *Moodle*.

O principal objetivo dos pesquisadores ao conduzir tal trabalho foi o de investigar a viabilidade de obtenção de informações sobre desempenho de alunos, que possibilitassem o monitoramento do desempenho escolar desde a fase inicial do processo ensino-aprendizagem, com vistas à tomada de ações pró-ativas. De acordo com os autores, os resultados obtidos foram promissores, com inferências feitas pelo sistema com precisões em torno de 75%, mesmo em períodos iniciais do curso.

O principal interesse ao conduzir essa revisão está relacionado à identificação dos critérios utilizados pelos pesquisadores para identificar tanto as disciplinas quanto os dados considerados relevantes para tal previsão.

2.3.1 Descrição dos Critérios para Escolha da Disciplina e Dados Utilizados

O trabalho teve por foco uma única disciplina, que foi escolhida utilizando os seguintes critérios:

(1) Maior número de alunos que concluíram a disciplina;

(2) Maior número de oferta da disciplina para turmas diferentes;

(3) Disponibilidade dos resultados de avaliações dos alunos;

(4) Maior número de recursos do AVA utilizados.

A disciplina escolhida (não é citada) teve um total de 140 alunos concluintes, em duas turmas diferentes. Essa quantidade não abrange alunos desistentes, tendo em vista a indisponibilidade de resultados intermediários ou nota final.

Os atributos que descrevem um aluno foram escolhidos de acordo com modelo proposto em [Gottardo *et al.* 2012a] e são apresentados e descritos na Tabela 2.7 agrupados em três categorias: (1) Perfil de uso do AVA; (2) Interação Estudante-Estudante e (3) Interação Estudante-Professor. A última linha da tabela é a relativa ao atributo classe (no contexto de técnicas de AM), atributo cujo respectivo valor indica a qual classe o aluno em questão pertence.

No estudo de caso em questão, a classe é representada pela nota final obtida pelo aluno. Com vistas a viabilizar um tratamento mais abrangente, por técnicas de MD, o valor contínuo do atributo classe foi discretizado. Por meio da análise das notas das turmas selecionadas, foi observado que seguiam uma distribuição normal, em que a média era 82, com desvio padrão de 5, a menor nota igual a 67 e a maior 97.

Os autores optaram por dividir os alunos em três classes, tendo como critério a nota obtida. Assim sendo, foram definidas as classes: B, a primeira classe a ser definida, agrupando alunos com desempenho intermediário, *i.e.*, aqueles com nota final entre 77 e 88; classe C, agrupando alunos com notas inferiores a 77 e finalmente, classe A, alunos com notas superiores a 88. A distribuição de alunos por classe resultou em: A: 16; B: 109 e C: 15.

Tabela 2.7 Atributos utilizados para representação de alunos, agrupados nas três dimensões sugeridas em [Gottardo *et al.* 2012a].

Dimensão 1 – PERFIL GERAL DO USO DO AVA	
ATRBUTO	DESCRIÇÃO
(1.1) nr_acessos	Nro. total de acessos ao AVA
(1.2) nr_post_foruns	Nro. total de postagens em fóruns
(1.3) nr_post_resp_foruns	Nro. total de respostas postadas em fóruns referindo-se a postagem de outros participantes
(1.4) nr_post_rev_foruns	Nro. total de revisões em postagens anteriores realizadas em fóruns
(1.5) nr_sessao_chat	Nro. de sessões de chat que o estudante participou
(1.6) nr_msg_env_chat	Nro. de mensagens enviadas ao chat
(1.7) nr_questoes_resp	Nro. de questões respondidas
(1.8) nr_questoes_acert	Nro. de questões respondidas corretamente
(1.9) freq_media_acesso	Frequência que o estudante acessa o AVA
(1.10) tempo_medio_acesso	Tempo médio de acesso ao sistema
(1.11) nr_dias_prim_acesso	Nro. de dias transcorridos entre o início do curso e o primeiro acesso ao AVA
(1.12) tempo_total_acesso	Tempo total conectado ao AVA
Dimensão 2 – INTERAÇÃO ESTUDANTE-ESTUDANTE	
ATRBUTO	DESCRIÇÃO
(2.1) nr_post_rec_foruns	Nro. de postagens do estudante que tiveram respostas feitas por outros estudantes
(2.2) nr_post_resp_foruns	Nro. de respostas que o estudante realizou
(2.3) nr_msg_rec	Nro. de mensagens recebidas de outros participantes durante a realização do curso
(2.4) nr_msg_env	Nro. de mensagens enviadas a outros participantes durante a realização do curso
Dimensão 3 – INTERAÇÃO ESTUDANTE-PROFESSOR	
ATRBUTO	DESCRIÇÃO
(3.1) nr_post_resp_prof_ foruns	Nro. de postagens de estudantes que tiveram respostas feitas por professores
(3.2) nr_post_env_prof_ foruns	Nro. de postagens de professores que tiveram respostas feitas por estudantes
(3.3) nr_msg_env_prof	Nro. de mensagens enviadas ao professor/tutor durante a realização do curso
(3.4) nr_msg_rec_prof	Nro. de mensagens recebidas do professor/tutor durante a realização do curso
OBJETIVO DA PREVISÃO	
ATRBUTO	DESCRIÇÃO
(4.1) resultado_final	Nota final obtida pelo estudante. Representa a classe objetivo da técnica de classificação.

2.3.2 Metodologia e Experimentos Realizados

Para predizer o desempenho de alunos foram utilizados dois algoritmos de MDE, o *RandomForrest* e o *MultilayerPerceptron* [Witten, Frank & Hall 2011]. Tendo por base o conjunto de 140 registros de dados descritos pelos atributos listados na Tabela 2.7, dois experimentos (referenciados como (I) e (II) no que segue) foram conduzidos, com o objetivo de avaliar as possibilidades de inferir sobre o desempenho (final) de alunos na disciplina. Para a avaliação dos resultados dos experimentos foi usada a acurácia, medida como a proporção entre o número de alunos corretamente classificados pelos algoritmos, em sua respectiva classe e o número total de alunos considerados no estudo. Em cada um dos experimentos foi empregada a 10-validação cruzada e a avaliação de cada algoritmo calculada como a média aritmética das 10 execuções.

No experimento (I) o período total do curso foi dividido em 3, considerando que o período total da disciplina é composto por três módulos sequenciais. O conjunto de dados relativo a cada módulo continha informações da data de início e data de fim do módulo. A Tabela 2.8 apresenta os dados do experimento (I), levando em consideração os dados obtidos relativos a cada um dos módulos. Os resultados mostrados na tabela evidenciam que o classificador induzido pelo algoritmo *RandomForrest*, por exemplo, com o conjunto de dados relativos ao Módulo 1, classificou aproximadamente 107 (76,2% ± 4,7) dos 140 estudantes na classe correta (A, B ou C).

Tabela 2.8 Experimento I: Acurácia média e desvio padrão das 100 execuções para os dois algoritmos [Gottardo *et al.* 2012b].

Classificador	Módulo 1	Módulo 2	Módulo 3
RandomForest	76,2 % ± 4,7	77,8 % ± 2,1	77,2 % ± 2,9
MultilayerPerceptron	76,5 % ± 4,4	77,8 % ± 2,1	77,2 % ± 2,9

Para o experimento (II) o período da disciplina foi dividido em 6 partes iguais. A divisão não levou em consideração datas de início e fim de módulos. A Tabela 2.9 mostra os resultados do experimento em cada uma das partes. Os valores apresentados são o percentual de acurácia médio e o desvio padrão de 100 execuções/algoritmo. Uma inspeção da tabela permite identificar que o classificador *MultilayerPerceptron*, por exemplo, classificou aproximadamente 104 (74,4% ± 6,2) dos 140 estudantes na classe correta (A, B ou C).

Tabela 2.9 Experimento II: Acurácia média e desvio padrão das 100 execuções para os dois algoritmos [Gottardo *et al.* 2012b]. RF: RandomForest e MP: MultilayerPerceptron.

ALG.	Períodos de Tempo					
	1	2	3	4	5	6
RF	75,2 % ± 5,7	75,2 % ± 7,0	81,5 % ± 5,9	77,2 % ± 2,9	77,2% ± 2,9	77,2 % ± 2,9
MP	74,4 % ± 6,2	74,0 % ± 6,7	81,3 % ± 6,1	77,2 % ± 2,9	77,2 % ± 2,9	77,2 % ± 2,9

Os resultados apresentados nas Tabelas 2.8 e 2.9 evidenciam que não existe uma grande variação entre os percentuais em cada módulo do experimento I ou período de tempo do experimento II.

A técnica "T pareado" foi utilizada para testar a significância estatística dos resultados. Este teste possuí nível de significância de 5%. O teste foi realizado no ambiente WEE – *Weka Experiment Enviroment*. A partir do resultado do teste, baseando-se no módulo I do experimento 1 e no período 1 do experimento II, os autores observaram que:

- Não existe diferença estatisticamente significativa entre o módulo 1 e os demais módulos, com relação às taxas de acurácia obtidas no experimento I. O resultado se repete para ambos os algoritmos.
- No experimento II, apenas o período 3 apresenta diferença estatisticamente significativa em relação ao período 1. Em ambos algoritmos utilizados no experimento II, este resultado se repete.

Os resultados obtidos nos experimentos mostraram que é possível inferir o desempenho de estudantes com taxas de acurácia acima de 74%, mesmo em períodos iniciais do curso. Estes dados, se disponibilizados aos professores durante o desenvolvimento da disciplina, são uteis para que estratégias de acompanhamento individual sejam aplicadas aos alunos em questão, com o objetivo de minimizar reprovações.

Um outro trabalho de Gottardo e colaboradores[Gottardo *et al.* 2012a] tem também como foco a avaliação de desempenho de alunos em cursos a distância, usando MD e, particularmente, investe na definição de um conjunto de atributos, amplo e generalizável, para a descrição dos dados de alunos, com vistas à inferência de desempenho discente. É nesse trabalho que as três categorias de dados, nomeadas

dimensões, comentadas na Seção 2.3.1 (Tabela 2.7) são propostas, voltadas a disciplinas ministradas via AVA, cada uma delas descrita como segue.

(1) *Primeira Dimensão*: Contempla o perfil geral do uso do ambiente. O objetivo aqui é identificar dados que representem aspectos do planejamento, organização e gestão do tempo do estudante. E outros atributos que representam atividades rotineiras.
(2) *Segunda Dimensão*: Representa a interação estudante-estudante. Pretende-se verificar se existe interação entre os estudantes utilizando as ferramentas disponíveis no ambiente, como fóruns, chat, envio ou recebimento de mensagens.
(3) *Terceira Dimensão*: Representa a interação estudante-professor. Pretende-se averiguar como professores e tutores interagem com estudantes em ambientes virtuais.

Já os três experimentos (I, II e III) apresentados e discutidos em [Gottardo *et al.* 2012a], foram realizados com o objetivo de verificar a adequação do conjunto de atributos proposto (Tabela 2.7). Para a coleta de dados os autores usaram o mesmo critério descrito no início da Seção 2.3.1 para a seleção de uma disciplina. A disciplina em questão tinha um total de 155 alunos concluintes, em quatro turmas distintas. Para os experimentos, o último atributo da Tabela 2.7 (classe) foi novamente discretizado, usando o método *equal-width* disponibilizado pela ferramenta *Weka* [Witten, Frank & Hall 2011]. Para o número de intervalos foi escolhido o valor 3 e, portanto, três classes foram criadas, com a seguinte distribuição (alunos/classe): A: 22; B: 109 e C: 24. Os algoritmos usados foram os mesmos descritos na Seção 2.3.1 i.e., *RandomForest* e *MultilayerPerceptron*.

No experimento I, o conjunto de dados foi descrito utilizando todos os atributos da Tabela 2.7; o último atributo (classe), foi discretizado como discutido na Seção 2.3.1.

No experimento II foram utilizados todos os atributos da Tabela 2.7, discretizados (usando o algoritmo padrão do Weka). A principal justificativa para a discretização se deve ao fato de alguns algoritmos trabalharem melhor com dados discretos.

O experimento III foi realizado usando apenas um subconjunto dos atributos da Tabela 2.7. A seleção dos atributos relevantes considerados foi feita usando o algoritmo

disponibilizado via Weka. O conjunto de atributos selecionado foi: {nr_acessos, nr_msg_rec_prof, nr_sessao_cht, nr_msg_env_chat}.

Na Tabela 2.10 são apresentadas as médias da acurácia e do desvio padrão de 100 execuções para cada experimento. O teste estatístico T pareado, foi utilizado para verificar a significância estatística dos valores encontrados nos experimentos. Tomando como base o experimento II, e utilizando os resultados do teste estatístico, observou-se que:

- Os resultados de ambos classificadores no experimento II foram, respectivamente, superiores àqueles obtidos pelos classificadores nos experimentos I e III.
- Os resultados do classificador induzido com o *MultilayerPerceptron*, no experimento II, foram superiores àqueles obtido pelo mesmo algoritmo no experimento I porém não apresentaram diferença estatisticamente significativa para o algoritmo *RandomForest.*

Tabela 2.10 Acurácia média e desvio padrão de 100 execuções dos classificadores nos experimentos [Gottardo *et al.* 2012a].

Classificador	Experimento I	Experimento II	Experimento III
RandomForest	70,6 % ± 9,9	76,6 % ± 8,5	64,9 % ± 10,8
MultilayerPerceptron	66,3 % ± 10,5	76,2 % ± 8,8	69,4 % ± 8,8

2.4 Previsão de Desempenho Discente Usando Sistemas de Recomendação e Acoplamento de Classificadores

De uma maneira geral um Sistema de Recomendação pode ser abordado como um sistema que, a partir de sua interação com um determinado usuário, infere um perfil de interesses do usuário em questão e, em próximas interações, além de refinar tal perfil, usualmente oferece sugestões que, potencialmente, poderiam ser de interesse do usuário. Esta seção focaliza o trabalho descrito em [Gotardo *et al.* 2013], que faz uso de um sistema de recomendação, articulado a um conjunto de classificadores, para a predição de desempenho acadêmico de alunos de curso a distância.

2.4.1 Sobre Sistemas de Recomendação

Como apontado em [Adomavicius & Tuzhilin 2005], sistemas de recomendação emergiram como uma área de pesquisa independente em meados dos anos 90, quando pesquisadores se voltaram a problemas de recomendação que explicitamente dependiam de uma estrutura de avaliação (*rating*).

Em sua abordagem mais simplista, um problema de recomendação se resume ao problema de estimar avaliações para itens que ainda não foram 'vistos' pelo usuário. Intuitivamente, essa estimativa é baseada em avaliações prévias feitas pelos usuários, sobre outros itens. A partir do momento em que se consegue estimar avaliações para itens ainda não avaliados, itens que tenham avaliações com altas estimativas podem ser recomendados ao usuário. Segue uma abordagem formal do problema, como proposta em [Adomavicius & Tuzhilin 2005]. Considere:

U: conjunto de todos os usuários e

I: conjunto de todos os possíveis itens que podem ser recomendados.

Considere a definição de uma função *utilidade* que mede a utilidade de um item *i* para um usuário *u*, ou seja, *utilidade*: $U \times I \rightarrow R$, em que R é um conjunto totalmente ordenado (*e.g.*, inteiros não negativos). Para cada usuário $u \in U$ deseja-se escolher um item $i^* \in I$ que maximiza a função *utilidade* daquele usuário, como formalmente dada pela Eq. (2.1).

$$\forall\, u \in U,\ i_u^* = \text{argmax}_{i \in I}\ \textit{utilidade}(u,i) \qquad (2.1)$$

É importante lembrar que:

(1) em algumas aplicações ambos os conjuntos, U e I, podem ser vastos, da ordem de milhares ou mesmo milhões de usuários ou itens, respectivamente.

(2) em sistemas de recomendação a utilidade de um item é, usualmente, representada por uma avaliação, a qual indica como um determinado usuário avaliou um determinado item. Por exemplo, Maria Silva avaliou o livro Memórias de Adriano com o valor 9, extraído do conjunto de possíveis valores de avaliação {0, 1, 2, 3, 4, 5, 6, 7, 8, 9, 10}, em que o valor 0 se refere à pior avaliação possível e 10 à melhor.

(3) em geral, a função utilidade pode ser uma função arbitrária, incluindo uma função de lucro. Na dependência da aplicação a função utilidade pode ser especificada pelo usuário, como frequentemente é feito para avaliações definidas pelo usuário ou, então, calculada pela aplicação, como pode ser o caso da função de utilidade baseada em lucro.

(4) cada elemento do conjunto de usuários U pode ser definido por um perfil, que pode incluir várias características do usuário, tais como: idade, sexo, estado civil, etc. O caso mais simples é aquele em que o usuário é definido apenas por uma única identificação (i.e., o *userid*).

(5) como em (4), os itens do conjunto I dos itens também podem ser representados não apenas por um identificador, mas também, por meio de um conjunto de suas características.

O problema central de sistemas de recomendação está na função *utilidade* que, via de regra, não está definida em todo seu domínio, i.e. $U \times I$, mas sim em apenas um seu subconjunto. Isso implica a necessidade de extrapolar *utilidade* para o domínio todo i.e., $U \times I$.

Em sistemas de recomendação a utilidade é tipicamente representada por avaliações e é inicialmente definida apenas com relação àqueles itens previamente avaliados pelos usuários. Em um sistema de recomendação de livros, por exemplo, usuários inicialmente avaliam algum subconjunto de livros que já leram. A Tabela 2.11 ilustra um possível subconjunto, em que avaliações são extraídas do conjunto {1,2,3,4,5} e o caractere hífen '-' indica que o usuário não leu um determinado livro. Um sistema de predição, portanto, deve ser capaz de estimar (predizer) as avaliações das combinações livros-não-avaliados/usuários e emitir recomendações com base nessas predições.

Tabela 2.11 Uma parte da matriz de avaliações de um sistema de recomendação de livros, em que avaliação $\in$ {1,2,3,4,5} e '–' significa não ter sido avaliado ainda.

Usuário	LIVROS				
	Memórias de Adriano	A Rainha Vermelha	O Perfume	Outlander	Tempo entre Costuras
Ana	5	4	–	3	2
Maria	4	–	3	2	4
Pedro	–	3	4	3	–
Alice	4	4	3	4	3
José	3	4	2	1	3

Extrapolações feitas a partir de avaliações conhecidas, para avaliações desconhecidas, geralmente são baseadas: (1) na especificação de heurísticas que definem a função *utilidade* e empiricamente validam o desempenho de tal função; (2) na estimativa da função *utilidade* que otimiza certos critérios de desempenho, tal como o erro médio quadrático. Uma vez que as avaliações desconhecidas são estimadas, as recomendações de um item a um usuário são feitas por meio da seleção da avaliação com o maior valor entre todas as avaliações estimadas para aquele usuário, de acordo como (1). Alternativamente, podem ser recomendados os N melhores itens a um usuário.

As avaliações de itens ainda não avaliados podem ser estimadas de diferentes maneiras, usando algoritmos de AM, teoria da aproximação e, também, várias heurísticas. Como comentado em [Adomavicius & Tuzhilin 2005], dependendo das estratégias usadas para construir novas recomendações, os sistemas de recomendação são classificados naqueles que fazem:

(1) *recomendações baseadas em conteúdo*: ao usuário serão recomendados itens similares àqueles que preferiu em situações anteriores. Em métodos de recomendação baseados em conteúdo, a função *utilidade(u,i)* do item *i* para o usuário *u* é estimada com base nos valores atribuídos pelo mesmo usuário *u* de *utilidade(u,*k_j*)*, para aqueles itens $k_j \in$ I que forem similares ao item *i*.

(2) *recomendações colaborativas*: ao usuário serão recomendados itens que foram recomendados a pessoas com gostos e preferências similares à do usuário. Sistemas de recomendação colaborativos buscam predizer a utilidade de itens para um determinado usuário, com base em itens previamente avaliados por outros usuários. Formalmente, o valor de *utilidade(u,i)* do item *i* para o usuário *u* é estimado com base nos valores de *utilidade(*u_j*,i)*, em que $u_j \in$ U são aqueles usuários que são 'similares' ao usuário *u*.

(3) *recomendações híbridas:* são baseadas em recomendações obtidas via combinação de abordagens baseadas em conteúdo e abordagens colaborativas. As diferentes maneiras de combinar ambas abordagens podem ser:

(a) implementando métodos baseados em conteúdo e métodos colaborativos separadamente e combinando suas predições;
(b) incorporando algumas características baseadas em conteúdo em uma abordagem colaborativa;

(c) incorporando algumas características colaborativas em uma abordagem baseada em conteúdo e

(d) construindo um modelo unificado que incorpora características colaborativas e baseadas em conteúdo.

Em [Ricci *et al.* 2010] três novos tipos de sistemas de recomendação foram mencionados: o *demográfico*, baseado em informações demográficas do perfil do usuário; o *baseado em conhecimento* que usa casos anteriores e o *baseado em comunidades*, que utiliza informações sobre relações entre participantes de comunidades virtuais. Nessa mesma referência, os autores chamam de *híbridos* aquele sistema de recomendação que agrega características de mais de um tipo de tais sistemas.

2.4.2 Articulação de um Sistema de Recomendação com Acoplamento de Classificadores

A pesquisa realizada por Gotardo em [Gotardo *et al.* 2013] usa uma variante da abordagem colaborativa, em que dados demográficos referentes a alunos são também utilizados. A motivação para o uso da abordagem foi, principalmente, a de suprir a falta de dados existente no estudo de caso abordado, considerando que a proposta tinha foco, também, no uso de algoritmos de aprendizado supervisionado.

No estudo de caso os dados referentes a um aluno são, além de seu *userid*, um conjunto de descritores sobre o seu comportamento no uso de um ambiente virtual de aprendizagem (implementado via *Moodle*) com relação à uma disciplina; o valor de avaliação de um usuário é dado pela aprovação ou não na referida disciplina. Os dados são referentes a 1 disciplina de 1 turma, com 252 alunos matriculados em 5 pólos (contendo dados inclusive de alunos desistentes e com a matrícula cancelada). Para aumentar o tamanho da base de dados disponível os autores fizeram uso de *bootstrapping*, que implica o conjunto de treinamento ser incrementado com a escolha das melhores instâncias classificadas do conjunto de teste.

Um efeito colateral no uso do *bootstrapping* é a propagação de erros aos dados rotulados (com a inserção dos novos dados rotulados que, muitas vezes, não estão descritos usando o conjunto total de descritores). Com vistas a minimizar tal efeito, foram usados mais de um algoritmo funcionando em conjunto e rotulando os dados de outro algoritmo, de forma cruzada. Segundo os autores, esse acoplamento de algoritmos é

baseado no conceito de *co-training* [Blum & Mitchell 1998], e permite um melhor desempenho. Foram usados o J48, o Naïve Bayes, a combinação J48+Naïve Bayes e a combinação Naïve-Bayes+Naïve-Bayes. O processo de aprendizado por acoplamento articulado a um sistema de recomendação é resumidamente descrito pelos autores do trabalho como:

(1) O conjunto de treinamento (composto por dados históricos, já rotulados, referentes a alunos que já tinham cursado a disciplina) é dividido entre os classificadores;
(2) Individualmente cada classificador gera um modelo e classifica as instâncias não rotuladas (dados referentes a alunos que recentemente ingressaram na disciplina) (conjunto de testes);
(3) Os dados rotulados são ordenados de acordo com a probabilidade de acerto dada pelo classificador Bayesiano;
(4) As instâncias melhores ranqueadas são escolhidas para serem inseridas no conjunto de treinamento do outro classificador (ou seja, um classificador insere as melhores instâncias no outro, e vice-versa);
(5) O processo é repetido a cada execução e a base aumenta o seu tamanho.

A análise dos resultados obtidos permitiu aos autores do trabalho concluir que o uso do acoplamento para gerar recomendações sobre o desempenho do aluno, incrementa a base inicial de dados (conjunto de treinamento) e oferece recomendações com dados próximos àqueles de uma base de dados real; resultados que não seriam alcançados usando somente o *bootstrapping*.

2.5 Previsão de Desempenho de Alunos de Pós-Graduação Usando Técnicas de Aprendizado de Máquina

Esta seção aborda brevemente o trabalho descrito em [Koutina & Kermanidis 2011] que, diferentemente dos anteriores, cujo foco são alunos de graduação, examina o desempenho de alunos de pós-graduação. O objetivo, entretanto, é o mesmo daquele dos trabalhos considerados anteriormente, i.e., predição de desempenho de alunos, usando algoritmos de aprendizado automático e, no caso, com particular interesse em evidenciar o algoritmo de AM mais eficiente.

2.5.1 Descrição dos Dados Utilizados

Os dados utilizados nos experimentos foram extraídos junto ao Programa de Pós-graduação da Ionian University (PPIU), Grécia, iniciado em 2009, chamado Diploma de Pós-Graduação em Especialização em Informática. O processo de seleção de candidatos para ingresso no Programa aceitou candidatos com graduações em qualquer área (*e.g.* Psicologia, Física, Economia, etc.). O principal objetivo do Programa foi o de formar alunos em nível de pós-graduação, em áreas de conhecimento especializadas e, também, em pesquisa, de maneira a capacitá-los na condução de pesquisa científica e na execução de tarefas envolvendo Tecnologia da Informação.

As instâncias de dados utilizadas nos experimentos são relativas aos alunos/desempenhos em três disciplinas do primeiro semestre de 2009-2010, do PPIU, como mostra a Tabela 2.12 e duas disciplinas do primeiro semestre de 2010-2011.

Tabela 2.12 Disciplinas do PPIU consideradas na pesquisa, em que RD: número de registro de dados.

DISCIPLINAS	SIGLA	RD
Tecnologia de Linguagem Avançada 2009-2010	TLA1	11
Redes de Computadores 2009-2010	RC1	35
Gerenciamento de Sistemas de Informação 2009-2010	GSI	35
Tecnologia de Linguagem Avançada 2010-2011	TLA2	8
Redes de Computadores 2010-2011	RC2	28
TOTAL (registro de dados)		117

Em algumas das disciplinas era esperado do estudante entregar um trabalho intermediário, enquanto em outras, o estudante deveria entregar mais de um trabalho intermediário. É importante ressaltar que alguns trabalhos entregues, deveriam ser realizados em grupos de alunos, enquanto outros, não. A quantidade de trabalhos por disciplina, não foi definida na pesquisa, nem o esquema de avaliação.

O padrão adotado para lidar com os dados foi o mesmo daquele utilizado em [Kotsiantis *et al.* 2003], discutido na Seção 2.2 deste capítulo. Os atributos que descrevem os conjuntos de dados, foram divididos em dois subconjuntos disjuntos: (1) aquele referente às informações gerais chamados dados demográficos (DD) de cada estudante e (2) aquele referente às informações do desempenho individual (D_IND) em cada disciplina, e o conceito final, como mostrados na Tabela 2.13.

Tabela 2.13 Conjuntos de atributos (DD & D_IND) que descrevem cada registro de dado e respectivos valores que cada um dos atributos pode assumir.

ATRIBUTOS ASSOCIADOS AOS DADOS DEMOGRÁFICOS (DD)	
NOME DO ATRIBUTO	VALORES POSSÍVEIS
(1) Sexo	masculino, feminino
(2) Grupo da Idade	a) [21-25] b) [26-30] c) [31-35] d) [36- ..]
(3) Estado Civil	solteiro, casado
(4) Número de filhos	0, 1, 2, mais
(5) Ocupação	não, parcial, integral
(6) Trabalho associado com computação	não, sim
(7) Graduação	Universidade, Técnica Instituto Educacional
(8) Outro Mestrado	não, sim
(9) Alfabetizado Computacionalmente	não, sim
(10) Graduação em Informática	não, sim

ATRIBUTOS ASSOCIADOS AO DESEMPENHO INDIVIDUAL (D_IND)		
NOME DO ATRIBUTO	SIGLA	VALORES POSSÍVEIS
(11) 1º. trabalho escrito	TE1	0-10
(12) 2º. trabalho escrito	TE2	0-10
(13) 3º. trabalho escrito	TE3	0-10
(14) Presença em sala de aula	PSA	não, médio, bom
(15) Nota Final	NF	ruim, bom, muito bom

Os dados demográficos (DD) foram extraídos de um questionário aplicado aos alunos do PPIU, e os dados de desempenho individual (D_IND) foram coletados por tutores com base nas notas dos trabalhos escritos de cada estudante e em sua presença ou não em sala de aula. Os resultados da nota final foram discretizados nos três grupos:

(1) *ruim*: nota $\in$ [0,0 4,9];

(2) *bom*: nota $\in$ [5,0 7,9];

(3) *muito bom*: nota $\in$ [8,0 10];

2.5.2 Algoritmos de Aprendizado de Máquina Utilizados

Os algoritmos utilizados nos experimentos estão listados na Tabela 2.14; as implementações usadas foram aquelas disponibilizadas pelo ambiente WEKA (*Waikato Environment for Knowledge Analysis*) [Weka 2012].

Tabela 2.14 Algoritmos utilizados nos experimentos descritos em [Koutina & Kermanidis 2011].

ALGORITMO/SISTEMA	SIGLA	Comentários
(1) C4.5	J48	Executado com o parâmetro de poda *on* e *off* e outros parâmetros com seus respectivos valores *default*.
(2) k-NN	k-NN	Executado nas versões 1-NN, 3-NN e 5-NN e parâmetros com seus respectivos valores *default.*
(3) Naïve Bayes	NB	Conceito aprendido representado como uma rede Bayesiana [Domingos and Pazzani 1997].
(4) RIPPER	J-Rip	O algoritmo RIPPER (*Repeated Incremental Pruning to Produce Error Reduction*) [Cohen 1995] induz a representação do conceito como um conjunto de regras de decisão. Esse algoritmo está implementado sob o Weka como J-Ripp.
(5) Random Forest	RF	O Random Forrest induz a representação do conceito pelo resultado de várias árvores de decisão aleatórias.
(6) SVMs (*Support Vector Machines*)	SMO	Usa o algoritmo SMO (*Sequential Minimal Optimization*) para treinamento e uma função kernel polinomial.

De acordo com as autoras, a pesquisa contribuiu com os seguintes aspectos:

(1) avaliação de alguns dos mais conhecidos algoritmos de AM, usados em um contexto de MDE, para prever o desempenho de estudantes do programa de mestrado, utilizando dados demográficos bem como desempenhos e comportamentos em sala de aula.

(2) avaliação dos AM e análise da variação.

(3) tratamento de classes desbalanceadas em dados de treinamento por meio do uso da função (filtro) *resample*, disponibilizado pelo Weka.

Durante a execução dos testes as autoras perceberam que as classes estavam desbalanceadas, problema que geralmente interfere com a qualidade do conceito aprendido. Como já mencionado, os conjuntos de instâncias são individuais, o menor possuí apenas 8 instâncias e os maiores possuem 35. O desbalanço também acontece com relação ao atributo nota final, em que a quantidade de instâncias com notas entre zero e

quatro e, entre oito e dez, é menor o de instâncias com notas entre cinco e sete. Isso de certa forma interfere com a precisão de classificação associada às classes minoritárias.

2.5.3 Descrição da Metodologia Aplicada nos Experimentos

O treinamento dos algoritmos foi dividido em cinco passos consecutivos.

(1) O primeiro passo realizou o treinamento com os dados demográficos, junto com os dados de desempenho individual, e com a classe resultante, para todos os conjuntos de dados. A Tabela 2.15 mostra a precisão das previsões ao treinar os dados iniciais.

Tabela 2.15 Precisão (%) considerando individualmente os dados relativos a cada disciplina.

Disciplinas	J48 (com poda)	J48 (sem poda)	1-NN	3-NN	5-NN	NB	RF	SMO	J-Rip
TLA1	54,54	54,54	72,72	63,63	54,54	72,72	72,72	72,72	54,54
RC1	57,41	40,00	54,28	62,85	62,85	71,42	57,14	71,42	51,42
GSI	51,42	54,28	57,14	60,00	57,14	60,00	51,42	51,42	51,42
TLA2	25,00	25,00	25,00	37,50	25,00	37,50	37,50	25,00	25,00
RC2	57,14	60,71	60,71	50,00	53,57	60,71	57,14	57,14	67,82

(2) No segundo passo os dados foram pré-processados por meio da aplicação da função *resample*, do Weka. O uso dessa função tenta estabelecer um equilíbrio entre as classes, para minimizar o problema causado pelo desbalanceamento entre classes. A Tabela 2.16 mostra a precisão das previsões em conjuntos com os dados pré-processados pela função *resample.*

Tabela 2.16 Precisão (%) com os dados iniciais pré-processados pela *resample*.

Disciplinas	J48 (com poda)	J48 (sem poda)	1-NN	3-NN	5-NN	NB	RF	SMO	J-Rip
TLA1	63,63	81,81	90,90	54,54	45,45	72,72	90,90	72,72	54,54
RC1	65,71	71,42	85,71	80,00	74,28	80,00	88,51	82,85	71,42
GSI	80,00	82,85	85,71	60,00	42,85	85,71	88,57	82,85	80,00
TLA2	62,50	62,50	87,50	62,50	67,50	87,50	87,50	87,50	37,50
RC2	82,14	82,14	100,00	64,28	67,85	85,71	96,42	89,28	71,42

(3) No terceiro passo foi utilizado o algoritmo *OneR* [Forman & Cohen 20014] para avaliar os atributos, identificando aqueles que têm o maior impacto sobre a classe, em todos os conjuntos de dados. A contribuição do OneR, quando do uso

do Naïve Bayes, Nearest Neighbor e SMO, é abordada no estudo descrito em [Wasikowski & Chen 2010].

Neste passo foi identificado que atributos tais como Graduação em Informática, Presença em Sala de Aula, Sexo e Grupo da Idade tiveram grande impacto no valor da classe, enquanto outros, tais como Estado Civil e Outro Mestrado foram considerados redundantes.

(4) O quarto passo utilizou os atributos relevantes identificados em (3) e os algoritmos que obtiveram os melhores resultados de precisão (1) foram executados novamente, considerando os dados descritos apenas por atributos relevantes. A Tabela 2.17 mostra a precisão do J48 (sem poda), 1-NN, NB, *Random Forest* (RF) e SMO em todos os conjuntos de dados.

Tabela 2.17 Precisão (%) dos algoritmos com melhor desempenho no passo (1) (Tabela 2.15), usando os dados descritos pelos atributos relevantes, identificados no passo (3).

Disciplinas	J48 (sem poda)	1-NN	NB	RF	SMO
TLA1	54,54	72,72	81,81	63,63	72,70
RC1	48,57	57,14	74,28	57,14	68,57
GSI	48,57	51,42	62,85	54,28	62,85
TLA2	25,00	12,50	50,00	12,50	25,00
RC2	42,85	53,57	57,14	42,85	67,85

(5) No quinto e último passo foi aplicada a função *resample* combinada com os 7 atributos que mostraram grande influência para a determinação da classe, em todos os conjuntos de dados. Os resultados estão explícitos na Tabela 2.18.

Tabela 2.18 Precisão total (%) com os dados alterados pela função *resample*, descritos com os 7 atributos que se mostraram relevantes.

Disciplinas	J48 (sem poda)	1-NN	NB	RF	SMO
TLA1	63,63	90,90	90,90	90,90	90,90
RC1	65,71	71,42	82,85	74,28	60,00
GSI	68,57	80,00	80,00	80,00	74,28
TLA2	62,50	100,00	100,00	87,50	87,50
RC2	67,85	71,42	71,42	71,42	64,28

Com base nos resultados obtidos, uma das conclusões do trabalho é a de que a acurácia de previsão do J48 é bem mais baixa do que aquela do NB e 1-NN; essa constatação evidencia que, para pequenos conjuntos de dados, o NB e o 1-NN podem, eventualmente, obter melhores resultados que árvores de decisão. Uma outra constatação foi a de que a acurácia da predição, usando dados filtrados por *resample* apenas ou, então,

dados filtrados combinados com seleção de atributos, melhora quando o conjunto original é menor em volume de instâncias (TLA1, TLA2).

Comparando os valores da Tabela 2.15 com os da Tabela 2.16, as previsões induzidas usando os conjuntos com dados filtrados são significativamente mais precisas; Existe também uma melhoria significativa nas acurácias dos conceitos induzidos pelos NB, 1-NN, *Random Forest* e SMO. O conceito inferido utilizando o 1-NN com dados do TLA2 é de 25%, enquanto que, usando os dados filtrados via *resample*, sobe para 85,7%.

Os resultados da Tabela 2.17 evidenciam que o uso de dados descritos apenas pelos atributos relevantes se mostrou útil para o NB, mas não colaborou para a melhoria de acurácia dos demais classificadores. Adicionalmente, comparando os dados mostrados na Tabela 2.18 e Tabela 2.16, fica evidente que o NB é o único classificador que tem o desempenho melhorado com o processo de seleção de atributos. Considerando o conjunto de dados TLA2, a acurácia do classificador é melhorada usando o *resample* e seleção de atributos.

Capítulo 3
Ambiente Acadêmico, Curso, Disciplina e Descrição dos Dados

3.1 Descrição do Ambiente Acadêmico e Curso

O curso alvo desta pesquisa é denominado Bacharelado em Ciência da Computação (BCC), de um Centro Universitário situado na cidade de São Paulo. Está em funcionamento desde 01/02/2000 e possui reconhecimento do MEC por meio da Portaria número 2.116 de 16/07/2004 (Publicado em 19/07/2004). A última renovação do reconhecimento do curso está registrada na Portaria 286 de 21/12/2012 (publicada em 27/12/2012).

O BCC é oferecido em regime semestral, podendo ser concluído em um período mínimo de 8 semestres, e em um máximo de 12 semestres. As aulas são no período noturno, com o número máximo de 100 vagas anuais. O curso possuí carga horária total de 3.340 horas. A admissão se dá sempre por meio de processo seletivo público (vestibular). Em caso de vagas ociosas a admissão se estende para transferências, para interessados que sejam portadores de diplomas de ensino superior e para interessados em cursar disciplinas isoladas. Após o vestibular, o ingresso do aluno pode ser dar por meio do Programa Universidade para todos – PROUNI ou bolsa institucional.

As informações que a instituição de ensino armazena sobre os alunos do Centro Universitário estão listadas na Tabela 3.1. Na tabela tais informações estão divididas em dois grupos. O primeiro grupo agrega informações de identificação do aluno (tais como nome, CPF, etc.) combinadas com informações pessoais (tais como cor e religião); tais informações são referenciadas coletivamente como *Registro do Estudante* (informações identificadas pelos números (1) a (18) na parte superior da Tabela 3.1). O segundo grupo agrega informações direta ou indiretamente relacionadas ao curso e ao desempenho do aluno, desde que ingressa no curso, e são coletivamente referenciadas como *Registro de Dados Acadêmicos.*

Tabela 3.1 Informações discentes junto ao Centro Universitário.

REGISTRO DO ESTUDANTE	COMENTÁRIOS E/OU OPÇÕES DISPONÍVEIS
(1) Nome	
(2) Endereço	
(3) Telefone	
(4) E_mail	
(5) Data de nascimento	
(6) Idade	
(7) Sexo	
(8) Estado Civil	solteiro(a), casado(a), divorciado(a), viúvo(a).
(9) RG	
(10) CPF	
(11) Título de eleitor	
(12) RNE	para estrangeiros apenas.
(13) Cor	amarela, branca, indígena, negra, parda, não declarado.
(14) Religião	61 possíveis religiões disponíveis para escolha.
(15) Maior grau de instrução	fundamental incompleto, fundamental completo, médio completo, superior incompleto, superior completo, especialista, mestre, doutor.
(16) Nome do pai	
(17) Nome da mãe	
(18) Nome do responsável	
DADOS ACADÊMICOS	**COMENTÁRIOS E/OU OPÇÕES DISPONÍVEIS**
(1) Registro Acadêmico (RA)	
(2) Data de Conclusão Ensino Médio	
(3) Data do vestibular	
(4) Nota no vestibular	
(5) Classificação no vestibular	
(6) Data ingresso no curso	
(7) Data término do curso	
(8) Status na turma	Reserva de vaga, cursando, concluído.
(9) Lista de Disciplinas Cursadas	Cada disciplina cursada tem as seguintes informações a ela associadas: [Nome,Turma,Sem.,Ano,NroFaltas,Notas,Média].
(10) Histórico escolar	Aproveitamentos, disciplinas cursadas, disciplinas a cursar, pendências.

3.2 Descrição da Disciplina Construção de Algoritmos e Programação I (CAP1)

Dentre os cursos oferecidos pelo Centro Universitário, o curso de Bacharelado em Ciência da Computação foi o curso escolhido para a pesquisa, devido à maior familiaridade da autora deste trabalho com a sua grade curricular. Alunos ingressam no BCC via exame de seleção promovido pela instituição, duas vezes ao ano. Para alunos que seguem a matriz curricular vigente, a duração do curso é de 4 anos. Via de regra a população estudantil provêm de regiões relativamente próximas ao Centro Universitário, sendo um pequeno número dessa população dividido entre estudantes proveniente de outros estados e alguns poucos estrangeiros.

Como comentado na Seção 3.1 as informações sobre alunos do Centro Universitário, são armazenadas em dois tipos de registros: o *Registro do Estudante* e o *Registro de Dados Acadêmicos*, detalhados na Tabela 3.1.

A disciplina escolhida para a investigação do uso de técnicas de MDE para a predição de desempenho é a *Construção de Algoritmos e Programação I* (CAP1), oferecida no 1º semestre do curso de BCC, com 4 créditos, que correspondem a uma carga horária de 72h semestrais. A ementa da disciplina CAP1, de acordo com o Projeto Pedagógico do Curso de Ciência da Computação, do Centro Universitário, contempla os seguintes tópicos:

- Comandos sequenciais;
- Entrada e Saída;
- Comando condicional.
- Laços de Repetição: incrementais, interrupção no início e interrupção no final;
- Sub-rotinas.

A maneira como o docente responsável pela disciplina, em um determinado ano, realizou as avaliações de alunos é considerada, neste trabalho, um *esquema de avaliação*. Nos dados levantados e utilizados nos experimentos, foram identificados os esquemas listados na Tabela 3.2. Note que os diferentes esquemas foram identificados apenas com relação ao uso de determinados atributos para a composição do atributo avaliação final.

Tabela 3.2 Esquemas de avaliação aplicados durante o semestre, associados à disciplina CAP1 em que: VE – nota do vestibular, AV – instância de avaliação (*i.e.*, Prova), LE – instância de Lista de Exercícios, PI – Prova Interdisciplinar, e AS – Avaliação Substitutiva.

Esquema de Avaliação	Composto por	Ano em que foi Empregado
E_1	VE, AV_1, AV_2, AV_3, LE_1, LE_2, LE_3, LE_4	2009
E_2	VE, AV_1, AV_2, AV_3	2010
E_3	VE, AV_1, AV_2, AV_3, PI, AS_1, AS_2	2011
E_4	VE, AV_1, AV_2, AV_3, AV_4	2012
E_5	VE, AV_1, AV_2, AV_3, PI	2013 e 2014
E_6	VE, AV_1, AV_2, PI	2015

Como resultado do levantamento de dados referente à disciplina CAP1 e considerando as variáveis envolvidas na caracterização das ofertas de tal disciplina, bem como os esquemas de avaliação empregados nela, a nomenclatura referente aos conjuntos de dados disponíveis, adotada neste trabalho, visando uma melhor organização dos dados,

foi a de referenciar cada conjunto de dados por um nome composto da identificação da Disciplina, Ano e Esquema de Avaliação empregado no ano em questão.

Por exemplo, uma instanciação da nomenclatura utilizada para organizar os dados referentes aos alunos que cursaram a disciplina CAP1, quando foi oferecida em 2012, oferta na qual o esquema de avaliação E_4 foi utilizado, é CAP1_2012_E4. Na Tabela 3.3 estão listadas as informações pessoais de alunos, extraídas do *Registro de Estudante* (Tabela 3.1), que participaram dos experimentos conduzidos neste trabalho (ver Capítulo 5) e na Tabela 3.4 estão listadas as abreviações das informações do *Registro de Dados Acadêmicos* que participaram dos experimentos conduzidos neste trabalho (ver Capítulo 5). As abreviações listadas nas tabelas 3.3 e 3.4 são usadas nas tabelas que as seguem. A Tabela 3.5 mostra as fórmulas utilizadas para o cálculo de AF (Avaliação Final).

Tabela 3.3 Informações do *Registro de Estudante* e respectivas abreviações e possíveis valores, envolvidas nos experimentos.

Informações Pessoais	Abreviação	Valores
(1) Data de Nascimento (Ano)	DN	Quatro dígitos (ano válido).
(2) Estado Civil	ES	S(olteiro),C(casado),D(ivorciado),V(iúvo)
(3) Sexo	S	M(asculino),F(eminino)
(4) Grau de Instrução	GI	MC (Ensino Médio Completo) SI (Ensino Superior Incompleto) SC (Ensino Superior Completo) ES (Especialista) ME (Mestre) DO (Doutor)

Tabela 3.4 Convenções de Nomenclatura das informações *do Registro de Dados Acadêmicos* Relativas à disciplina CAP1, envolvidas nos experimentos.

Abreviação	Significado e valores
(1) VE	Nota do vestibular (de 0 a 10)
(2) AV_i (i=1, ... , 4)	Notas em instância de avaliação (de 0 a 10)
(3) LE_i (i=1, ..., 4)	Notas em instâncias de listas de exercícios (de 0 a 10)
(4) AS_i (i=1, 2)	Notas em instâncias de avaliação substitutiva (de 0 a 10)
(5) PI	Nota na prova interdisciplinar (de 0 a 1)
(6) F	Número de faltas na disciplina (de 0 a 72)
(7) LD	Nota na disciplina de Lógica e Matemática Discreta (de 0 a 10)
(8) FC	Nota na disciplina Física Computacional (de 0 a 10)
(9) LP	Nota na disciplina Laboratório de Programação (de 0 a 10)
(10) AF	Nota na avaliação final (valor entre 0 e 10)
(11) G	Informação artificialmente inserida indicando se o aluno foi Aprovado (A) ou Reprovado (R)

Tabela 3.5 Fórmula utilizada em cada ano para gerar o valor de AF.

ANO	Avaliação Final (AF) = Fórmula
(1) 2009	AF = ((AV_1*0,27)+(AV_2*0,31)+(AV_3*0,36)+((LE_1+LE_2+LE_3+LE_4)/4*0,06))
(2) 2010	AF = (AV_1+AV_2+AV_3)/3
(3) 2011	AF = (AV_1+AV_2+AV_3+AS_1-AS_2)/3 + PI
(4) 2012	AF = ((AV_1*30)+(AV_2*30)+(AV_3*30)+(AV_4*10))/100
(5) 2013	AF = ((AV_1 + AV_2 + AV_3)/3) * 0,9 + PI
(6) 2014	AF = ((AV_1*2) + (AV_2*2) + (AV_3*1))/5 * 0,9 + PI
(7) 2015	AF = ((AV_1 + AV_2)/2) * 0.9 + PI

Note que na Tabela 3.5, o atributo AF é o que efetivamente indica aprovação ou reprovação. A composição do valor de AF em qualquer dos esquemas de avaliação é composta por todas as avaliações aplicadas naquele ano, com mais alguma variação anual de atributos. No ano de 2009 as quatro listas são consideradas com um peso menor em relação às avaliações, já em outros quatro anos 2011, 2013, 2014 e 2015, a PI é considerada. Existem sete fórmulas diferentes para o cálculo do AF, as fórmulas 5 e 6 embora diferentes correspondem ao esquema E5, pois envolvem os mesmos atributos, isto é, AV1, AV2, AV3 e PI.

As sete tabelas, 3.6 – 3.12, exemplificam as informações contidas em um registro de dados, de um determinado ano, referente à disciplina CAP1, que serão consideradas para os experimentos com técnicas de AM, com vistas à predição de desempenho. Cada uma de tais tabelas identifica a informação e mostra um exemplo real, extraído dos dados coletados. O valor de G foi artificialmente inserido nos registros, de maneira a discretizar o valor de AF e facilitar a indução de classificadores.

Tabela 3.6 Informações utilizadas nos experimentos, relativas aos alunos do curso de BCC, disciplina Construção de Algoritmos e Programação 1 (CAP1), ano de 2009, na qual foi utilizada o Esquema de Avaliação E_1

CAP1_2009_E1																
DN	EC	S	GI	VE	AV_1	AV_2	AV_3	LE_1	LE_2	LE_3	LE_4	F	AF	LD	FC	G
1992	S	M	MC	3,0	7,0	4,5	5,5	10,0	10,0	10,0	10,0	0	6,0	6,0	6,0	A

Tabela 3.7 Informações utilizadas nos experimentos, relativas aos alunos do curso de BCC, disciplina Construção de Algoritmos e Programação 1 (CAP1), ano de 2010, na qual foi utilizada o Esquema de Avaliação E_2

CAP1_2010_E2												
DN	EC	S	GI	VE	AV_1	AV_2	AV_3	F	AF	LD	FC	G
1992	S	M	MC	5,5	8,0	7,5	7,0	0	7,5	8,0	7,0	A

Tabela 3.8 Informações utilizadas nos experimentos, relativas aos alunos do curso de BCC, disciplina Construção de Algoritmos e Programação 1 (CAP1), ano de 2011, na qual foi utilizada o Esquema de Avaliação E_3

CAP1_2011_E3															
DN	EC	S	GI	VE	AV_1	AV_2	AV_3	PI	AS_1	AS_2	F	AF	LD	FC	G
1984	S	M	MC	6,0	7,0	9,5	8,0	0	0	0	4	8,0	6,0	7,0	A

Tabela 3.9 Informações utilizadas nos experimentos, relativas aos alunos do curso de BCC, disciplina Construção de Algoritmos e Programação 1 (CAP1), ano de 2012, na qual foi utilizada o Esquema de Avaliação E_4.

CAP1_2012_E4													
DN	EC	S	GI	VE	AV_1	AV_2	AV_3	AV_4	F	AF	LD	FC	G
1992	S	M	MC	4,860	8,5	7,0	8,0	10,0	0	8,0	9,0	6,0	A

Tabela 3.10 Informações utilizadas nos experimentos, relativas aos alunos do curso de BCC, disciplina Construção de Algoritmos e Programação 1 (CAP1), ano de 2013, na qual foi utilizada o Esquema de Avaliação E_5.

CAP1_2013_E5													
DN	EC	S	GI	VE	AV_1	AV_2	AV_3	PI	F	AF	LD	FC	G
1978	C	M	MC	5,500	5,5	1,0	4,0	0,5	0	3,5	6,0	7,0	R

Tabela 3.11 Informações utilizadas nos experimentos, relativas aos alunos do curso de BCC, disciplina Construção de Algoritmos e Programação 1 (CAP1), ano de 2014, na qual foi utilizada o Esquema de Avaliação E_5.

CAP1_2014_E5													
DN	EC	S	GI	VE	AV_1	AV_2	AV_3	PI	F	AF	LD	FC	G
1996	S	M	MC	7,710	0,56	9,0	8,5	10,0	0	8,5	7,5	7,0	A

Tabela 3.12 Informações utilizadas nos experimentos, relativas aos alunos do curso de BCC, disciplina Construção de Algoritmos e Programação 1 (CAP1), ano de 2015, na qual foi utilizada o Esquema de Avaliação E_6.

CAP1_2015_E6												
DN	EC	S	GI	VE	AV_1	AV_2	PI	F	AF	LD	LP	G
1997	S	F	MC	-	7,5	2,5	0,5	0	5,0	7,5	7,0	R

A Tabela 3.13 fornece os números de alunos/ano que se inscreveram na disciplina CAP1; a tabela também traz o número dos que, por alguma razão, não terminaram a disciplina e o número efetivo de alunos que a finalizaram (tendo sido aprovados ou não).

Tabela 3.13 Alunos da disciplina CAP1 por ano (período: 2009–2015). TI: total de alunos inscritos, TD: total de alunos desistentes, TF: total de alunos que concluíram a disciplina (aprovados ou não), G=A: total de alunos aprovados e G=R: total de alunos reprovados.

ANO	TI	TD	TF	G=A	G=R
2009	62	12	50	40	10
2010	89	48	41	37	4
2011	56	13	43	31	12
2012	70	13	57	44	13
2013	78	18	60	32	28
2014	52	15	37	23	14
2015	98	29	69	41	28
TOTAL	**505**	**148**	**357**	**248**	**109**

É importante comentar que é requisito obrigatório que o estudante tenha registro de presença em 75% ou mais nas aulas de uma disciplina, para poder ser considerado para aprovação (na dependência da nota obtida). Um percentual maior que 25% implica reprovação independente da nota, o que corresponde a, no máximo, 18 faltas no semestre.

Capítulo 4
Algoritmos e Métodos Utilizados no Trabalho de Pesquisa

4.1 Introdução

No que segue são apresentadas breves descrições tanto dos algoritmos de MDE utilizados na parte experimental deste trabalho, quanto dos métodos empregados para a pré-seleção de atributos relevantes, dentre aqueles que descrevem os dados.

A Seção 4.2 aborda um classificador probabilístico, o Naïve Bayes [Hruschka & Nicoletti 2013]. A Seção 4.3 introduz árvores de decisão e comenta sobre o algoritmo ID3 [Quinlan 1991], que subsidiou o desenvolvimento do C4.5 e da versão J48, disponibilizada no ambiente Weka (versão 3.6.12) e utilizada nos experimentos. A Seção 4.4 descreve um dos classificadores mais simples, o NN (*Nearest Neighbour*) [Cover & Hart 1967], que foi o algoritmo que deu origem a uma família de algoritmos conhecida como algoritmos baseados em instâncias (ABI). A Seção 4.5 aborda e discute dois aspectos importantes quando do uso de algoritmos de AM/MD: a representatividade do conjunto de atributos que descreve os dados de treinamento e a presença de instâncias de treinamento com valores de atributo ausentes. Ambos os problemas são abordados e tratados quando da realização dos experimentos, nos capítulos seguintes. E finalmente, a Seção 4.6 apresenta as considerações finais deste capítulo.

4.2 Classificador Probabilístico - o Naïve Bayes

O algoritmo *Naïve Bayes* (NB) é considerado um classificador probabilístico simples, que tem uma clara semântica associada. O NB tem por base o Teorema de Bayes (também referenciado como Regra de Bayes). A caracterização do NB como "ingênuo" (*naïve*) se deve ao fato que tal classificador assume que todos os atributos (que descrevem as instâncias do conjunto de treinamento) são condicionalmente independentes entre si, dada a variável classe e, também, que todos os atributos são diretamente dependentes da

variável classe [Hruschka & Nicoletti 2013]. A Seção 4.2.1 introduz alguns conceitos básicos, extraídos de [Russel & Norvig 1995], [Santos & Nicoletti 1996], [Murphy 2006] e [Ng 1990], necessários ao entendimento da formalização associada à apresentação e discussão do NB, feita na Seção 4.2.2.

4.2.1 Definições Básicas

Em um contexto probabilístico considere:

- A um evento;
- Ω a coleção de todos os possíveis eventos elementares, denominado *espaço amostral*.

A probabilidade de um evento A é denotada por p(A) e toda função de probabilidade p deve satisfazer os três axiomas:

(1) a probabilidade de todo o espaço amostral é 1, i.e., $p(\Omega) = 1$;

(2) a probabilidade de qualquer evento elementar é não negativa ou seja, $\forall A \in \Omega$, $p(A) \geq 0$;

(3) se n eventos A_1, A_2, ..., A_n são mutuamente exclusivos, ou seja, eles não podem ocorrer simultaneamente, então a probabilidade de que pelo menos um destes n eventos ocorra é dada pela soma das probabilidades de cada um deles, como mostra a Eq. (4.1).

$$p(A_1 \cup A_2 \cup ... \cup A_n) = \Sigma_{i=1,n}\, p(A_i) \tag{4.1}$$

Os axiomas (1) e (2) podem ser combinados na expressão em Eq. (4.2).

$$\forall\ A \in \Omega,\ 0 \leq p(A) \leq 1 \tag{4.2}$$

O complemento de A, ou negação do evento A ($\neg$A) contém a coleção de todos os eventos elementares no espaço amostral Ω, exceto o evento A. Considerando que A e $\neg$A são mutuamente exclusivos e $A \cup \neg A = \Omega$, pelos axiomas (1) e (3) tem-se que:

$$\begin{aligned} p(A \cup \neg A) = \ & p(A) + p(\neg A) = p(\Omega) = 1, \text{ i.e,} \\ & p(\neg A) = 1 - p(A) \end{aligned} \tag{4.3}$$

A Eq. (4.3) viabiliza o cálculo de $p(\neg A)$ a partir de p(A).

Proposições sempre verdadeiras têm probabilidade 1 e proposições sempre falsas têm probabilidade 0, ou seja p(verdade) = 1 e p(falso) = 0. A notação p(A) é usada para designar a *probabilidade a priori* ou *probabilidade incondicional* de que a proposição A seja verdade. Por exemplo, se Hipertireoidismo representa uma proposição que determinado indivíduo tem Hipertireoidismo, então p(Hipertireoidismo) = 0,1 significa que "na ausência de qualquer outra informação" será atribuída a probabilidade de 0,1 (chance de 10%) ao evento do indivíduo em questão ter Hipertireoidismo. É importante lembrar que p(A) pode ser usada apenas quando não existe outra informação disponível. Também p(Hipertireoidismo) pode ser visto como uma simplificação para p(Hipertireoidismo=verdade) e de maneira análoga, p(¬Tireoidismo) como uma abreviação para p(Hipertireoidismo=falso).

A proposição à qual se refere a probabilidade pode ser representada por um símbolo, por exemplo A, como em p(A). Proposições podem incluir também igualdades envolvendo as chamadas variáveis randômicas (aleatórias). Para a variável aleatória Temperatura_Verão, por exemplo, pode-se ter: p(Temperatura_Verão = 0^o) = 0,01; p(Temperatura_Verão = 15^o) = 0,09; p(Temperatura_Verão = 25^o) = 0,4; p(Temperatura_Verão = 35^o) = 0,5.

Variáveis aleatórias podem assumir um conjunto de valores. Neste caso pode-se usar uma notação tal como **p**(Temperatura_Verão) para referência a um vetor de valores para as probabilidades de cada estado individual de Temperatura_Verão i.e., **p**(Temperatura_Verão) = <0,01 0,09 0,4 0,5>, que define *a distribuição de probabilidades para a variável Temperatura_Verão*.

Uma vez obtida alguma evidência com relação às proposições que descrevem um domínio de conhecimento, a probabilidade a priori não é mais aplicável; ao invés dela usa-se a *probabilidade a posteriori* ou *probabilidade condicional*. A notação p(A|B) expressa a probabilidade de A dado que tudo o que se sabe é B. Por exemplo p(Hipertireoidismo|taquicardia) = 0,7 indica que foi observado que um paciente é taquicardíaco e ainda não existe disponível qualquer outra informação; então a probabilidade de que o paciente esteja com hipertireoidismo é 0,7.

Para estabelecer a definição formal, considere A, B $\in \Omega$. A probabilidade que o evento A ocorra dado que o evento B ocorreu, denotada p(A|B), é chamada *probabilidade condicional de A dado B* e definida pela Eq. (4.4),

$$p(A|B) = \frac{p(A \cap B)}{p(B)} \tag{4.4}$$

em que: $p(A \cap B)$ é a interseção probabilística ou seja, a probabilidade que ambos, tanto o evento A quanto o B vão ocorrer e p(B) é a probabilidade *a priori* de B, desde que essa probabilidade seja diferente de 0. A Eq. (4.4) pode ser reescrita como a Eq. (4.5), chamada de *regra do produto*.

$$p(A \cap B) = p(A|B) \times p(B) \tag{4.5}$$

De maneira análoga, a probabilidade condicional de B dado A, p(B|A), é dada pela Eq. (4.6).

$$p(B|A) = \frac{p(B \cap A)}{p(A)} \quad \rightarrow \quad p(B \cap A) = p(B|A) \times p(A) \tag{4.6}$$

Como a intersecção probabilística é comutativa, a Eq. (4.7), pode ser derivada a partir da Eq. (4.5).

$$p(A \cap B) = p(B \cap A) = p(B|A) \times p(A) \tag{4.7}$$

Por meio da substituição da Eq. (4.6) na Eq. (4.4) chega-se à Eq. (4.8), conhecida como *regra de Bayes* (ou *teorema de Bayes*).

$$p(A|B) = \frac{p(B|A) \times p(A)}{p(B)} \tag{4.8}$$

4.2.2 O Naïve Bayes (NB)

Como apontado em [Domingos & Pazzani 1996], o teorema de Bayes formaliza e substancia o desenvolvimento de um procedimento, no contexto de AM, que pode ser usado para a predição da classe associada à uma nova instância, dado um conjunto de instâncias de treinamento. A classe escolhida deve ser aquela que maximiza a Eq. (4.9),

$$p(C_i|E) = \frac{p(E|C_i) \times p(C_i)}{p(E)} \quad (4.9)$$

em que C_i é a i-ésima classe, E é uma instância e p(Y|X) representa a probabilidade condicional de Y dado X; probabilidades são estimadas a partir das instâncias de treinamento. A Eq. (4.9) é uma representação da Eq. (4.8) que substitui os eventos pelas variáveis da instância e da classe.

Considere que uma instância é representada por um vetor de n atributos. Considerando que os atributos sejam independentes, dada a classe, a probabilidade $p(E|C_i)$ pode ser decomposta no produto $p(v_1|C_i) \times p(v_2|C_i) \times ... \times p(v_n|C_i)$, em que v_j é o valor do j-ésimo atributo da instância E. Portanto a classe que deve ser predita é aquela que maximiza a Eq. (4.10)

$$p(C_i|E) = \frac{p(C_i)}{p(E)} \prod_{j=1}^{n} p(v_j|C_i) \quad (4.10)$$

4.2.3 Exemplo do Uso do Naïve Bayes

A situação de aprendizado descrita nesta seção foi extraída de [Witten & Frank 2005] e adaptada à notação empregada nesta dissertação. Essa situação, entretanto, é recorrente em várias publicações associadas a aprendizado de máquina e foi descrita, pela primeira vez, quando da proposta do algoritmo ID3 por Quinlan, em [Quinlan 1991], para exemplificar os conceitos básicos associados a *conjuntos de treinamento* utilizados por algoritmos de aprendizado automático. Essa situação é, também, empregada para ilustrar pedagogicamente, o funcionamento de inúmeros algoritmos de aprendizado automático. Ela volta a ser usada na Seção 4.3 deste capítulo para ilustrar o processo de geração de árvores de decisão.

O exemplo, muitas vezes referenciado como o *problema do tempo*, é descrito como um conjunto de 14 instâncias de dados que, supostamente, descrevem condições que são adequadas (ou não) para se jogar um determinado jogo, em um local aberto. Em geral, instâncias de dados são descritas por uma sequência de valores de atributos, que caracterizam determinados aspectos da situação que a instância descreve. No caso específico do exemplo sendo considerado, os quatro atributos são: *aspecto*, *temperatura*,

umidade e *ventoso* e, para cada instância, dependendo dos valores dos seus atributos, o valor de seu atributo final, atributo esse conhecido como *classe* (que representa a atividade pretendida ao ar livre *i.e.*, jogo), informa se a instância em questão descreve uma situação climática propícia (ou não), a que o jogo em questão possa ser realizado.

No exemplo sendo considerado todos os quatro atributos, assim como a classe, têm valores simbólicos (*i.e.*, não numéricos) associados a eles, como segue:

(1) *aspecto* assume valores do conjunto: {ensolarado, nublado, chuvoso};

(2) *temperatura* assume valores do conjunto: {quente, confortável, fresca};

(3) *umidade* assume valores do conjunto: {alta, normal};

(4) *ventoso* assume valores do conjunto: {verdade, falso};

e a *classe* assume valores no conjunto {sim, não}.

Tabela 4.1 Conjunto de treinamento do tempo, que indica se é possível jogar tênis ou não em determinadas condições climáticas [Witten & Frank 2005].

aspecto	*temperatura*	*umidade*	*ventoso*	*classe*
ensolarado	quente	alta	falso	não(n)
ensolarado	quente	alta	verdade	não(n)
nublado	quente	alta	falso	sim(s)
chuvoso	confortável	alta	falso	sim(s)
chuvoso	fresca	normal	falso	sim(s)
chuvoso	fresca	normal	verdade	não(n)
nublado	fresca	normal	verdade	sim(s)
ensolarado	confortável	alta	falso	não(n)
ensolarado	fresca	normal	falso	sim(s)
chuvoso	confortável	normal	falso	sim(s)
ensolarado	confortável	normal	verdade	sim(s)
nublado	confortável	alta	verdade	sim(s)
nublado	quente	normal	falso	sim(s)
chuvoso	confortável	alta	verdade	não(n)

A Tabela 4.2 mostra como o algoritmo NB contabiliza os dados do conjunto de treinamento descrito na Tabela 4.1. O NB conta quantas vezes cada par *atributo-valor* ocorre, associado a cada valor (sim/não) do atributo classe. Na Tabela 4.2, por exemplo, o par *aspecto = ensolarado* comparece em 5 instâncias; em 2 delas *classe = (s)* e nas 3 restantes, *classe = (n).*

Os valores que comparecem nas primeiras linhas de dados da tabela contabilizam essas ocorrências, para todos os possíveis valores de cada atributo e os números sob a coluna final (classe), são referentes ao número total de ocorrências de (s) e (n). Na parte

inferior da tabela a mesma informação está reescrita na forma de probabilidades observadas; por exemplo, das 9 instâncias em que *classe = (s)*, *aspecto = ensolarado* comparece em 2 delas, produzindo a fração 2/9. Já quando classe é considerada, a fração representa a proporção de instâncias em que classe é (s) e (n), respectivamente.

O sumário apresentado pela Tabela 4.2 evidencia que o NB usa todos os atributos e permite que façam contribuições à classe, que são igualmente importantes e independentes umas das outras, dada a classe. Como comentam Witten & Frank, apesar dessa abordagem irrealística, o NB é um procedimento simples e que funciona surpreendentemente bem na prática.

Tabela 4.2 Sumário obtido pelo NB a partir do conjunto da Tabela 4.1. Na tabela (s) e (n) representam os dois possíveis valores do atributo classe.

aspecto			*temperatura*			*umidade*			*ventoso*			*classe*	
	(s)	(n)		(s)	(n)		(s)	(n)		(s)	(n)	(s)	(n)
ensolarado	2	3	quente	2	2	alta	3	4	falso	6	2	9	5
nublado	4	0	confortável	4	2	normal	6	1	verdade	3	3		
chuvoso	3	2	fresca	3	1								
ensolarado	2/9	3/5	quente	2/9	2/5	alta	3/9	4/5	falso	6/9	2/5	9/14	5/14
nublado	4/9	0/5	confortável	4/9	2/5	normal	6/9	1/5	verdade	3/9	3/5		
chuvoso	3/9	2/5	fresca	3/9	1/5								

Suponha agora que uma instância como a representada na Tabela 4.3 seja disponibilizada, em que o valor de classe é desconhecido (representado por ?).

Tabela 4.3 Exemplo de instância a ser classificada, o valor da classe está representado por (?).

aspecto	*temperatura*	*umidade*	*ventoso*	*classe*
ensolarado	fresca	alta	verdade	?

Os cinco descritores da Tabela 4.2 – *aspecto, temperatura, umidade, ventoso* e a verossimilhança total que *classe* seja (s) ou (n) – tratados como peças de evidência igualmente importantes e independentes, produzem:

verossimilhança de *classe = (s)* = $2/9 \times 3/9 \times 3/9 \times 3/9 \times 9/14 = 0{,}0052$.

verossimilhança de *classe = (n)* = $3/5 \times 1/5 \times 4/5 \times 3/5 \times 5/14 = 0{,}0206$.

Esses valores indicam que para a nova instância descrita na Tabela 4.3, o valor da classe ser (n) é mais (quatro vezes mais) provável do que ser (s).

Tais valores podem ser transformados em probabilidades por meio de normalização, de maneira que somem 1, como segue.

$$\text{probabilidade de (s)} = \frac{0{,}0053}{0{,}0053+0{,}0206} = 20{,}5\%$$

probabilidade de (n) = $\frac{0,0206}{0,0053+0,0206}$ = 79,5%

Como mostrado no exemplo, esse método é baseado na regra de Bayes da probabilidade condicional, discutida nas seções 4.2.1 e 4.2.2. Lembrando, a regra de Bayes estabelece que dada uma hipótese H e uma evidência E que suporta a hipótese, então

$$p(H|E) = \frac{p(E|H) \times p(H)}{p(E)}$$

em que p(A) denota a probabilidade de um evento A e p(A|B) denota a probabilidade de A, condicional a um outro evento B.

No caso do exemplo sendo apresentado, pode se considerar que a hipótese H seja a de que a classe tenha como valor, por exemplo, (s), e portanto, p(H|E) será o valor 20,5%, como determinado anteriormente. A evidência E é a combinação específica de valores de atributos descrita na Tabela 4.3 ou seja, *aspecto = ensolarado, temperatura = fresca, umidade = alta e ventoso = verdade.* Assumindo que essas peças de evidências são independentes (dada a classe), a probabilidade combinada delas é obtida multiplicando as probabilidades:

$$p((s)|E) = \frac{p(E_1|(s)) \times p(E_2|(s)) \times p(E_3|(s)) \times p(E_4|(s)) \times p((s))}{p(E)}$$

O denominador da fração acima será ignorado e eliminado no passo de normalização final, quando as probabilidades de (s) e (n) devem somar 1, como feito previamente. A p((s)) ao final do produto é a probabilidade da classe ser (s) sem o conhecimento de qualquer evidência E, isto é, sem conhecer qualquer informação a respeito de um dia específico – que é a *probabilidade a priori* da hipótese (no caso (s)). No caso do exemplo, o seu valor é 9/14 uma vez que 9 das 14 instâncias de treinamento têm o valor (s) associado à classe. Substituindo as probabilidades das evidências pelos respectivos valores mostrados na Tabela 4.2, tem-se:

$$p((s)|E) = \frac{2/9 \times 3/9 \times 3/9 \times 3/9 \times 9/14}{p(E)},$$

justamente como foi calculado anteriormente.

4.3 Árvore de Decisão – o Algoritmo *ID3*

O termo TDIDT (*Top Down Induction of Decision Trees*) [Quinlan 1986] é utilizado para agrupar um conjunto de algoritmos de aprendizado automático que generalizam um conjunto de treinamento em uma estrutura conhecida como árvore de decisão.

O algoritmo ID3, introduzido por Quinlan [Quinlan 1986], faz parte da família TDIDT. O algoritmo constrói árvores de decisão a partir de um conjunto de instâncias (conjunto de treinamento), descritos por atributos (e respectivos valores) e uma classe associada (conjunto de instâncias como aquele mostrado na Tabela 4.1). No programa original os atributos eram categóricos i.e., os valores associados a cada atributo deveriam ser elementos de um determinado conjunto finito de valores discretos. Os valores associados ao atributo *classe* eram dois possíveis valores categóricos, e.g. (s) ou (n).

As implementações do ID3 e de sistemas considerados evoluções do ID3 (i.e., C4.5 e J48) permitem mais do que dois valores de classe e, também, permitem atributos que possam ter como valores números inteiros ou números reais. O pseudocódigo do algoritmo ID3 está descrito no Algoritmo 4.1.

A árvore de decisão (AD) é construída de cima para baixo (*top-down*), com a escolha, a cada passo, do atributo mais relevante que deve comparecer como um nó da árvore – a escolha do atributo é feita por meio das medidas de entropia e ganho de informação associadas aos atributos candidatos. Escolhido o melhor atributo para representar a raiz da árvore, o algoritmo parte para a construção do próximo nível da árvore. O processo se repete até que a árvore classifique as instâncias do conjunto de treinamento ou, então, até que todas (as instâncias) alcancem uma folha (ou classe).

O ID3 é iterativo em sua estrutura básica. Como comenta Quinlan em [Quinlan 1986], a escolha do teste em cada nó da árvore sendo construída é crucial para que a AD seja simples. Considere que C seja o conjunto de todas as instâncias e que C contenha p instâncias de classe (S) e n instâncias de classe (N). Duas suposições são feitas, no contexto do ID3:

procedure ID3*(Exemplos, Atributo_Alvo, Atributos)*
Input:
Exemplos são os exemplos de treinamento.
Atributo_Alvo é o atributo cujo valor será predito pela árvore (i.e., *classe*).
Atributos é uma lista de outros atributos que podem ser testados pela AD sendo construída.
Output:
Árvore de decisão (AD) que classifica corretamente *Exemplos*.

- Crie o nó *Raiz* para a árvore.
- Se todos *Exemplos* forem positivos, retorne a árvore de nó único (*Raiz)*, com rótulo = +
- Se todos *Exemplos* forem negativos, retorne a árvore de nó único (*Raiz)*, com rótulo = –
- Se *Atributos*= ∅, **return**: a árvore de nó único *Raiz,* com rótulo = valor mais frequente do *Atributo_Alvo* em *Exemplos*
- Caso contrário

begin

- $A \leftarrow$ o atributo de *Atributos* que melhor(*) classifica *Exemplos*
- *Raiz* $\leftarrow A$
- Para cada possível valor, v_i, de *A,*
  - Adicione um novo ramo da árvore abaixo da *Raiz*, correspondente ao teste $A = v_i$
  - Faça $Exemplos_{v_i}$ser o subconjunto de *Exemplos* que tem valor v_i para *A.*
  - Se $Exemplos_{v_i} = \emptyset$
    - Então embaixo deste novo ramo, adicione uma folha com rótulo = valor mais frequente do *Atributo_Alvo* em *Exemplos*
    - Senão embaixo deste novo ramo adicione a subárvore ID3(*Exemplos*, *Atributo_Alvo*, *Atributos – {A})*

end

- **return** *Raiz*

(*)O melhor atributo é aquele com o maior ganho de informação.

Algoritmo 4.1 Pseudocódigo do algoritmo ID3 traduzido de Mitchell [Mitchell 1997].

(1) qualquer AD que represente C corretamente deverá classificar instâncias na mesma proporção como comparecem em C. Uma instância arbitrária irá pertencer à classe S com probabilidade p/(p+n) e à classe N com probabilidade n/(p+n).

(2) quando uma AD é usada para classificar uma instância, ela retorna uma classe. Assim sendo, a AD pode ser considerada como a fonte de uma mensagem 'S' ou 'N', com a esperada informação necessária para gerar essa mensagem dada pela Eq. (4.11).

$$I(p,n) = -\frac{p}{p+n}\log_2\frac{p}{p+n} - \frac{n}{p+n}\log_2\frac{n}{p+n} \qquad (4.11)$$

Se o atributo A com valores $\{A_1, A_2, ..., A_v\}$ for escolhido para ser a raiz da árvore de decisão, irá particionar o conjunto de instâncias C em $\{C_1, C_2, ..., C_v\}$, em que C_i irá conter as instâncias de C cujo valor para seu atributo A for A_i. Considere que C_i, contenha p_i instâncias da classe (S) e n_i instâncias da classe (N). A informação requerida para a definição da árvore com raiz A é obtida como a média ponderada (entropia) dada pela Eq. (4.12)

$$\text{entropia}(A) = \sum_{i=1}^{v} \frac{p_i + n_i}{p+n} I(p_i, n_i) \quad (4.12)$$

na qual o peso associado ao i-ésimo ramo é dado pela proporção de instâncias de C que pertencem a C_i. O ganho de informação obtido pela ramificação em A é dado pela Eq. (4.13).

$$\text{ganho}(A) = I(p, n) - E(A) \quad (4.13)$$

Uma boa opção para a identificação do atributo no qual ramificar é a de escolher aquele que tem o maior ganho de informação. Como ilustração do processo de escolha do atributo durante a construção da AD, considere novamente as instâncias descritas na Tabela 4.1. Das 14 instâncias, 9 são da classe (s) e 5 são da classe (n), o que faz com que a informação (medida em bits) requerida para classificação seja:

$$I\big((s), (n)\big) = -\frac{9}{9+5}\log_2\frac{9}{9+5} - \frac{5}{9+5}\log_2\frac{5}{9+9} = 0{,}940 \text{ bits}$$

Considere agora o atributo *aspecto*, que assume valores no conjunto {ensolarado, nublado, chuvoso}. Dentre as 14 instâncias 5 têm *aspecto* = *ensolarado*; 2 das 5 são da classe (s) e as 3 restantes são da classe (n). Portanto, $p_1 = 2$, $n_1 = 3$, o que faz com que $I(p_1, n_1) = 0{,}971$. Similarmente, $p_2 = 4$, $n_2 = 0$ resultando em $I(p_2, n_2) = 0$ e $p_3 = 3$, $n_3 = 2$ resultando em $I(p_3, n_3) = 0{,}971$.

A expectativa de informação necessária após testar esse atributo é portanto: E(aspecto) = 5/14 × $I(p_1, n_1)$ + 4/14 × $I(p_2, n_2)$ + 5/14 × $I(p_3, n_3)$ = 0,694 bits. O ganho desse atributo é portanto: ganho(*aspecto*) = 0,940 – E(*aspecto*) = 0,246 bits. Procedimentos similares vão produzir: ganho(*temperatura*) = 0,029; ganho(*umidade*) = 0,151; ganho(*chuvoso*) = 0,048. O processo construtivo realizado pelo ID3 irá então escolher o atributo *aspecto* como raiz da AD. Tendo como entrada o conjunto de instâncias da Tabela 4.1, o ID3 produz como saída a árvore mostrada na Figura 4.1.

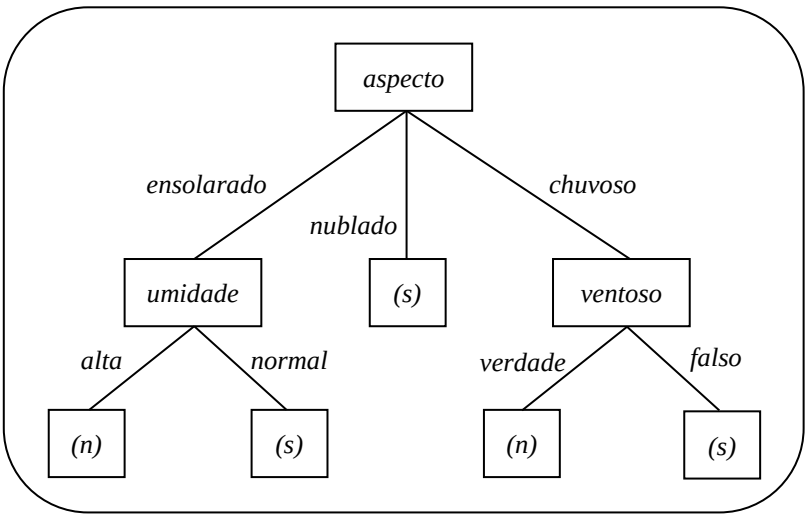

Figura 4.1 Árvore de decisão induzida pelo ID3 a partir do conjunto de instâncias descritas na Tabela 4.1.

O sistema C4.5, também proposto por Quinlan e descrito em [Quinlan 1993], implementa o algoritmo ID3 e acrescenta algumas funções para suprir limitações presentes no mesmo. Já a versão conhecida como J48 é a implementação do algoritmo C4.5 disponível na ferramenta de MD Weka [Weka 2012]. O J48 foi escrito na linguagem multiplataforma Java e, quando de sua finalização, como a última versão disponibilizada do C4.5 era a C4.8, a versão Weka foi nomeada como J48.

A Figura 4.2 apresenta um exemplo induzido pelo classificador J48, usando o mesmo conjunto de instâncias descritas na Tabela 4.1. O conjunto de dados possui 14 instâncias descritas por quatro atributos mais a classe, com dois possíveis resultados s(im) e n(ão).

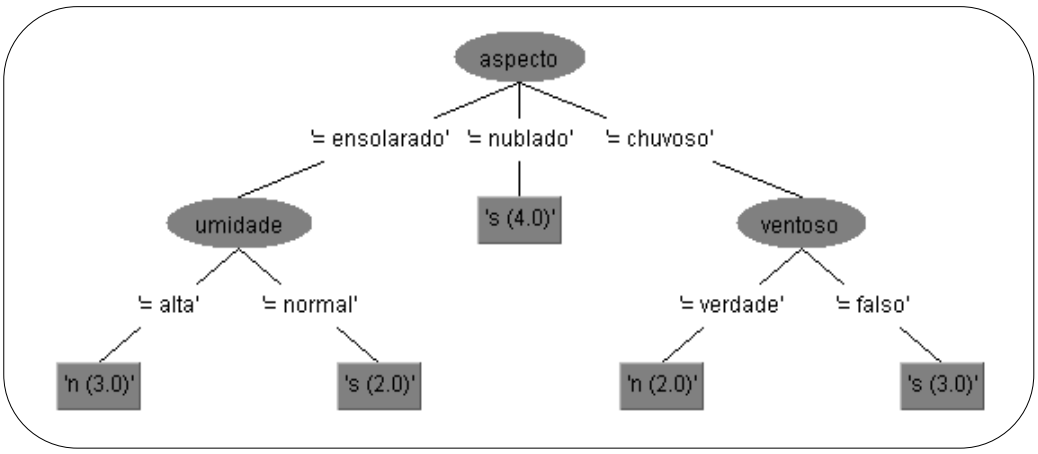

Figura 4.2 Árvore de decisão induzida pelo J48 a partir do conjunto de instâncias descritas na Tabela 4.1.

A Figura 4.3, traz a saída produzida pelo classificador: Resumo (Summary), Precisão detalhada por classe (Detailed Accuracy By Class) e Matriz de confusão (Confusion Matrix).

```
[…]
=== Summary ===

Correctly Classified Instances      7            50      %
Incorrectly Classified Instances    7            50      %
Kappa statistic                     -0.0426
Mean absolute error                 0.4167
Root mean squared error             0.5984
Relative absolute error             87.5         %
Root relative squared error         121.2987     %
Total Number of Instances           14

=== Detailed Accuracy By Class ===

               TP Rate   FP Rate   Precision   Recall  F-Measure   ROC Area  Class
               0.556     0.6       0.625       0.556   0.588       0.633     s
               0.4       0.444     0.333       0.4     0.364       0.633     n
Weighted Avg.  0.5       0.544     0.521       0.5     0.508       0.633

=== Confusion Matrix ===

 a  b   < --  classified as
 5  4 | a  = s
 3  2 | b  = n
```

Figura 4.3 Saída do classificador J48: resumo, precisão detalhada por classe e matriz de confusão.

No que segue são detalhadas as principais informações produzidas pelo Weka, para um maior entendimento dos resultados produzidos; várias dessas medidas são úteis para comparar os resultados obtidos pelos classificadores.

(1) *Matriz de Confusão (MC)*, comumente conhecida como *tabela de contingência*, é uma tabela com um determinado *layout* que permite a visualização do desempenho de um algoritmo, tipicamente um algoritmo supervisionado. O elemento MC[i,j] representa o número de instâncias de classe i, às quais foram atribuídas a classe j.

O nome da matriz é uma referência ao fato que, por meio da inspeção da matriz, pode ser verificado se o algoritmo de aprendizado está confundindo as classes envolvidas. A dimensão dessa matriz é função do número de classes envolvidas no problema. Para um problema envolvendo n classes, por exemplo, tal matriz tem dimensão n × n. Segue um exemplo de uma matriz de confusão 2 × 2, da maneira como o Weka a apresenta ao usuário do sistema:

```
=== Confusion Matrix ===
a  b <-- classified as
7 2 | a = yes
3 2 | b = no
```

O número de instâncias corretamente classificadas pelo classificador induzido sob o Weka é dado pela soma dos elementos da diagonal principal da matriz que, no exemplo, é 9. A soma de todos os outros números que comparecem na matriz são de instâncias que foram incorretamente classificadas, no caso, 5. Duas instâncias da classe "a" são classificadas erradamente como "b" e três instâncias da classe "b" são classificadas erradamente como "a".

(2) *TP* (*True Positive* – Positivo Verdadeiro) TP representa a proporção de instâncias que foram classificadas como classe x, entre todas as instâncias que efetivamente têm classe x. Pode ser extraída da matriz de confusão; considerando uma determinada linha da matriz, TP é o elemento da diagonal principal na linha considerada, dividido pela soma dos elementos da linha. No exemplo, para a classe "yes", TP = 7/(7+2) = 0,778 e para a classe "no", TP = 2/(3+2) = 0,4. TP é equivalente ao Recall.

(2) *FP* (*False Positive* – Positivo Falso) FP representa a proporção de instâncias que foram classificadas como classe x, mas que pertencem a uma classe diferente de x entre todos as instâncias que não são da classe x. Para o cálculo de FP, a partir da Matriz de Confusão, soma-se os elementos associados à coluna da classe x (i.e., classe que foi predita), subtraindo o elemento da diagonal; o resultado é dividido pela soma das linhas de todas as outras classes. No exemplo, FP = 3/5 = 0,6 para a classe "yes" e 2/9 = 0,222 para a classe "no".

(3) *Precisão* (*Precision*) é a proporção de instâncias que realmente têm classe x, considerando todas que foram classificadas como x. Usando a MC, esse valor é obtido pela divisão do elemento da diagonal pela soma dos elementos da coluna em questão. Precisão para a classe "yes" é dada por 7/(7+3) = 0,7 e para a classe "no" por 2/2+2 = 0,5.
(4) *F-Measure* (Medida F) é uma medida combinada para *Precisão* e *Recall*, definida como *F-Measure* = 2 × Precisão × *Recall*/(Precisão+*Recall*).

4.4 Aprendizado Baseado em Instâncias – Algoritmo Nearest Neighbor

O *Nearest Neighbor* (NN) [Cover & Hart 1967] é caracterizado como um algoritmo de *aprendizado baseado em instâncias* (ABI) (também conhecido como aprendizado preguiçoso). Considerando um conjunto de treinamento X, algoritmos ABI simplesmente armazenam X como a representação do conceito sendo aprendido; diferentemente da maioria dos algoritmos de AM, algoritmos ABI não 'trabalham' o conjunto de treinamento de maneira a gerar uma representação mais concisa do conceito que, também, seja ampla e mais geral.

O processo de 'generalização' feito por algoritmos ABI é 'adiado' (daí o uso do termo preguiçoso, para caracterizá-los) até o momento em que uma nova instância, de classe desconhecida, precisa ser classificada. Portanto, ao invés de generalizar o conceito a partir do conjunto de treinamento recebido, algoritmos ABI estimam o conceito localmente, para cada nova instância a ser classificada, durante a fase de classificação.

Particularmente, o NN é o algoritmo mais básico de ABI e, de acordo com Gates em [Gates 1972], sua popularidade se deve "... à sua simplicidade conceitual, a qual leva a uma implementação direta, embora não necessariamente, mais eficiente." O NN considera que todas as instâncias armazenadas (descritas por M atributos e uma classe associada) correspondem a pontos no espaço M-dimensional R^M e, então, uma função de distância é usada para determinar qual instância, dentre aquelas armazenadas, é a mais próxima de uma dada nova instância (a ser classificada). Uma vez determinada a instância mais próxima, sua classe é atribuída à nova instância considerada. No NN a fase de aprendizado simplesmente armazena as instâncias de dados que participam do conjunto de treinamento fornecido ao NN.

A regra $d(x, x_i) \leq d(x,x_j)$, como estabelecida no Algoritmo 4.2, é chamada de regra 1-NN, uma vez determina a distância da nova instância a cada uma das instâncias do conjunto de treinamento; para determinar a classe da nova instância, considera a classe de apenas *uma* instância, aquela mais próxima da nova instância. Via de regra a métrica d é implementada por meio do cálculo da distância euclidiana.

Considere:

- espaço M-dimensional de atributos
- K classes, numeradas 1,2,3,...,K
- N instâncias de treinamento, cada uma representada como um par (x_i,θ_i), $1 \le i \le N$, tal que:

(a) x_i é uma instância representada por um vetor de M valores de atributos:

$$x_{i=(}x_{i_1}, x_{i_2}, \ldots, x_{i_M})$$

(b) $\theta_i \in \{1,2,...,K\}$ denota a classe correta da instância x_i

Seja $T_{NN} = \{(x_1, \theta_1), (x_2, \theta_2), ..., (x_N, \theta_N)\}$ o conjunto de treinamento do NN.

Dada uma instância desconhecida x, a regra de decisão do algoritmo decide que x está na classe θ_j se

$$d(x, x_j) \le d(x,x_i) \qquad 1 \le i \le N$$

em que d é alguma métrica M-dimensional de distância.

Algoritmo 4.2 Algoritmo NN como originalmente proposto em [Cover & Hart 1967].

Na literatura podem ser encontradas inúmeras variantes do NN, muitas delas propostas com o intuito de contornar alguma das deficiências do algoritmo (para uma discussão detalhada ver [Aha *et al.* 1991] [Aha 1992], [Beale & Jackson 1994], [Gates 1972], [Hart 1968], [Angiulli 2007]).

Considere que uma instância arbitrária x_i seja descrita pelo vetor M-dimensional de valores de atributos notado por $<x_{i_1}, x_{i_2}, \ldots, x_{i_M}>$, em que x_{i_r} representa o valor do r-ésimo atributo da instância x_i, $1 \le r \le M$). A distância euclidiana entre duas instâncias x_i e x_j é mostrada na Eq. (4.14).

$$d(x_i, x_j) = \sqrt{\sum_{r=1}^{M} (x_{i_r} - x_{j_r})^2} \qquad (4.14)$$

Para exemplificar o uso do algoritmo NN, considere um espaço bidimensional contendo 25 instâncias, cada uma delas descritas por dois atributos numéricos, x e y, sendo 10 delas da classe c_1 e 15 da classe c_2, com os valores listados na Tabela 4.4, e com representação gráfica mostrada na Figura 4.4.

Tabela 4.4 Conjunto de treinamento T_{NN}, para o NN. T_{NN} tem 25 instâncias, descritas por dois atributos numéricos (x e y). Das 25, 10 são da classe c_1 e 15 são da classe c_2.

(x y classe)	(x y classe)	(x y classe)	(x y classe)	(x y classe)
(1,9 1,1 c_1)	(1,1 1,4 c_1)	(2,0 4,0 c_2)	(2,9 3,9 c_2)	(3,5 3,2 c_2)
(1,8 1,9 c_1)	(1,3 1,6 c_1)	(2,8 3,4 c_2)	(2,5 3,5 c_2)	(3,2 3,0 c_2)
(1,0 1,5 c_1)	(1,2 1,0 c_1)	(2,6 3,7 c_2)	(2,8 3,1 c_2)	(3,1 3,3 c_2)
(1,7 1,2 c_1)	(1,5 1,3 c_1)	(2,4 3,2 c_2)	(2,4 3,0 c_2)	(3,5 3,7 c_2)
(1,6 1,8 c_1)	(1,4 1,7 c_1)	(2,1 3,4 c_2)	(3,3 4,0 c_2)	(3,3 3,3 c_2)

A Figura 4.4 exibe a expressão do conceito sob o NN i.e., o conjunto de treinamento fornecido, constituído de 25 instâncias de dados.

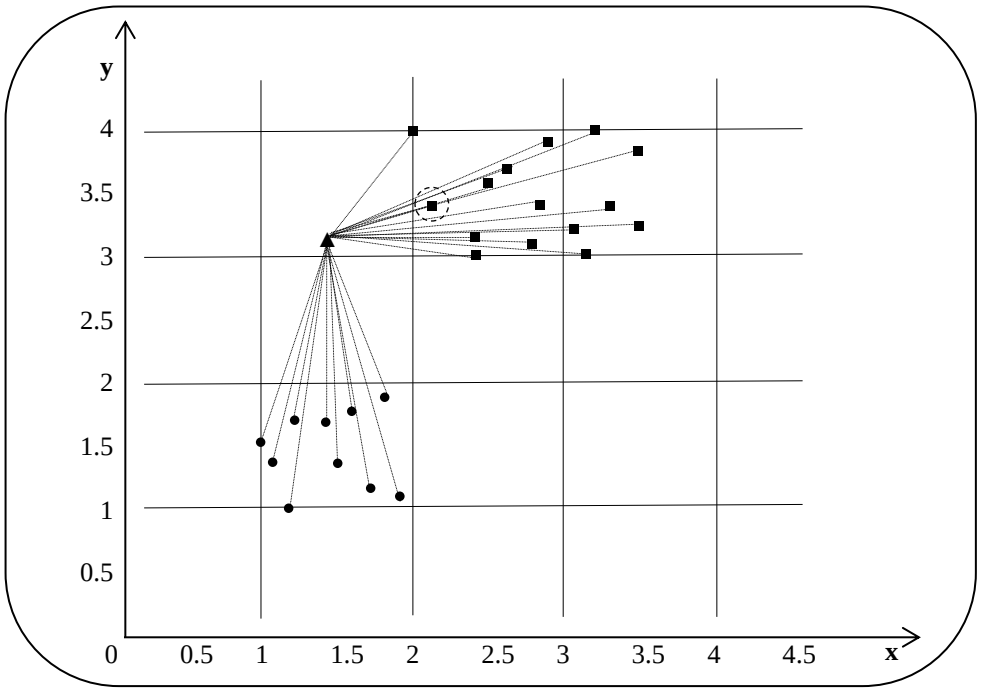

Figura 4.4 Uso do NN para a classificação de uma nova instância, com classe desconhecida, identificada por ▲. Os pontos identificados com • representam instâncias da classe c_1, e aqueles identificados com ■, da classe c_2. Como a instância ■ (dentro do círculo tracejado) é a instância que está mais próxima da nova instância, à nova instância será atribuída a classe c_2.

Uma nova instância, de classe desconhecida, identificada pelo triângulo, é classificada pelo NN como de classe c_2 uma vez que, dentre todas as distâncias calculadas ente a nova instância e as 25 instâncias que representam o conceito, a instância de classe c_2 (dentro de um círculo tracejado) é a que está mais próxima da nova instância.

4.5 Seleção de Atributos e Imputação de Dados

Esta seção apresenta e discute dois aspectos relevantes no uso de algoritmos de AM/MD: o da representatividade do conjunto de treinamento para a indução do conceito, discutido na Seção 4.6.1, e o da ausência de valores de atributos, em instâncias que são parte do conjunto de treinamento, tratado na Seção 4.6.2. Tais aspectos direcionaram parte dos experimentos descritos no Capítulo 5.

4.5.1 Seleção de Subconjunto de Atributos Relevantes

Como já comentado anteriormente e novamente enfatizado nesta seção, algoritmos de MD/AM aprendem a partir de um conjunto de instâncias de dados (referenciadas coletivamente como conjunto de treinamento) que, supostamente, são representativas do conceito a ser aprendido. Muitos aspectos podem afetar o sucesso de tais algoritmos. A qualidade do conjunto de treinamento, entretanto, é um dos mais impactam o resultado do aprendizado [Liu & Setiono 1996].

Em ambientes de aprendizado supervisionado, instâncias de treinamento são, geralmente, representadas como vetores de pares atributo-valor e de uma classe associada. Dado um conjunto de treinamento, não necessariamente todos os atributos que comparecem na descrição das instâncias são efetivamente relevantes para a determinação da classe associada à instâncias.

Como discutido em [Santoro et al. 2007], se os dados disponibilizados falham em descrever e representar apropriadamente o conceito que, supostamente deveriam representar ou, então, em exibir as propriedades estatísticas esperadas pelo algoritmo, o resultado de um processo de aprendizado tem grandes chances de não ter qualquer significado. Em muitos problemas que envolvem o uso de algoritmos de AM/MD, nem sempre os atributos que realmente caracterizam um determinado problema são conhecidos. Uma prática comum é usar o maior número possível de atributos para descrever o conjunto de treinamento, em uma tentativa de fornecer o máximo possível de informação. Muitos dos atributos incluídos, entretanto, podem ser parcialmente ou completamente irrelevantes/redundantes à descrição do conceito sendo aprendido. Um conjunto de treinamento descrito por atributos redundantes ou irrelevantes, por exemplo,

pode direcionar o algoritmo de aprendizado a construir uma representação ruim para o conceito que está embutido nos dados de treinamento.

O problema caracterizado como o da seleção de atributos relevantes *i.e.*, da seleção de atributos que representam um papel importante na caracterização de um determinado conceito, é um problema recorrente na área de AM e, consequentemente, na área de MD [Blum & Langley 1997] [Kohavi & John 1997] [Yang & Honavar 1998] [Zhang *et al.* 2015]. Processos de seleção de subconjuntos de atributos relevantes, via de regra, investem na identificação e remoção do máximo possível de informações irrelevantes ou redundantes. Tais processos, geralmente, conduzem uma busca heurística no espaço de busca definido por todos os possíveis subconjuntos do conjunto original de atributos, na tentativa de identificar aquele mais relevante para a tarefa de aprendizado automático sendo (ou a ser) realizada Um método de seleção de atributos busca identificar aqueles atributos que efetivamente contribuem para o estabelecimento das classes associadas às instâncias e, consequentemente, acabam eliminando aqueles atributos que são irrelevantes ou, então, redundantes Métodos de seleção de atributos relevantes contribuem para a redução da dimensionalidade do espaço no qual o conceito é representado, diminuindo, dessa forma, sua complexidade representacional. Uma redução em dimensionalidade, também, promove a precisão e a compreensibilidade dos conceitos induzidos [Santoro *et al.* 2007].

Um número substancial de trabalhos de pesquisa relacionados ao problema da seleção de atributos relevantes tem por objetivo melhorar a acurácia do classificador induzido por um algoritmo de AM/DM. Em algumas circunstâncias tal melhoria pode, também, promover outras características, tais como a redução nas exigências de mensuração e armazenamento de instâncias de treinamento bem como uma descrição de conceito mais compacta e mais facilmente entendida. A seleção de atributos relevantes é particularmente importante em situações em que as instâncias de treinamento são descritas por um conjunto amplo de atributos e a relevância dos atributos, na descrição do conceito, é desconhecida (ver [Dash & Liu, 1997; Liu & Motoda 2012]).

4.5.2 Imputação de Valores Ausentes de Atributos

Um grande número de algoritmos de AM/MD assume que os dados disponibilizados já estão prontos para uso, ou seja, já estão normalizados (quando necessário), não têm ruídos, erros de digitação e, também, não tem valores de atributos que estão ausente.

A razão disso se deve ao fato de usuários de algoritmos de AM/DM geralmente, submeterem os dados disponibilizados a uma fase inicial de pré-processamento, que pode envolver o uso de vários procedimentos, com o objetivo de tornar tais dados prontos para serem usados por algoritmos de AM/DM. Além de submeterem os dados a processos para a seleção de subconjuntos de atributos relevantes, um outro processamento comumente utilizado na fase de pré-processamento, é o do tratamento daquelas instâncias de dados que têm valores de atributos não informados (caracterizados como *ausentes*).

Como apontado em [Bertini *et al.* 2014], provavelmente a prática mais popular para o tratamento de instâncias com valor de atributo ausente é simplesmente eliminá-las do conjunto de treinamento. Dependendo da aplicação assim como do volume disponível de instâncias, eliminar esse tipo de instância pode ser uma escolha conveniente, pois é fácil e rápida de ser implementada e não interfere com os dados, no sentido de introduzir dados artificialmente gerados. Entretanto, se o volume de instâncias disponível for baixo, essa não é uma opção a ser considerada e, portanto, um processo para a imputação dos valores faltantes deve ser implementado.

Um processo de imputação basicamente detecta valores ausentes em instâncias de dados e os substitui por valores plausíveis [Schafer 1999]. Na literatura podem ser encontrados vários algoritmos que implementam processos de imputação, tais como aqueles comparados em [Grzymala-Busse & Hu 2000]. Geralmente algoritmos de imputação calculam os valores ausentes com base em outros valores do mesmo atributo que estão presentes nas instâncias de dados, levando em consideração a classe (ou não) à qual a instância pertence.

4.6 Considerações Finais

Este capítulo apresentou brevemente três algoritmos de aprendizado de máquina e discutiu, de uma maneira geral, dois métodos utilizados em pré-processamento de instâncias de treinamento, com vistas a melhorar a qualidade de tais instâncias: aquela

fornecida por métodos de seleção de atributos relevantes e aquela fornecida por métodos de imputação de valores ausentes de atributos. Os três algoritmos e as duas técnicas fazem parte da metodologia de avaliação do uso de algoritmos de aprendizado de máquina para a predição de desempenho acadêmico em dados reais, abordada no Capítulo 5.

Capítulo 5
Metodologia Adotada

5.1 Introdução

Este capítulo está organizado em três seções principais. A Seção 5.2 revisita os dados coletados, com foco nos experimentos, com o objetivo de recuperar a nomenclatura adotada. A Seção 5.3 define a metodologia que foi utilizada para a realização dos experimentos relacionados ao uso de algoritmos de AM na predição de desempenho acadêmico de alunos universitários, descritos nos próximos dois capítulos. Com o objetivo de promover a organização do capítulo, a Seção 5.3 é abordada dividida em três subseções: a Seção 5.3.1 contempla a descrição do tratamento de valores ausentes de atributos, a Seção 5.3.2 comenta sobre o processo de formatação de arquivos de dados utilizado no ambiente Weka e, na Seção 5.3.3, é feita uma breve apresentação de alguns dos possíveis métodos para seleção de atributos, existentes junto ao ambiente Weka.

5.2 Sobre as Instâncias de Dados Utilizadas

Como informado no Capítulo 3, o curso alvo desta pesquisa é o Bacharelado em Ciência da Computação (BCC), do Centro Universitário, que é oferecido em regime semestral. Dentre as disciplinas oferecidas pelo BCC, esse trabalho usa dados relativos à disciplina de Construção de Algoritmos e Programação I (CAP1), parte da grade curricular do curso, que foram extraídos do banco de dados do Centro Universitário. Como informado no Capítulo 3, os dados de cada estudante são divididos em dois conjuntos: (1) o *Registro de Estudante*, que reúne informações pessoais e de identificação de um determinado estudante (Tabela 3.1 – parte superior) e (2) o *Registro de Dados Acadêmicos*, que reúne informações do desempenho acadêmico do estudante (Tabela 3.1 – parte inferior).

Embora o Registro do Estudante tenha 18 campos relativos a informações gerais sobre o estudante, tais como nome, endereço, filiação, religião, etc., para os experimentos foram consideradas apenas as informações dos seguintes campos: (5) Data de

Nascimento, (7) Sexo, (8) Estado Civil e (15) Maior Grau de Instrução. Já do Registro de Dados Acadêmicos, foram utilizadas as informações contidas nos campos (4) Nota no vestibular, (9) Lista de Disciplinas Cursadas e (10) Histórico Escolar. Ainda assim, esse conjunto de atributos inicial sofreu uma redução, por meio do uso de métodos de seleção de atributos, como discutido na Seção 4.6.1.

O impacto de atributos, bem como o de determinadas combinações de atributos, nos resultados (*i.e.*, aluno aprovado ou reprovado), foi também um dos objetivos dos experimentos conduzidos. A escolha da disciplina Construção de Algoritmos e Programação I (CAP1) foi também motivada pelo fato de ser uma disciplina do primeiro ano de curso, o que tornou os conjuntos de dados utilizados mais apropriados (em termos de volume de dados), quando da realização dos experimentos. As convenções de nomenclatura mostradas na Tabela 5.1 são usadas nas tabelas mostradas no Apêndice, em que cada uma delas apresenta o conjunto de dados utilizado relativo a uma oferta da disciplina CAP1. É importante mencionar que as informações AF e G são equivalentes, considerando que o valor de G é gerado como resultado da regra *Se* $AF \geq 6{,}0 \rightarrow$ *então A, caso contrário R.*

Nos experimentos foram utilizados os dados relacionados a alunos que efetivamente cursaram a disciplina CAP1 até sua finalização (tendo sido aprovados ou não), em sete anos que foi oferecida, especificamente, 2009, 2010, 2011, 2012, 2013, 2014 e 2015. Portanto, o número total de conjuntos de dados utilizados foi 7. Cada conjunto de dados brutos foi reciclado, por meio da remoção das instâncias relativas a alunos desistentes. Foi removido um total de 140 instâncias associadas a alunos desistentes nos vários anos considerados.

Tabela 5.1 Convenções de Nomenclatura sobre Informações Pessoais, Graus de Instrução e Esquemas de Avaliação.

Abreviação	Significado

(1) DN	Data de Nascimento (Ano)
(2) EC	Estado Civil (S, C, D, V)
(3) S	Sexo (M, F)
(4) GI	Grau de Instrução (MC, SI, SC, ES, ME, DO)
(5) VE	Nota do Vestibular (valor entre 0 e 10)
(6) AV_i ($i \in \{1,2,3,4\}$)	Instância de Avaliação (valor entre 0 e 10)
(7) LE_i ($i \in \{1,2,3,4\}$)	Instâncias de Listas de Exercícios (valor entre 0 e 10)
(8) AS_i ($i \in \{1,2\}$)	Instâncias de Avaliação Substitutiva (valor entre 0 e 10)
(9) PI	Prova Interdisciplinar (valor entre 0 e 1)
(10) F	Número de faltas na disciplina (valor entre 0 e 72)
(11) AF	Avaliação Final (valor entre 0 e 10 – ver Tabela 3.5)
(12) LD	Nota de Lógica e Mat. Discreta (valor entre 0 e 10)
(13) FC	Física Computacional (valor entre 0 e 10)
(14) LP	Laboratório de Programação (valor entre 0 e 10)
(15) G	Informação artificialmente inserida indicando se o aluno foi Aprovado (A) ou Reprovado (R). $Se\ AF \geq 6{,}0 \rightarrow A;\ Se\ AF < 6{,}0 \rightarrow R$

A Tabela 5.2 é uma reapresentação da Tabela 3.13, na qual o número de instâncias de dados associadas a cada um dos conjuntos de dados utilizados nos experimentos está evidenciada em negrito, juntamente com a identificação do conjunto de dados que as armazena. Como pode ser visto na Tabela 5.2, os experimentos envolvem 357 instâncias de dados, que representam alunos que efetivamente cursaram a disciplina CAP1 até seu final, tendo sido aprovados ou reprovados, em um período de 7 anos.

Tabela 5.2 Alunos da disciplina CAP1 por ano (período: 2009–2015). TI: total de alunos inscritos, TD: total de alunos desistentes e TF: total de alunos que concluíram a disciplina (aprovados ou não)/identificação do conjunto de dados associados.

ANO	TI	TD	TF/CONJ_DADOS
2009	62	12	**50/ CAP1_2009_E1**
2010	89	48	**41/ CAP1_2010_E2**
2011	56	13	**43/ CAP1_2011_E3**
2012	70	11	**57/ CAP1_2012_E4**
2013	78	18	**60/ CAP1_2013_E5**
2014	52	14	**37/ CAP1_2014_E5**
2015	99	28	**69/ CAP1_2015_E6**
TOTAL	506	144	**357**

Apesar dos 7 conjuntos compartilharem inúmeras informações, muitas outras são compartilhadas apenas por alguns dos arquivos. É o caso, por exemplo, das listas de

exercícios, que comparecem no conjunto CAP1_2009_E1 e em mais nenhum, ou então, de PI, que comparece em quatro dos sete conjuntos CAP1_2011_E3, CAP1_2013_E5, CAP1_2014_E5 e CAP1_2015_E6. A Tabela 5.3 exibe as informações contidas em cada um dos 7 arquivos e ajuda na visualização das informações compartilhadas entre todos eles, entre apenas alguns, ou exclusivas de um único conjunto.

Tabela 5.3 Informações contidas em cada um dos sete conjuntos de instâncias utilizados nos experimentos.

Info	CAP1_2009_E1	CAP1_2010_E2	CAP1_2011_E3	CAP1_2012_E4	CAP1_2013_E5	CAP1_2014_E5	CAP1_2015_E6
DN	×	×	×	×	×	×	×
EC	×	×	×	×	×	×	×
S	×	×	×	×	×	×	×
GI	×	×	×	×	×	×	×
VE	×	×	×	×	×	×	×
AV_1	×	×	×	×	×	×	×
AV_2	×	×	×	×	×	×	×
AV_3	×	×	×	×	×	×	
AV_4				×			
LE_1	×						
LE_2	×						
LE_3	×						
LE_4	×						
AS_1			×				
AS_2			×				
PI			×		×	×	×
F	×	×	×	×	×	×	×
AF	×	×	×	×	×	×	×
LD	×	×	×	×	×	×	×
FC	×	×	×	×	×	×	
LP							×
G	×	×	×	×	×	×	×

5.3 Descrição da Metodologia Adotada para os Experimentos

A Figura 5.1, extraída de [Fayyad, Piatetsky-Shapiro & Smyth 1996], mostra um diagrama do fluxo de processos pelos quais dados brutos passam, até serem 'traduzidos' em conhecimento. No caso específico deste trabalho de pesquisa, tal diagrama pode ser abordado como uma metodologia geral de trabalho que, a partir de conjuntos de dados,

induz classificadores que permitem inferir o desempenho escolar de alunos de uma disciplina (CAP1), durante o período que a disciplina está sendo ministrada.

Em situações em que os dados disponibilizados para a experimentação com técnicas de MDE são dados brutos, é importante que passem por um processo de tratamento, considerado um pré-processamento, com vistas à remoção de alguns problemas típicos que podem ser encontrados em tais dados.

No trabalho realizado uma inspeção inicial dos dados brutos coletados permitiu a detecção de um problema típico frequentemente encontrado em tal tipo de dado *i.e.*, o de atributo com valor ausente. O problema de atributo com valor ausente, como brevemente abordado na Seção 4.6.2 pode ser tratado de várias maneiras. No contexto deste trabalho, foi tratado via imputação.

A seleção de atributos relevantes (que descrevem as instâncias de dados) é também um outro aspecto do pré-processamento dos dados brutos, uma vez que atributos redundantes e/ou irrelevantes interfere nos resultados obtidos por algoritmos de MDE. Esse assunto foi abordado na Seção 4.6.1 e, para os experimentos descritos neste trabalho, foi incorporado no pré-processamento, a possibilidade de selecionar atributos.

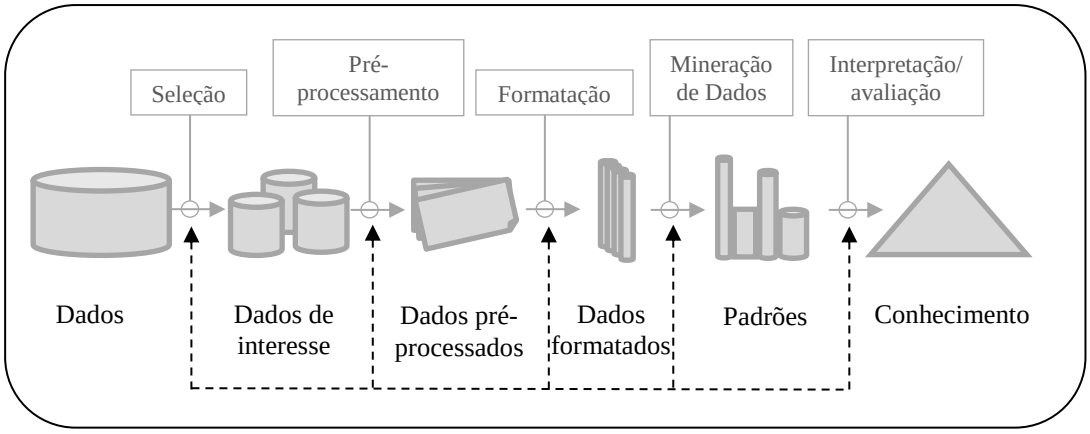

Figura 5.1 Passos para extração de conhecimento dos dados, envolvendo técnicas de mineração (figura extraída de [Fayyad, Piatetsky-Shapiro & Smyth 1996]).

Espelhada na Figura 5.1 e customizada ao problema em questão, um diagrama da metodologia adotada para os experimentos é apresentado na Figura 5.2. Nele, cada conjunto de dados é pré-processado por meio de um processo de imputação, realizado manualmente por meio da atribuição de valores àqueles atributos cujos respectivos valores estavam ausentes (ver Seção 5.3.1) e, em seguida, por um processo de formatação (ver Seção 5.3.2), também realizado manualmente, de maneira a formatar os dados de acordo com a sintaxe exigida pelo Weka. O conjunto de dados é então submetido a um

processo de seleção de atributos (ver Seção 5.3.3), via Weka e o conjunto de dados, reescrito considerando apenas os atributos relevantes, é entrada para um dos algoritmos de MDE (discutidos nas seções seguintes à 5.3.3). O classificador aprendido é então analisado com relação ao seu desempenho. Os blocos da Figura 5.2 têm diferentes formatos com o objetivo de distinguir dados de entrada, dos passos que foram realizados manualmente, por meio de edição de texto, como é o caso tanto da imputação quanto da formatação, e dos passos realizados automaticamente, quando do uso do Weka. Para a condução dos experimentos descritos na Seção 5.4 a metodologia geral descrita pelo fluxograma da Figura 5.2 foi adotada, para cada um dos sete conjuntos de instâncias existentes, *i.e.*, CAP1_2009_E1, CAP1_2010_E2, CAP1_2011_E3, CAP1_2012_E4, CAP1_2013_E5, CAP1_2014_E5 e CAP1_2015_E6.

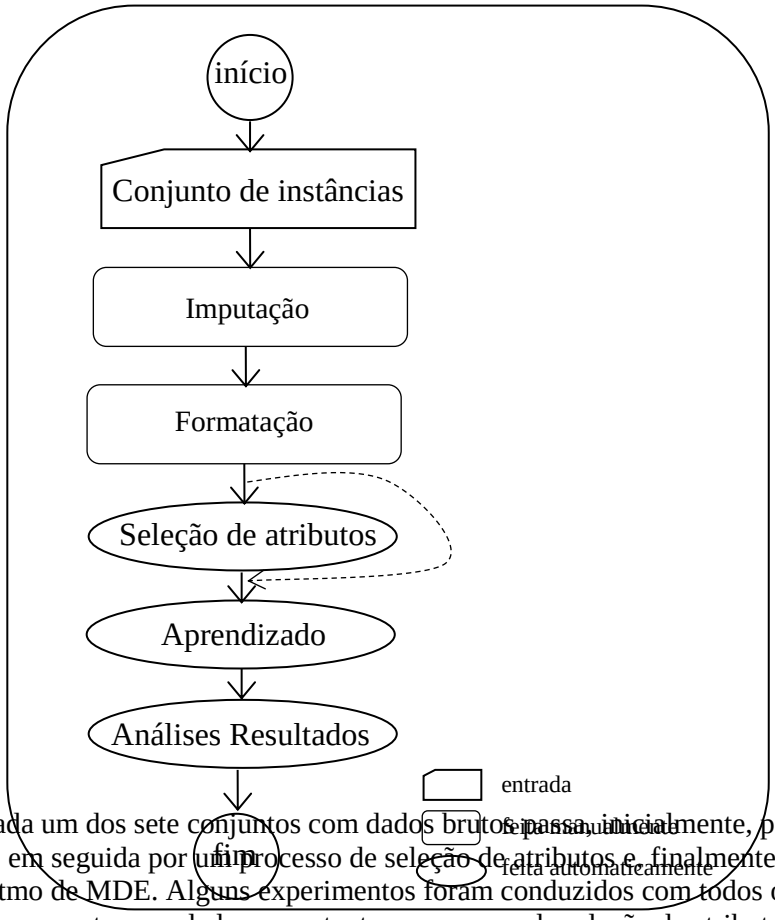

Figura 5.2 Cada um dos sete conjuntos com dados brutos passa, inicialmente, por um processo de imputação, em seguida por um processo de seleção de atributos e, finalmente, é entrada para um algoritmo de MDE. Alguns experimentos foram conduzidos com todos os atributos originalmente presentes nos dados e, portanto, o processo de seleção de atributos nem sempre foi realizado.

5.3.1 Tratamento de Valores Ausentes de Atributos

Várias das 357 instâncias de dados brutos selecionadas tinham valores ausentes de atributos e foram tratadas por meio de imputação de valores. Cada instância com valor ausente de atributo numérico teve o valor ausente imputado como a média dos valores do mesmo atributo, que comparecem nas outras instâncias. Já cada instância cujo atributo com valor ausente era alfanumérico, o valor ausente foi imputado como o valor mais frequente que comparece nas instâncias com valor de atributo. A Tabela 5.4 mostra o total de atributos com valores ausentes em cada um dos conjuntos de instâncias, descritos nas tabelas contidas nos apêndices A.4 a A.10. A Tabela 5.5 evidencia os números associados à falta de valor de atributo, para todos os conjuntos de dados considerados. A Tabela 5.6 descreve a técnica usada para a imputação de valores, quando de sua ausência, para cada um dos atributos com o problema.

Tabela 5.4 Nro. de atributos com valores ausentes em cada um dos conjuntos de dados.

CONJ_DADOS	Tabela	EC	GI	VE	LD	FC
CAP1_2009_E1	A.4	1	23	7	1	1
CAP1_2010_E2	A.5	6	20	6		
CAP1_2011_E3	A.6	1	28	1	1	
CAP1_2012_E4	A.7	1	7	8		
CAP1_2013_E5	A.8	6	42	27	1	
CAP1_2014_E5	A.9	5	24	9		
CAP1_2015_E6	A.10	23	27	25	3	
TOTAL	–	43	171	83	6	1

Tabela 5.5 Número de instâncias de dados de cada um dos conjuntos de dados, que tiveram algum(s) de seus atributos com valor imputado. TF: total de instâncias no ano; TII: total de valores de atributos imputados e CONJ_DADOS: conjunto de dados associado.

ANO	TF	TII	CONJ_DADOS
2009	50	33	CAP1_2009_E1
2010	41	32	CAP1_2010_E2
2011	43	31	CAP1_2011_E3
2012	57	16	CAP1_2012_E4
2013	60	76	CAP1_2013_E5
2014	37	38	CAP1_2014_E5
2015	69	78	CAP1_2015_E6
TOTAL	357	304	–

Os procedimentos descritos na Tabela 5.6 foram adotados para a atribuição de valor àqueles atributos com valores ausentes (*i.e;* evidencia a regra usada para a imputação do valor ausente).

Seguem algumas poucas informações pertinentes, relacionada aos dados utilizados. A instituição fornecedora dos dados, além de centro universitário, também oferece outras modalidades de ensino como, educação infantil, ensino fundamental I, ensino fundamental II, ensino médio, pós-graduação lato e *stricto sensu*.

Tabela 5.6 Atributos com valores ausentes em cada um dos conjuntos de dados.

Atributo	Nome	Tratamento do valor ausente
EC	Estado Civil	valor com maior frequência
GI	Grau Instrução	valor com maior frequência
VE	Nota Vestibular	média dos valores presentes
LD	Nota (Lógica e Mat. Discreta)	média dos valores presentes
FC	Nota (Física Computacional)	média dos valores presentes

Também, foram feitas algumas correções em valores do atributo *Grau de Instrução (GI)*, que estavam incorretamente preenchidos – foram corrigidos 10 registros de dados. Via inspeção manual dos dados percebeu-se que, por vezes, o cadastro apresentava o valor do atributo *GI* desatualizado, uma evidência que algumas das informações, relativas ao cadastro geral do aluno, não tinham sido atualizadas com frequência. O atributo *GI*, referente ao maior grau de instrução cursado pelo estudante até o momento, em alguns registros, estava desatualizado e continha o valor *Fundamental Incompleto (FI) ou Fundamental Completo (FC)*, em uma óbvia referência ao período em que o estudante ingressou na instituição. Apesar da instituição possuir a documentação obrigatória que comprova a conclusão do ensino médio, exigida pelo MEC para que o estudante ingresse no ensino superior, em algumas situações o atributo *GI* não refletia isso. Sendo assim, quando o valor mostrado no atributo *GI* fazia referência a algum grau de instrução anterior ao ensino médio, ou estava ausente, foi imputado o valor MC ou Ensino Médio Completo, que é o mínimo exigido para se cursar o ensino superior.

Com relação ao atributo *Nota do Vestibular (VE)*, além da nota no vestibular nacional da instituição, ocorre o aproveitamento da nota do ENEM – Exame Nacional do Ensino Médio, para os estudantes que realizaram o exame. Ambas as notas são consideradas para ingresso na instituição. Neste caso ocorreram duas alterações, uma de formatação, exibindo a nota com apenas uma casa decimal, e os valores ausentes, foram imputados com a média aritmética dos alunos que possuíam nota, nesta turma. A nota do vestibular é fator aprovador para que o estudante ingresse na instituição. O ingresso se dá pela classificação de notas da maior para a menor e pelo preenchimento da quantidade de

vagas disponíveis no curso. Valores do atributo *Estado Civil (EC)*, quando ausentes, foram imputados com o valor mais frequente de tal atributo na turma em questão.

É importante lembrar que o atributo *Número de Faltas (F)* pode ser determinante na aprovação ou reprovação do aluno. Como informado na Tabela 5.1 o valor do atributo F está entre 0 e 72. O estudante deve ter, pelo menos, 75% (54) de comparecimento às aulas, para ser considerado para aprovação (na dependência do desempenho obtido); caso sua frequência tenha sido menor que 25% (18), o estudante é sumariamente reprovado por faltas.

5.3.2 Formatação do Arquivo de Dados

A formatação do conjunto de instâncias é, basicamente, um processo de edição de texto, que foi feito manualmente, de maneira a deixar a descrição das instâncias de dados de acordo com a sintaxe exigida pelo ambiente Weka. O conjunto de 14 instâncias de treinamento, do conjunto *jogar tênis*, já apresentado no Capítulo 4 (Tabela 4.1), volta a ser apresentado nesta seção, com o objetivo de exemplificar o processo de formatação do arquivo de instâncias, requisitado pelo Weka. Nele, entretanto, o valor do atributo temperatura, ao invés de nominal, passou a ser numérico.

Tabela 5.7 Conjunto de treinamento do tempo, que indica se é possível jogar tênis ou não em determinadas condições climáticas [Witten & Frank 2005].

aspecto	*temperatura*	*umidade*	*ventoso*	*classe*
ensolarado	38	alta	falso	não(n)
ensolarado	37	alta	verdade	não(n)
nublado	39	alta	falso	sim(s)
chuvoso	27	alta	falso	sim(s)
chuvoso	18	normal	falso	sim(s)
chuvoso	17	normal	verdade	não(n)
nublado	19	normal	verdade	sim(s)
ensolarado	25	alta	falso	não(n)
ensolarado	20	normal	falso	sim(s)
chuvoso	27	normal	falso	sim(s)
ensolarado	26	normal	verdade	sim(s)
nublado	24	alta	verdade	sim(s)
nublado	33	normal	falso	sim(s)
chuvoso	25	alta	verdade	não(n)

Um arquivo de instâncias de dados para o Weka tem que estar em um arquivo ARFF (*Attribute-Relation File Format*), que é um arquivo texto em ASCII, que contém o conjunto de instâncias de dados, descritas pelo mesmo conjunto de atributos. Um arquivo ARFF deve ter duas seções distintas. A seção *Header*, que contém o nome da relação, a

lista dos atributos (*i.e.*, as colunas de uma descrição tabular das instâncias, tal como a mostrada na Tabela 5.7), e a informação sobre os seus respectivos tipos. A seção *Data* contém a descrição de todas as instâncias, uma por linha, com os valores de atributos separados por vírgulas e a classe comparecendo ao final da descrição da instância. Linhas inicializadas com o símbolo de porcentagem (%) são interpretadas como comentários. O arquivo ARFF associado aos dados da Tabela 5.7 está mostrado na Figura 5.3.

```
% ARQUIVO ARFF CORRESPONDENTE ÀS INSTÂNCIAS
% DA TABELA 5.7

@relation tempo

@attribute aspecto {ensolarado, nublado, chuvoso}
@attribute temperatura real
@attribute umidade {alta, normal}
@attribute ventoso {falso, verdade}
@attribute classe  {s, n}

@data
% segue a descrição das 14
% instancias de dados da
% Tabela 5.7

ensolarado, 38, alta, falso, n
ensolarado, 37, alta, verdade, n
nublado, 39, alta, falso, s
chuvoso, 27, alta, falso, s
chuvoso, 18, normal, falso, s
chuvoso, 17, normal, verdade, n
nublado, 19, normal, verdade, s
ensolarado, 25, alta, falso, n
ensolarado, 20, normal, falso, s
chuvoso, 27, normal, falso, s
ensolarado, 26, normal, verdade, s
nublado, 24, alta, verdade, s
nublado, 33, normal, falso, s
chuvoso, 25, alta, verdade, n
```

Figura 5.3 Arquivo ARFF para as instâncias de dados da Tabela 5.7.

5.3.3 Seleção de Atributos Relevantes

A Tabela 5.8 lista alguns dos métodos de avaliação de subconjuntos de atributo (e também de avaliação de atributos individuais) apresentados em Witten & Frank [Witten & Frank 2005], a serem usados em combinação com algum dos métodos de busca/ranqueadores, também disponibilizados naquele ambiente, como listados na Tabela

5.9. Note que estes não são os únicos métodos disponíveis no ambiente Weka (versão 3.6.12). Ambas as tabelas possuem duas partes, a parte superior corresponde aos métodos desenvolvidos para serem aplicados a conjuntos de atributos, e a parte inferior corresponde a métodos a serem aplicados a atributos individuais.

Tabela 5.8 Métodos para avaliação e seleção de atributos, retirados de [Witten & Frank 2005].

AVALIAM SUBCONJUNTOS DE ATRIBUTOS	
NOME	FUNÇÃO
(1) CfsSubsetEval	Considera o valor preditivo de cada atributo individualmente, junto com o grau de redundância entre eles.
(2) ClassifierSubsetEval	Usa um classificador para avaliar o subconjunto de atributos.
(3) ConsistencySubsetEval	Projeta o conjunto de treinamento sobre o conjunto de atributos e mede a consistência nos valores das classes.
(4) WrapperSubsetEval	Usa um classificador e validação cruzada.
AVALIAM UM ÚNICO ATRIBUTO	
NOME	FUNÇÃO
(5) ChiSquaredAttributeEval	Calcula o valor da medida chi-quadrado de cada atributo com relação à classe.
(6) GainRatioAttributeEval	Avalia o atributo com base no *gain ratio.*
(7) InfoGainAttributeEval	Avalia o atributo baseado no *information gain*.
(8) OneRAttributeEval	Utiliza a metodologia *OneR* para avaliar atributos.
(9) PrincipalComponents	Executa PCA (*Principal Components Analysis*)
(10) ReliefAttributeEval	Usa um valiador de atributo baseado em instâncias.
(11) SVMAttributeEval	Usa SVM (*Support Vector Machine*) linear para avaliar o valor do atributo.
(12) SymmetricalUncertAttributeEval	Avalia o atributo baseado na incerteza simétrica.

Tabela 5.9 Métodos de busca para a seleção de atributos, retirados de [Witten & Frank 2005].

MÉTODOS DE BUSCA PARA A SELEÇÃO DE ATRIBUTOS	
NOME	FUNÇÃO
(1) BestFirst	*Hill-climbing* com *backtracking*.
(2) ExhaustiveSearch	Busca exaustiva.
(3) GeneticSearch	Busca usando um algoritmo genético simples.
(4) GreedyStepwise	*Hill-climbing* sem *backtracking*. Opcionalmente gera uma lista ordenada de atributos.
(5) RaceSearch	Usa a metodologia *race search.*
(6) RandomSearch	Usa busca randômica.
(7) RankSearch	Classifica os atributos e 'ranqueia' os subconjuntos de atributos relevantes, usando um avaliador de subconjunto de atributos.
MÉTODOS DE *RANKING*	
NOME	FUNÇÃO
(8) Ranker	Ranqueia os atributos (individualmente) de acordo com a sua correspondente avaliação.

No contexto do trabalho realizado os experimentos conduzidos com os métodos de avaliação de subconjuntos de atributos (ou o seu ranqueamento) estão apresentados em detalhes no Capítulo 6 e são utilizados para a condução de parte dos experimentos apresentados no Capítulo 7, que trata do uso de algoritmos de MDE nos dados coletados, com vistas à predição de desempenho acadêmico.

Capítulo 6
Experimentos Relacionados à Seleção de Atributos Relevantes

6.1 Introdução

Este capítulo aborda experimentos relacionados ao processo de seleção de atributos relevantes dos dados considerados. De acordo com a metodologia adotada, como já comentado, tal passo tem como objetivo simplificar o processo de AM utilizando os algoritmos escolhidos, que será apresentado em detalhes no Capítulo 7. Os experimentos aqui descritos foram realizados adotando técnicas associadas às duas possíveis abordagens disponíveis à seleção de atributos relevantes, disponibilizadas no ambiente Weka: (1) a que seleciona subconjuntos relevantes, abordada na Seção 6.2 e (2) a que 'ranqueia' atributos individuais, por ordem de relevância na determinação da classe, tratada na Seção 6.3. A Seção 6.4 apresenta o resumo dos resultados relacionados a identificação dos atributos relevantes.

Note que em todos os conjuntos de instâncias o atributo AF (Nota da Avaliação Final, calculada de acordo com as regras descritas na Tabela 3.5) não foi incluído. Como esse atributo representa implicitamente a classe à qual a instância pertence *i.e.*, se AF ≥ 6,0 aluno aprovado, caso contrário reprovado, tal atributo foi discretizado e renomeado como G, que assume um dentre dois possíveis valores: A(provado) e R(eprovado). Também, os conjuntos de dados considerados já passaram pelo processo de tratamento de valor ausente de atributo, via imputação.

6.2 Seleção de Subconjuntos de Atributos Relevantes

Nesta seção são abordados os experimentos realizados considerando cada um dos sete conjuntos de dados disponíveis, dois dos métodos disponibilizados para a seleção de subconjuntos de atributos relevantes *i.e.*, os métodos *CfsSubsetEval* e *ConsistencySubsetEval*, listados na parte superior da Tabela 5.8. Como comentado na Seção 5.3.3 do Capítulo 5, no ambiente Weka os métodos de seleção são utilizados em

combinação com um dentre os métodos de busca disponibilizados. Para os experimentos foram escolhidos seis métodos de busca, *BestFirst*, *ExhaustiveSearch*, *GeneticSearch*, *GreedyStepwise*, *RandomSearch* e *RankSearch*. Cada uma das próximas sete subseções aborda o processo de seleção de subconjuntos de atributos associado a um particular conjunto de dados, aquele referenciado no título da seção.

6.2.1 Seleção em CAP1_2009_E1

O conjunto de dados do ano de 2009 (atributos na Tabela 6.1) foi usado como um protótipo para o estabelecimento de uma metodologia a ser seguida, para os demais testes (*i.e.*, aqueles relativos aos demais conjuntos de dados). Lembrando, tal conjunto tem 50 instâncias de dados descritas por 15 atributos e 1 classe (G) associada, sendo que em 40 instâncias, G=A(provado) e nas 10 restantes, G=R.

Tabela 6.1 Atributos que descrevem o conjunto de instâncias CAP1_2009_E1.

DN (Ano Nasc.)	VE (Nota Vest.)	LE_1 (Nota Exer. 1)	F (Nro. Faltas)
EC (Est. Civ.)	AV_1 (Nota Avaliação 1)	LE_2 (Nota Exer. 2)	LD (Log. Mat.Disc.)
S (Sexo)	AV_2 (Nota Avaliação 2)	LE_3 (Nota Exer. 3)	FC (Fis. Comp.)
GI (Grau Inst.)	AV_3 (Nota Avaliação 3)	LE_4 (Nota Exer. 4)	*G (Avaliação Final)*

A Tabela 6.2 mostra os resultados dos experimentos realizados com os dois métodos de seleção de subconjuntos de atributos e suas combinações com os seis métodos de busca escolhidos, utilizando o conjunto de dados CAP1_2009_E1.

Tabela 6.2 Resultados do processo de seleção de subconjuntos de atributos usando o conjunto de dados CAP1_2009_E1.

CAP1_2009_E1		
Método de avaliação	Método de busca	Atributos selecionados
CfsSubsetEval	*BestFirst*	S, AV1, AV2, AV3
	ExhaustiveSearch	S, AV1, AV2, AV3
	GeneticSearch	S, AV1, AV2, AV3
	GreedyStepwise	S, AV1, AV2, AV3
	RandomSearch	S, AV1, AV2, AV3
	RankSearch	S, AV1, AV2, AV3
ConsistencySubsetEval	*BestFirst*	EC, S, AV1
	ExhaustiveSearch	EC, S, AV1
	GeneticSearch	EC, AV1, AV2, LE1, LE3
	GreedyStepwise	AV1
	RandomSearch	EC, S, GI, AV1, LE3, LE4, F, FC
	RankSearch	AV1, AV2, AV3

Com base nos resultados mostrados na Tabela 6.2, os seguintes comentários e análises podem ser feitos:

(1) A articulação do método de avaliação *CfsSubsetEval,* com qualquer dos métodos de busca (*BestFirst, ExhaustiveSearch, GeneticSearch, GreedyStepwise, RandomSearch e RankSearch*) utilizados, gerou sempre o mesmo conjunto de atributos como resultado *i.e.*, {S, AV1, AV2, AV3}.

(2) O atributo *Sexo (S)* foi selecionado para compor o conjunto de atributos relevantes em 9 dos 12 experimentos realizados, o que corresponde a 75,00% dos resultados. Apesar de não ser um atributo relevante para o estabelecimento da classe, a seleção de tal atributo pode, eventualmente, ter sido consequência do desbalanceamento do conjunto de dados com relação aos dois possíveis valores que o atributo S pode assumir (F ou M) e, também, ao desbalanceamento entre as duas classes A & R. Do total de 50 instâncias, (a) 48 delas têm S=M, sendo que em 39 dessas 48, G=A e nas 9 restantes, G=R; (b) apenas 2 têm S=F, sendo uma delas com G=A e a outra com G=R.

(3) Situação ligeiramente semelhante acontece com o atributo *Estado Civil (EC),* que foi selecionado para compor o conjunto de atributos relevantes em 4 dos 12 experimentos, o que corresponde a 33,34% dos resultados. Sua seleção pode também ter sido consequência do desbalanceamento dos dados. Do total de 50 instâncias, (a) 46 delas têm EC=S, sendo que em 37, G=A e nas 9 restantes, G=R; (b) Apenas 4 têm EC=C, sendo três delas com G=A e apenas uma com G=R.

(4) O atributo *Número de Faltas (F),* que implica diretamente em reprovação só foi selecionado para compor 1 dos 12 conjuntos de atributos obtidos (representando 0,33%), muito embora em nenhuma instância do conjunto de dados sendo usado, a situação que implica reprovação por faltas *i.e., se* $F > 18 \rightarrow R$, tenha acontecido. No caso particular desse conjunto de dados, F foi um atributo que não contribuiu para a determinação da reprovação mas, em contrapartida, todas as instâncias em que G=A, o valor de F = 0.

(5) Os três atributos que representam os valores em três avaliações *i.e.*, AV1, AV2 e AV3, foram selecionados para compor o conjunto de atributos relevantes em 7 dos 12 experimentos (o que representa 58,33%). Os atributos AV1 e AV2 foram escolhidos para compor o conjunto de atributos relevantes em 1 dos 12 experimentos, representando 8,334% e o atributo AV1 foi escolhido para compor o conjunto dos atributos relevantes em 4 dos 12 experimentos, o que representa 33,33%. É interessante notar que o esquema de avaliação E1 (ver tabelas 3.2 e 3.5) regula a aprovação ou reprovação (excetuando o

caso em que faltas estão envolvidas), em função de três avaliações e quatro notas associadas à quatro listas de exercícios. As três avaliações são ponderadas, sendo a AV1 com 0,27, AV2 com 0,31 e a AV3 com 0,36; a terceira avaliação é ligeiramente mais relevante que as outras duas, sendo a segunda ligeiramente mais relevante que a primeira.

O método de seleção *CfsSubsetEval* (combinado com qualquer dos 6 métodos de busca utilizado), detectou os três atributos associados às avaliações como relevantes (muito embora tenha também selecionado S, como comentado anteriormente. O método de seleção *ConsistencySubsetEval*, quando usado com *GeneticSearch* ou *RandomSearch*, foi capaz de detectar a relevância das listas de exercícios, muito embora a seleção não tenha sido consistente, considerando que a ponderação das notas associadas às listas é quase insignificante; também não foi capaz de detectar a importância dos outros atributos associados à avaliação (o AV3, no caso da *GeneticSearch* e o AV1 e AV2 no caso da *RandomSearch*).

(6) O método *ConsistencySubsetEval* usando a *RandomSearch* selecionou, como relevante, o atributo GI, talvez devido também ao desbalanceamento de instâncias, com relação aos valores assumidos por esse atributo.

(7) O atributo FC (nota associada à disciplina *Física Computacional*) aparece selecionado em apenas um experimento (*ConsistencySubsetEval + RandomSearch).*

6.2.2 Seleção em CAP1_2010_E2

Os experimentos realizados com os dois métodos de seleção de subconjuntos de atributos e suas combinações com métodos de busca, utilizando todo o conjunto de dados CAP1_2010_E2 (atributos informados na Tabela 6.3) obtiveram os resultados mostrados na Tabela 6.4. Lembrando, tal conjunto tem 41 instâncias de dados descritas por 11 atributos e 1 classe (G) associada, sendo que em 37 instâncias, G=A(provado) e nas 4 restantes, G=R.

Tabela 6.3 Atributos que descrevem o conjunto de instâncias CAP1_2010_E2.

DN (Ano Nasc.)	VE (Nota Vest.)	F (Nro. faltas)
EC (Est. Civ.)	AV_1 (Nota Avaliação 1)	LD (Log. Mat. Disc.)
S (Sexo)	AV_2 (Nota Avaliação 2)	FC (Fis. Comp.)
GI (Grau Inst.)	AV_3 (Nota Avaliação 3)	*G (Avaliação Final)*

A Tabela 6.4 mostra os resultados dos experimentos realizados com os dois métodos de seleção de subconjuntos de atributos e suas combinações com os seis métodos

de busca escolhidos, utilizando o conjunto de dados CAP1_2010_E2.

Tabela 6.4 Resultados do processo de seleção de subconjuntos de atributos usando o conjunto de dados CAP1_2010_E2.

CAP1_2010_E2		
Método de avaliação	Método de busca	Atributos selecionados
CfsSubsetEval	*BestFirst*	S, AV1, AV2, AV3, LD, FC
	ExhaustiveSearch	S, AV1, AV2, AV3, LD, FC
	GeneticSearch	S, AV1, AV2, AV3, LD, FC
	GreedyStepwise	S, AV1, AV2, AV3, LD, FC
	RandomSearch	S, AV1, AV2, AV3, LD, FC
	RankSearch	S, AV1, AV2, AV3, LD, FC
ConsistencySubsetEval	*BestFirst*	GI, AV2
	ExhaustiveSearch	GI, AV2
	GeneticSearch	GI, AV2
	GreedyStepwise	GI, AV2
	RandomSearch	DN, S, AV1, AV2, AV3, FC
	RankSearch	AV1, AV2, AV3, LD, FC

Os resultados mostrados na Tabela 6.4, de certa forma, seguem alguns padrões evidenciados no experimento anterior. Com relação aos resultados, os seguintes comentários podem ser feitos:

(1) Como ocorreu no experimento anterior, o resultado da combinação do método de avaliação *CfSubsetEval*, com qualquer dos 6 métodos de busca *i.e.*, *BestFirst, ExhastiveSearch, GeneticSearch, GreedyStepwise, RandomSearch e RankSearch*, gerou sempre o mesmo resultado *i.e.*, o conjunto de atributos {S, AV1, AV2, AV3, LD, FC}. Vale aqui o mesmo comentário anterior, relativo à seleção do atributo S para compor o conjunto de atributos relevantes. Interessante, nesse experimento, foi a seleção dos atributos LD e FC (notas associadas às disciplinas *Lógica e Matemática Discreta* e *Física Computacional*, respectivamente) que, obviamente, não participam do esquema E2 de cálculo da nota final, que caracteriza aprovação ou reprovação. O Esquema E2 é o mais simples dos 5 esquemas utilizados e consiste na média aritmética das notas associadas às três avaliações, AV1, AV2 e AV3. Por inspeção do conjunto de dados pode ser evidenciado que os valores associados com LD e FC correlacionam com o valor da classe e pode ter sido essa a razão da escolha de ambos os atributos para compor o conjunto de atributos relevantes.

A Figura 6.1 apresenta graficamente, em um espaço bidimensional, as instâncias do conjunto de dados representadas apenas pelos valores do atributo LD (nota na disciplina *Lógica e Matemática Discreta*) e AF (*Avaliação Final*), do conjunto de 41

instâncias, no gráfico estão registrados 29 pontos (LD,AF), removidos os 12 valores duplicados.

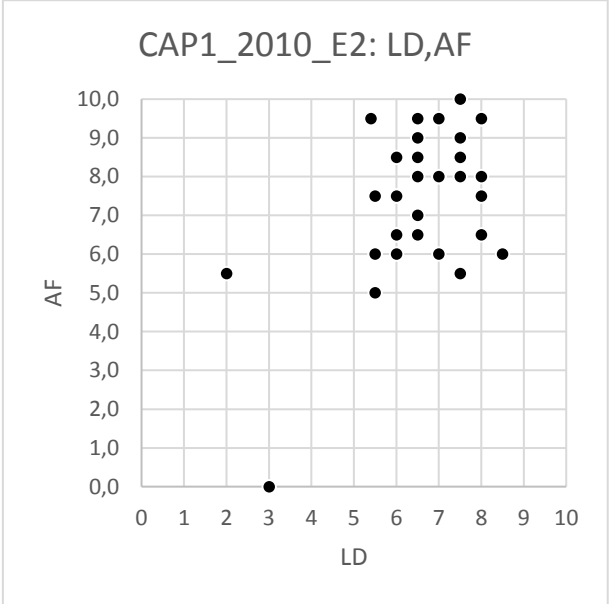

Figura 6.1 Representação gráfica do conjunto de dados CAP1_2010_E2, em que cada instância de dados distinta é descrita por um ponto bidimensional definido pelos atributos (LD,AF). São 41 instâncias no conjunto, com 12 valores duplicados, resultando em 29 pontos com valores exclusivos representados no gráfico.

(5) O atributo S comparece em 7 dos conjuntos de atributos relevantes selecionados; considerando que 12 desses conjuntos foram obtidos, esse valor representa 58,34%. O comentário feito anteriormente com relação à seleção do atributo S quando do uso do conjunto CAP1_2009_E1, continua pertinente, com relação ao experimento discutido nesta seção.

A Figura 6.2 apresenta graficamente, em um espaço bidimensional, as instâncias do conjunto de dados representadas apenas pelos valores do atributo FC (nota na disciplina *Física Computacional*) e AF (*Avaliação Final*), do conjunto de 41 instâncias, no gráfico estão registrados 31 pontos (LD,AF), removidos os 10 valores duplicados.

(6) Além das seis ocorrências de comparecimento de ambos os atributos, LD e FC, em todos os conjuntos de atributos relevantes selecionados usando o método *CfsSubsetEval*, ambos foram também selecionados quando a combinação *ConsistencySubsetEval* + *RankSearch* foi usada e apenas o FC (junto com mais 5 outros), quando o mesmo método de avaliação foi usado com o método de busca *RandomSearch.*

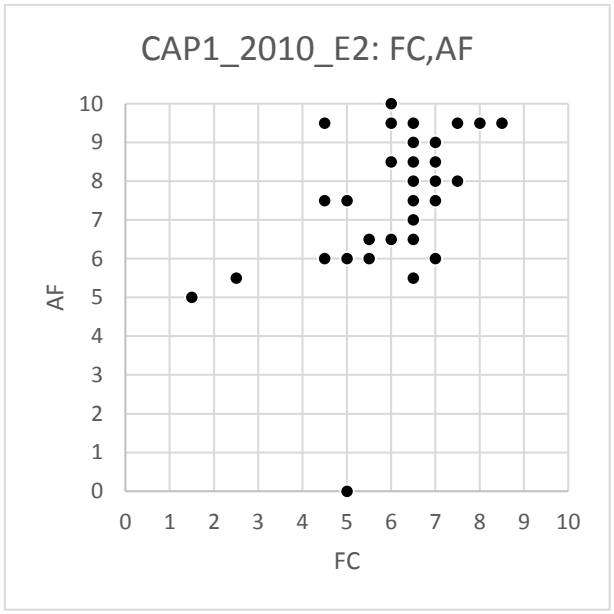

Figura 6.2 Representação gráfica do conjunto de dados CAP1_2010_E2, em que cada instância de dados distinta é descrita por um ponto bidimensional definido pelos atributos (FC,AF). São 41 instâncias no conjunto, com 10 valores duplicados, resultando em 31 pontos com valores exclusivos representados no gráfico.

(7) O método de avaliação *ConsistencySubsetEval*, combinado com qualquer um dos métodos de busca, *BestFirst, ExhaustiveSearch, GeneticSearch e GreedyStepwise*, gerou o mesmo conjunto de atributos, {GI, AV2}.

Do total de 41 instâncias, (a) 40 delas têm GI=MC, sendo que em 37, G=A e nas 3 restantes, G=R; (b) Apenas 1 têm GI=SI, com G=R. É possível que a seleção do atributo GI tenha se dado pelo fato das 37 instâncias com G=A, terem GI=MC.

6.2.3 Seleção em CAP1_2011_E3

Os experimentos realizados com os vários processos de seleção de subconjuntos de atributos e suas combinações com métodos de busca, utilizando o conjunto de dados CAP1_2011_E3 (Tabela 6.5) obtiveram os resultados mostrados na Tabela 6.6. Lembrando, tal conjunto tem 43 instâncias de dados descritas por 14 atributos e 1 classe (G) associada, sendo que em 31 instâncias, G=A(provado) e nas 12 restantes, G=R.

Tabela 6.5 Atributos que descrevem o conjunto de instâncias CAP1_2011_E3.

DN (Ano Nasc.)	VE (Nota Vest.)	PI (Nota Prov. Inter)	LD (Log. Mat. Disc.)
EC (Est. Civ.)	AV_1 (Nota Avaliação 1)	AS_1 (Nota Aval. Sub1)	FC (Fis. Comp.)
S (Sexo)	AV_2 (Nota Avaliação 2)	AS_2 (Nota Aval. Sub2)	*G (Avaliação Final)*
GI (Grau Inst.)	AV_3 (Nota Avaliação 3)	F (Nro. faltas)	

Com base nos resultados mostrados na Tabela 6.6, os seguintes comentários a respeito do conjunto de dados de 2011, podem ser feitos:

Tabela 6.6 Resultados do processo de seleção de subconjuntos de atributos usando o conjunto de dados CAP1_2011_E3.

CAP1_2011_E3		
Método de avaliação	Método de busca	Atributos selecionados
CfsSubsetEval	*BestFirst*	EC, S, AV1, AV2, AS1
	ExhaustiveSearch	EC, S, AV1, AV2, AS1
	GeneticSearch	EC, S, AV1, AV2, AS1
	GreedyStepwise	EC, S, AV1, AV2, AS1
	RandomSearch	S, GI, AV1, AV2, AS1, FC
	RankSearch	EC, S, AV1, AV2, AS1
ConsistencySubsetEval	*BestFirst*	EC, AV2, AS1
	ExhaustiveSearch	EC, AV2, AS1
	GeneticSearch	EC, AV2, AS1, FC
	GreedyStepwise	EC, AV2, AS1
	RandomSearch	EC, GI, VE, AV2, PI, AS1, F
	RankSearch	S, AV1, AV2, AS1

(1) Os dados em CAP1_2011_E3 incorporam, além das notas de três avaliações, duas outras avaliações identificadas como substitutiva (AS1 e AS2). É o único, dentre os 7 conjuntos de dados coletados, que são descritos, também, por esses dois atributos. São relevantes uma vez que colaboram no cálculo da avaliação final (AF = AV1 + AV2 + AV3 + AS1 - AS2)/3 + PI.

Os atributos AV1, AV2, AV3, representam as avaliações regulares, se o aluno tirou uma nota inferior/ou não realizou uma das avaliações necessárias para concluir a disciplina, ele tem uma oportunidade para tentar recuperar esta nota, realizando a AS1 (Avaliação Substitutiva 1). A nota adquirida em AS1 é colocada na fórmula no atributo correspondente (AS1), e a menor nota tirada em uma das três avaliações (AV1, AV2, AV3), que será substituída pela nota de AS1, é lançada no atributo AS2, para que este valor seja removido da avaliação e sejam considerados apenas três valores, para as avaliações.

(2) O método de avaliação *CfSubsetEval*, quando combinado com cinco dos seis métodos de busca *i.e.*, *BestFirst*, *ExhaustiveSearch*, *GeneticSearch*, *GreedyStepwise* e *RankSearch*, gerou o mesmo subconjunto de atributos {EC, S, AV1, AV2, AS1}.

(3) O atributo S comparece em 7 dos 12 resultados gerados como conjuntos de atributos, o que representa 58,22%, mantendo a mesma proporção de comparecimento que a evidenciada em CAP1_2010_E2. Isso de certa forma reforça o comentário realizado sobre o desbalanceamento de valores de atributo no conjunto de dados, feito na Seção 6.2.1.

(4) O atributo EC comparece em 10 dos 12 conjuntos de atributos resultado dos testes realizados, o que corresponde a 83,33%. Esse resultado reforça o comentário feito no item (3) na Seção 6.2.1, o de que o atributo EC pode ter sido selecionado pelo desbalanceamento da classe em relação aos valores do atributo. Do total de 43 instâncias, (a) 37 delas têm EC=S, sendo que, em 26, G = A e nas 11 restantes, G = R; (b) 6 delas têm EC=C, sendo que, em 5, G = A e em 1, G = R.

(5) O atributo GI aparece apenas em 2 dos 12 subconjuntos de atributos, representando 16,66%. Nas duas ocasiões o atributo GI foi selecionado pelo método de busca *RandomSearch*.

(6) O atributo PI, que entra para o cálculo do valor de AF com peso 0,1 (ver item (1)), foi evidenciado como relevante e escolhido para compor o subconjunto de atributos relevantes apenas quando do uso da combinação *ConsistencySubsetEval* + *RandomSearch*.

(7) O atributo FC aparece em apenas 2 dos 12 subconjuntos selecionados e o atributo LD em nenhum deles. Duas plotagens do conjunto CAP1_2011_E3, representado por pontos (LD,AF) e (FC,AF), são mostradas nas Figura 6.3 e Figura 6.4, respectivamente.

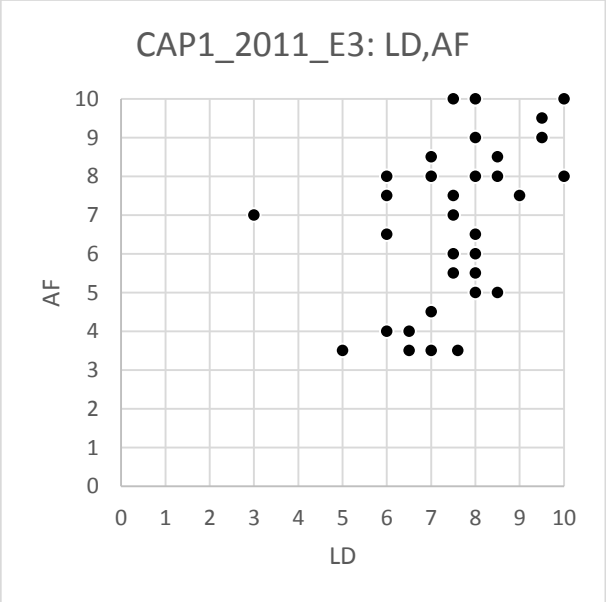

Figura 6.3 Representação gráfica do conjunto de dados CAP1_2011_E3, em que cada instância de dados distinta é descrita por um ponto bidimensional definido pelos atributos (LD,AF). São 43 instâncias no conjunto, com 10 valores duplicados, resultando em 33 pontos com valores exclusivos representados no gráfico.

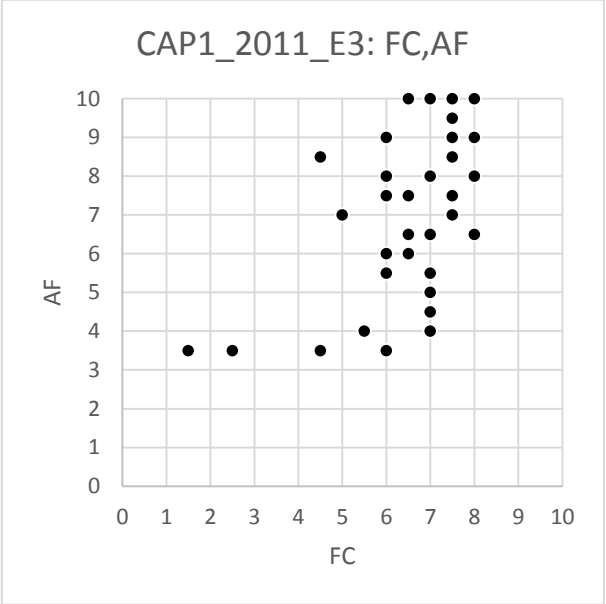

Figura 6.4 Representação gráfica do conjunto de dados CAP1_2011_E3, em que cada instância de dados distinta é descrita por um ponto bidimensional definido pelos atributos (FC,AF). São 43 instâncias no conjunto, com 10 valores duplicados, resultando em 33 pontos com valores exclusivos representados no gráfico.

(8) O atributo VE comparece em apenas 1 conjunto dentre os 12 subconjuntos de atributos selecionados, o que corresponde a 8,33%. Este atributo não participa no cálculo do AF. A Figura 6.5 mostra o conjunto CAP1_2011_E3 representado por suas instâncias descritas apenas pelos atributos VE e FC (i.e., pontos (VE,FC)) para uma visualização gráfica bidimensional do conjunto.

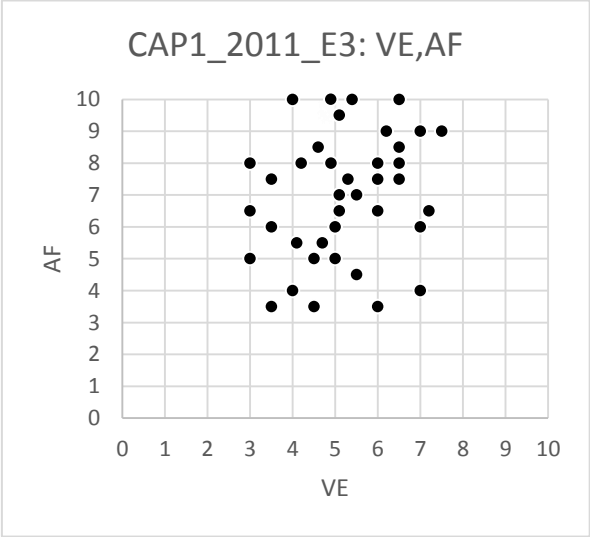

Figura 6.5 Representação gráfica do conjunto de dados CAP1_2011_E3, em que cada instância de dados distinta é descrita por um ponto bidimensional definido pelos atributos (VE,AF). São 43 instâncias no conjunto, com 4 valores duplicados, resultando em 39 pontos com valores exclusivos representados no gráfico.

6.2.4 Seleção em CAP1_2012_E4

Os experimentos realizados com os vários processos de seleção de subconjuntos de atributos e suas combinações com métodos de busca, utilizando o conjunto de dados CAP1_2012_E4 (Tabela 6.7) obtiveram os resultados mostrados na Tabela 6.8. Lembrando, tal conjunto tem 57 instâncias de dados descritas por 12 atributos e 1 classe (G) associada, sendo que em 44 instâncias, G=A(provado) e nas 13 restante, G=R.

Tabela 6.7 Atributos que descrevem o conjunto de instâncias CAP1_2012_E4.

DN (Ano Nasc.)	VE (Nota Vest.)	AV_4 (Nota Avaliação 4)	*G (Avaliação Final)*
EC (Est. Civ.)	AV_1 (Nota Avaliação 1)	F (Nro. faltas)	
S (Sexo)	AV_2 (Nota Avaliação 2)	LD (Log. Mat.Disc.)	
GI (Grau Inst.)	AV_3 (Nota Avaliação 3)	FC (Fis. Comp.)	

Tabela 6.8 Resultados do processo de seleção de subconjuntos de atributos usando o conjunto de dados CAP1_2012_E4.

CAP1_2012_E4		
Método de avaliação	Método de busca	Atributos selecionados
CfsSubsetEval	*BestFirst*	EC, AV1, AV2, AV3, AV4, LD
	ExhaustiveSearch	EC, AV1, AV2, AV3, AV4, LD
	GeneticSearch	EC, AV1, AV2, AV3, AV4, LD, FC
	GreedyStepwise	EC, AV1, AV2, AV3, AV4, LD
	RandomSearch	EC, S, AV1, AV2, AV3, AV4, LD, FC
	RankSearch	EC, AV1, AV2, AV3, AV4, LD, FC
ConsistencySubsetEval	*BestFirst*	EC, S, AV1, AV2, AV3, AV4, LD, FC
	ExhaustiveSearch	EC, AV1, AV3, AV4, LD, FC
	GeneticSearch	EC, AV1, AV3, AV4, LD, FC
	GreedyStepwise	EC, AV1, AV2, AV4
	RandomSearch	DN, EC, S, AV1, AV3, AV4, LD, FC
	RankSearch	EC, AV1, AV2, AV3, AV4, LD, FC

Com base nos resultados mostrados na Tabela 6.8, os seguintes comentários com relação aos dados relativos a 2012 podem ser feitos:

(1) Os atributos AV1, AV2, AV3 e AV4 comparecem em 8 dos 12 subconjuntos de atributos, o que corresponde a 66,66% dos resultados. Como pode ser evidenciado na Tabela 3.5, a fórmula de cálculo para o valor de AF em 2012, para a disciplina em questão, é dada por: $((AV_1 \times 30) + (AV_2 \times 30) + (AV_3 \times 30) + (AV_4 \times 10))/100$, em que as três primeiras avaliações têm peso 0,30 e a última, tem peso 0,10. Já os atributos AV1, AV3 e AV4, comparecem em 3 subconjuntos dos 12, correspondendo a 25% dos resultados e, por fim, os atributos AV1, AV2 e AV4 comparecem em 1 dos 12 subconjuntos, representando 8,33% do total. É interessante ressaltar que nos dois últimos conjuntos de atributos citados, os métodos de busca optaram por selecionar o atributo AV4 que possui

menor peso na nota final, no lugar dos atributos AV2 e AV3 respectivamente. Esta escolha talvez possa se justificar pelas altas notas conseguidas pelos estudantes nesta avaliação, a média dos valores do atributo AV4 corresponde a 8,535, enquanto que o atributo AV3, possuí média 6,982, o atributo AV1 possuí média 5,991 e o atributo AV2, possuí a menor média com 5,337.

(2) O atributo EC comparece em 12 dos 12 conjuntos de seleção de atributos, participando em 100% dos resultados. Como já comentado isso pode ocorrer devida à ambos, distribuição não balanceada das classes (44 em G=A e 13 em G=R) e distribuição não balanceada dos valores do atributo EC. Do total de 57 instâncias, (a) 52 delas têm EC=S, sendo que, em 40, G = A e em 12, G = R; (b) 4 delas têm EC=C e todas com G=A; (c) 1 tem EC = D, com G = R.

(3) O atributo LD comparece em 11 dos 12 subconjuntos de atributos selecionados, o que corresponde a 91,66% dos resultados e o atributo FC comparece em 8 dos 12 subconjuntos selecionados, representando 66,66% dos resultados. Ambos resultados são um indicativo que desempenhos em ambas as disciplinas estão relacionados com o desempenho em CAP1. As figuras 6.6 e 6.7 exibem uma representação do conjunto CAP1_2012_E4 cujas instâncias, representadas apenas por dois atributos, são os pontos bidimensionais (LD,AF) e (FC,AF), respectivamente.

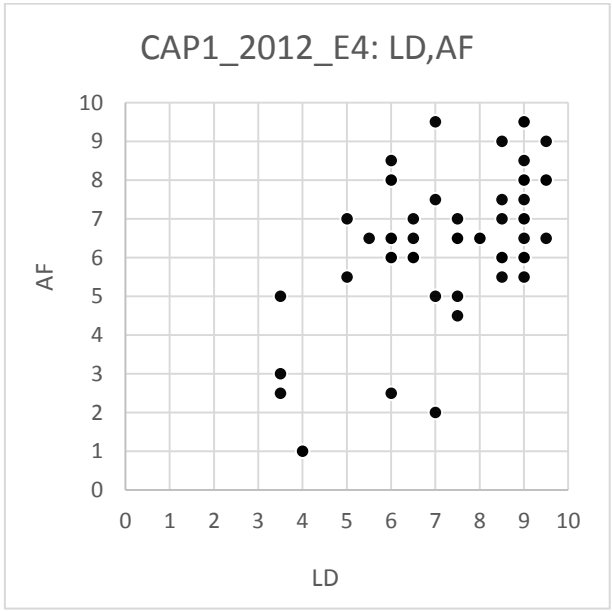

Figura 6.6 Representação gráfica do conjunto de dados CAP1_2012_E4, em que cada instância de dados distinta é descrita por um ponto bidimensional definido pelos atributos (LD,AF). São 57 instâncias no conjunto, com 17 valores duplicados, resultando em 40 pontos com valores exclusivos representados no gráfico.

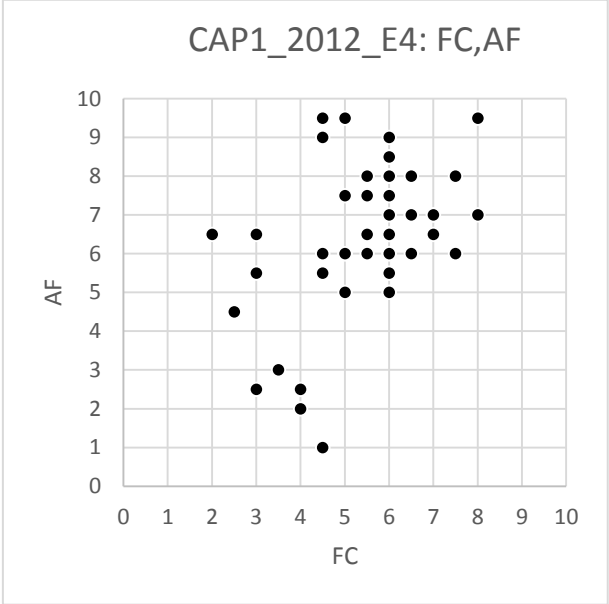

Figura 6.7 Representação gráfica do conjunto de dados CAP1_2012_E4, em que cada instância de dados distinta é descrita por um ponto bidimensional definido pelos atributos (FC,AF). São 57 instâncias no conjunto, com 18 valores duplicados, resultando em 39 pontos com valores exclusivos representados no gráfico.

(4) O atributo S comparece em apenas 3 (25%) dos 12 subconjuntos de atributos selecionados. Do total de 57 instâncias no conjuntos, (a) 55 delas têm S=M, sendo que, em 43 o valor de G = A e em 12, o valor G = R; (b) 2 delas têm S=F, sendo 1 instância com G = A e a outra, com G = R.

6.2.5 Seleção em CAP1_2013_E5

Os experimentos realizados com os vários processos de seleção de subconjuntos de atributos e suas combinações com métodos de busca, utilizando o conjunto de dados CAP1_2013_E5 (Tabela 6.9) obtiveram os resultados mostrados na Tabela 6.10. Lembrando, tal conjunto tem 60 instâncias de dados descritas por 12 atributos e 1 classe (G) associada, sendo que em 32 instâncias, G=A(provado) e nas 28 restantes G=R, o que evidencia um melhor balanceamento entre classes, inexistente nos conjunto de dados abordados anteriormente.

Com base nos resultados mostrados na Tabela 6.10, os seguintes comentários podem ser feitos:

(1) A combinação do método de seleção de atributos *CfsSubsetEval* com cada um dos seis métodos de busca (*BestFirst, ExhaustiveSearch, GeneticSearch, GreedyStepwise, RandomSearch* e *RankSearch*) gerou o mesmo conjunto de atributos i.e., {AV1, AV2,

AV3, PI}. Esse resultado era esperado, considerando que a fórmula do cálculo de AF para a disciplina CAP1, em 2013, envolve a média das três avaliações, AV1, AV2 e AV3, ponderadas por 0,9 somada à nota da prova interdisciplinar (PI), ponderada por 0,1.

Tabela 6.9 Atributos que descrevem o conjunto de instâncias CAP1_2013_E5.

DN (Ano Nasc.)	VE (Nota Vest.)	PI (Nota Prov. Inter)	*G (Avaliação Final)*
EC (Est. Civ.)	AV_1 (Nota Avaliação 1)	F (Nro. faltas)	
S (Sexo)	AV_2 (Nota Avaliação 2)	LD (Log. Mat.Disc.)	
GI (Grau Inst.)	AV_3 (Nota Avaliação 3)	FC (Fis. Comp.)	

Tabela 6.10 Resultados do processo de seleção de subconjuntos de atributos usando o conjunto de dados CAP1_2013_E5.

CAP1_2013_E5		
Método de avaliação	Método de busca	Atributos selecionados
CfsSubsetEval	*BestFirst*	AV1, AV2, AV3, PI
	ExhaustiveSearch	AV1, AV2, AV3, PI
	GeneticSearch	AV1, AV2, AV3, PI
	GreedyStepwise	AV1, AV2, AV3, PI
	RandomSearch	AV1, AV2, AV3, PI
	RankSearch	AV1, AV2, AV3, PI
ConsistencySubsetEval	*BestFirst*	AV1, AV2, AV3
	ExhaustiveSearch	AV1, AV2, AV3
	GeneticSearch	AV1, AV2, AV3, PI
	GreedyStepwise	AV1, AV2, AV3
	RandomSearch	DN, GI, AV1, AV2, AV3, PI, FC
	RankSearch	AV1, AV2, AV3

(2) Entretanto, a combinação *ConsistencySubsetEval* com 4 dos métodos de busca, a saber, *BestFirst, ExhaustiveSearch, GreedyStepwise,* e *RankSearch,* geraram o mesmo subconjunto de atributos, contendo apenas as avaliações *i.e.*, {AV1, AV2, AV3}. A combinação *ConsistencySubsetEval + GeneticSearch* evidenciou o conjunto dos atributos que participam do cálculo de AF i.e., {AV1, AV2, AV3, PI}. Já a combinação *ConsistencySubsetEval + RandomSearch,* além dos 4 atributos, acrescentou ao subconjunto de atributos relevantes o DN, GI e FC. Os atributos AV1, AV2 e AV3, portanto, comparecem nos 12 subconjuntos gerados pelas 12 combinações.

(3) O atributo PI comparece em 8 dos 12 subconjuntos de atributos selecionados, o que corresponde a 66,66%. Como comentado, este atributo participa do cálculo de AF mas com um peso comparativamente mais baixo que os atributos AV1, AV2 e AV3.

(4) O atributo que representa a nota na disciplina *Lógica e Matemática Discreta* (LD) não foi selecionado por qualquer das 12 combinações e a nota na disciplina *Física Computacional* foi selecionada apenas pela combinação *ConsistencySubsetEval + RandomSearch.* As figuras 6.8 e 6.9 exibem uma representação do conjunto

CAP1_2013_E5 cujas instâncias, representadas apenas por dois atributos, são os pontos bidimensionais (LD,AF) e (FC,AF), respectivamente.

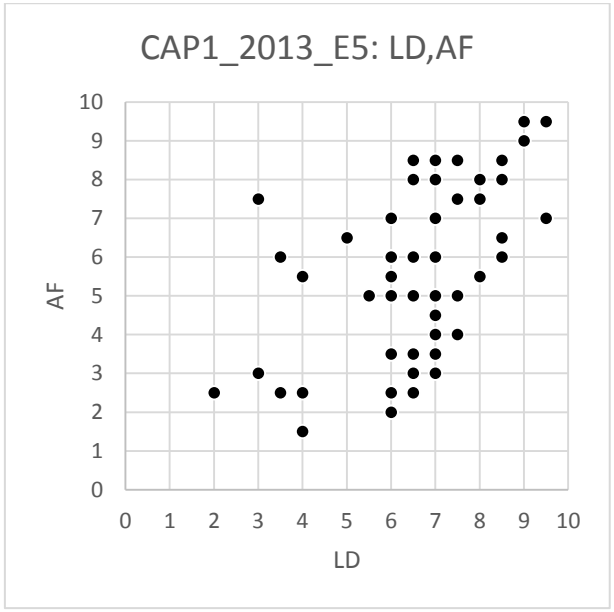

Figura 6.8 Representação gráfica do conjunto de dados CAP1_2013_E5, em que cada instância de dados distinta é descrita por um ponto bidimensional definido pelos atributos (LD,AF). São 60 instâncias no conjunto, com 12 valores duplicados, resultando em 48 pontos com valores exclusivos representados no gráfico.

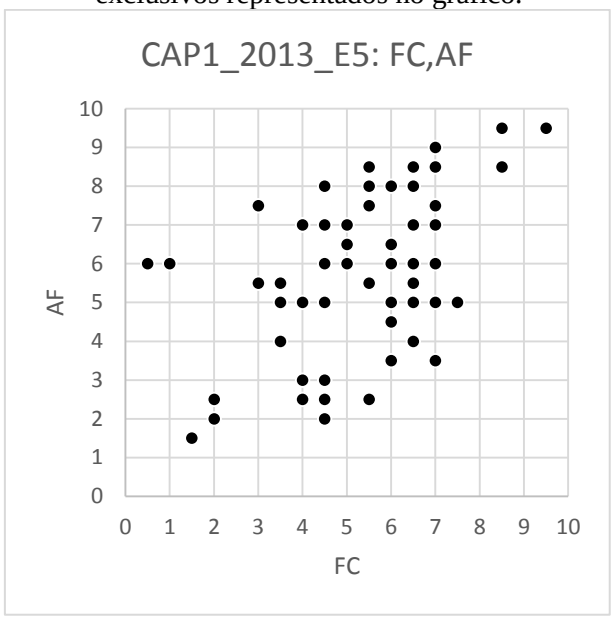

Figura 6.9 Representação gráfica do conjunto de dados CAP1_2013_E5, em que cada instância de dados distinta é descrita por um ponto bidimensional definido pelos atributos (FC,AF). São 60 instâncias no conjunto, com 7 valores duplicados, resultando em 53 pontos com valores exclusivos representados no gráfico.

6.2.6 Seleção em CAP1_2014_E5

Os experimentos realizados com os vários processos de seleção de subconjuntos de atributos e suas combinações com métodos de busca, utilizando o conjunto de dados CAP1_2014_E5 (Tabela 6.11) obtiveram os resultados mostrados na Tabela 6.12. Lembrando, tal conjunto tem 37 instâncias de dados descritas por 12 atributos e 1 classe (G) associada, sendo que em 23 instâncias, G=A(provado) e nas 14 restantes G=R.

Tabela 6.11 Atributos que descrevem o conjunto de instâncias CAP1_2014_E5.

DN (Ano Nasc.)	VE (Nota Vest.)	PI (Nota Prov. Inter)	*G (Avaliação Final)*
EC (Est. Civ.)	AV_1 (Nota Avaliação 1)	F (Nro. faltas)	
S (Sexo)	AV_2 (Nota Avaliação 2)	LD (Log. Mat.Disc.)	
GI (Grau Inst.)	AV_3 (Nota Avaliação 3)	FC (Fis. Comp.)	

Tabela 6.12 Resultados do processo de seleção de subconjuntos de atributos usando o conjunto de dados CAP1_2014_E5.

CAP1_2014_E5		
Método de avaliação	Método de busca	Atributos selecionados
CfsSubsetEval	*BestFirst*	S, PI, AV1, AV2, FC
	ExhaustiveSearch	S, PI, AV1, AV2, FC
	GeneticSearch	S, PI, AV1, AV2, FC
	GreedyStepwise	S, PI, AV1, AV2, FC
	RandomSearch	S, PI, AV1, AV2, FC
	RankSearch	S, PI, AV1, AV2, FC
ConsistencySubsetEval	*BestFirst*	AV2
	ExhaustiveSearch	PI, AV1, AV2
	GeneticSearch	S, PI, AV1, FC
	GreedyStepwise	AV2
	RandomSearch	EC, S, GI, PI, LD, FC
	RankSearch	PI, AV1, AV2, FC

Com base nos resultados mostrados na Tabela 6.12, os seguintes comentários podem ser feitos:

(1) A fórmula do cálculo de AF para a disciplina CAP1, em 2014, envolve a média ponderada por 0,9, das ponderações de três avaliações, AV1, AV2 e AV3, em que AV1 e AV2 são ponderados por 2 e AV3 por 1, somadas ao valor de PI, ponderado por 0,1.

(2) Como ocorrido anteriormente (nas subseções 6.2.1, 6.2.2 e 6.2.5) as seis combinações de *CfsSubsetEval* com cada um dos seis métodos de busca (*BestFirst, ExhaustiveSearch, GeneticSearch, GreedyStepwise, RandomSearch* e *RankSearch*) selecionaram o mesmo subconjunto de atributos i.e., {S, PI, AV1, AV2, FC}.

(3) Os atributos AV1 e AV2 comparecem em 8 dos 12 conjuntos de atributos selecionados, o que corresponde a 66,66%. Apesar do atributo AV3 participar do cálculo

de AF (muito embora com um peso menor que AV1 e AV2), não foi selecionado por qualquer uma das 12 combinações utilizadas.

(4) O atributo PI comparece em 10 dos 12 conjuntos de atributos selecionados, o que corresponde a 83,33%. Sua seleção é substanciada pela participação do valor de PI no cálculo de AF, apesar da baixa ponderação (1).

(5) Particularmente nesse conjunto de dados, os resultados levam a crer que os desempenhos nas disciplinas Física Computacional e Lógica e Matemática Discreta estão diretamente relacionadas, uma vez que o atributo FC comparece em 9 dos 12 subconjuntos de atributos selecionados, correspondendo a 75% do total. A Figura 6.10 mostra a plotagem das instâncias do conjunto CAP1_2014_E5 representadas por pontos, cujas coordenadas são dadas por (FC,AF) no gráfico.

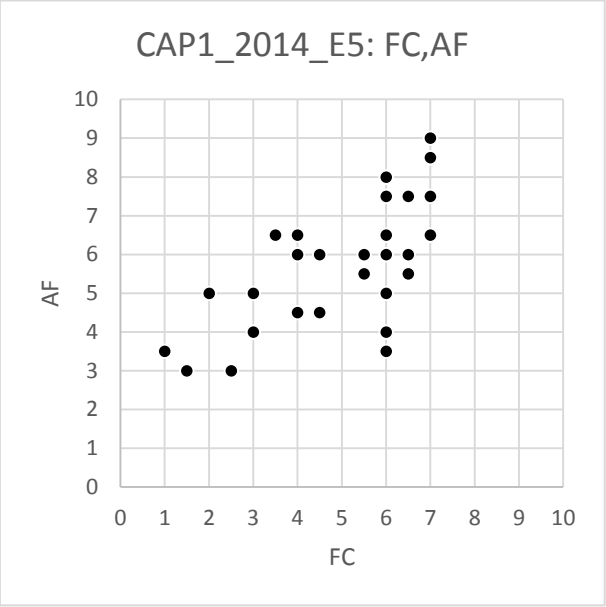

Figura 6.10 Representação gráfica do conjunto de dados CAP1_2014_E5, em que cada instância distinta de dados é descrita por um ponto bidimensional definido pelos atributos (FC,AF). São 37 instâncias no conjunto, com 9 valores duplicados, resultando em 28 pontos com valores exclusivos representados no gráfico.

(6) O atributo S comparece em 8 dos 12 conjuntos de seleção de atributos representando 66,66%. Do total de 37 instâncias, (a) 29 delas tem S=M, sendo que, em 21 o valor de G = A e em 8, o valor de G = R; (b) 8 delas têm S=F, sendo que em 2, o valor de G=A e nas 6 restantes, G = R.

(7) A combinação *ConsistencySubsetEval* + *RandomSearch*, selecionou o conjunto de atributos {EC, S, GI, PI, LD, FC}. Do total de 37 instâncias, 35 têm EC=S, sendo que 23

têm o valor G = A e as 12 restantes, G = R; (b) 1 instância tem EC=C com o valor G = A e (c) 1 instância tem o valor EC=D com G = D.

6.2.7 Seleção em CAP1_2015_E6

Os experimentos realizados com os vários processos de seleção de subconjuntos de atributos e suas combinações com métodos de busca, utilizando o conjunto de dados CAP1_2015_E6 (Tabela 6.13) obtiveram os resultados mostrados na Tabela 6.14. Lembrando, tal conjunto tem 69 instâncias de dados descritas por 11 atributos e 1 classe (G) associada, sendo que em 41 instâncias, G=A(provado) e nas 28 restantes G=R.

Tabela 6.13 Atributos que descrevem o conjunto de instâncias CAP1_2015_E6.

DN (Ano Nasc.)	VE (Nota Vest.)	F (Nro. faltas)
EC (Est. Civ.)	AV_1 (Nota Avaliação 1)	LD (Log. Mat.Disc.)
S (Sexo)	AV_2 (Nota Avaliação 2)	LP (Lab. Program.)
GI (Grau Inst.)	PI (Nota Prov. Inter)	*G (Avaliação Final)*

Tabela 6.14 Resultados do processo de seleção de subconjuntos de atributos usando o conjunto de dados CAP1_2015_E6.

CAP1_2015_E6		
Método de avaliação	Método de busca	Atributos selecionados
CfsSubsetEval	*BestFirst*	DN, AV1, AV2, PI, LP
	ExhaustiveSearch	DN, AV1, AV2, PI, LP
	GeneticSearch	DN, AV1, AV2, PI, LP
	GreedyStepwise	DN, AV1, AV2, PI, LP
	RandomSearch	DN, AV1, AV2, PI, LD, LP
	RankSearch	DN, AV1, AV2, PI, LP
ConsistencySubsetEval	*BestFirst*	AV1, AV2, PI
	ExhaustiveSearch	AV1, AV2, PI
	GeneticSearch	AV1, AV2, PI, F
	GreedyStepwise	AV1, AV2, PI
	RandomSearch	DN, S, AV1, AV2, PI, LP
	RankSearch	AV1, AV2, PI, LP

Com base nos resultados mostrados na Tabela 6.14, os seguintes comentários podem ser feitos:

(1) O cálculo de AF associada à disciplina CAP1, em 2015, é média aritmética ponderada por 0,9 das duas avaliações, AV1 e AV2, somada ao valor de PI, ponderado por 0,1.
(2) O atributo LP (que representa a nota na disciplina *Laboratório de Programação*) participa em 8 dos 12 subconjuntos obtidos pelas 12 combinações, enquanto que LD aparece em apenas um dos 12 subconjuntos obtidos. Duas plotagens do conjunto de dados, em que cada instância distinta é representada por um ponto descrito pelos atributos (LD,AF) e (LC,AF) estão mostradas nas Figura 6.11 e Figura 6.12, respectivamente.

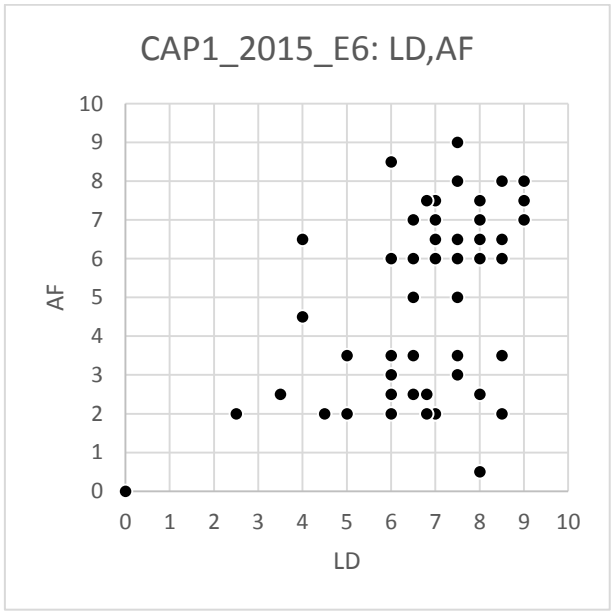

Figura 6.11 Representação gráfica do conjunto de dados CAP1_2015_E5, em que cada instância de dados distinta é descrita por um ponto bidimensional definido pelos atributos (LD,AF). São 69 instâncias no conjunto, com 21 valores duplicados, resultando em 48 pontos com valores exclusivos representados no gráfico.

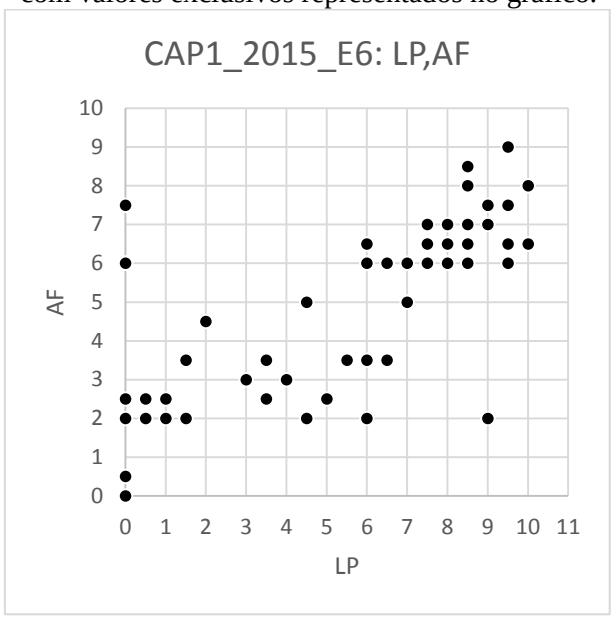

Figura 6.12 Representação gráfica do conjunto de dados CAP1_2015_E5, em que cada instância de dados distinta é descrita por um ponto bidimensional definido pelos atributos (LP,AF). São 69 instâncias no conjunto, com 20 valores duplicados, resultando em 49 pontos com valores exclusivos representados no gráfico.

(3) A combinação do método de avaliação de atributos *CfsSubsetEval* com cada um dos cinco dos métodos de busca i.e., *BestFirst, ExhaustiveSearch, GeneticSearch, GreedyStepwise* e *RankSearch*, obteve como resultado o mesmo conjunto de atributos, {DN, AV1, AV2, PI, LP}.

6.3 Experimentos Realizados com Ranqueamento de Atributos

A abordagem de seleção individual de atributos é mais rápida (comparada às da Seção 6.2). Tal abordagem avalia atributos individualmente e, então, classifica-os por ordem de relevância, descartando aqueles que estão abaixo de um determinado limite pré-estabelecido pelo usuário. Esse processo pode ser realizado por meio da seleção de um dos 8 possíveis métodos de avaliação de um único atributo, descrito na parte inferior da Tabela 5.8 (extraída de [Witten & Frank 2005]), combinado com o método de ordenação *Ranker*, descrito na parte inferior da Tabela 5.9. Para os experimentos foram escolhidos os métodos (5), (6) e (7) da Tabela 5.8, a saber (5) *ChiSquaredAttributeEval* (*Chi*), (6) *GainRatioAttributeEval* (*GainR*) e o (7) *InfoGainAttributeEval* (*InfoGain*). As possíveis combinações de cada um deles, com o 'ranqueador' *Ranker* (item (8) da Tabela 5.9), estão descritas na sequência.

Nas próximas sete subseções, cada um dos sete conjuntos de instâncias que já passaram por um processo de imputação, de maneira a não terem mais atributos com valores ausentes, será abordado individualmente e submetido a três diferentes processos de avaliação de atributos, especificamente, o *Chi*, *GainR* e o (*InfoGain*. Como comentado no parágrafo anterior, cada um dos processos é usado em conjunto com o *Ranker* que, lembrando, não é um método de busca de subconjuntos de atributos, mas sim, um esquema de 'ranqueamento' para atributos individuais. Os resultados obtidos nos experimentos foram truncados a partir da quarta casa decimal.

6.3.1 Ranqueamento de Atributos em CAP1_2009_E1

As 50 instâncias de dados do conjunto CAP1_2009_E1 estão descritas pelos atributos mostrados na Tabela 6.15 (15 atributos + classe (*G*) associada).

Tabela 6.15 Atributos que descrevem o conjunto de instâncias CAP1_2009_E1.

(1) DN (Ano Nasc.)	(5) VE (Nota Vest.)	(9) LE_1 (Nota Exer. 1)	(13) F (Nro. faltas)
(2) EC (Est. Civ.)	(6) AV_1 (Nota Avaliação 1)	(10) LE_2 (Nota Exer. 2)	(14) LD (Log. Mat.Disc.)
(3) S (Sexo)	(7) AV_2 (Nota Avaliação 2)	(11) LE_3 (Nota Exer. 3)	(15) FC (Fis. Comp.)
(4) GI (Grau Inst.)	(8) AV_3 (Nota Avaliação 3)	(12) LE_4 (Nota Exer. 4)	(16) *G (Avaliação Final)*

Quando o CAP1_2009_E1 é carregado no ambiente Weka, estatísticas associadas a cada um dos atributos podem ser mostradas (mediante solicitação do usuário). Para o CAP1_2009_E1 elas são:

```
Selected attributes

(1) Name: DN
Statistic     Value
Minimum       1972
Maximum       1992
Mean          1989.16
StdDev        3.113

(2) Name: EC
Nro     Label     Count
1         S        46
2         C        4
3         D        0

(3) Name: S
Nro     Label     Count
1       M         48
2       F         2

(4) Name: GI
Nro     Label     Count
1         MC        49
2         SI        1
3         SC        0
4         ME        0
5         DO        0

(5) Name: VE
Statistic     Value
Minimum       1
Maximum       7.6
Mean          4.13
StdDev        1.503

(6) Name: AV1
Statistic     Value
Minimum       0
Maximum       10
Mean          7.712
StdDev        3.722

(7) Name: AV2
Statistic     Value
Minimum       0
Maximum       10
Mean          7.076
StdDev        3.246

(8) Name: AV3
Statistic     Value
Minimum       0
Maximum       10
Mean          6.224
StdDev        3.749

(9) Name: LE1
Statistic     Value
Minimum       0
Maximum       10
Mean          9.4
StdDev        2.399

(10) Name: LE2
Statistic     Value
Minimum       0
Maximum       10
Mean          8.9
StdDev        3.079

(11) Name: LE3
Statistic     Value
Minimum       0
Maximum       10
Mean          7.2
StdDev        4.536

(12) Name: LE4
Statistic     Value
Minimum       0
Maximum       10
Mean          8.988
StdDev        3.028

(13) Name: F
Statistic     Value
Minimum       0
Maximum       0
Mean          0
StdDev        0

(14) Name: LD
Statistic     Value
Minimum       2
Maximum       9
Mean          6.732
StdDev        0.982

(15) Name: FC
Statistic     Value
Minimum       0.5
Maximum       7.5
Mean          5.756
StdDev        1.488

(16) Name: G
Nro     Label     Count
1         A         40
2         R         10
```

O Weka permite, também, uma visualização gráfica da distribuição dos valores associados a cada atributo, como mostra a Figura 6.13. Os resultados do processo de ranqueamento dos atributos estão mostrados na Tabela 6.16 e resumidos na Tabela 6.17.

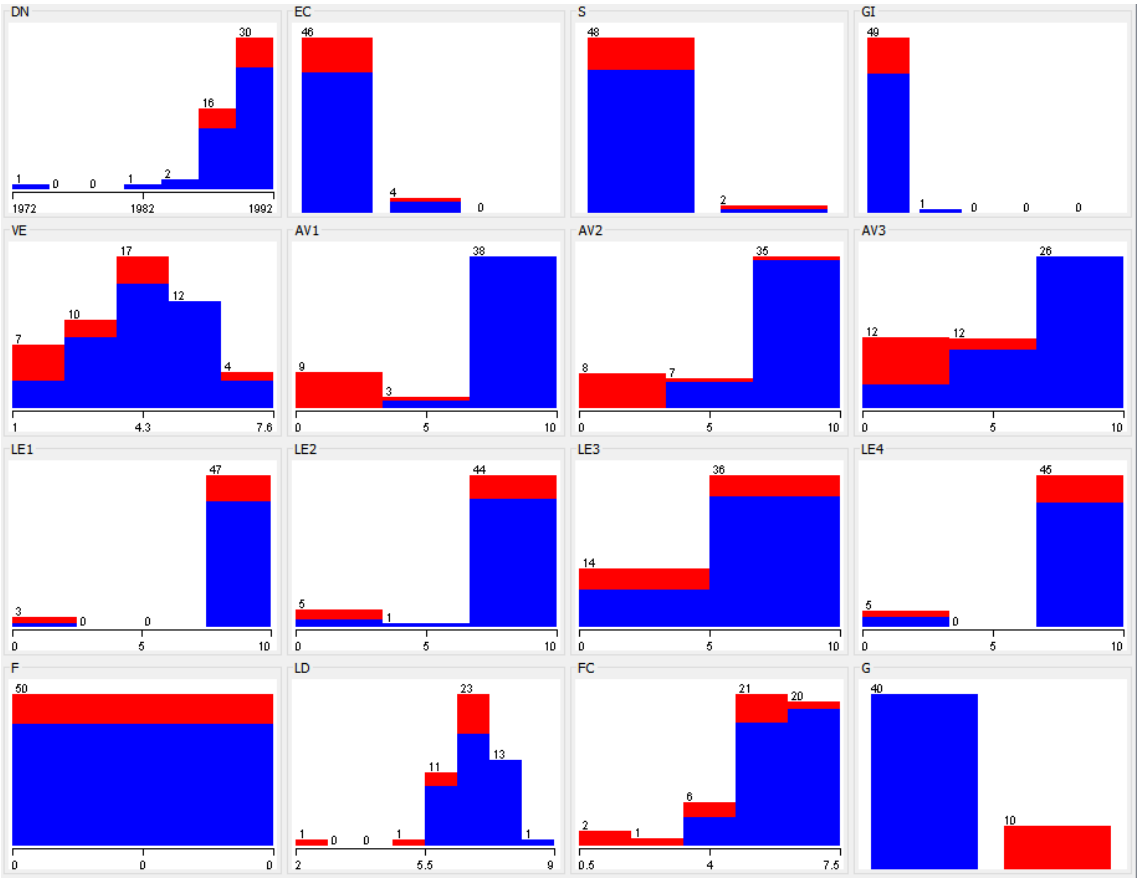

Figura 6.13 Gráficos associados à distribuição dos valores dos 15 atributos + classe G, que descrevem o conjunto de dados CAP1_2009_E1, obtidos via Weka.

Tabela 6.16 Ranqueamento dos 15 atributos que descrevem as instâncias em CAP1_2009_E1.

Chi+Ranker			*GainR+Ranker*			*InfoGain+Ranker*		
43.9024	(6)	AV1	0.8620	(6)	AV1	0.5862	(6)	AV1
38.2813	(7)	AV2	0.6831	(7)	AV2	0.4932	(7)	AV2
23.8343	(8)	AV3	0.3820	(8)	AV3	0.3268	(8)	AV3
12.5	(15)	FC	0.2113	(15)	FC	0.1525	(15)	FC
1.1719	(3)	S	0.0559	(3)	S	0.0135	(3)	S
0.2551	(4)	GI	0.0460	(4)	GI	0.0065	(4)	GI
0.0679	(2)	EC	0.0023	(2)	EC	0.0009	(2)	EC
0	(5)	VE	0	(5)	VE	0	(5)	VE
0	(11)	LE3	0	(11)	LE3	0	(11)	LE3
0	(13)	F	0	(13)	F	0	(13)	F
0	(10)	LE2	0	(10)	LE2	0	(10)	LE2
0	(9)	LE1	0	(9)	LE1	0	(9)	LE1
0	(14)	LD	0	(14)	LD	0	(14)	LD
0	(12)	LE4	0	(12)	LE4	0	(12)	LE4
0	(1)	DN	0	(1)	DN	0	(1)	DN

Tabela 6.17 Ranqueamento dos 15 atributos usando três métodos & *Ranker* (Weka).

Chi+Ranker	6, 7, 8, 15, 3, 4, 2, 5, 11, 13, 10, 9, 14, 12, 1	AV1, AV2, AV3, FC, S, GI, EC, VE, LE3, F, LE2, LE1, LD, LE4, DN
GainR+Ranker	6, 7, 8, 15, 3, 4, 2, 5, 11, 13, 10, 9, 14, 12, 1	AV1, AV2, AV3, FC, S, GI, EC, VE, LE3, F, LE2, LE1, LD, LE4, DN
InfoGain+Ranker	6, 7, 8, 15, 3, 4, 2, 5, 11, 13, 10, 9, 14, 12, 1	AV1, AV2, AV3, FC, S, GI, EC, VE, LE3, F, LE2, LE1, LD, LE4, DN

6.3.2 Ranqueamento de Atributos em CAP1_2010_E2

As 41 instâncias de dados do conjunto CAP1_2010_E2 estão descritas pelos atributos mostrados na Tabela 6.18 (11 atributos + classe (*G*) associada).

Tabela 6.18 Atributos que descrevem o conjunto de instâncias CAP1_2010_E2.

(1) DN (Ano Nasc.)	(5) VE (Nota Vest.)	(9) F (Nro. faltas)
(2) EC (Est. Civ.)	(6) AV_1 (Nota Avaliação 1)	(10) LD (Log. Mat.Disc.)
(3) S (Sexo)	(7) AV_2 (Nota Avaliação 2)	(11) FC (Fis. Comp.)
(4) GI (Grau Inst.)	(8) AV_3 (Nota Avaliação 3)	(12) *G (Avaliação Final)*

As estatísticas associadas a cada um dos atributos, para o CAP1_2010_E2, são:

Selected attributes

(1) Name: DN

Statistic	Value
Minimum	1974
Maximum	1992
Mean	1989.561
StdDev	3.755

(2) Name: EC

Nro	Label	Count
1	S	39
2	C	1
3	D	1

(3) Name: S

Nro	Label	Count
1	M	38
2	F	3

(4) Name: GI

Nro	Label	Count
1	MC	40
2	SI	1
3	SC	0
4	ME	0
5	DO	0

(5) Name: VE

Statistic	Value
Minimum	1
Maximum	7.5
Mean	2.69
StdDev	2.08

(6) Name: AV1

Statistic	Value
Minimum	0
Maximum	10
Mean	8.895
StdDev	1.82

(7) Name: AV2

Statistic	Value
Minimum	0
Maximum	10
Mean	7.559
StdDev	2.042

(8) Name: AV3

Statistic	Value
Minimum	0
Maximum	9.5
Mean	6.085
StdDev	2.337

(9) Name: F

Statistic	Value
Minimum	0
Maximum	8
Mean	1.561
StdDev	3.21

(10) Name: LD

Statistic	Value
Minimum	2
Maximum	8.5
Mean	6.644
StdDev	1.293

(11) Name: FC

Statistic	Value
Minimum	1.5
Maximum	8.5
Mean	6.146
StdDev	1.338

(12) Name: G

Nro	Label	Count
1	A	37
2	R	4

A Figura 6.14 mostra a visualização gráfica da distribuição dos valores associados a cada atributo.

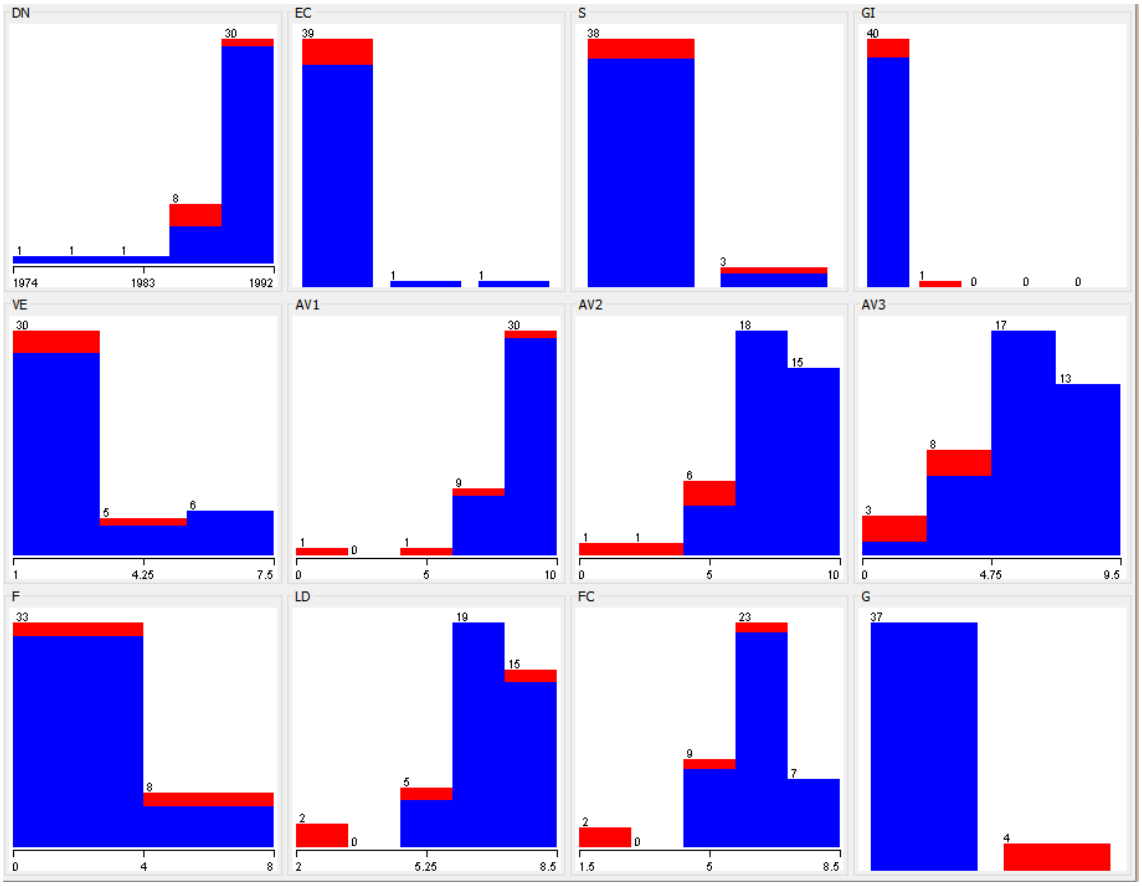

Figura 6.14 Gráficos associados à distribuição dos valores dos 11 atributos + classe G, que descrevem o conjunto de dados CAP1_2010_E2, obtidos via Weka.

Os resultados do processo de ranqueamento dos atributos estão mostrados na Tabela 6.19 e resumidos na Tabela 6.20.

Tabela 6.19 Ranqueamento dos 11 atributos que descrevem as instâncias em CAP1_2010_E2.

Chi+Ranker			*GainR+Ranker*			*InfoGain+Ranker*		
31.9135	(7)	AV2	0.6976	(7)	AV2	0.3731	(7)	AV2
19.4487	(11)	FC	0.653	(11)	FC	0.2244	(8)	AV3
19.4487	(10)	LD	0.653	(10)	LD	0.1836	(11)	FC
19.4487	(6)	AV1	0.653	(6)	AV1	0.1836	(10)	LD
13.7405	(8)	AV3	0.5215	(4)	GI	0.1836	(6)	AV1
9.4813	(4)	GI	0.28	(8)	AV3	0.0862	(4)	GI
2.0437	(3)	S	0.0655	(3)	S	0.0247	(3)	S
0.2273	(2)	EC	0.0225	(2)	EC	0.0074	(2)	EC
0	(5)	VE	0	(5)	VE	0	(9)	F
0	(9)	F	0	(9)	F	0	(5)	VE
0	(1)	DN	0	(1)	DN	0	(1)	DN

Tabela 6.20 Ranqueamento dos 11 atributos usando três métodos & *Ranker* (Weka).

Chi+Ranker	7, 11, 10, 6, 8, 4, 3, 2, 5, 9, 1	AV2, FC, LD, AV1, AV3, GI, S, EC, VE, F, DN
GainR+Ranker	7, 11, 10, 6, 4, 8, 3, 2, 5, 9, 1	AV2, FC, LD, AV1, GI, AV3, S, EC, VE, F, DN
InfoGain+Ranker	7, 8, 11, 10, 6, 4, 3, 2, 9, 5, 1	AV2, AV3, FC, LD, AV1, GI, S, EC, F, VE, DN

6.3.3 Ranqueamento de Atributos em CAP1_2011_E3

As 43 instâncias de dados do conjunto CAP1_2011_E3 estão descritas pelos atributos mostrados na Tabela 6.21 (14 atributos + classe (*G*) associada).

Tabela 6.21 Atributos que descrevem o conjunto de instâncias CAP1_2011_E3.

(1) DN (Ano Nasc.)	(5) VE (Nota Vest.)	(9) PI (Nota Prov. Inter)	(13) LD (Log. Mat.Disc.)
(2) EC (Est. Civ.)	(6) AV_1 (Nota Avaliação 1)	(10) AS_1 (Nota Aval. Sub1)	(14) FC (Fis. Comp.)
(3) S (Sexo)	(7) AV_2 (Nota Avaliação 2)	(11) AS_2 (Nota Aval. Sub2)	(15) *G (Avaliação Final)*
(4) GI (Grau Inst.)	(8) AV_3 (Nota Avaliação3)	(12) F (Nro. faltas)	

As estatísticas associadas a cada um dos atributos, para o CAP1_2011_E3, são:

Selected attributes

(1) Name: DN

Statistic	Value
Minimum	1981
Maximum	1993
Mean	1990.535
StdDev	3.05

(2) Name: EC

Nro	Label	Count
1	S	37
2	C	5
3	D	1

(3) Name: S

Nro	Label	Count
1	M	42
2	F	1

(4) Name: GI

Nro	Label	Count
1	MC	40
2	SI	3
3	SC	0
4	ME	0
5	DO	0

(5) Name: VE

Statistic	Value
Minimum	3
Maximum	7.5
Mean	5.228
StdDev	1.206

(6) Name: AV1

Statistic	Value
Minimum	2.5
Maximum	10
Mean	6.93
StdDev	2.089

(7) Name: AV2

Statistic	Value
Minimum	0
Maximum	10
Mean	6.895
StdDev	2.444

(8) Name: AV3

Statistic	Value
Minimum	0
Maximum	10
Mean	5.535
StdDev	3.401

(9) Name: PI

Statistic	Value
Minimum	0
Maximum	0.86
Mean	0.29
StdDev	0.284

(10) Name: AS1

Statistic	Value
Minimum	0
Maximum	10
Mean	1.43
StdDev	2.665

(11) Name: AS2

Statistic	Value
Minimum	0
Maximum	8
Mean	1.012
StdDev	1.863

(12) Name: F

Statistic	Value
Minimum	0
Maximum	36
Mean	6.14
StdDev	6.61

(13) Name: LD

Statistic	Value
Minimum	3
Maximum	10
Mean	7.607
StdDev	1.27

(14) Name: FC

Statistic	Value
Minimum	1.5
Maximum	8
Mean	6.488
StdDev	1.316

(15) Name: G

Nro	Label	Count
1	A	31
2	R	12

A Figura 6.15 mostra a visualização gráfica da distribuição dos valores associados a cada atributo. Os resultados do processo de ranqueamento dos atributos estão mostrados na Tabela 6.22 e resumidos na Tabela 6.23.

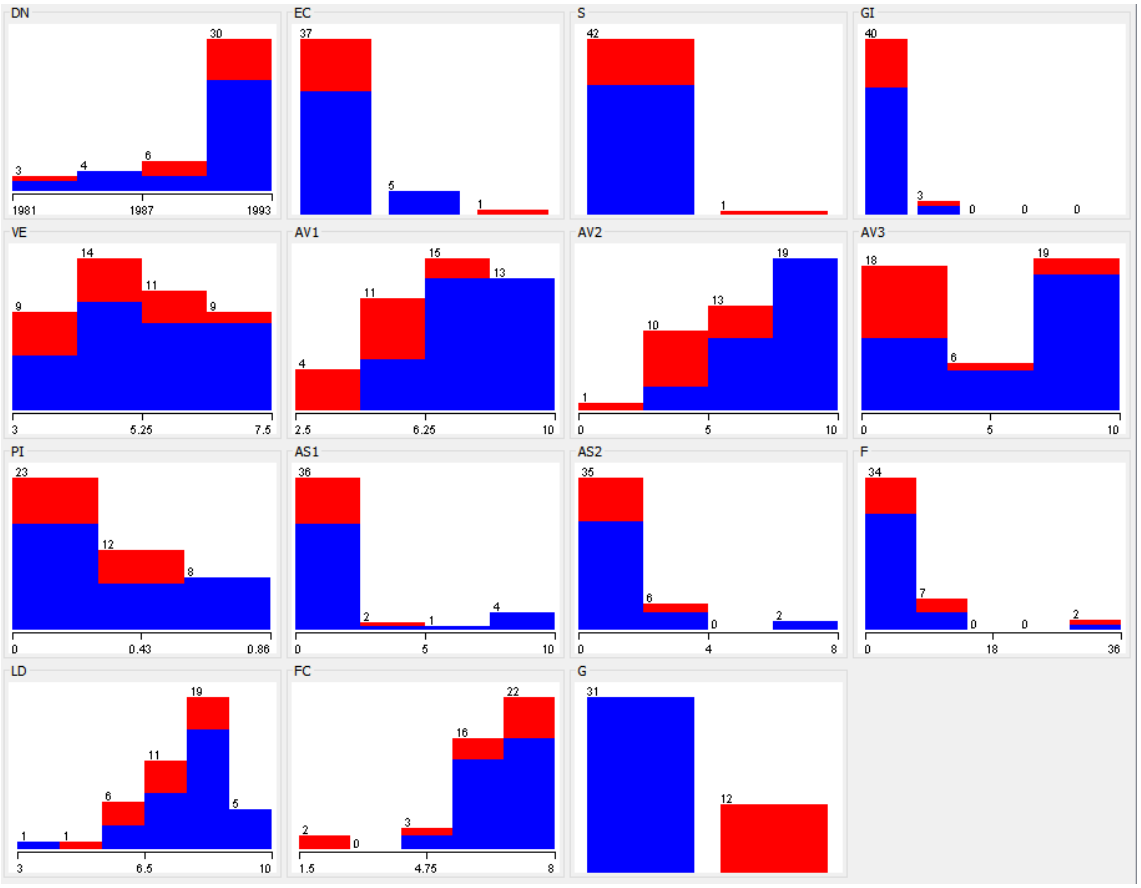

Figura 6.15 Gráficos associados à distribuição dos valores dos 14 atributos + classe G, que descrevem o conjunto de dados CAP1_2011_E3, obtidos via Weka.

Tabela 6.22 Ranqueamento dos 14 atributos que descrevem as instâncias em CAP1_2011_E3.

Chi+Ranker			*GainR+Ranker*			*InfoGain+Ranker*		
28.0887	(7)	AV2	0.5799	(7)	AV2	0.5523	(7)	AV2
25.1065	(6)	AV1	0.3627	(6)	AV1	0.5155	(6)	AV1
22.4173	(10)	AS1	0.2792	(10)	AS1	0.3875	(10)	AS1
5.7218	(14)	FC	0.2752	(3)	S	0.1439	(14)	FC
4.5799	(2)	EC	0.1754	(14)	FC	0.0987	(2)	EC
2.6448	(3)	S	0.1465	(2)	EC	0.0438	(3)	S
0.0472	(4)	GI	0.0021	(4)	GI	0.0007	(4)	GI
0	(9)	PI	0	(9)	PI	0	(9)	PI
0	(11)	AS2	0	(12)	F	0	(11)	AS2
0	(5)	VE	0	(5)	VE	0	(5)	VE
0	(12)	F	0	(8)	AV3	0	(12)	F
0	(13)	LD	0	(11)	AS2	0	(13)	LD
0	(8)	AV3	0	(13)	LD	0	(8)	AV3
0	(1)	DN	0	(1)	DN	0	(1)	DN

Tabela 6.23 Ranqueamento dos 14 atributos usando três métodos & *Ranker* (Weka).

Chi+Ranker	7, 6, 10, 14, 2, 3, 4, 9, 11, 5, 12, 13, 8, 1	AV2, AV1, AS1, FC, EC, S, GI, PI, AS2, VE, F, LD, AV3, DN
GainR+Ranker	7, 6, 10, 3, 14, 2, 4, 9, 12, 5, 8, 11, 13, 1	AV2, AV1, AS1, S, FC, EC, GI, PI, F, VE, AV3, AS2, LD, DN
InfoGain+Ranker	7, 6, 10, 14, 2, 3, 4, 9, 11, 5, 12, 13, 8, 1	AV2, AV1, AS1, FC, EC, S, GI, PI, AS2, VE, F, LD, AV3, DN

6.3.4 Ranqueamento de Atributos em CAP1_2012_E4

As 57 instâncias de dados do conjunto CAP1_2012_E4 estão descritas pelos atributos mostrados na Tabela 6.24 (12 atributos + classe (*G*) associada).

Tabela 6.24 Atributos que descrevem o conjunto de instâncias CAP1_2012_E4.

(1) DN (Ano Nasc.)	(5) VE (Nota Vest.)	(9) AV_4 (Nota Avaliação 4)	(13) *G (Avaliação Final)*
(2) EC (Est. Civ.)	(6) AV_1 (Nota Avaliação 1)	(10) F (Nro. faltas)	
(3) S (Sexo)	(7) AV_2 (Nota Avaliação 2)	(11) LD (Log. Mat.Disc.)	
(4) GI (Grau Inst.)	(8) AV_3 (Nota Avaliação 3)	(12) FC (Fis. Comp.)	

As estatísticas associadas a cada um dos atributos, para o CAP1_2014_E4, são:

Selected attributes

(1) Name: DN

Statistic	Value
Minimum	1970
Maximum	1995
Mean	1991.246
StdDev	4.763

(2) Name: EC

Nro	Label	Count
1	S	52
2	C	4
3	D	1

(3) Name: S

Nro	Label	Count
1	M	55
2	F	2

(4) Name: GI

Nro	Label	Count
1	MC	57
2	SI	0
3	SC	0
4	ME	0
5	DO	0

(5) Name: VE

Statistic	Value
Minimum	1
Maximum	8
Mean	4.442
StdDev	1.31

(6) Name: AV1

Statistic	Value
Minimum	0
Maximum	10
Mean	5.991
StdDev	2.028

(7) Name: AV2

Statistic	Value
Minimum	0
Maximum	10
Mean	5.337
StdDev	2.671

(8) Name: AV3

Statistic	Value
Minimum	0
Maximum	10
Mean	6.982
StdDev	2.952

(9) Name: AV4

Statistic	Value
Minimum	0
Maximum	10
Mean	8.535
StdDev	2.841

(10) Name: F

Statistic	Value
Minimum	0
Maximum	28
Mean	2.667
StdDev	6.817

(11) Name: LD

Statistic	Value
Minimum	3.5
Maximum	9.5
Mean	7.368
StdDev	1.749

(12) Name: FC

Statistic	Value
Minimum	2
Maximum	8
Mean	5.482
StdDev	1.292

(13) Name: *G*

Nro	Label	Count
1	A	44
2	R	13

A Figura 6.16 mostra a visualização gráfica da distribuição dos valores associados a cada atributo.

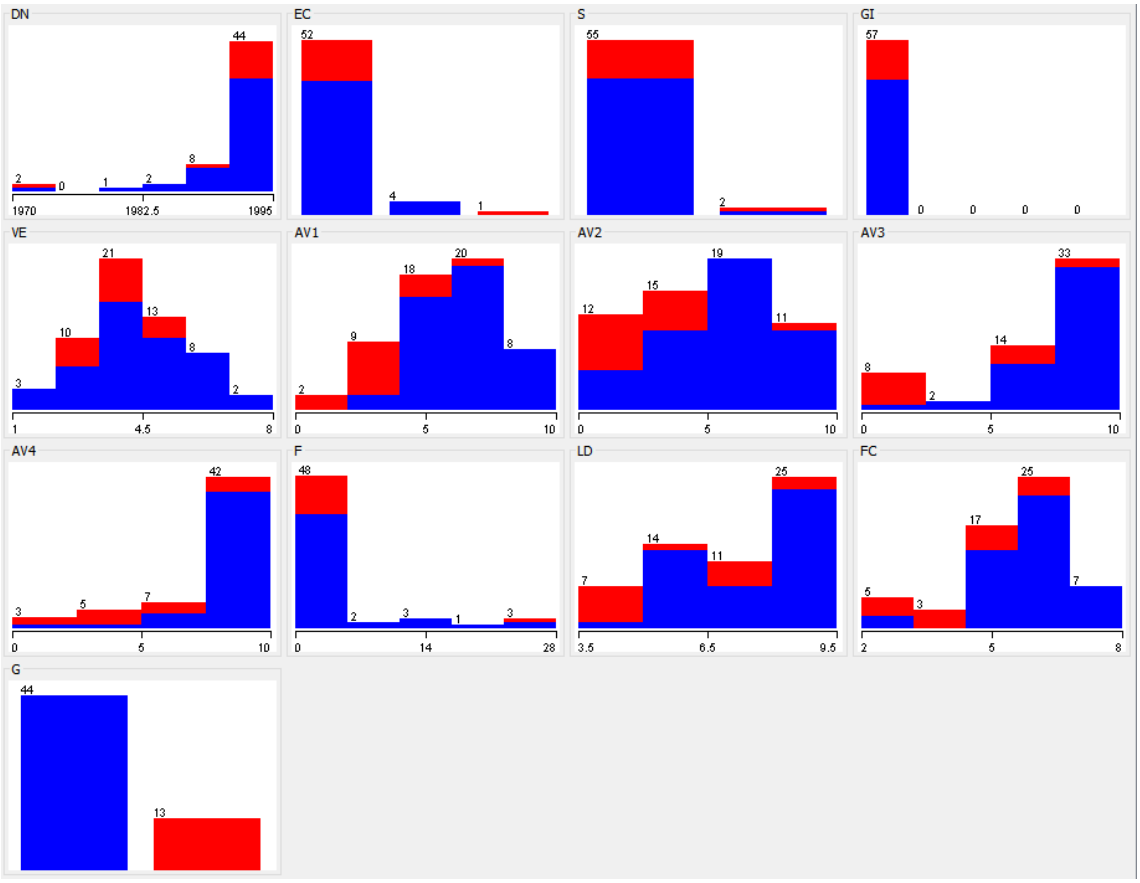

Figura 6.16 Gráficos associados à distribuição dos valores dos 12 atributos + classe G, que descrevem o conjunto de dados CAP1_2012_E4, obtidos via Weka.

Os resultados do processo de ranqueamento dos atributos estão mostrados na Tabela 6.25 e resumidos na Tabela 6.26.

Tabela 6.25 Ranqueamento dos 12 atributos que descrevem as instâncias em CAP1_2012_E4.

Chi+Ranker			*GainR+Ranker*			*InfoGain+Ranker*		
26.669	(6)	AV1	0.4888	(11)	LD	0.3208	(6)	AV1
22.123	(8)	AV3	0.4055	(8)	AV3	0.2705	(7)	AV2
21.924	(7)	AV2	0.3746	(6)	AV1	0.2372	(8)	AV3
18.55	(11)	LD	0.3007	(7)	AV2	0.2095	(11)	LD
15.995	(12)	FC	0.2223	(12)	FC	0.1848	(12)	FC
15.995	(9)	AV4	0.2223	(9)	AV4	0.1848	(9)	AV4
4.569	(2)	EC	0.1293	(2)	EC	0.0636	(2)	EC
0.871	(3)	S	0.0423	(3)	S	0.0092	(3)	S
0	(4)	GI	0	(4)	GI	0	(4)	GI
0	(10)	F	0	(10)	F	0	(10)	F
0	(5)	VE	0	(5)	VE	0	(5)	VE
0	(1)	DN	0	(1)	DN	0	(1)	DN

Tabela 6.26 Ranqueamento dos 12 atributos usando três métodos & *Ranker* (Weka).

Chi+Ranker	6, 8, 7, 11, 12, 9, 2, 3, 4, 10, 5, 1	AV1, AV3, AV2, LD, FC, AV4, EC, S, GI, F, VE, DN
GainR+Ranker	11, 8, 6, 7, 12, 9, 2, 3, 4, 10, 5, 1	LD, AV3, AV1, AV2, FC, AV4, EC, S, GI, F, VE, DN
InfoGain+Ranker	6, 7, 8, 11, 12, 9, 2, 3, 4, 10, 5, 1	AV1, AV2, AV3, LD, FC, AV4, EC, S, GI, F, VE, DN

6.3.5 Ranqueamento de Atributos em CAP1_2013_E5

As 60 instâncias de dados do conjunto CAP1_2013_E5 estão descritas pelos atributos mostrados na Tabela 6.27 (12 atributos + classe (*G*) associada).

Tabela 6.27 Atributos que descrevem o conjunto de instâncias CAP1_2013_E5.

(1) DN (Ano Nasc.)	(5) VE (Nota Vest.)	(9) PI (Nota Prov. Inter)	(13) *G (Avaliação Final)*
(2) EC (Est. Civ.)	(6) AV_1 (Nota Avaliação 1)	(10) F (Nro. faltas)	
(3) S (Sexo)	(7) AV_2 (Nota Avaliação 2)	(11) LD (Log. Mat.Disc.)	
(4) GI (Grau Inst.)	(8) AV_3 (Nota Avaliação 3)	(12) FC (Fis. Comp.)	

As estatísticas associadas a cada um dos atributos, para o CAP1_2013_E5, são:

```
Selected attributes

(1) Name: DN
Statistic     Value
Minimum       1978
Maximum       1995
Mean          1992.65
StdDev        3.236

(2) Name: EC
Nro     Label     Count
1         S         59
2         C          1
3         D          0

(3) Name: S
Nro     Label     Count
1         M         57
2         F          3

(4) Name: GI
Nro     Label     Count
1         MC        60
2         SI         0
3         SC         0
4         ME         0
5         DO         0

(5) Name: VE
Statistic     Value
Minimum       1.5
Maximum       9.5
Mean          5.212
StdDev        1.346

(6) Name: AV1
Statistic     Value
Minimum       3
Maximum       10
Mean          6.558
StdDev        1.69

(7) Name: AV2
Statistic     Value
Minimum       0.5
Maximum       10
Mean          5.45
StdDev        2.948

(8) Name: AV3
Statistic     Value
Minimum       0.5
Maximum       10
Mean          5.308
StdDev        3.18

(9) Name: PI
Statistic     Value
Minimum       0
Maximum       0.8
Mean          0.465
StdDev        0.16

(10) Name: F
Statistic     Value
Minimum       0
Maximum       0
Mean          0
StdDev        0

(11) Name: LD
Statistic     Value
Minimum       2
Maximum       9.5
Mean          6.508
StdDev        1.598

(12) Name: FC
Statistic     Value
Minimum       0.5
Maximum       9.5
Mean          5.25
StdDev        1.8

(13) Name: G
Nro     Label     Count
1         A         32
2         R         28
```

A Figura 6.17 mostra a visualização gráfica da distribuição dos valores associados a cada atributo.

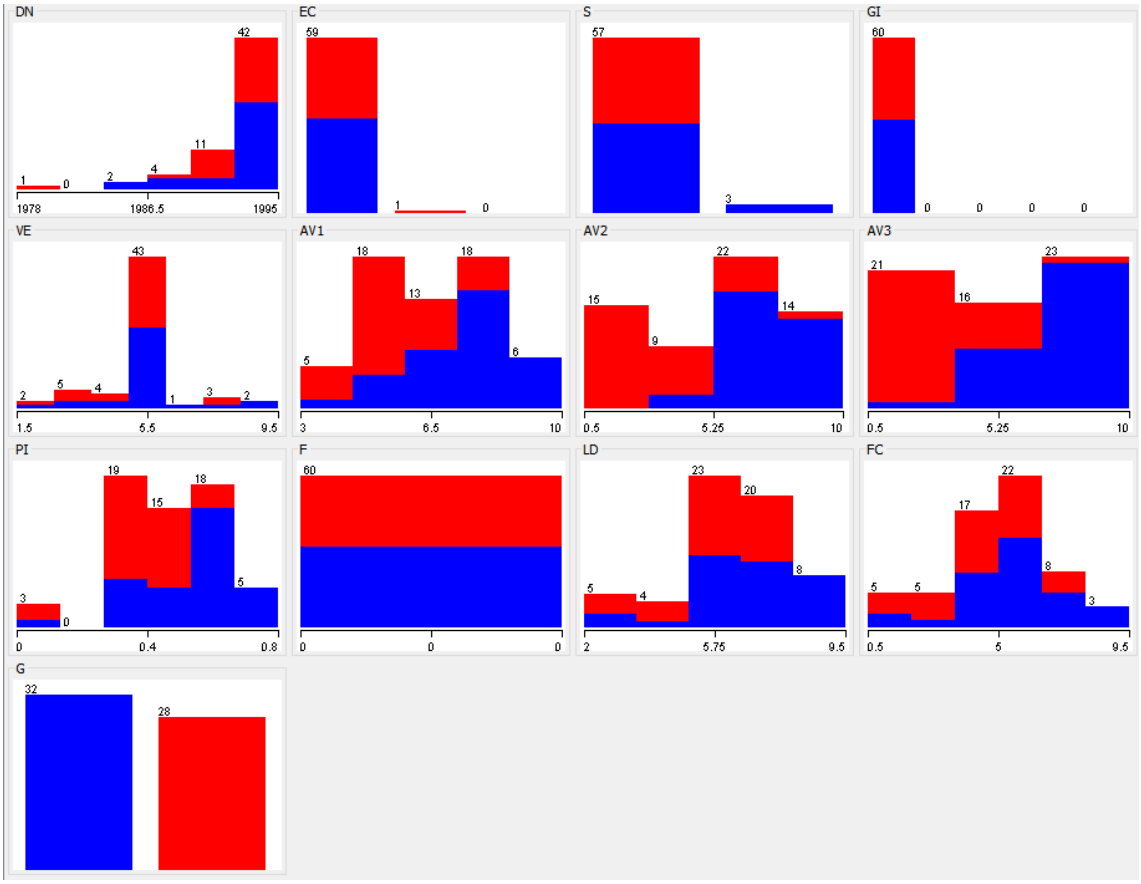

Figura 6.17 Gráficos associados à distribuição dos valores dos 12 atributos + classe G, que descrevem o conjunto de dados CAP1_2012_E4, obtidos via Weka.

Os resultados do processo de ranqueamento dos atributos estão mostrados na Tabela 6.28 e resumidos na Tabela 6.29.

Tabela 6.28 Ranqueamento dos 12 atributos que descrevem as instâncias em CAP1_2013_E5.

Chi+Ranker			*GainR+Ranker*			*InfoGain+Ranker*		
33.2207	(7)	AV2	0.488	(7)	AV2	0.4625	(7)	AV2
32.4107	(8)	AV3	0.437	(8)	AV3	0.4373	(8)	AV3
21.4716	(6)	AV1	0.277	(6)	AV1	0.2757	(6)	AV1
16.9414	(9)	PI	0.236	(11)	LD	0.2221	(9)	PI
8.0769	(11)	LD	0.231	(9)	PI	0.1338	(11)	LD
2.7632	(3)	S	0.164	(3)	S	0.047	(3)	S
1.1622	(2)	EC	0.152	(2)	EC	0.0186	(2)	EC
0	(4)	GI	0	(4)	GI	0	(4)	GI
0	(12)	FC	0	(12)	FC	0	(12)	FC
0	(5)	VE	0	(5)	VE	0	(5)	VE
0	(10)	F	0	(10)	F	0	(10)	F
0	(1)	DN	0	(1)	DN	0	(1)	DN

Tabela 6.29 Ranqueamento dos 12 atributos usando três métodos & *Ranker* (Weka).

Chi+Ranker	7, 8, 6, 9, 11, 3, 2, 4, 12, 5, 10, 1	AV2, AV3, AV1, PI, LD, S, EC, GI, FC, VE, F, DN
GainR+Ranker	7, 8, 6, 11, 9, 3, 2, 4, 12, 5, 10, 1	AV2, AV3, AV1, LD, PI, S, EC, GI, FC, VE, F, DN
InfoGain+Ranker	7, 8, 6, 9, 11, 3, 2, 4, 12, 5, 10, 1	AV2, AV3, AV1, PI, LD, S, EC, GI, FC, VE, F, DN

6.3.6 Ranqueamento de Atributos em CAP1_2014_E5

As 37 instâncias de dados do conjunto CAP1_2014_E5 estão descritas pelos atributos mostrados na Tabela 6.30 (12 atributos + classe (*G*) associada).

Tabela 6.30 Atributos que descrevem o conjunto de instâncias CAP1_2014_E5.

(1) DN (Ano Nasc.)	(5) VE (Nota Vest.)	(9) PI (Nota Prov. Inter)	(13) *G (Avaliação Final)*
(2) EC (Est. Civ.)	(6) AV_1 (Nota Avaliação 1)	(10) F (Nro. faltas)	
(3) S (Sexo)	(7) AV_2 (Nota Avaliação 2)	(11) LD (Log. Mat.Disc.)	
(4) GI (Grau Inst.)	(8) AV_3 (Nota Avaliação 3)	(12) FC (Fis. Comp.)	

As estatísticas associadas a cada um dos atributos, para o CAP1_2014_E5, são:

```
Selected attributes

(1) Name: DN
Statistic     Value
Minimum       1975
Maximum       1997
Mean          1993.405
StdDev        4.272

(2) Name: EC
Nro     Label     Count
1         S         35
2         C          1
3         D          1

(3) Name: S
Nro     Label     Count
1         M         29
2         F          8

(4) Name: GI
Nro     Label     Count
1         MC        37
2         SI         0
3         SC         0
4         ME         0
5         DO         0

(5) Name: VE
Statistic     Value
Minimum       3
Maximum       8.5
Mean          5.224
StdDev        1.265

(6) Name: AV1
Statistic     Value
Minimum       2.5
Maximum       10
Mean          6.716
StdDev        1.891

(7) Name: AV2
Statistic     Value
Minimum       0
Maximum       10
Mean          4.716
StdDev        2.493

(8) Name: AV3
Statistic     Value
Minimum       3.3
Maximum       10
Mean          9.141
StdDev        1.604

(9) Name: PI
Statistic     Value
Minimum       0
Maximum       0.62
Mean          0.274
StdDev        0.141

(10) Name: F
Statistic     Value
Minimum       0
Maximum       20
Mean          1.108
StdDev        3.688

(11) Name: LD
Statistic     Value
Minimum       1.5
Maximum       8
Mean          5.986
StdDev        1.521

(12) Name: FC
Statistic     Value
Minimum       1
Maximum       7
Mean          5.203
StdDev        1.734

(13) Name: G
Nro     Label     Count
1         A         23
2         R         14
```

A Figura 6.18 mostra a visualização gráfica da distribuição dos valores associados a cada atributo.

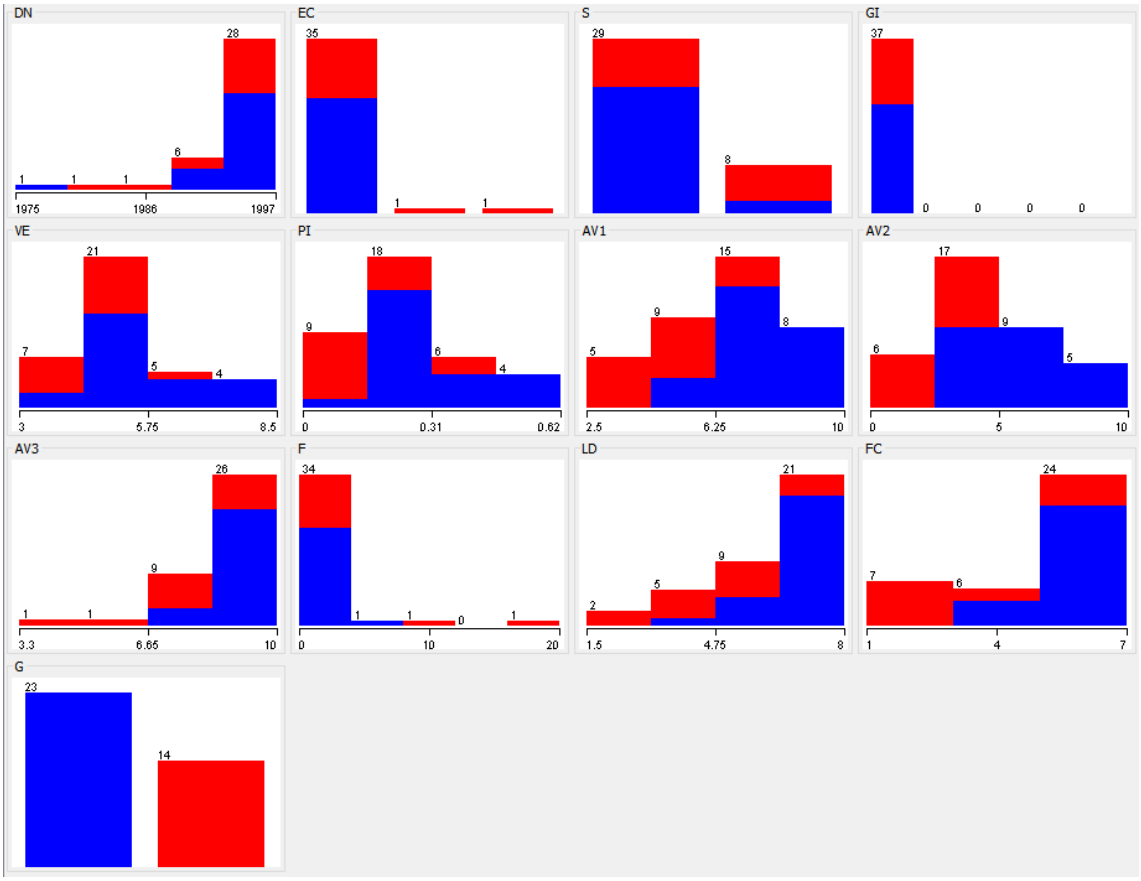

Figura 6.18 Gráficos associados à distribuição dos valores dos 12 atributos + classe G, que descrevem o conjunto de dados CAP1_2014_E4, obtidos via Weka.

Os resultados do processo de ranqueamento dos atributos estão mostrados na Tabela 6.31. e resumidos na Tabela 6.32.

Tabela 6.31 Ranqueamento dos 12 atributos que descrevem as instâncias em CAP1_2014_E5.

Chi+Ranker			*GainR+Ranker*			*InfoGain+Ranker*		
25.572	(8)	AV2	0.584	(8)	AV2	0.5686	(8)	AV2
18.745	(7)	AV1	0.459	(12)	FC	0.391	(7)	AV1
14.183	(12)	FC	0.445	(7)	AV1	0.3214	(12)	FC
13.178	(6)	PI	0.334	(6)	PI	0.2672	(6)	PI
11.453	(11)	LD	0.237	(11)	LD	0.2336	(11)	LD
5.993	(3)	S	0.222	(2)	EC	0.1155	(3)	S
3.473	(2)	EC	0.153	(3)	S	0.0795	(2)	EC
0	(9)	AV3	0	(9)	AV3	0	(9)	AV3
0	(4)	GI	0	(4)	GI	0	(4)	GI
0	(10)	F	0	(10)	F	0	(10)	F
0	(5)	VE	0	(5)	VE	0	(5)	VE
0	(1)	DN	0	(1)	DN	0	(1)	DN

Tabela 6.32 Ranqueamento dos 12 atributos usando três métodos & *Ranker* (Weka).

Chi+Ranker	8, 7, 12, 6, 11, 3, 2, 9, 4, 10, 5, 1	AV2, AV1, FC, PI, LD, S, EC, AV3, GI, F, VE, DN
GainR+Ranker	8, 12, 7, 6, 11, 2, 3, 9, 4, 10, 5, 1	AV2, FC, AV1, PI, LD, EC, S, AV3, GI, F, VE, DN
InfoGain+Ranker	8, 7, 12, 6, 11, 3, 2, 9, 4, 10, 5, 1	AV2, AV1, FC, PI, LD, S, EC, AV3, GI, F, VE, DN

6.3.7 Ranqueamento de Atributos em CAP1_2015_E6

As 69 instâncias de dados do conjunto CAP1_2015_E6 estão descritas pelos atributos mostrados na Tabela 6.33 (11 atributos + classe (*G*) associada).

Tabela 6.33 Atributos que descrevem o conjunto de instâncias CAP1_2015_E6.

(1) DN (Ano Nasc.)	(5) VE (Nota Vest.)	(9) F (Nro. faltas)
(2) EC (Est. Civ.)	(6) AV_1 (Nota Avaliação 1)	(10) LD (Log. Mat.Disc.)
(3) S (Sexo)	(7) AV_2 (Nota Avaliação 2)	(11) LP (Lab. Program.)
(4) GI (Grau Inst.)	(8) PI (Nota Prov. Inter)	(12) *G (Avaliação Final)*

As estatísticas associadas a cada um dos atributos, para o CAP1_2015_E6, são:

Selected attributes

(1) Name: DN

Statistic	Value
Minimum	1979
Maximum	1998
Mean	1995
StdDev	2.951

(2) Name: EC

Nro	Label	Count
1	S	68
2	C	1
3	D	0

(3) Name: S

Nro	Label	Count
1	M	62
2	F	7

(4) Name: GI

Nro	Label	Count
1	MC	68
2	SI	0
3	SC	1
4	ME	0
5	DO	0

(5) Name: VE

Statistic	Value
Minimum	1.6
Maximum	9.4
Mean	5.339
StdDev	1.433

(6) Name: AV1

Statistic	Value
Minimum	0
Maximum	9
Mean	5.493
StdDev	2.05

(7) Name: AV2

Statistic	Value
Minimum	0
Maximum	10
Mean	4.93
StdDev	3.002

(8) Name: PI

Statistic	Value
Minimum	0
Maximum	0.7
Mean	0.349
StdDev	0.158

(9) Name: F

Statistic	Value
Minimum	0
Maximum	16
Mean	2.275
StdDev	4.253

(10) Name: LD

Statistic	Value
Minimum	0
Maximum	9
Mean	6.825
StdDev	1.575

(11) Name: LP

Statistic	Value
Minimum	0
Maximum	10
Mean	5.688
StdDev	3.399

(12) Name: *G*

Nro	Label	Count
1	A	41
2	R	28

A Figura 6.19 mostra a visualização gráfica da distribuição dos valores associados a cada atributo.

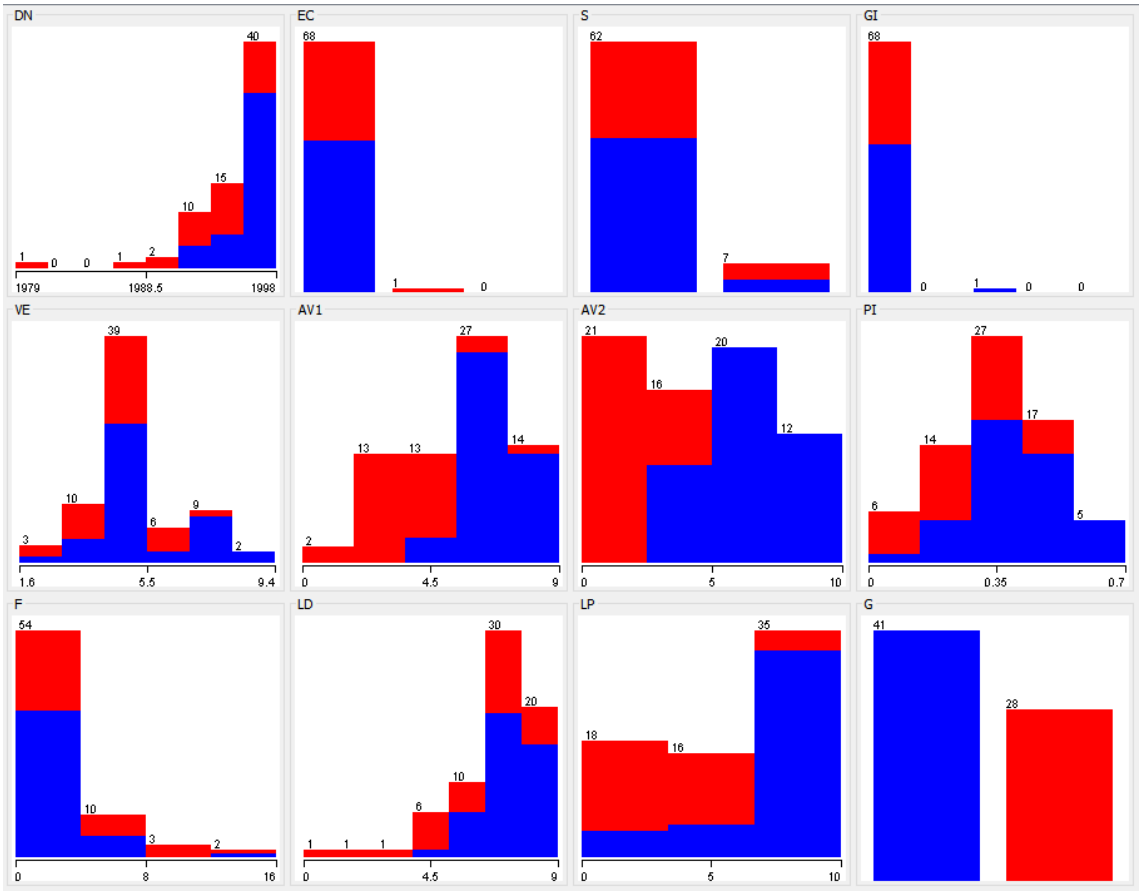

Figura 6.19 Gráficos associados à distribuição dos valores dos 11 atributos + classe G, que descrevem o conjunto de dados CAP1_2015_E6, obtidos via Weka.

Os resultados do processo de ranqueamento dos atributos estão mostrados na Tabela 6.34 e resumidos na Tabela 6.35.

Tabela 6.34 Ranqueamento dos 11 atributos que descrevem as instâncias em CAP1_2015_E6.

Chi+Ranker			*GainR+Ranker*			*InfoGain+Ranker*		
55.6697	(7)	AV2	0.6886	(6)	AV1	0.7834	(7)	AV2
53.4387	(6)	AV1	0.5191	(7)	AV2	0.6581	(6)	AV1
34.7189	(11)	LP	0.4314	(11)	LP	0.4297	(11)	LP
25.8261	(8)	PI	0.291	(8)	PI	0.2856	(8)	PI
15.9749	(1)	DN	0.1971	(10)	LD	0.1846	(1)	DN
10.0174	(10)	LD	0.1912	(1)	DN	0.1100	(10)	LD
1.4858	(2)	EC	0.1746	(2)	EC	0.0190	(2)	EC
0.8863	(3)	S	0.1005	(4)	GI	0.0109	(4)	GI
0.693	(4)	GI	0.0192	(3)	S	0.0090	(3)	S
0	(5)	VE	0	(9)	F	0	(5)	VE
0	(9)	F	0	(5)	VE	0	(9)	F

Tabela 6.35 Ranqueamento dos 11 atributos usando três métodos & *Ranker* (Weka).

Chi+Ranker	7, 6, 11, 8, 1, 10, 2, 3, 4, 5, 9	AV2, AV1, LP, PI, DN, LD, EC, S, GI, VE, F
GainR+Ranker	6, 7, 11, 8, 10, 1, 2, 4, 3, 9, 5	AV1, AV2, LP, PI, LD, DN, EC, GI, S, F, VE
InfoGain+Ranker	7, 6, 11, 8, 1, 10, 2, 4, 3, 5, 9	AV2, AV1, LP, PI, DN, LD, EC, GI, S, VE, F

6.4 Resumo dos Resultados Relacionados à Identificação de Atributos Relevantes

A Tabela 6.36 apresenta, para cada um dos conjuntos de dados considerados, o subconjunto de atributos relevantes escolhido (com base nos resultados obtidos dos 12 experimentos e descritos na Seção 6.2). Tal subconjunto é composto pelos atributos selecionados em 7 ou mais dos 12 experimentos realizados para cada conjunto de dados.

Tabela 6.36 Subconjuntos de atributos escolhidos, dentre aqueles identificados na Seção 6.2. O valor entre parênteses associado a cada atributo indica o número de experimentos (dentre os 12) que selecionou o atributo em questão.

Conjunto de Dados	Atributos
CAP1_2009_E1	AV1(12), S(9), AV2(8), AV3(7)
CAP1_2010_E2	AV2(12), AV1(8), AV3(8), FC(8), S(7), LD(7)
CAP1_2011_E3	AV2(12), AS1(12), EC(10), S(7), AV1(7)
CAP1_2012_E4	EC(12), AV1(12), AV4(12), AV3(11) LD(11), AV2(9), FC(8)
CAP1_2013_E5	AV1(12), AV2(12), AV3(12), PI(8)
CAP1_2014_E5	PI(10), AV1(9), AV2(9), FC(9), S(8)
CAP1_2015_E6	AV1(12), AV2(12), PI(12), LP(8), DN(7)

A Tabela 6.37 apresenta, para cada um dos conjuntos de dados considerados, os atributos comuns dentre os sete primeiros, nos três experimentos de ranqueamento realizados, com cada arquivo de dados.

Tabela 6.37 Subconjunto de atributos escolhido, dentre aqueles ranqueados na Seção 6.3.

Conjunto de Dados	Atributos
CAP1_2009_E1	AV1, AV2, AV3, FC, S, GI, EC
CAP1_2010_E2	AV2, FC, LD, AV1, AV3, GI, S
CAP1_2011_E3	AV2, AV1, AS1, FC, EC, S, GI
CAP1_2012_E4	AV1, AV3, AV2, LD, AV4, EC
CAP1_2013_E5	AV2, AV3, AV1, PI, LD, S, EC
CAP1_2014_E5	AV2, AV1, FC, PI, LD, S, EC
CAP1_2015_E6	AV2, AV1, LP, PI, DN, LD

Os conjuntos de atributos mostrados na Tabela 3.36 e Tabela 6.37 serão ainda combinados quando dos experimentos de AM descritos no Capítulo 7 (especificamente, Seção 7.2.3 Terceiro Experimento).

Capítulo 7
Experimentos de MDE utilizando os Algoritmos Naïve Bayes, NN &J48

7.1 Introdução

Este capítulo descreve os experimentos de aprendizado de máquina utilizando os sete conjuntos de dados educacionais coletados e três dos algoritmos de ML disponibilizados junto ao ambiente Weka, a saber, o Naïve Bayes (NB) [Domingos & Pazzani 1997], o J48 (versão Weka do C4.5 [Quinlan 1993]) e o 1-Nearest Neighbor (1Bk) [Cover & Hart 1967]).

7.2 Experimentos de Aprendizado Automático

Lembrando novamente, cada instância de dado é descrita por um conjunto de atributos que, administrativamente, é abordado dividido em dois grupos: (1) Informações Pessoais (IP) do estudante; e (2) Desempenho acadêmico (DE) do estudante. O conjunto IP é composto por: (1) Data de nascimento (ano), (2) Estado civil {S, C, D}, (3) Sexo {M, F} e (4) Grau de Instrução {FI, FC, MC, SI, SC, ES, ME, DO}, como descritos na Tabela 3.3.

O conjunto DE contém informações que fazem parte do registro de dados acadêmicos. Para a abordagem utilizada neste trabalho, o DE associado à disciplina CAP1 foi composto pela nota do vestibular (VE) e diversos outros atributos, bem como a nota associada à avaliação final na disciplina (como descrito na Tabela 3.4). O número de atributos que descrevem DE varia anualmente, de acordo com o Esquema de Avaliação (EA) utilizado. Os esquemas de avaliação foram tratados no Capítulo 3 e estão listados na Tabela 3.2. No que segue estão descritos vários experimentos de aprendizado automático realizado utilizando dados dos arquivos de dados coletados, e os algoritmos: (1) Naïve Bayes – NB; (2) k-NN (IBk) - NN; (3) C4.5 (J48) – AD; associado aos experimentos, foram coletados e calculados os índices descritos na Tabela 7.1.

Tabela 7.1 Informações estatísticas coletadas junto aos resultados dos experimentos sob o Weka usando 4-validação cruzada e, também, calculadas.

ESTATÍSTICAS GERAIS	
SIGLA	DESCRIÇÃO
CC (%CC)	Número de instâncias classificadas corretamente (% Instâncias classificadas corretamente)
CI (%ICI)	Número de instâncias classificadas incorretamente (% Instâncias classificadas incorretamente)
MÉTRICAS DE AVALIAÇÃO DE DESEMPENHO (POR CLASSE)	
SIGLA	DESCRIÇÃO
TP	True Positive
FP	False Positive
TN	True Negative
FN	False Negative
TP Rate (TPR) (*sensitivity* ou *recall*)	TP/(TP+FN)
FP Rate (FPR) (*fall-out*)	FP/(FP+TN)
Precisão (P)	TP/(TP+FP)
Success Rate (SR) (*Accuracy*)	(TP+TN)/(TP+FP+TN+FN)

Como comentado anteriormente no Capítulo 4, quando do uso de classificadores, o Weka, ao final do processo, apresenta várias informações, sendo uma delas um relatório (Summary) que informa, entre outros, o número de instâncias classificadas corretamente e o respectivo percentual (CC e CC%) bem como o número de instâncias classificadas incorretamente assim como o respectivo percentual (CI e CI%). Os valores TP, FP, TN, FN podem ser facilmente extraídos da matriz de confusão, informada ao final, sob o título Confusion Matrix. Segue um exemplo (Witten & Frank 2011). Considere o aprendizado de uma árvore de decisão a partir de um conjunto com 14 instâncias, com duas classes (yes e no) e que parte do relatório do Weka seja:

```
=== Detailed Accuracy By Class ===

TP Rate   FP Rate   Precision   Recall   Class
  0.778     0.6       0.7       0.778    yes
  0.4       0.222     0.5       0.4      no

=== Confusion Matrix ===
a       b   <--- classified as
7       2   |  a = yes
3       2   |  b = no
```

As classes yes(a) e no(b) podem estar associadas aos rótulos *positive* e *negative*.

- O número de instâncias que são *true positive* (TP) (realmente yes (a)) é 7 (linha a e coluna a na matriz). TP = 7.
- O número de instâncias que são *true negative* (TN) (realmente no (b)) é 2 (linha b e coluna b na matriz). TN = 2.
- O número de instâncias *false positive* (FP) (são da classe b mas foram classificadas pela árvore induzida como a) é 3 (linha b e coluna a na matriz). FP = 3.
- O número de instâncias *false negative* (FN) (são da classe a mas foram classificadas pela árvore induzida como b) é 2 (linha a e coluna b na matriz). FN = 2.

Para o exemplo em questão, os valores obtidos são pois:

SIGLA	VALORES
TP	7
FP	3
TN	2
FN	2
TP Rate (TPR) (*sensitivity* ou *recall*)	Classe yes(a): TP/(TP+FN) = 7/(7+2) = 7/9 = 0,778 Classe no(b): TN/(TN+FP) = 2/(2+3) = 2/5 = 0,40
FP Rate (FPR) (*fall-out*)	Classe yes(a): FP/(FP+TN) = 3/(3+2) = 3/5 = 0,60 Classe no(b): FN/(FN+TP) = 2/(2+7) = 2/9 = 0,222
Precisão (P)	Classe yes(a): TP/(TP+FP) = 7/(7+3) = 7/10 = 0,7 Classe no(b): TN/(TN + FN) = 2/(2 + 2) = 2/4 = 0,50
Success Rate (SR) (*Accuracy*)	(TP+TN)/(TP+FP+TN+FN) = (7+2)/(7+3+2+2)=9/14= 0,6428 (instâncias corretamente classificadas)

7.2.1 Primeiro Experimento

A metodologia do experimento descrito nessa seção, identificado como *Primeiro Experimento*, reflete aquela de um dos experimentos descritos em [Kotsiantis *et al.* 2003] (ver Seção 2.2), que usou um conjunto de informações acadêmicas relacionadas a desempenho de alunos, em uma disciplina ministrada durante o ano acadêmico. O trabalho em questão aborda os dados também divididos em dois grupos: (1) informações gerais do aluno e (2) informações associadas ao desempenho acadêmico durante o período da disciplina, em quatro instâncias de encontro presencial e quatro instâncias de entrega de trabalho escrito.

A abordagem adotada pelos autores foi a de, em uma tentativa de simular uma situação de uso das informações acadêmicas coletadas até um determinado ponto do desenvolvimento da disciplina, avaliar a possibilidade de inferir o resultado final de desempenho do aluno na referida disciplina. Ou seja, considerando que a disciplina tem um período de 11 meses de duração, será que com os valores de desempenho do aluno durante os dois primeiros meses, ou então com apenas o resultado da primeira avaliação, por exemplo, seria possível prever o desempenho final do estudante, na disciplina em questão? O aspecto dinâmico da condução da disciplina CAP1 e das correspondentes avaliações feitas durante o seu desenrolar, foi simulado por meio de abordagens da descrição parcial (em termos de atributos envolvidos), do conjunto dos registros acadêmicos relativos à disciplina finalizada.

7.2.1.1 Primeiro Experimento (CAP1_2009_E1)

A simulação do aspecto dinâmico incorporado ao conjunto CAP1_2009_E1 (conjunto original, com 15 atributos, sendo 4 atributos em IP e a classe G), por exemplo, fraciona o conjunto original em 12 conjuntos, como mostra a Figura 7.1. As instâncias do conjunto identificado como CAP1_2009_E1_1, por exemplo, são descritas apenas pelos quatro atributos que definem IP {DN, EC, S, GI} e pela classe G associada à cada uma das instâncias; as do CAP1_2009_E2_2 são descritas por {DN, EC, S, GI, VE, G} e assim por diante.

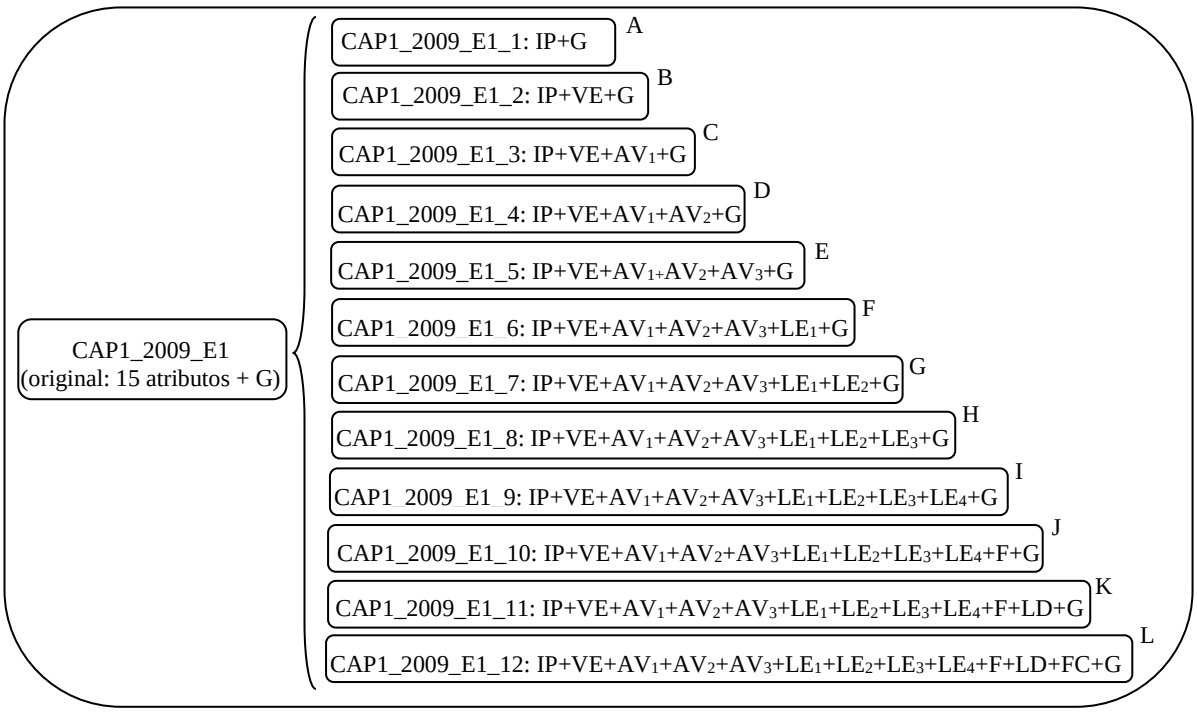

Figura 7.1 Simulação da dinâmica de avaliação durante a disciplina CAP1 (CAP1_2009_E1).

As três tabelas 7.2, 7.3 e 7.4 mostram os valores coletados dos índices mostrados na Tabela 7.1, com o uso dos algoritmos NB, NN e AD, tendo cada um dos arquivos de dados descrito na Figura 7.1, como input.

Tabela 7.2 Resultados do processo de 4-validação cruzada usando o NB, para os conjuntos de dados gerados a partir de CAP1_2009_E1.

ID	*Atrib*	*CC(%)*	*CI(%)*	*TP*	*FP*	*TN*	*FN*	*TPR*	*FPR*	*P*	*SR*
1	A	38(76)	12(24)	38	10	0	2	A=0,950 R=0	A=1 R=0,050	A=0,792 R=0	0,76
2	B	37(74)	13(26)	35	8	2	5	A=0,875 R=0,200	A=0,800 R=0,125	A=0,814 R=0,286	0,74
3	C	48(96)	2(4)	39	1	9	1	A=0,975 R=0,900	A=0,100 R=0,025	A=0,975 R=0,900	0,96
4	D	48(96)	2(4)	39	1	9	1	A=0,975 R=0,900	A=0,100 R=0,025	A=0,975 R=0,900	0,96
5	E	49(98)	1(2)	40	1	9	0	A=1 R=0,900	A=0,100 R=0	A=0,976 R=1	0,98
6	F	49(98)	1(2)	40	1	9	0	A=1 R=0,900	A=0,100 R=0	A=0,976 R=1	0,98
7	G	49(98)	1(2)	40	1	9	0	A=1 R=0,900	A=0,100 R=0	A=0,976 R=1	0,98
8	H	49(98)	1(2)	40	1	9	0	A=1 R=0,900	A=0,100 R=0	A=0,976 R=1	0,98
9	I	49(98)	1(2)	40	1	9	0	A=1 R=0,900	A=0,100 R=0	A=0,976 R=1	0,98
10	J	49(98)	1(2)	40	1	9	0	A=1 R=0,900	A=0,100 R=0	A=0,976 R=1	0,98
11	K	49(98)	1(2)	40	1	9	0	A=1 R=0,900	A=0,100 R=0	A=0,976 R=1	0,98
12	L	49(98)	1(2)	40	1	9	0	A=1 R=0,900	A=0,100 R=0	A=0,976 R=1	0,98

Tabela 7.3 Resultados do processo de 4-validação cruzada usando o NN, para os conjuntos de dados gerados a partir de CAP1_2009_E1.

ID	*Atrib*	*CC(%)*	*CI(%)*	*TP*	*FP*	*TN*	*FN*	*TPR*	*FPR*	*P*	*SR*
1	A	40(80)	10(20)	40	10	0	0	A=1 R=0	A=1 R=0	A=0,800 R=0	0,80
2	B	42(84)	8(16)	38	6	4	2	A=0,950 R=0,400	A=0,600 R=0,050	A=0,864 R=0,667	0,84
3	C	47(94)	3(6)	39	2	8	1	A=0,975 R=0,800	A=0,200 R=0,025	A=0,951 R=0,889	0,94
4	D	49(98)	1(2)	40	1	9	0	A=1 R=0,900	A=0,100 R=0	A=0,976 R=1	0,98
5	E	49(98)	1(2)	40	1	9	0	A=1 R=0,900	A=0,100 R=0	A=0,976 R=1	0,98
6	F	49(98)	1(2)	40	1	9	0	A=1 R=0,900	A=0,100 R=0	A=0,976 R=1	0,98
7	G	48(96)	2(4)	39	1	9	1	A=1 R=0,900	A=0,100 R=0	A=0,976 R=1	0,96
8	H	50(100)	0(0)	40	0	10	0	A=1 R=1	A=0 R=0	A=1 R=1	1,00
9	I	50(100)	0(0)	40	0	10	0	A=1 R=1	A=0 R=0	A=1 R=1	1,00
10	J	50(100)	0(0)	40	0	10	0	A=1 R=1	A=0 R=0	A=1 R=1	1,00
11	K	50(100)	0(0)	40	0	10	0	A=1 R=1	A=0 R=0	A=1 R=1	1,00
12	L	50(100)	0(0)	40	0	10	0	A=1 R=1	A=0 R=0	A=1 R=1	1,00

Tabela 7.4 Resultados do processo de 4-validação cruzada usando o AD, para os conjuntos de dados gerados a partir de CAP1_2009_E1.

ID	*Atrib*	*CC(%)*	*CI(%)*	*TP*	*FP*	*TN*	*FN*	*TPR*	*FPR*	*P*	*SR*
1	A	40(80)	10(20)	40	10	0	0	A=1 R=0	A=1 R=0	A=0,800 R=0	0,80
2	B	40(80)	10(20)	40	10	0	0	A=1 R=0	A=1 R=0	A=0,800 R=0	0,80
3	C	46(92)	4(8)	39	3	7	1	A=0,975 R=0,700	A=0,300 R=0,025	A=0,929 R=0,875	0,92
4	D	46(92)	4(8)	39	3	7	1	A=0,975 R=0,700	A=0,300 R=0,025	A=0,929 R=0,875	0,92
5	E	46(92)	4(8)	39	3	7	1	A=0,975 R=0,700	A=0,300 R=0,025	A=0,929 R=0,875	0,92
6	F	46(92)	4(8)	39	3	7	1	A=0,975 R=0,700	A=0,300 R=0,025	A=0,929 R=0,875	0,92
7	G	46(92)	4(8)	39	3	7	1	A=0,975 R=0,700	A=0,300 R=0,025	A=0,929 R=0,875	0,92
8	H	46(92)	4(8)	39	3	7	1	A=0,975 R=0,700	A=0,300 R=0,025	A=0,929 R=0,875	0,92
9	I	46(92)	4(8)	39	3	7	1	A=0,975 R=0,700	A=0,300 R=0,025	A=0,929 R=0,875	0,92
10	J	46(92)	4(8)	39	3	7	1	A=0,975 R=0,700	A=0,300 R=0,025	A=0,929 R=0,875	0,92
11	K	46(92)	4(8)	39	3	7	1	A=0,975 R=0,700	A=0,300 R=0,025	A=0,929 R=0,875	0,92
12	L	46(92)	4 (8)	39	3	7	1	A=0,975 R=0,700	A=0,300 R=0,025	A=0,929 R=0,875	0,92

7.2.1.2 Primeiro Experimento (CAP1_2010_E2)

A simulação do aspecto dinâmico, incorporado ao conjunto CAP1_2010_E2 (conjunto original, com 11 atributos, sendo 4 atributos em IP e a classe G), por exemplo, fraciona o conjunto original como mostra a Figura 7.2.

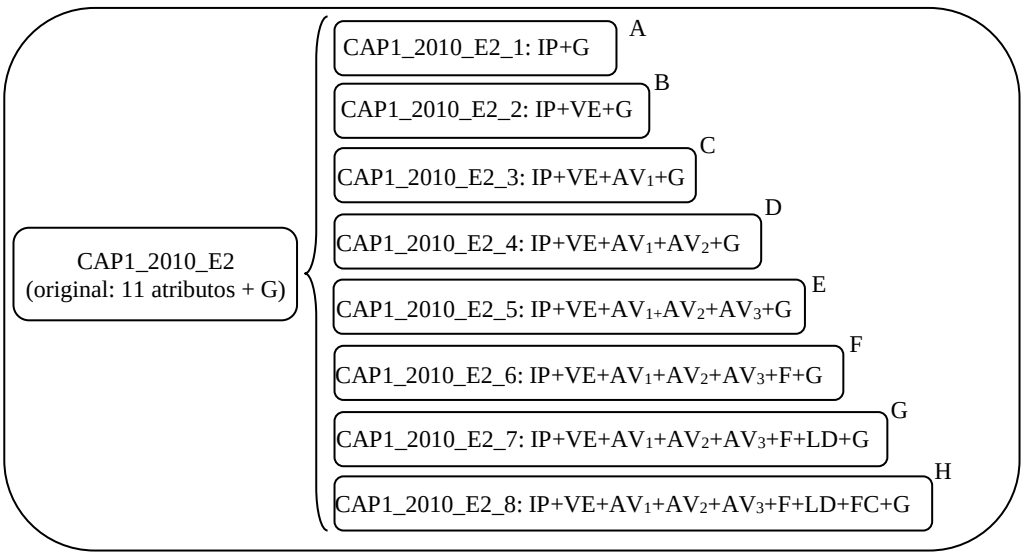

Figura 7.2 Simulação da dinâmica de avaliação durante a disciplina CAP1 (CAP1_2010_E2).

As três tabelas 7.5, 7.6 e 7.7 mostram os valores coletados dos índices mostrados na Tabela 7.1, com o uso dos algoritmos NB, NN e AD, tendo cada um dos arquivos de dados descrito na Figura 7.2, como input.

Tabela 7.5 Resultados do processo de 4-validação cruzada usando o NB, para os conjuntos de dados gerados a partir de CAP1_2010_E2.

ID	*Atrib*	*CC(%)*	*CI(%)*	*TP*	*FP*	*TN*	*FN*	*TPR*	*FPR*	*P*	*SR*
1	A	37(90,2)	4(9,8)	37	4	0	0	A=1 R=0	A=1 R=0	A=0,902 R=0	0,90
2	B	37(90,2)	4(9,8)	37	4	0	0	A=1 R=0	A=1 R=0	A=0,902 R=0	0,90
3	C	39(95,1)	2(4,9)	37	2	2	0	A=1 R=0,500	A=0,500 R=0	A=0,949 R=1	0,95
4	D	40(97,6)	1(2,4)	37	1	3	0	A=1 R=0,750	A=0,250 R=0	A=0,974 R=1	0,97
5	E	39(95,1)	2(4,9)	36	1	3	1	A=0,973 R=0,750	A=0,250 R=0,027	A=0,973 R=0,750	0,95
6	F	39(95,1)	2(4,9)	36	1	3	1	A=0,973 R=0,750	A=0,250 R=0,027	A=0,973 R=0,750	0,95
7	G	39(95,1)	2(4,9)	36	1	3	1	A=0,973 R=0,750	A=0,250 R=0,027	A=0,973 R=0,750	0,95
8	H	39(95,1)	2(4,9)	36	1	3	1	A=0,973 R=0,750	A=0,250 R=0,027	A=0,973 R=0,750	0,95

Tabela 7.6 Resultados do processo de 4-validação cruzada usando o NN, para os conjuntos de dados gerados a partir de CAP1_2010_E2.

ID	*Atrib*	*CC(%)*	*CI(%)*	*TP*	*FP*	*TN*	*FN*	*TPR*	*FPR*	*P*	*SR*
1	A	35(85,4)	6(14,6)	34	3	1	3	A=0,919 R=0,250	A=0,750 R=0,081	A=0,919 R=0,250	0,85
2	B	34(82,9)	7(17,1)	33	3	1	4	A=0,892 R=0,250	A=0,750 R=0,108	A=0,917 R=0,200	0,82
3	C	35(85,4)	6(14,6)	35	4	0	2	A=0,946 R=0	A=1 R=0,054	A=0,897 R=0	0,85
4	D	36 (87,8)	5(12,2)	35	3	1	2	A=0,946 R=0,250	A=0,750 R=0,054	A=0,921 R=0,333	0,87
5	E	37(90,2)	4(9,8)	36	3	1	1	A=0,973 R=0,250	A=0,750 R=0,027	A=0,923 R=0,500	0,90
6	F	36 (87,8)	5(12,2)	36	4	0	1	A=0,973 R=0	A=1 R=0,027	A=0,900 R=0	0,87
7	G	36(87,8)	5(12,2)	36	4	0	1	A=0,973 R=0	A=1 R=0,027	A=0,900 R=0	0,87
8	H	37(90,2)	9(9,8)	36	3	1	1	A=0,973 R=0,250	A=0,750 R=0,027	A=0,923 R=0,500	0,90

Tabela 7.7 Resultados do processo de 4-validação cruzada usando o AD, para os conjuntos de dados gerados a partir de CAP1_2010_E2.

ID	*Atrib*	*CC(%)*	*CI(%)*	*TP*	*FP*	*TN*	*FN*	*TPR*	*FPR*	*P*	*SR*
1	A	36(87,8)	5(12,2)	35	3	1	2	A=0,946 R=0,250	A=0,750 R=0,054	A=0,921 R=0,333	0,87
2	B	36(87,8)	5(12,2)	35	3	1	2	A=0,946 R=0,250	A=0,750 R=0,054	A=0,921 R=0,333	0,87
3	C	37(90,2)	4(9,8)	37	4	0	0	A=1 R=0	A=1 R=0	A=0,902 R=0	0,90
4	D	39(95,1)	2(4,9)	36	1	3	1	A=0,973 R=0,750	A=0,250 R=0,027	A=0,973 R=0,750	0,95
5	E	39(95,1)	2(4,9)	36	1	3	1	A=0,973 R=0,750	A=0,250 R=0,027	A=0,973 R=0,750	0,95
6	F	39(95,1)	2(4,9)	36	1	3	1	A=0,973 R=0,750	A=0,250 R=0,027	A=0,973 R=0,750	0,95
7	G	38(92,7)	3(7,3)	36	2	2	1	A=0,973 R=0,500	A=0,500 R=0,027	A=0,947 R=0,667	0,92
8	H	37(90,2)	4(9,8)	36	3	1	1	A=0,973 R=0,250	A=0,750 R=0,027	A=0,923 R=0,500	0,90

7.2.1.3 Primeiro Experimento (CAP1_2011_E3)

A simulação do aspecto dinâmico, incorporado ao conjunto CAP1_2011_E3 (conjunto original, com 14 atributos, sendo 4 atributos em IP e a classe G), por exemplo, fraciona o conjunto original como mostra a Figura 7.3.

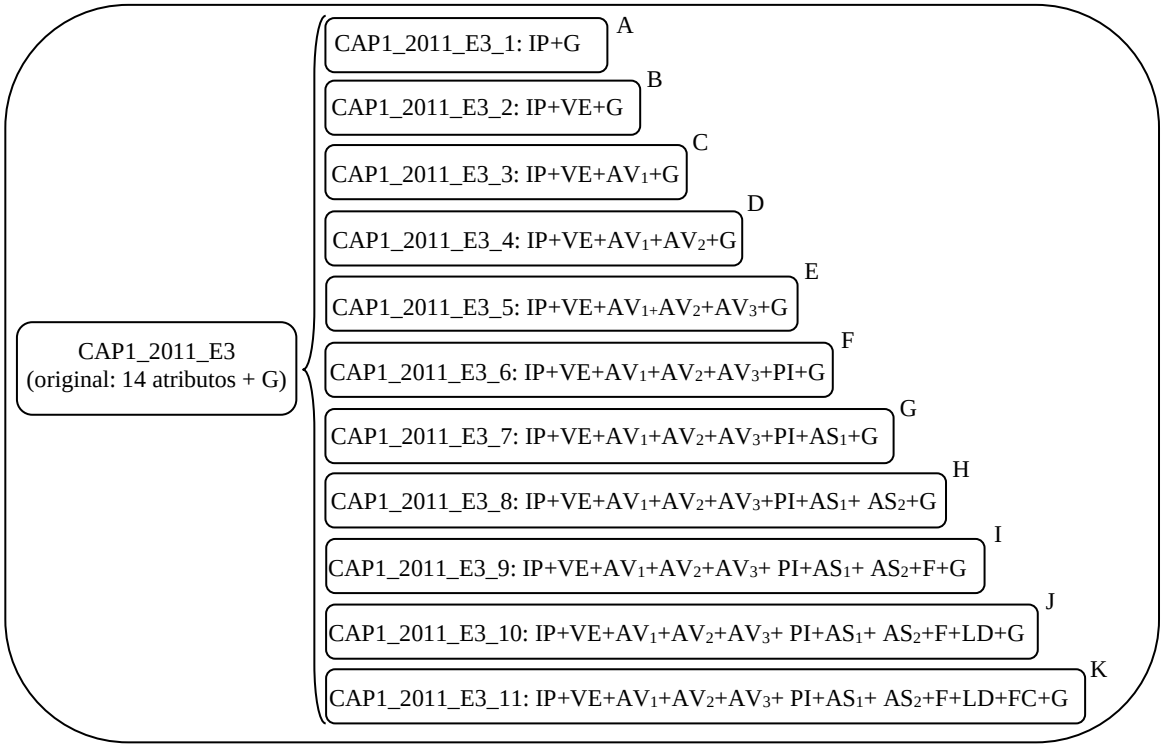

Figura 7.3 Simulação da dinâmica de avaliação durante a disciplina CAP1 (CAP1_2011_E3).

As três tabelas 7.8, 7.9 e 7.10 mostram os valores coletados dos índices mostrados na Tabela 7.1, com o uso dos algoritmos NB, NN e AD, tendo cada um dos arquivos de dados descrito na Figura 7.3, como input.

Tabela 7.8 Resultados do processo de 4-validação cruzada usando o NB, para os conjuntos de dados gerados a partir de CAP1_2011_E3.

ID	*Atrib*	*CC(%)*	*CI(%)*	*TP*	*FP*	*TN*	*FN*	*TPR*	*FPR*	*P*	*SR*
1	A	31(72)	12(28)	31	12	0	0	A=1 R=0	A=1 R=0	A=0,721 R=0	0,72
2	B	30(69,8)	13(30,2)	29	11	1	2	A=0,935 R=0,083	A=0,917 R=0,065	A=0,725 R=0,333	0,69
3	C	34(79)	9(21)	28	6	6	3	A=0,903 R=0,500	A=0,500 R=0,097	A=0,824 R=0,667	0,79
4	D	36(83,7)	7(16,3)	27	3	9	4	A=0,871 R=0,750	A=0,250 R=0,129	A=0,900 R=0,692	0,83
5	E	35(81,4)	8(18,6)	27	4	8	4	A=0,871 R=0,667	A=0,333 R=0,129	A=0,871 R=0,667	0,81
6	F	36(83,7)	7(16,3)	28	4	8	3	A=0,903 R=0,667	A=0,333 R=0,097	A=0,875 R=0,727	0,83
7	G	38(88,4)	5(11,6)	30	4	8	1	A=0,968 R=0,667	A=0,333 R=0,032	A=0,882 R=0,889	0,88
8	H	37(86)	6(14)	29	4	8	2	A=0,935 R=0,667	A=0,333 R=0,065	A=0,879 R=0,800	0,86
9	I	36(83,7)	7(16,3)	28	4	8	3	A=0,903 R=0,667	A=0,333 R=0,097	A=0,875 R=0,097	0,83
10	J	36(83,7)	7(16,3)	28	4	8	3	A=0,903 R=0,667	A=0,333 R=0,097	A=0,875 R=0,097	0,83
11	K	36(83,7)	7(16,3)	28	4	8	3	A=0,903 R=0,667	A=0,333 R=0,097	A=0,875 R=0,097	0,83

Tabela 7.9 Resultados do processo de 4-validação cruzada usando o NN, para os conjuntos de dados gerados a partir de CAP1_2011_E3.

ID	*Atrib*	*CC(%)*	*CI(%)*	*TP*	*FP*	*TN*	*FN*	*TPR*	*FPR*	*P*	*SR*
1	A	31(72)	12(28)	31	12	0	0	A=1 R=0	A=1 R=0	A=0,721 R=0	0,72
2	B	24(55,8)	19(44,2)	21	9	3	10	A=0,677 R=0,250	A=0,750 R=0,323	A=0,700 R=0,231	0,55
3	C	33(76,7)	10(23,3)	26	5	7	5	A=0,839 R=0,583	A=0,417 R=0,161	A=0,839 R=0,583	0,76
4	D	35(81,4)	8(18,6)	28	5	7	3	A=0,903 R=0,583	A=0,417 R=0,097	A=0,848 R=0,700	0,81
5	E	39(90,7)	4(9,3)	29	2	10	2	A=0,935 R=0,833	A=0,167 R=0,065	A=0,935 R=0,833	0,90
6	F	37(86)	6(14)	29	4	8	2	A=0,935 R=0,667	A=0,333 R=0,065	A=0,879 R=0,800	0,86
7	G	38(88,4)	5(11,6)	29	3	9	2	A=0,935 R=0,750	A=0,250 R=0,065	A=0,906 R=0,818	0,88
8	H	38(88,4)	5(11,6)	29	3	9	2	A=0,935 R=0,750	A=0,250 R=0,065	A=0,906 R=0,818	0,88
9	I	37(86)	6(14)	28	3	9	3	A=0,903 R=0,750	A=0,250 R=0,097	A=0,903 R=0,750	0,86
10	J	37(86)	6(14)	28	3	9	3	A=0,903 R=0,750	A=0,250 R=0,097	A=0,903 R=0,750	0,86
11	K	37(86)	6(14)	28	3	9	3	A=0,903 R=0,750	A=0,250 R=0,097	A=0,903 R=0,750	0,86

Tabela 7.10 Resultados do processo de 4-validação cruzada usando o AD, para os conjuntos de dados gerados a partir de CAP1_2011_E3.

ID	*Atrib*	*CC(%)*	*CI(%)*	*TP*	*FP*	*TN*	*FN*	*TPR*	*FPR*	*P*	*SR*
1	A	31(72)	12(28)	31	12	0	0	A=1 R=0	A=1 R=0	A=0,721 R=0	0,72
2	B	31(72)	12(28)	31	12	0	0	A=1 R=0	A=1 R=0	A=0,721 R=0	0,72
3	C	35(81,4)	8(18,6)	25	2	10	6	A=0,806 R=0,833	A=0,167 R=0,194	A=0,926 R=0,625	0,81
4	D	36(83,7)	7(16,3)	27	3	9	4	A=0,871 R=0,750	A=0,250 R=0,129	A=0,900 R=0,692	0,83
5	E	39(90,7)	4(9,3)	24	2	10	2	A=0,935 R=0,833	A=0,167 R=0,065	A=0,935 R=0,833	0,90
6	F	39(90,7)	4(9,3)	24	2	10	2	A=0,935 R=0,833	A=0,167 R=0,065	A=0,935 R=0,833	0,90
7	G	39(90,7)	4(9,3)	24	2	10	2	A=0,935 R=0,833	A=0,167 R=0,065	A=0,935 R=0,833	0,90
8	H	39(90,7)	4(9,3)	24	2	10	2	A=0,935 R=0,833	A=0,167 R=0,065	A=0,935 R=0,833	0,90
9	I	39(90,7)	4(9,3)	24	2	10	2	A=0,935 R=0,833	A=0,167 R=0,065	A=0,935 R=0,833	0,90
10	J	39(90,7)	4(9,3)	24	2	10	2	A=0,935 R=0,833	A=0,167 R=0,065	A=0,935 R=0,833	0,90
11	K	39(90,7)	4(9,3)	24	2	10	2	A=0,935 R=0,833	A=0,167 R=0,065	A=0,935 R=0,833	0,90

7.2.1.4 Primeiro Experimento (CAP1_2012_E4)

A simulação do aspecto dinâmico, incorporado ao conjunto CAP1_2012_E4 (conjunto original, com 12 atributos, sendo 4 atributos em IP e a classe G), por exemplo, fraciona o conjunto original como mostra a Figura 7.4.

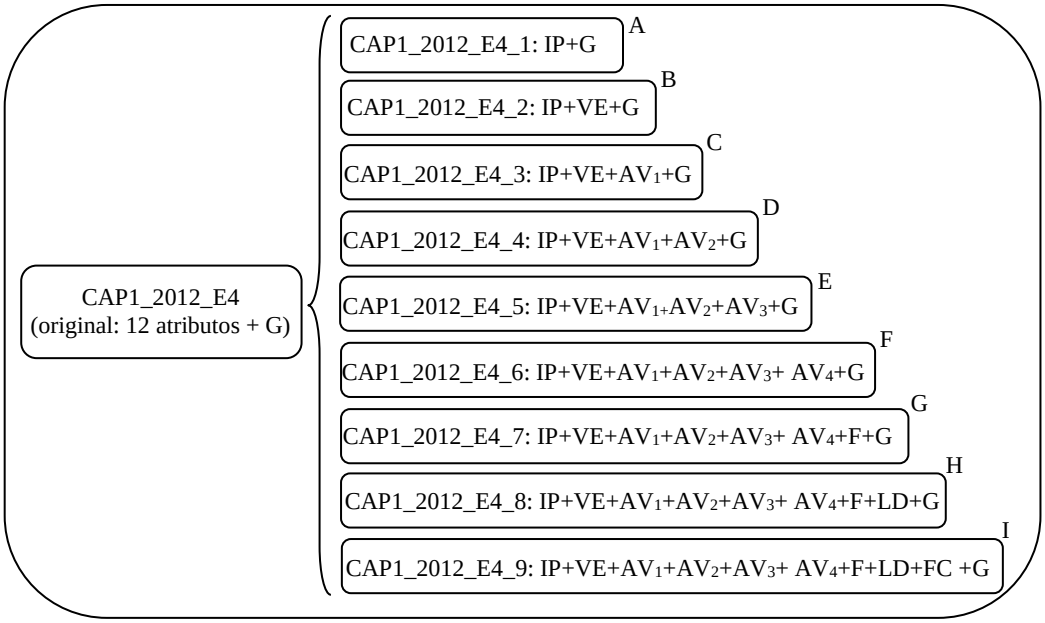

Figura 7.4 Simulação da dinâmica de avaliação durante a disciplina CAP1 (CAP1_2012_E4).

As três tabelas 7.11, 7.12 e 7.13 mostram os valores coletados dos índices mostrados na Tabela 7.1, com o uso dos algoritmos NB, NN e AD, tendo cada um dos arquivos de dados descrito na Figura 7.4, como input.

Tabela 7.11 Resultados do processo de 4-validação cruzada usando o NB, para os conjuntos de dados gerados a partir de CAP1_2012_E4.

ID	*Atrib*	*CC(%)*	*CI(%)*	*TP*	*FP*	*TN*	*FN*	*TPR*	*FPR*	*P*	*SR*
1	A	42(73,7)	15(26,3)	42	13	0	2	A=0,955 R=0	A=1 R=0,045	A=0,764 R=0	0,73
2	B	41(72)	16(28)	39	11	2	5	A=0,886 R=0,154	A=0,846 R=0,114	A=0,780 R=0,286	0,71
3	C	48(84,2)	9(15,8)	41	6	7	3	A=0,932 R=0,538	A=0,462 R=0,068	A=0,872 R=0,700	0,84
4	D	50(87,7)	7(12,3)	41	4	9	3	A=0,932 R=0,692	A=0,308 R=0,068	A=0,911 R=0,750	0,87
5	E	48(84,2)	9(15,8)	40	5	8	4	A=0,909 R=0,615	A=0,385 R=0,091	A=0,889 R=0,667	0,84
6	F	47(82,5)	10(17,5)	38	4	9	6	A=0,864 R=0,692	A=0,308 R=0,136	A=0,905 R=0,600	0,82
7	G	47(82,5)	10(17,5)	38	4	9	6	A=0,864 R=0,692	A=0,308 R=0,136	A=0,905 R=0,600	0,82
8	H	48(84,2)	9(15,8)	39	4	9	5	A=0,886 R=0,692	A=0,308 R=0,114	A=0,907 R=0,643	0,84
9	I	47(82,5)	10(17,5)	39	5	8	5	A=0,886 R=0,615	A=0,385 R=0,114	A=0,886 R=0,615	0,82

Tabela 7.12 Resultados do processo de 4-validação cruzada usando o NN, para os conjuntos de dados gerados a partir de CAP1_2012_E4.

ID	*Atrib*	*CC(%)*	*CI(%)*	*TP*	*FP*	*TN*	*FN*	*TPR*	*FPR*	*P*	*SR*
1	A	38(66,7)	19(33,3)	38	13	0	6	A=0,864 R=0	A=1 R=0,136	A=0,746 R=0	0,66
2	B	33(57,9)	24(42,1)	33	13	0	11	A=0,750 R=0	A=1 R=0,250	A=0,777 R=0	0,57
3	C	41(72)	16(28)	36	8	5	8	A=0,818 R=0,385	A=0,615 R=0,182	A=0,818 R=0,385	0,72
4	D	49(86)	8(14)	41	5	8	3	A=0,932 R=0,615	A=0,385 R=0,068	A=0,891 R=0,727	0,85
5	E	52(91,2)	5(8,8)	42	3	10	2	A=0,955 R=0,769	A=0,231 R=0,045	A=0,933 R=0,833	0,91
6	F	49(86)	8(14)	41	5	8	3	A=0,932 R=0,615	A=0,385 R=0,068	A=0,891 R=0,727	0,85
7	G	49(86)	8(14)	41	5	8	3	A=0,932 R=0,615	A=0,385 R=0,068	A=0,891 R=0,727	0,85
8	H	50(87,7)	7(12,3)	41	4	9	3	A=0,932 R=0,692	A=0,308 R=0,068	A=0,911 R=0,750	0,87
9	I	51(89,5)	6(10,5)	42	4	9	2	A=0,955 R=0,692	A=0,308 R=0,045	A=0,913 R=0,818	0,89

Tabela 7.13 Resultados do processo de 4-validação cruzada usando o AD, para os conjuntos de dados gerados a partir de CAP1_2012_E4.

ID	*Atrib*	*CC(%)*	*CI(%)*	*TP*	*FP*	*TN*	*FN*	*TPR*	*FPR*	*P*	*SR*
1	A	44(77,2)	13(22,8)	44	13	0	0	A=1 R=0	A=1 R=0	A=0,772 R=0	0,77
2	B	44(77,2)	13(22,8)	44	13	0	0	A=1 R=0	A=1 R=0	A=0,772 R=0	0,77
3	C	47(82,5)	10(17,5)	38	4	9	6	A=0,864 R=0,692	A=0,308 R=0,136	A=0,905 R=0,600	0,82
4	D	52(91,2)	5(8,8)	42	3	10	2	A=0,955 R=0,769	A=0,231 R=0,045	A=0,933 R=0,833	0,91
5	E	51(89,5)	6(10,5)	41	3	10	3	A=0,932 R=0,769	A=0,231 R=0,068	A=0,932 R=0,769	0,89
6	F	52(91,2)	5(8,8)	43	4	9	1	A=0,977 R=0,692	A=0,308 R=0,023	A=0,915 R=0,900	0,91
7	G	52(91,2)	5(8,8)	43	4	9	1	A=0,977 R=0,692	A=0,308 R=0,023	A=0,915 R=0,900	0,91
8	H	49(86)	8(14)	42	6	7	2	A=0,955 R=0,538	A=0,462 R=0,045	A=0,875 R=0,778	0,85
9	I	49(86)	8(14)	42	6	7	2	A=0,955 R=0,538	A=0,462 R=0,045	A=0,875 R=0,778	0,85

7.2.1.5 Primeiro Experimento (CAP1_2013_E5)

A simulação do aspecto dinâmico, incorporado ao conjunto CAP1_2013_E5 (conjunto original, com 12 atributos, sendo 4 atributos em IP e a classe G), por exemplo, fraciona o conjunto original como mostra a Figura 7.5.

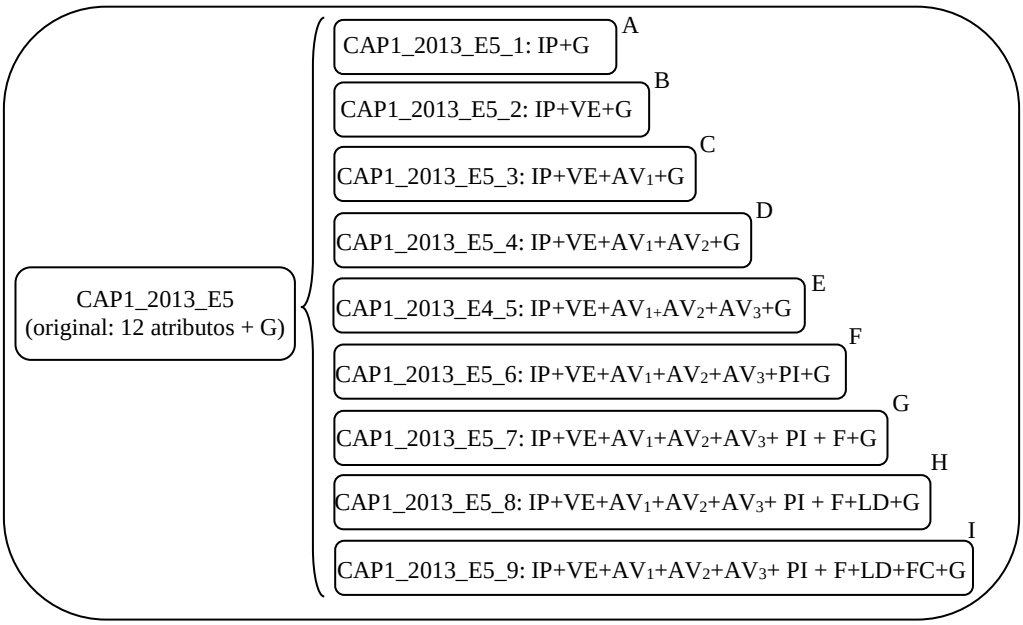

Figura 7.5 Simulação da dinâmica de avaliação durante a disciplina CAP1 (CAP1_2013_E5).

As três tabelas 7.14, 7.15 e 7.16 mostram os valores coletados dos índices mostrados na Tabela 7.1, com o uso dos algoritmos NB, NN e AD, tendo cada um dos arquivos de dados descrito na Figura 7.5, como input.

Tabela 7.14 Resultados do processo de 4-validação cruzada usando o NB, para os conjuntos de dados gerados a partir de CAP1_2013_E5.

ID	*Atrib*	*CC(%)*	*CI(%)*	*TP*	*FP*	*TN*	*FN*	*TPR*	*FPR*	*P*	*SR*
1	A	30(50)	30(50)	23	21	7	9	A=0,719 R=0,250	A=0,750 R=0,281	A=0,523 R=0,438	0,50
2	B	28(46,7)	32(53,3)	15	15	13	17	A=0,469 R=0,464	A=0,536 R=0,531	A=0,500 R=0,433	0,46
3	C	41(68,3)	19(31,7)	22	9	19	10	A=0,688 R=0,679	A=0,321 R=0,313	A=0,710 R=0,655	0,68
4	D	50(83,3)	10(16,7)	27	5	23	5	A=0,844 R=0,821	A=0,179 R=0,156	A=0,844 R=0,821	0,83
5	E	53(88,3)	7(11,7)	27	2	26	5	A=0,844 R=0,929	A=0,071 R=0,156	A=0,031 R=0,539	0,88
6	F	54(90)	6(10)	28	2	26	4	A=0,875 R=0,929	A=0,071 R=0,125	A=0,933 R=0,867	0,90
7	G	54(90)	6(10)	28	2	26	4	A=0,875 R=0,929	A=0,071 R=0,125	A=0,933 R=0,867	0,90
8	H	54(90)	6(10)	28	2	26	4	A=0,875 R=0,929	A=0,071 R=0,125	A=0,933 R=0,867	0,90
9	I	53(88,3)	7(11,7)	27	2	26	5	A=0,844 R=0,929	A=0,071 R=0,156	A=0,931 R=0,839	0,88

Tabela 7.15 Resultados do processo de 4-validação cruzada usando o NN, para os conjuntos de dados gerados a partir de CAP1_2013_E5.

ID	*Atrib*	*CC(%)*	*CI(%)*	*TP*	*FP*	*TN*	*FN*	*TPR*	*FPR*	*P*	*SR*
1	A	32(53,3)	28(46,7)	19	15	13	13	A=0,594 R=0,464	A=0,536 R=0,406	A=0,559 R=0,464	0,53
2	B	28(46,7)	32(53,3)	18	18	10	14	A=0,563 R=0,357	A=0,643 R=0,438	A=0,500 R=0,417	0,46
3	C	31(51,7)	29(48,3)	18	15	13	14	A=0,563 R=0,464	A=0,536 R=0,438	A=0,545 R=0,481	0,51
4	D	41(68,3)	19(31,7)	23	10	18	9	A=0,719 R=0,643	A=0,357 R=0,281	A=0,697 R=0,697	0,68
5	E	48(80)	12(20)	25	5	23	7	A=0,781 R=0,821	A=0,179 R=0,219	A=0,833 R=0,767	0,80
6	F	52(86,7)	8(13,3)	26	2	26	6	A=0,813 R=0,929	A=0,071 R=0,813	A=0,929 R=0,813	0,86
7	G	52(86,7)	8(13,3)	26	2	26	6	A=0,813 R=0,929	A=0,071 R=0,813	A=0,929 R=0,813	0,86
8	H	50(83,3)	10(16,7)	26	4	24	6	A=0,813 R=0,857	A=0,143 R=0,188	A=0,867 R=0,800	0,83
9	I	50(83,3)	10(16,7)	26	4	24	6	A=0,813 R=0,857	A=0,143 R=0,188	A=0,867 R=0,800	0,83

Tabela 7.16 Resultados do processo de 4-validação cruzada usando o AD, para os conjuntos de dados gerados a partir de CAP1_2013_E5.

ID	*Atrib*	*CC(%)*	*CI(%)*	*TP*	*FP*	*TN*	*FN*	*TPR*	*FPR*	*P*	*SR*
1	A	32(53,3)	28(46,7)	32	28	0	0	A=1 R=0	A=1 R=0	A=0,533 R=0	0,53
2	B	32(53,3)	28(46,7)	28	24	4	4	A=0,875 R=0,143	A=0,857 R=0,125	A=0,538 R=0,500	0,53
3	C	47(73,3)	13(21,7)	25	6	22	7	A=0,781 R=0,786	A=0,214 R=0,219	A=0,806 R=0,759	0,78
4	D	43(71,7)	17(28,3)	22	7	21	10	A=0,688 R=0,750	A=0,250 R=0,313	A=0,759 R=0,677	0,71
5	E	47(78,3)	13(21,7)	27	8	20	5	A=0,844 R=0,714	A=0,286 R=0,156	A=0,771 R=0,800	0,78
6	F	47(78,3)	13(21,7)	27	8	20	5	A=0,844 R=0,714	A=0,286 R=0,156	A=0,771 R=0,800	0,78
7	G	47(78,3)	13(21,7)	27	8	20	5	A=0,844 R=0,714	A=0,286 R=0,156	A=0,771 R=0,800	0,78
8	H	47(78,3)	13(21,7)	27	8	20	5	A=0,844 R=0,714	A=0,286 R=0,156	A=0,771 R=0,800	0,78
9	I	47(78,3)	13(21,7)	27	8	20	5	A=0,844 R=0,714	A=0,286 R=0,156	A=0,771 R=0,800	0,78

7.2.1.6 Primeiro Experimento (CAP1_2014_E5)

A simulação do aspecto dinâmico, incorporado ao conjunto CAP1_2014_E5 (conjunto original, com 12 atributos, sendo 4 atributos em IP e a classe G), por exemplo, fraciona o conjunto original como mostra a Figura 7.6.

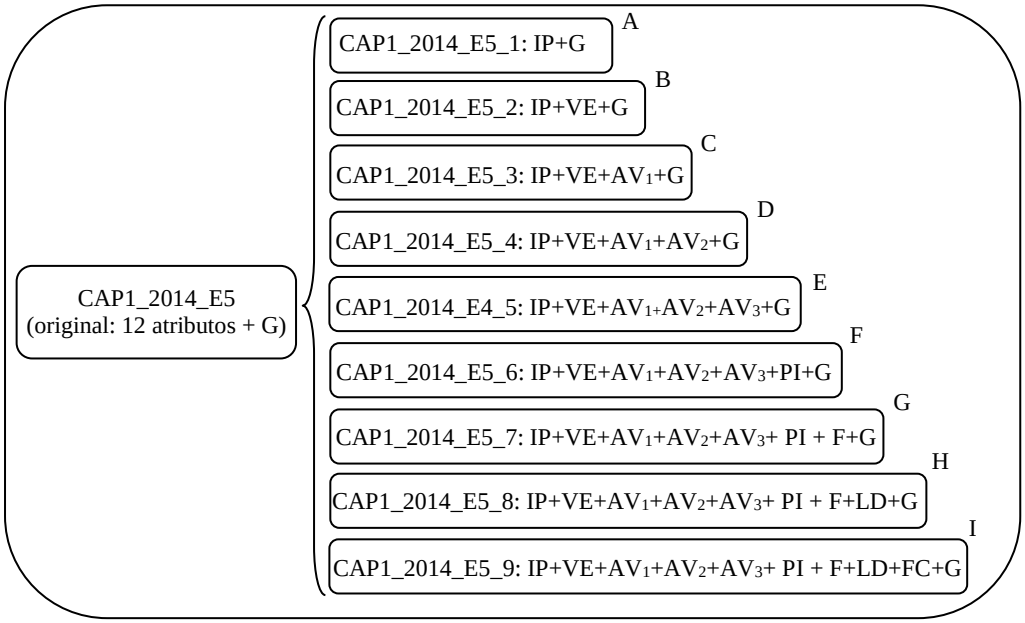

Figura 7.6 Simulação da dinâmica de avaliação durante a disciplina CAP1 (CAP1_2014_E5).

As três tabelas 7.17, 7.18 e 7.19 mostram os valores coletados dos índices mostrados na Tabela 7.1, com o uso dos algoritmos NB, NN e AD, tendo cada um dos arquivos de dados descrito na Figura 7.6, como input.

Tabela 7.17 Resultados do processo de 4-validação cruzada usando o NB, para os conjuntos de dados gerados a partir de CAP1_2014_E5.

ID	*Atrib*	*CC(%)*	*CI(%)*	*TP*	*FP*	*TN*	*FN*	*TPR*	*FPR*	*P*	*SR*
1	A	27(73)	10(27)	20	7	7	3	A=0,870 R=0,500	A=0,500 R=0,130	A=0,741 R=0,700	0,72
2	B	27(73)	10(27)	20	7	7	3	A=0,870 R=0,500	A=0,500 R=0,130	A=0,741 R=0,700	0,72
3	C	29(78,4)	8(21,6)	20	5	9	3	A=0,870 R=0,643	A=0,357 R=0,130	A=0,800 R=0,750	0,78
4	D	34(91,9)	3(8,1)	22	2	12	1	A=0,957 R=0,857	A=0,143 R=0,043	A=0,917 R=0,923	0,91
5	E	34(91,9)	3(8,1)	22	2	12	1	A=0,957 R=0,857	A=0,143 R=0,043	A=0,917 R=0,923	0,91
6	F	34(91,9)	3(8,1)	21	1	13	2	A=0,913 R=0,929	A=0,071 R=0,087	A=0,955 R=0,867	0,91
7	G	33(89,1)	4(10,9)	21	2	12	2	A=0,919 R=0,857	A=0,143 R=0,087	A=0,913 R=0,857	0,89
8	H	32(86,5)	5(13,5)	20	2	12	3	A=0,870 R=0,857	A=0,143 R=0,130	A=0,909 R=0,800	0,86
9	I	32(86,5)	5(13,5)	20	2	12	3	A=0,870 R=0,857	A=0,143 R=0,130	A=0,909 R=0,800	0,86

Tabela 7.18 Resultados do processo de 4-validação cruzada usando o NN, para os conjuntos de dados gerados a partir de CAP1_2014_E5.

ID	*Atrib*	*CC(%)*	*CI(%)*	*TP*	*FP*	*TN*	*FN*	*TPR*	*FPR*	*P*	*SR*
1	A	18(48,6)	19(51,4)	15	11	3	8	A=0,625 R=0,214	A=0,786 R=0,273	A=0,577 R=0,273	0,48
2	B	25(67,6)	12(32,4)	17	6	8	6	A=0,739 R=0,571	A=0,429 R=0,261	A=0,739 R=0,571	0,67
3	C	22(59,5)	15(40,5)	17	9	5	6	A=0,739 R=0,357	A=0,643 R=0,261	A=0,654 R=0,455	0,59
4	D	26(70,3)	11(29,7)	17	5	9	6	A=0,739 R=0,643	A=0,357 R=0,261	A=0,773 R=0,600	0,70
5	E	29(78,4)	8(21,6)	20	5	9	3	A=0,870 R=0,643	A=0,357 R=0,130	A=0,800 R=0,750	0,78
6	F	30(81)	7(19)	20	4	10	3	A=0,870 R=0,714	A=0,286 R=0,130	A=0,833 R=0,769	0,81
7	G	31(83,8)	6(16,2)	20	3	11	3	A=0,870 R=0,786	A=0,214 R=0,130	A=0,870 R=0,786	0,83
8	H	28(75,7)	9(24,3)	21	7	7	2	A=0,913 R=0,500	A=0,500 R=0,687	A=0,750 R=0,778	0,75
9	I	27(73)	10(27)	20	7	7	3	A=0,870 R=0,500	A=0,500 R=0,130	A=0,741 R=0,700	0,72

Tabela 7.19 Resultados do processo de 4-validação cruzada usando o AD, para os conjuntos de dados gerados a partir de CAP1_2014_E5.

ID	*Atrib*	*CC(%)*	*CI(%)*	*TP*	*FP*	*TN*	*FN*	*TPR*	*FPR*	*P*	*SR*
1	A	27(73)	10(27)	21	8	6	2	A=0,913 R=0,429	A=0,571 R=0,087	A=0,724 R=0,750	0,72
2	B	25(67,5)	12(32,4)	18	7	7	5	A=0,783 R=0,500	A=0,500 R=0,217	A=0,720 R=0,583	0,67
3	C	27(73)	10(27)	19	6	8	4	A=0,826 R=0,571	A=0,429 R=0,174	A=0,760 R=0,667	0,72
4	D	28(75,7)	9(24,3)	20	6	8	3	A=0,870 R=0,571	A=0,429 R=0,130	A=0,769 R=0,727	0,75
5	E	28(75,7)	9(24,3)	19	5	9	4	A=0,826 R=0,643	A=0,357 R=0,174	A=0,792 R=0,692	0,75
6	F	28(75,7)	9(24,3)	19	5	9	4	A=0,826 R=0,643	A=0,357 R=0,174	A=0,792 R=0,692	0,75
7	G	28(75,7)	9(24,3)	19	5	9	4	A=0,826 R=0,643	A=0,357 R=0,174	A=0,792 R=0,692	0,75
8	H	28(75,7)	9(24,3)	19	5	9	4	A=0,826 R=0,643	A=0,357 R=0,174	A=0,792 R=0,692	0,75
9	I	28(75,7)	9(24,3)	19	5	9	4	A=0,826 R=0,643	A=0,357 R=0,174	A=0,792 R=0,692	0,75

7.2.1.7 Primeiro Experimento (CAP1_2015_E6)

A simulação do aspecto dinâmico, incorporado ao conjunto CAP1_2015_E6 (conjunto original, com 11 atributos, sendo 4 atributos em IP e a classe G), por exemplo, fraciona o conjunto original como mostra a Figura 7.7.

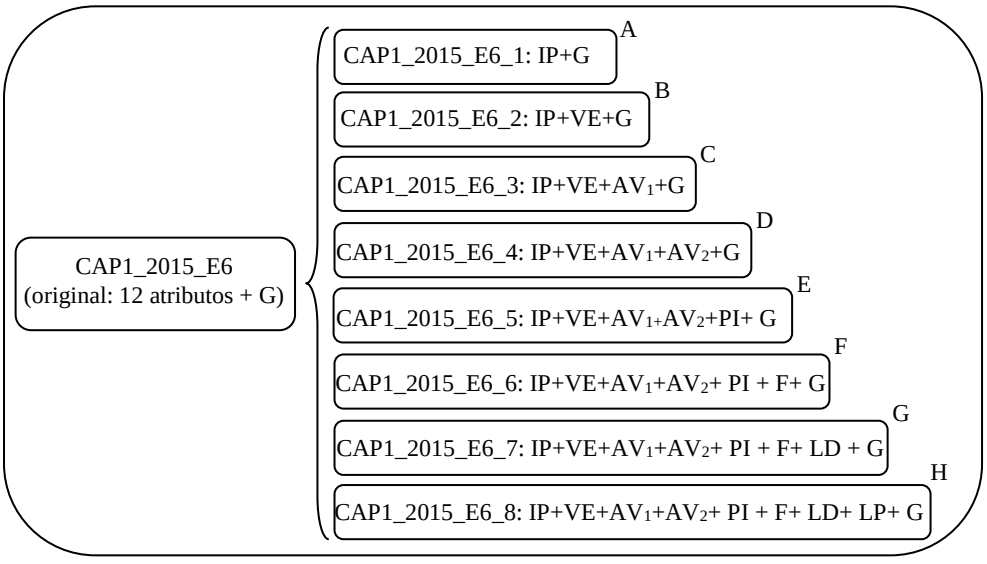

Figura 7.7 Simulação da dinâmica de avaliação durante a disciplina CAP1 (CAP1_2015_E6).

As três tabelas 7.20, 7.21 e 7.22 mostram os valores coletados dos índices mostrados na Tabela 7.1, com o uso dos algoritmos NB, NN e AD, tendo cada um dos arquivos de dado descrito na Figura 7.7, como input.

Tabela 7.20 Resultados do processo de 4-validação cruzada usando o NB, para os conjuntos de dados gerados a partir de CAP1_2015_E6.

ID	*Atrib*	*CC(%)*	*CI(%)*	*TP*	*FP*	*TN*	*FN*	*TPR*	*FPR*	*P*	*SR*
1	A	46(66,7)	23(33,3)	36	18	10	5	A=0,878 R=0,357	A=0,643 R=0,122	A=0,667 R=0,667	0,66
2	B	44(63,8)	25(36,2)	34	18	10	7	A=0,829 R=0,357	A=0,643 R=0,171	A=0,654 R=0,588	0,63
3	C	63(91,3)	6(8,7)	38	3	25	3	A=0,927 R=0,893	A=0,107 R=0,073	A=0,927 R=0,893	0,91
4	D	64(92,8)	5(7,2)	40	4	24	1	A=0,976 R=0,857	A=0,143 R=0,024	A=0,909 R=0,960	0,92
5	E	65(94,2)	4(5,8)	40	3	25	1	A=0,976 R=0,893	A=0,107 R=0,024	A=0,930 R=0,962	0,94
6	F	63(91,3)	6(8,7)	39	4	24	2	A=0,951 R=0,857	A=0,143 R=0,049	A=0,907 R=0,923	0,91
7	G	64(92,8)	5(7,2)	39	3	25	2	A=0,951 R=0,893	A=0,107 R=0,049	A=0,929 R=0,926	0,92
8	H	64(92,8)	5(7,2)	39	3	25	2	A=0,951 R=0,893	A=0,107 R=0,049	A=0,929 R=0,926	0,92

Tabela 7.21 Resultados do processo de 4-validação cruzada usando o NN, para os conjuntos de dados gerados a partir de CAP1_2015_E6.

ID	*Atrib*	*CC(%)*	*CI(%)*	*TP*	*FP*	*TN*	*FN*	*TPR*	*FPR*	*P*	*SR*
1	A	36(52,2)	33(47,8)	29	21	7	12	A=0,707 R=0,250	A=0,750 R=0,293	A=0,580 R=0,368	0,52
2	B	40(58)	29(42)	30	18	10	11	A=0,732 R=0,357	A=0,643 R=0,268	A=0,625 R=0,476	0,57
3	C	63(91,3)	6(8,7)	37	2	26	4	A=0,902 R=0,929	A=0,071 R=0,098	A=0,949 R=0,867	0,91
4	D	67(97,1)	2(2,9)	40	1	27	1	A=0,976 R=0,964	A=0,036 R=0,024	A=0,976 R=0,964	0,97
5	E	67(97,1)	2(2,9)	41	2	26	0	A=1 R=0,929	A=0,071 R=0	A=0,953 R=1	0,97
6	F	66(95,7)	3(4,3)	41	3	25	0	A=1 R=0,893	A=0,107 R=0	A=0,932 R=1	0,95
7	G	65(94,2)	4(5,8)	41	4	24	0	A=1 R=0,857	A=0,143 R=0	A=0,911 R=1	0,94
8	H	63(91,3)	6(8,7)	39	4	24	2	A=0,951 R=0,857	A=0,143 R=0,049	A=0,907 R=0,923	0,91

Tabela 7.22 Resultados do processo de 4-validação cruzada usando o AD, para os conjuntos de dados gerados a partir de CAP1_2015_E6.

ID	*Atrib*	*CC(%)*	*CI(%)*	*TP*	*FP*	*TN*	*FN*	*TPR*	*FPR*	*P*	*SR*
1	A	44(63,8)	25(36,2)	26	10	18	15	A=0,634 R=0,643	A=0,357 R=0,366	A=0,722 R=0,545	0,63
2	B	41(59,4)	28(40,6)	27	14	14	14	A=0,659 R=0,500	A=0,500 R=0,341	A=0,659 R=0,500	0,59
3	C	64(92,7)	5(7,3)	39	3	25	2	A=0,951 R=0,893	A=0,107 R=0,049	A=0,929 R=0,926	0,92
4	D	61(88,4)	8(11,6)	39	6	22	2	A=0,951 R=0,786	A=0,214 R=0,049	A=0,867 R=0,917	0,88
5	E	61(88,4)	8(11,6)	39	6	22	2	A=0,951 R=0,786	A=0,214 R=0,049	A=0,867 R=0,917	0,88
6	F	61(88,4)	8(11,6)	39	6	22	2	A=0,951 R=0,786	A=0,214 R=0,049	A=0,867 R=0,917	0,88
7	G	61(88,4)	8(11,6)	39	6	22	2	A=0,951 R=0,786	A=0,214 R=0,049	A=0,867 R=0,917	0,88
8	H	61(88,4)	8(11,6)	39	6	22	2	A=0,951 R=0,786	A=0,214 R=0,049	A=0,867 R=0,917	0,88

7.2.1.8 Resumo e Análise dos Resultados do Primeiro Experimento

7.2.1.8.1 *Primeiro Experimento* – NB

No conjunto CAP1_2009_E1 a melhor precisão obtida pelo algoritmo Naïve Bayes foi de 98%, como mostra a Figura 7.8. Note que, muito embora os conjuntos de atributos utilizados em cada uma das 12 instâncias do conjunto original variam, o gráfico pretende apenas exibir o comportamento do NB associado a cada um dos 12 conjuntos de atributos considerados. Este resultado foi alcançado na instância do conjunto original de dados identificada por CAP1_2009_E1_5 (E), (ver Tabela 7.1), e se manteve até a instância do conjunto (L). A 5ª instância do conjunto (E) é descrita pelo conjunto de atributos {DN, EC, S, GI, VE, AV_1, AV_2, AV_3}. Pode ser verificado que o melhor resultado alcançado para este conjunto de dados foi obtido quando foram inseridos os três atributos de avaliação AV_1, AV_2 e AV_3; O resultado não se alterou quando foram acrescentados ao conjunto os atributos restantes LE_1, LE_2, LE_3, LE_4, F, LD e FC.

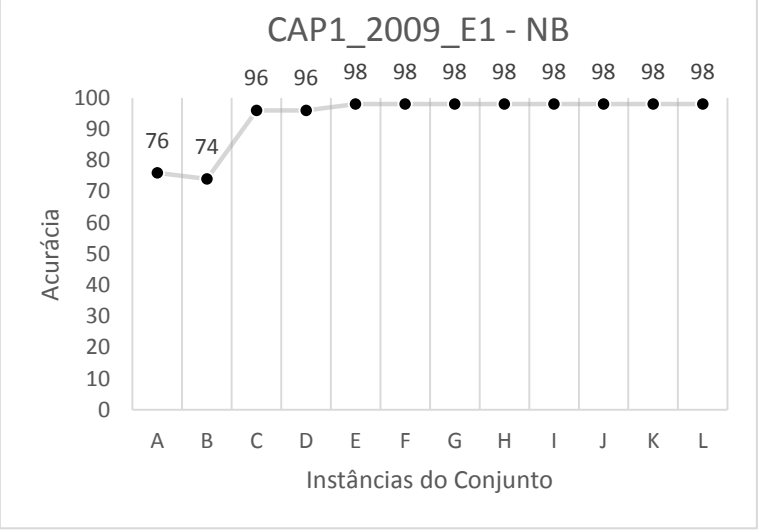

Figura 7.8 Gráfico da acurácia obtida pelo classificador NB usando as 12 instâncias do conjunto CAP1_2009_E1.

Com relação ao conjunto de dados CAP1_2010_E2, o NB obteve sua melhor acurácia na instância CAP1_2010_E2_4 (D) (ver Tabela 7.5), quando foi inserido o atributo AV_2, compondo o conjunto {DN, EC, S, GI, VE, AV_1, AV_2}. A porcentagem de classificação caiu de 97%, para 95% quando os demais atributos AV_3, F, LD, FC também passaram a fazer parte do conjunto. O esperado seria que o melhor resultado aparecesse no fim/ou perto do fim, quando todos ou quase todos os atributos participassem do

conjunto de dados, o que se mostrou diferente para este conjunto de dados neste experimento.

Quando do uso dos dados em CAP1_2011_E3, o NB obteve 72% de precisão na predição, na primeira instância desse conjunto CAP1_2011_E3 _1(A) e, na última instância, obteve 83% (ver Tabela 7.8), sendo que seu maior valor de acurácia foi obtido quando do uso da instância do conjunto identificada como CAP1_2011_E3_7 (G), quando o atributo AS_1 foi agregado ao conjunto de atributos anterior, resultando em DN, EC, S, GI, VE, AV_1, AV_2, AV_3, PI, AS_1, G. Esta situação é similar às ocorridas quando os conjuntos CAP1_2012_E4 (ver Tabela 7.11), CAP1_2014_E5 (ver Tabela 7.17) e CAP1_2015_E6 (ver Tabela 7.20), foram utilizados.

A Tabela 7.23, mostra a média aritmética dos sete conjuntos de dados, levando em conta o conjunto de instâncias do conjunto original, associado a cada um dos sete conjuntos. Uma instância do conjunto original é caracterizada pelos atributos que descrevem as instâncias do conjunto; no caso do CAP1_2009_E1 existem 12 instâncias associadas ao conjunto (de A até L). Foi calculada a média aritmética do desempenho obtido pelo classificador NB por instância do conjunto, considerando todos os conjuntos de dados usados.

Tabela 7.23 Média aritmética dos resultados de acurácia do processo de 4-validação cruzada usando o NB, para todos os conjuntos de dados, para cada instância do conjunto original.

	2009	**2010**	**2011**	**2012**	**2013**	**2014**	**2015**	**MÉDIA**
A	76	90	72	73	50	72	66	71,2
B	74	90	69	71	46	72	63	69,2
C	96	95	79	84	68	78	91	84,4
D	96	97	83	87	83	91	92	89,8
E	98	95	81	84	88	91	94	90,1
F	98	95	83	82	90	91	91	90
G	98	95	88	82	90	89	92	90,5
H	98	95	86	84	90	86	92	90,1
I	98		83	82	88	86		87,4
J	98		83					90,5
K	98		83					90,5
L	98							98

Os atributos contidos em cada instância do conjunto de cada um dos conjuntos usados no cálculo dos resultados da Tabela 7.23, estão listados na Tabela 7.24.

Tabela 7.24 Conjunto de atributos presentes em cada instância do conjunto original, dos citados nas tabelas 7.23, 7.25 e 7.26.

	2009	2010	2011	2012	2013	2014	2015
A	IP+G	IP+G	IP+G	IP+G	IP+G	IP+G	IP+G
B	IP+VE+G	IP+VE+G	IP+VE+G	IP+VE+G	IP+VE+G	IP+VE+G	IP+VE+G
C	IP+VE+AV_1+G	IP+VE+AV_1+G	IP+VE+AV_1+G	IP+VE+AV_1+G	IP+VE+AV_1+G	IP+VE+AV_1+G	IP+VE+AV_1+G
D	IP+VE+AV_1+AV_2+G	IP+VE+AV_1+AV_2+G	IP+VE+AV_1+AV_2+G	IP+VE+AV_1+AV_2+G	IP+VE+AV_1+AV_2+G	IP+VE+AV_1+AV_2+G	IP+VE+AV_1+AV_2+G
E	IP+VE+AV_1+AV_2+AV_3+G	IP+VE+AV_1+AV_2+AV_3+G	IP+VE+AV_1+AV_2+AV_3+G	IP+VE+AV_1+AV_2+AV_3+G	IP+VE+AV_1+AV_2+AV_3+G	IP+VE+AV_1+AV_2+AV_3+G	IP+VE+AV_1+AV_2+PI+G
F	IP+VE+AV_1+AV_2+AV_3+LE_1+G	IP+VE+AV_1+AV_2+AV_3+F+G	IP+VE+AV_1+AV_2+AV_3+PI+G	IP+VE+AV_1+AV_2+AV_3+AV_4+G	IP+VE+AV_1+AV_2+AV_3+PI+G	IP+VE+AV_1+AV_2+AV_3+PI+G	IP+VE+AV_1+AV_2+PI+F+G
G	IP+VE+AV_1+AV_2+AV_3+LE_1+LE_2+G	IP+VE+AV_1+AV_2+AV_3+F+LD+G	IP+VE+AV_1+AV_2+AV_3+PI+AS_1+G	IP+VE+AV_1+AV_2+AV_3+AV_4+F+G	IP+VE+AV_1+AV_2+AV_3+PI+F+G	IP+VE+AV_1+AV_2+AV_3+PI+F+G	IP+VE+AV_1+AV_2+PI+F+LD+G
H	IP+VE+AV_1+AV_2+AV_3+LE_1+LE_2+LE_3+G	IP+VE+AV_1+AV_2+AV_3+F+LD+FC+G	IP+VE+AV_1+AV_2+AV_3+PI+AS_1+AS_2+G	IP+VE+AV_1+AV_2+AV_3+AV_4+F+LD+G	IP+VE+AV_1+AV_2+AV_3+PI+F+LD+G	IP+VE+AV_1+AV_2+AV_3+PI+F+LD+G	IP+VE+AV_1+AV_2+PI+F+LD+LP+G
I	IP+VE+AV_1+AV_2+AV_3+LE_1+LE_2+LE_3+LE_4+G		IP+VE+AV_1+AV_2+AV_3+PI+AS_1+AS_2+F+G	IP+VE+AV_1+AV_2+AV_3+AV_4+F+LD+FC+G	IP+VE+AV_1+AV_2+AV_3+PI+F+LD+FC+G	IP+VE+AV_1+AV_2+AV_3+PI+F+LD+FC+G	
J	IP+VE+AV_1+AV_2+AV_3+LE_1+LE_2+LE_3+LE_4+F+G		IP+VE+AV_1+AV_2+AV_3+PI+AS_1+AS_2+F+LD+G				
K	IP+VE+AV_1+AV_2+AV_3+LE_1+LE_2+LE_3+LE_4+F+LD+G		IP+VE+AV_1+AV_2+AV_3+PI+AS_1+AS_2+F+LD+FC+G				
L	IP+VE+AV_1+AV_2+AV_3+LE_1+LE_2+LE_3+LE_4+F+LD+FC+G						

7.2.1.8.2 *Primeiro Experimento* – NN

Quando do uso do conjunto CAP1_2009_E1, o classificador NN obteve 80% de acurácia na primeira instância de tal conjunto *i.e.*, CAP1_2009_E1(A), e alcançou 100% de acurácia em CAP1_2009_E1_8 (H), mantendo este resultado até a última instância (L). Já quando do uso do conjunto CAP1_2010_E2, o NN obteve 85% de acurácia quando do uso da primeira instância do conjunto i.e., CAP1_2010_E2_1(A). Nas instâncias seguintes os resultados de acurácia do NN variaram no intervalo 82 a 90: B=82%, C=85%, D=87%, E=90%, F=87%, G=87% e H=90%. O maior valor de acurácia obtido

para este conjunto de dados com o NN foi 90%, quando foi utilizada a instância CAP1_2010_E2_5(E), cujo grupo de atributos associado é DN, EC, S, GI, VE, AV_1, AV_2, AV_3 e G. O valor de acurácia de 90% foi obtido quando do uso da última instância do conjunto i.e., CAP1_2010_E2_8(H), cujo conjunto de atributos associado é DN, EC, S, GI, VE, AV_1, AV_2, AV_3 F, LD, FC, G. A Tabela 7.25, mostra a média aritmética da acurácia do NN dos sete conjuntos de dados, levando em conta o conjunto de instâncias do conjunto original, associado a cada um dos sete conjuntos.

Tabela 7.25 Média aritmética dos resultados de acurácia do processo de 4-validação cruzada usando o NN, para todos os conjuntos de dados, considerando cada instância do conjunto original.

	2009	**2010**	**2011**	**2012**	**2013**	**2014**	**2015**	**MÉDIA**
A	80	85	72	66	53	48	52	65,1
B	84	82	55	57	46	67	57	64
C	94	85	76	72	51	59	91	75,4
D	98	87	81	85	68	70	97	83,7
E	98	90	90	91	80	78	97	89,1
F	98	87	86	85	86	81	95	88,3
G	96	87	88	85	86	83	94	88,4
H	100	90	88	87	83	75	91	87,7
I	100		86	89	83	72		86
J	100		86					93
K	100		86					93
L	100							100

Os atributos contidos em cada instância do conjunto de cada um dos conjuntos usados no cálculo dos resultados da Tabela 7.25, estão listados na Tabela 7.24.

7.2.1.8.3 *Primeiro Experimento* – J48(AD)

A primeira instância do conjunto de dados CAP1_2009_E1, CAP1_2009_E1_1 (A), é descrita pelos atributos DN, EC, S, GI e G. Quando do uso dessa instância do conjunto, o classificador AD gerou a árvore exibida em A, da Figura 7.9, com Tamanho da Árvore = 1 e Número de Folhas = 1. Quando do uso da segunda instância do conjunto de dados, CAP1_2009_E1_2 (B), o classificador gerou a árvore exibida em B da Figura 7.9, com Tamanho da Árvore = 7 e Número de Folhas = 4. Para as dez instâncias restantes

do conjunto i.e., aquelas referenciadas de C a L, o algoritmo gerou a árvore exibida em C da Figura 7.9, com Tamanho da Árvore = 3 e Número de Folhas = 2.

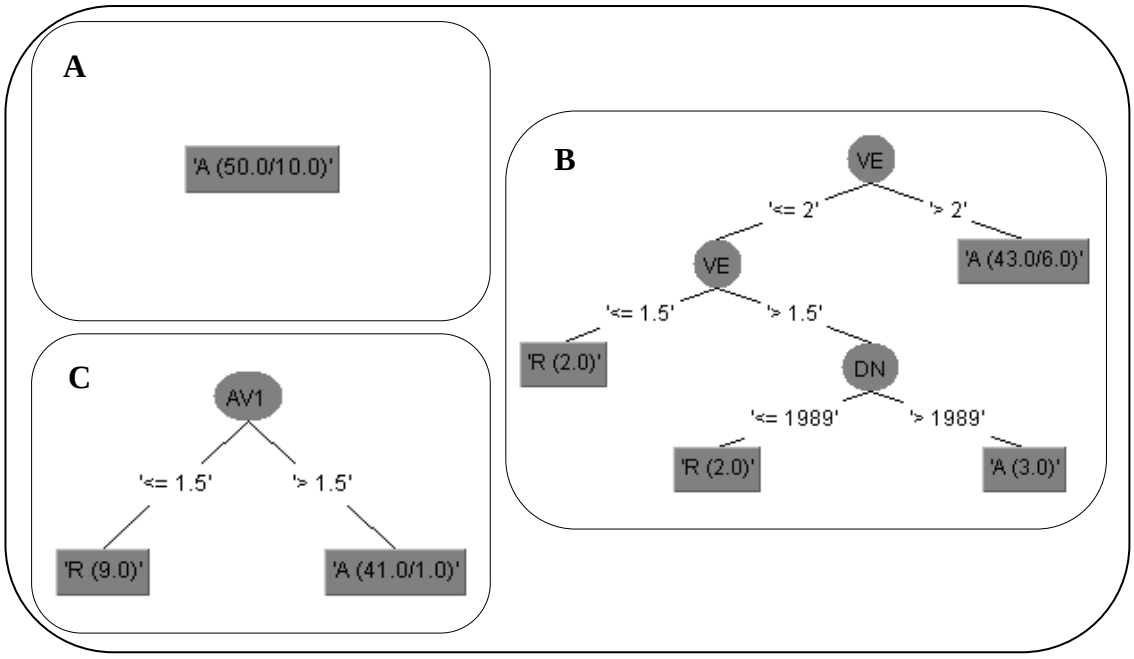

Figura 7.9 Árvores A, B e C geradas a partir do classificador J48, com o conjunto CAP1_2009_E1. A árvore A foi induzida a partir da instância CAP1_2009_E1_1(A), a árvore B foi induzida a partir da instância CAP1_2009_E1_2(B), e a árvore C foi induzida para cada uma das 10 últimas instâncias do conjunto (de C até L).

A Tabela 7.26 mostra a média aritmética da acurácia do AD(J48) dos sete conjuntos de dados, levando em conta o conjunto de instâncias do conjunto original, associado a cada um dos sete conjuntos.

Tabela 7.26 Média aritmética dos resultados de acurácia do processo de 4-validação cruzada usando o AD, para todos os conjuntos de dados, para cada instância do conjunto original.

	2009	**2010**	**2011**	**2012**	**2013**	**2014**	**2015**	**MÉDIA**
A	80	87	72	77	53	72	63	72
B	80	87	72	77	53	67	59	70,7
C	92	90	81	82	78	72	92	83,9
D	92	95	83	91	71	75	88	85
E	92	95	90	89	78	75	88	86,7
F	92	95	90	91	78	75	88	87
G	92	92	90	91	78	75	88	86,6
H	92	90	90	85	78	75	88	85,4
I	92		90	85	78	75		84
J	92		90					91
K	92		90					91
L	92							92

Os atributos contidos em cada instância do conjunto de cada um dos conjuntos usados no cálculo dos resultados da Tabela 7.26, estão listados na Tabela 7.24.

7.2.1.8.4 Resumo dos Resultados do *Primeiro Experimento* – Visão Geral

Quando da análise dos resultados dos três classificadores utilizados, pode se dizer que tais resultados, associados à primeira instância dos 7 conjuntos utilizados (para cada um dos conjuntos, sua respectiva primeira instância é descrita pelo mesmo conjunto de atributos que as demais primeiras instâncias) {DN, EC, S, GI, G}, teve valores relativamente razoáveis, com uma média de 69% de acurácia.

A Tabela 7.27 mostra os valores de acurácia obtidos pelos três classificadores, quando treinados com a primeira instância de cada um dos sete conjuntos de instâncias. Observando os valores de acurácia obtidos nos primeiros quatro anos (2009–2012) nota-se que, tais valores drasticamente caíram em 2013, para os três classificadores, à taxa de aproximadamente 50%, o que torna as suas respectivas contribuições como classificadores, totalmente dispensáveis. Em 2014 teve uma ligeira melhora, associada aos resultados do NB e AD, enquanto que o resultado do NN piorou mais ainda. Já em 2015, quando ambos, NB e AD tiveram novamente uma queda em acurácia, o NN se recuperou, muito embora muito pouco. Os desempenhos bons-razoáveis nos primeiros quatro anos talvez tenha sido provocado pelo desbalanceamento entre valores de atributos, particularmente de S e de EC (algumas vezes).

Tabela 7.27 Resultados do processo de 4-validação cruzada usando os três classificadores (NB, NN, AD), na primeira instância do conjunto {DN, EC, S, GI, G} de todos os conjuntos.

Class./ Cnj.D	**2009**	**2010**	**2011**	**2012**	**2013**	**2014**	**2015**
NB	76%	90%	72%	73%	50%	72%	66%
NN	80%	85%	72%	66%	53%	48%	52%
AD	80%	87%	72%	77%	53%	72%	63%

Cada um dos gráficos da Figura 7.10 traz um panorama da distribuição de valores associados a cada um dos atributos que compõem o grupo de informações pessoais (IP), em relação às classes. O total de dados distribuídos nos sete conjuntos de dados, totaliza 357. Em cada um dos gráficos da Figura 7.10, as 357 instâncias de dados estão

representadas por um dos quatro atributos de IP = {DN, EC, S e GI}, em relação às classes: A(provado) e R(eprovado).

Como comentado anteriormente, cogita-se que o desbalanceamento dos valores de alguns atributos (particularmente S e EC), também levando em conta os valores de classe, possam ter interferido pesadamente nos resultados. Como exemplo do desbalanceamento, o gráfico da Figura 7.10, chamado Atributo Sexo (S), mostra que das 357 instâncias, apenas 26 (7,28%) são do sexo F(eminino) e menos de 50% foram aprovadas.

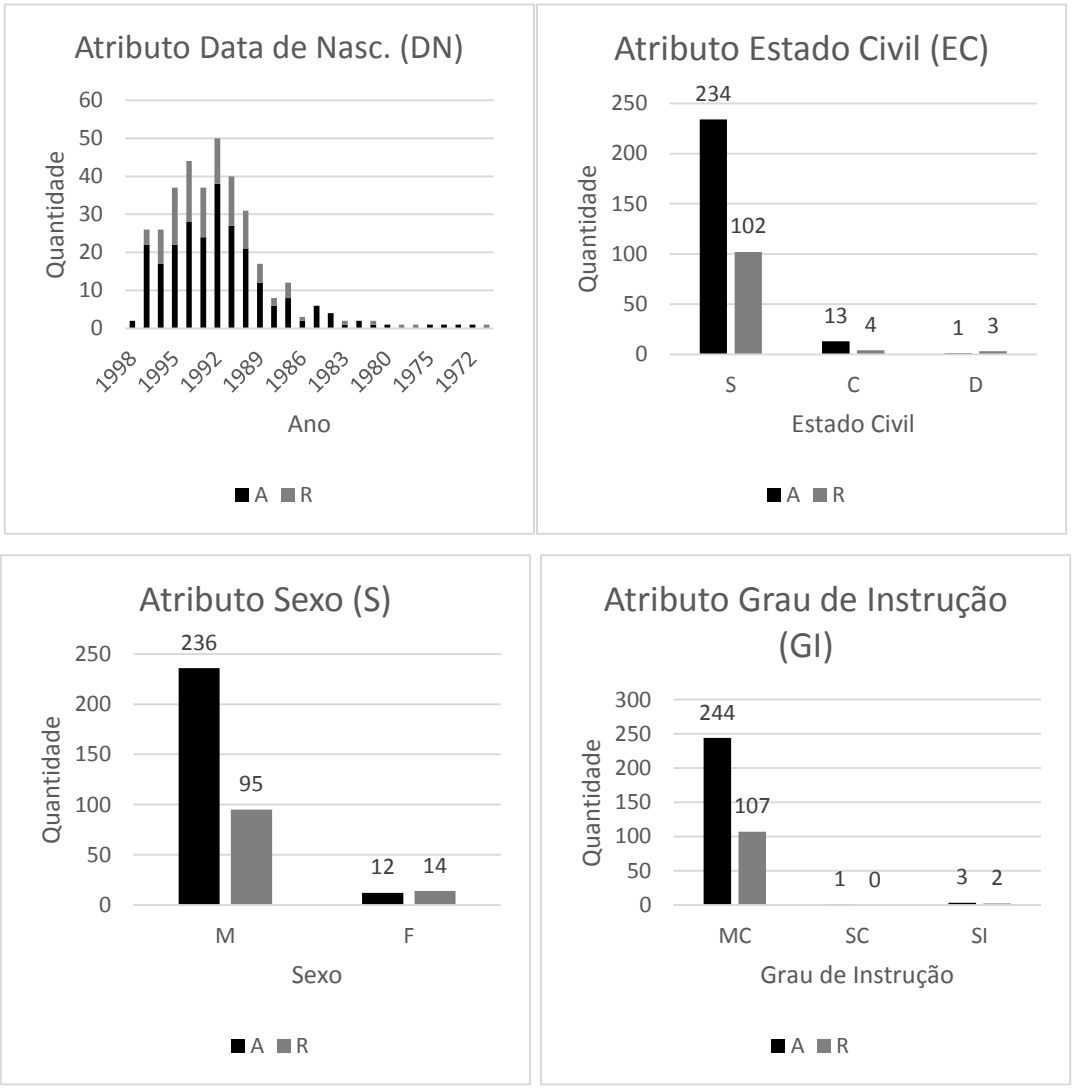

Figura 7.10 Gráficos da distribuição de valores dos atributos que compõem IP(Informações Pessoais) = {DN, EC, S, GI}, considerando as 357 instâncias extraídas de todos os conjuntos de dados.

7.2.2 Segundo Experimento

O experimento descrito nessa seção, identificado como Segundo Experimento, focaliza apenas o aprendizado utilizando árvores de decisão e considerou o maior subconjunto de atributos comuns a todos os sete conjuntos de dados coletados ao longo dos sete anos em que a disciplina CAP1 foi ministrada. Como pode ser visualizado na Tabela 5.3 do Capítulo 5, esse conjunto de atributos é o conjunto {DN, EC, S, GI, VE, AV_1, AV_2, F, LD, G} (note que o atributo AF, referente à nota final da disciplina, não está sendo considerado, uma vez que tal atributo foi discretizado e representado pelo atributo (classe) G, cujos possíveis valores são A(provado) ou R(eprovado)).

Note que os atributos: (1) F (nro. total de faltas) é relevante apenas ao final da disciplina, quando são contabilizados o número de faltas durante o período da disciplina e (2) LD (referente à nota que o aluno obteve na disciplina Lógica e Matemática Discreta), que também é disponibilizada apenas ao final do período letivo. Portanto, não existe interesse em que ambos os atributos participem da indução do conceito porque o objetivo final é ter a possibilidade de antever um fracasso (G=R), antes que ele aconteça.

No experimento descrito nessa seção foi utilizada a metodologia mostrada no diagrama da Figura 7.11. Todos os arquivos de dados envolvidos foram descritos apenas pelos atributos do conjunto {DN, EC, S, GI, VE, AV_1, AV_2, G}. Como discutido anteriormente quando dos experimentos relacionados à seleção de atributos relevantes, os atributos DN, EC, S, GI e VE foram, em alguns dos experimentos, elencados como relevantes devido ao desbalanceamento de classes e/ou de valores de determinados atributos. A decisão metodológica de deixá-los como atributos que descrevem os conjuntos de dados preparados para o Segundo Experimento, foi baseada no conhecimento de que o algoritmo de indução de árvores de decisão já implementa um processo interno de seleção de atributos, direcionado pela medida de entropia. Para melhor organizar o Segundo Experimento, ele foi abordado dividido em três situações distintas (I, II e III), tratadas nas próximas três subseções que seguem.

7.2.2.1 Segundo Experimento (Situação I)

Na situação (I) a metodologia mostrada na Figura 7.11 busca evidenciar se, utilizando os dados pessoais, nota do vestibular e a primeira instância de avaliação, é

possível inferir um modelo robusto que possa ser usado quando da nova oferta da mesma disciplina, no ano seguinte, para predizer o desempenho, logo após a primeira avaliação.

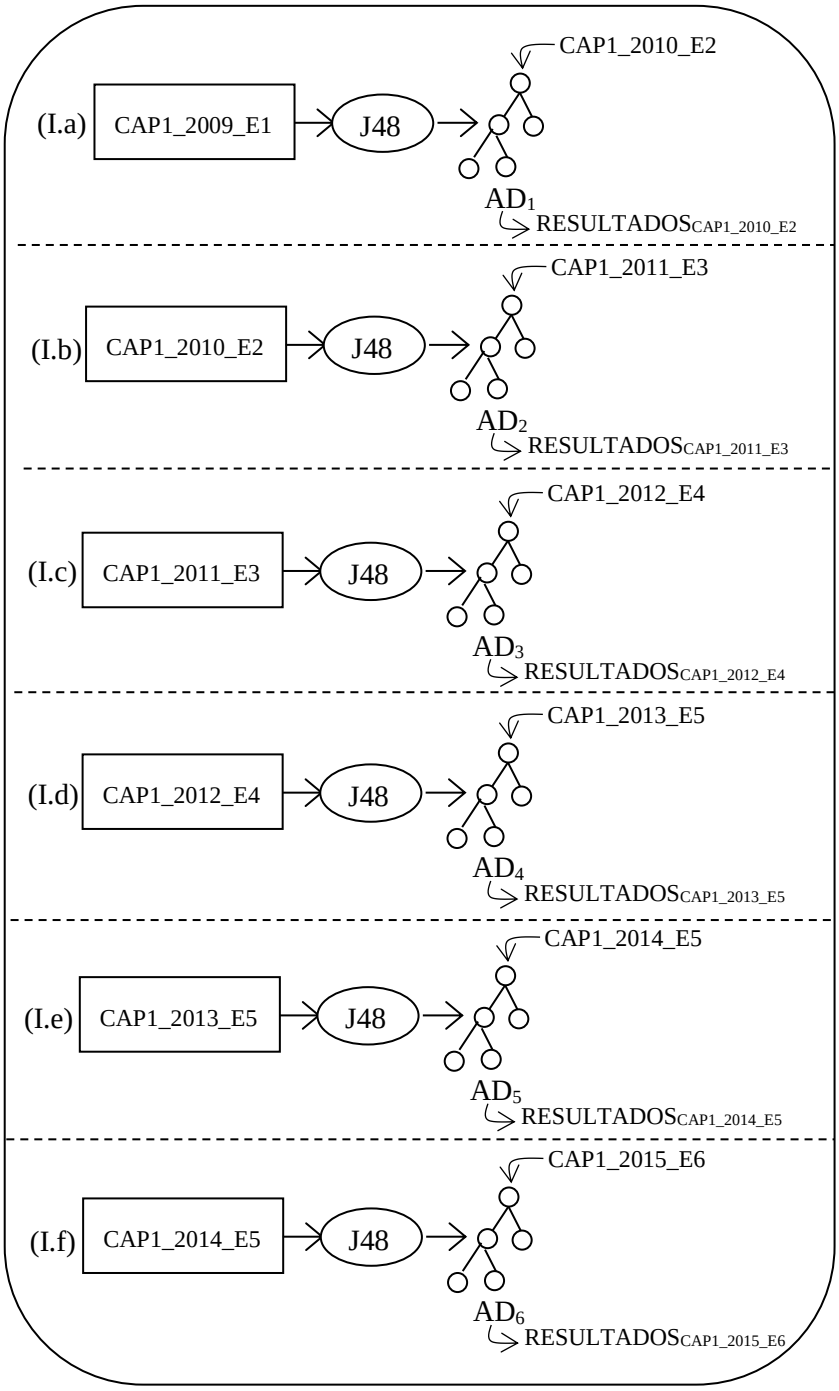

Figura 7.11 Metodologia para o Segundo Experimento (Situação I).

A Tabela 7.28 informa, na sua primeira coluna, o identificador (ID) utilizado no diagrama da Figura 7.11 (de I.a até I.f). Cada linha da tabela, portanto, traz os resultados relacionados à situação (na Figura 7.11) identificada pelo identificador. Por exemplo, a

linha associada ao ID = I.a exibe os resultados gerados pelo classificador induzido usando o conjunto de dados CAP1_2009_E1 (classificador representado pela pequena árvore no diagrama da Figura 7.8), avaliado no conjunto de dados CAP1_2010_E2.

Tabela 7.28 Resultados do processo proposto no diagrama da Figura 7.11, usando o J48 (AD).

ID	*CC(%)*	*CI(%)*	*TP*	*FP*	*TN*	*FN*	*TPR*	*FPR*	*P*	*SR*
I.a	38(92,7)	3(7,3)	37	3	1	0	A=1 R=0,250	A=0,750 R=0	A=0,925 R=1	0,92
I.b	27(62,8)	16(37,2)	15	0	12	16	A=0,484 R=1	A=0 R=0,516	A=1 R=0,429	0,62
I.c	36(63,2)	21(36,8)	23	0	13	21	A=0,523 R=1	A=0 R=0,477	A=1 R=0,382	0,63
I.d	36(60,0)	24(40,)	32	24	4	0	A=1 R=0,143	A=0,857 R=0	A=0,571 R=1	0,60
I.e	31(83,8)	6(16,2)	20	3	11	3	A=0,870 R=0,786	A=0,214 R=0,130	A=0,870 R=0,786	0,83
I.f	62(89,9)	7(10,1)	35	1	27	6	A=0,854 R=0,964	A=0,036 R=0,146	A=0,972 R=0,818	0,89

Como mostra o diagrama do experimento identificado por I.a na Figura 7.11, o conjunto CAP1_2009_E1 é input para o J48 que, com base nos dados do conjunto, induz um classificador (representado pela árvore de decisão nomeada de AD1 no diagrama). O desempenho desse classificador é, então, avaliado usando o conjunto de dados CAP1_2010_E2. O experimento busca evidenciar quão preciso é um classificador induzido a partir de dados de um determinado ano (no caso 2009), para classificar dados do ano seguinte, no caso, 2010. Os demais experimentos da Figura 7.11 (i.e., I.b, I.c, I.d, I.e, I.f) seguem esquema semelhante: o classificador induzido com os dados relativos a um determinado ano é avaliado usando dados do ano seguinte.

A Figura 7.12, mostra os seis classificadores (árvores de AD_1 até AD_6) gerados a partir da metodologia apresentada no diagrama da Figura 7.11. A acurácia na classificação (ou *Success Rate*) da coluna SR da Tabela 7.28, mostra que a melhor classificação corresponde a 92%, está registrada no índice I.a, obtido pelo classificador (AD_1 da Figura 7.12) induzido pelo conjunto de dados CAP1_2009_E1, com 50 instâncias e avaliado pelo conjunto CAP1_2010_E2, com 41 instâncias.

A segunda melhor classificação corresponde a 89% de acurácia na classificação, está registrada no índice I.f da Tabela 7.28, obtido pelo classificador (AD6 da Figura 7.12) induzido pelo conjunto de dados CAP1_2014_E5, com 37 instâncias e avaliado pelo conjunto CAP1_2015_E6, com 69 instâncias. A terceira melhor classificação descrita no índice I.e, com 83% de acurácia na classificação. A sequência de classificações é descrita

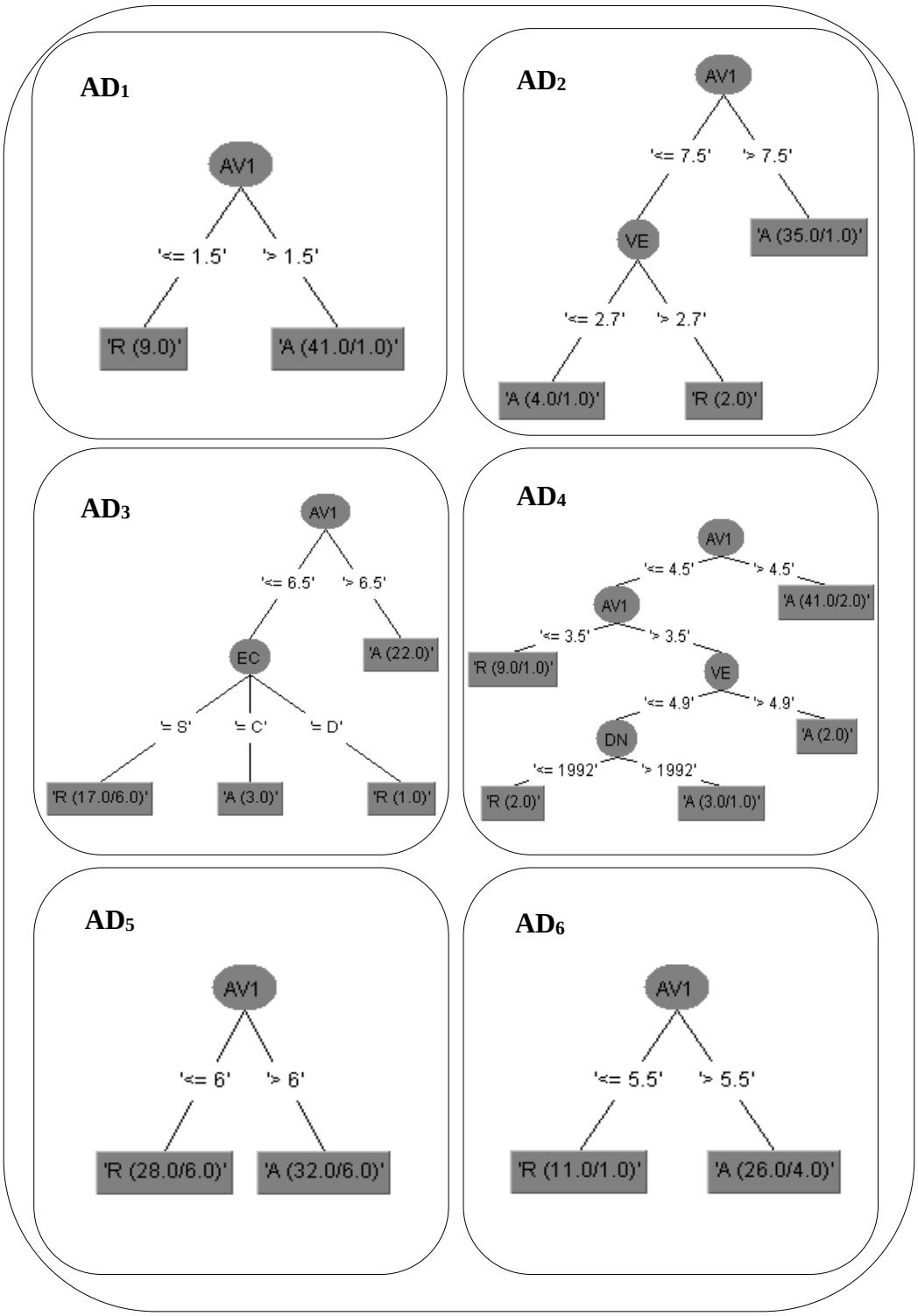

Figura 7.12 Classificadores (árvores de AD_1 até AD_6) gerados a partir da metodologia apresentada para a situação I, na Figura 7.11. Cada classificador foi induzido com os dados relativos a um determinado ano e avaliado usando dados do ano seguinte.

pelos índices I.c (63%), I.b (62%) e I.d (60%). Todos os classificadores gerados alcançaram resultados iguais ou superiores a 60% na classificação do conjunto de instâncias do ano seguinte.

O classificador AD_1 da Figura 7.12, foi induzido pelo conjunto CAP1_2009_E1 que tem 50 instâncias divididas em classes pelo atributo G em A=50 e R=10; O classificador AD_2 foi induzido pelo conjunto CAP1_2010_E2, que tem 41 instâncias divididas em classes pelo atributo G em A=37 e R=4; O classificador AD_3 foi induzido pelo conjunto CAP1_2011_E3, que tem 43 instâncias divididas em classes pelo atributo G em A=31 e R=12; O classificador AD_4 foi induzido pelo conjunto CAP1_2012_E4, que tem 57 instâncias divididas em classes pelo atributo G em A=44 e R=13; O classificador AD_5 foi induzido pelo conjunto CAP1_2013_E4, que tem 60 instâncias divididas em classes pelo atributo G em A=31 e R=12; O classificador AD_6 foi induzido pelo conjunto CAP1_2014_E4, que tem 37 instâncias divididas em classes pelo atributo G em A=23 e R=14. De uma maneira geral pode-se afirmar que os classificadores gerados utilizando os seis atributos, em que apenas um deles é relativo à avaliação, podem colaborar na previsão de desempenho escolar em situações futuras.

7.2.2.2 Segundo Experimento (Situação II)

Na situação (II) a metodologia na Figura 7.13 busca evidenciar se, utilizando os dados pessoais, nota do vestibular e as duas primeiras instâncias de avaliação, é possível inferir um modelo que possa ser usado quando da nova oferta da mesma disciplina, no ano seguinte, para predizer o desempenho, logo após a segunda avaliação.

A Tabela 7.29 informa, na sua primeira coluna, o identificador (ID) utilizado no diagrama da Figura 7.13 (de II.a até II.f). Cada linha da tabela, portanto, traz os resultados relacionados à situação (na Figura 7.13) identificada pelo identificador. Por exemplo, a linha associada ao ID = II.a exibe os resultados gerados pelo classificador induzido usando o conjunto de dados CAP1_2009_E1 (classificador representado pela pequena árvore no diagrama da Figura 7.13), avaliado no conjunto de dados CAP1_2010_E2.

A melhor classificação alcançada pela Situação II foi no experimento II.a, com 92% de classificação correta, em que o classificador foi induzido com o conjunto de dados CAP1_2009_E1, com 50 instâncias e avaliado no conjunto CAP1_2010_E2, com 41 instâncias. A segunda melhor classificação, de 91%, foi obtida no experimento II.f, em que o classificador foi induzido a partir dos dados CAP1_2014_E5, com 37 instâncias e avaliado nas 69 instâncias de CAP1_2015_E6. A terceira melhor classificação aconteceu no experimento II.b, com 83% na acurácia, seguida pelos índices II.e (75%), II.d (63%)

e II.c (49%). Apenas um classificador (II.c) alcançou resultado abaixo de 50% na classificação do conjunto de instâncias do ano seguinte.

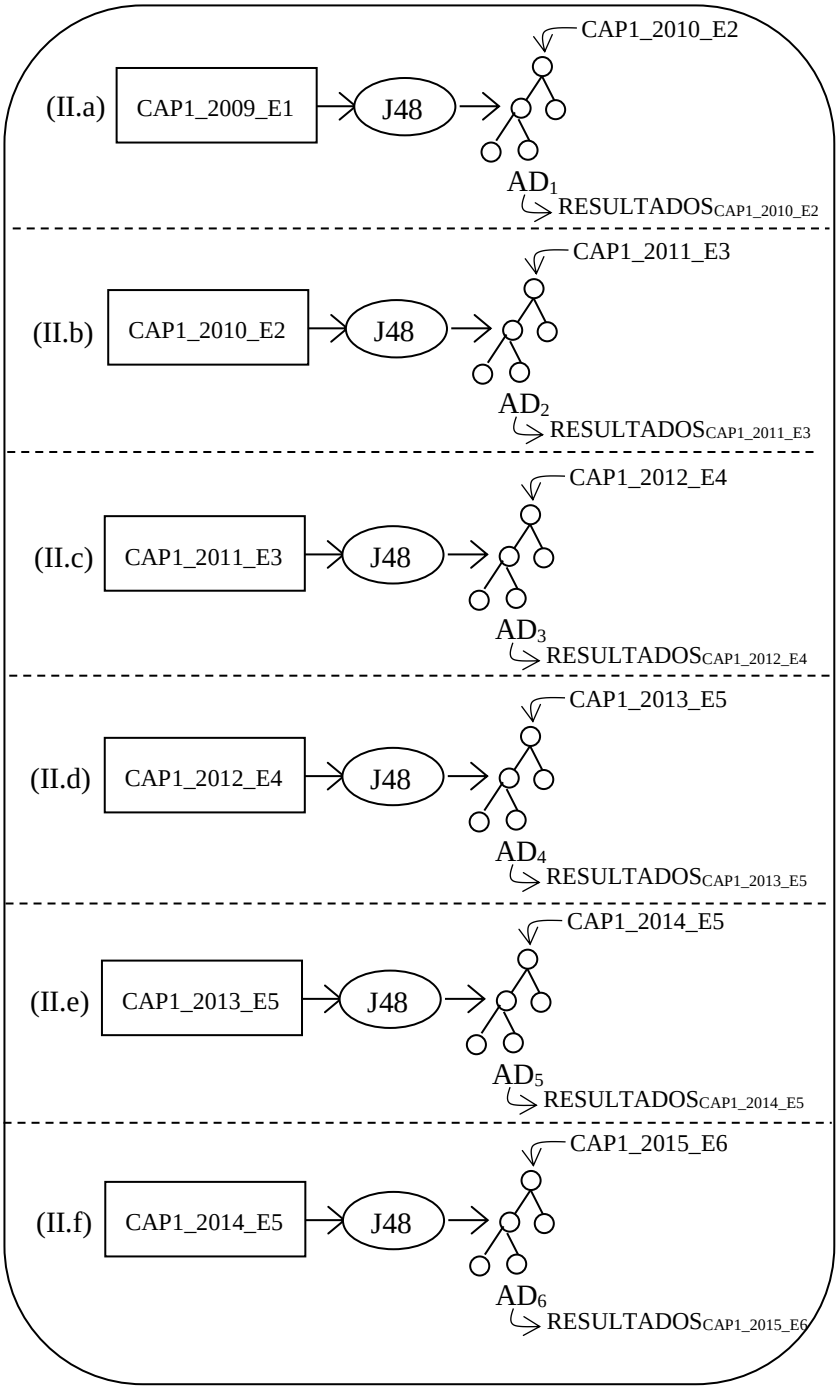

Figura 7.13 Metodologia para o Segundo Experimento (Situação II).

Tabela 7.29 Resultados do processo proposto na Figura 7.13, usando o J48 (AD). O resultado segue o mesmo índice correspondente ao citado na figura, assim como os conjuntos de dados utilizados.

ID	*CC(%)*	*CI(%)*	*TP*	*FP*	*TN*	*FN*	*TPR*	*FPR*	*P*	*SR*
II.a	38(92,7)	3(7,3)	37	3	1	0	A=1 R=0,250	A=0,750 R=0	A=0,925 R=1	0,92
II.b	36(83,7)	7(16,3)	28	4	8	3	A=0,903 R=0,667	A=0,33 R=0,097	A=0,875 R=0,727	0,83
II.c	28(49,1)	29(50)	16	1	12	28	A=0,364 R=0,923	A=0,077 R=0,636	A=0,941 R=0,300	0,49
II.d	38(63,3)	22(36,7)	32	22	6	0	A=1 R=0,214	A=0,786 R=0	A=0,593 R=1	0,63
II.e	28(75,7)	9(24,3)	14	0	14	9	A=0,609 R=1	A=0 R=0,391	A=1 R=0,609	0,75
II.f	63(91,3)	6(8,7)	41	6	22	0	A=1 R=0,786	A=0,214 R=0	A=0,872 R=1	0,91

A Figura 7.14, mostra as seis árvores (de AD_1 até AD_6) geradas a partir da metodologia apresentada na Figura 7.13. Quatro das seis árvores de decisão induzida nos seis experimentos usam apenas um dos dois atributos relacionados à avaliação: A AD_1 usa apenas o AV_1 e as AD_2, AD_3 e AD_4 usam apenas o atributo AV_2. Já as árvores AD_4 e AD_6 usam ambas as avaliações, AV_1 e Av_2, para caracterizar o classificador.

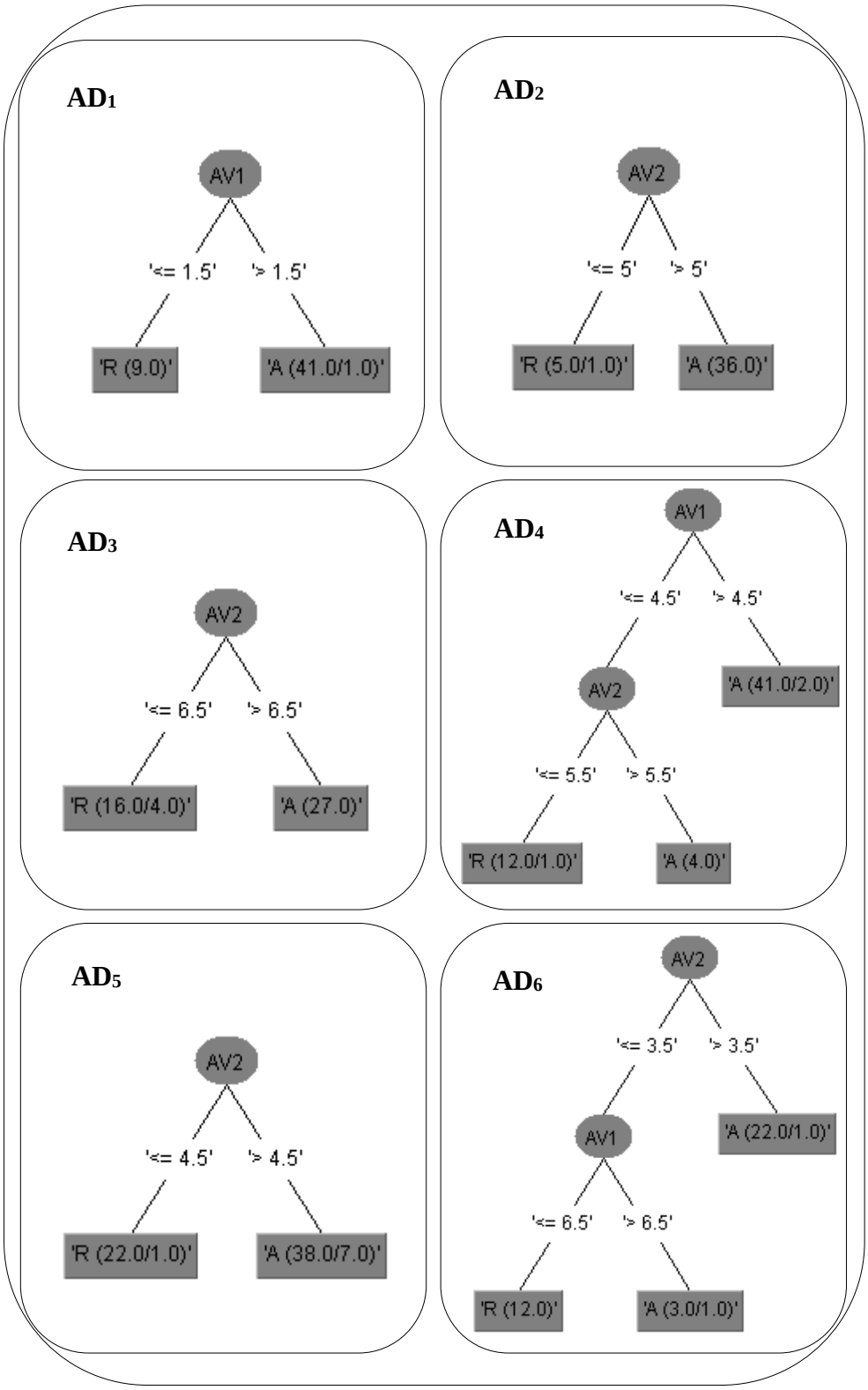

Figura 7.14 Classificadores (árvores de AD_1 até AD_6) gerados a partir da metodologia apresentada para a situação II, na Figura 7.13. Cada classificador foi induzido com os dados relativos a um determinado ano e avaliado usando dados do ano seguinte.

7.2.2.3 Segundo Experimento (Situação III)

Na situação (III) a metodologia utilizada foi a mostrada na Figura 7.15. O experimento associado à situação III busca evidenciar se, utilizando os dados pessoais, nota do vestibular e as duas primeiras instâncias de avaliação, considerando todos os dados históricos relativo à disciplina, é possível inferir um modelo que possa ser usado quando da nova oferta da mesma disciplina, no ano seguinte, para predizer o desempenho, logo após a segunda avaliação.

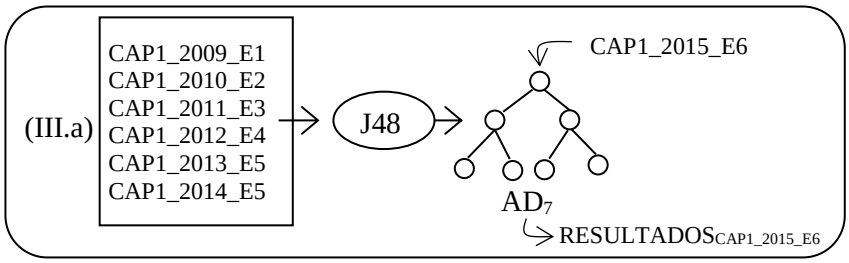

Figura 7.15 Metodologia para o Segundo Experimento (Situação III).

A Tabela 7.30 mostra o resultado da avaliação realizada pelo conjunto de dados CAP1_2015_E6, no classificador induzido por um conjunto maior de dados com 288 instâncias gerado a partir dos 6 conjuntos de dados, cronologicamente seus anteriores i.e., CAP1_2009_E1 até CAP1_2014_E5. Esse experimento busca investigar, nos dados disponibilizados, o quanto dados históricos são relevantes para refletir uma situação atual e, com base nos valores obtidos de CC%, pode-se dizer que dados acumulados de instâncias anteriores de uma mesma disciplina, desde que customizadas de maneira a serem descritas pelo menos conjunto de atributos, podem colaborar na previsão de desempenho em situações futuras.

Tabela 7.30 Resultado do processo proposto no diagrama da Figura 7.15, usando o J48 (AD). O resultado apresentado na tabela, assim como os conjuntos de dados utilizados correspondem aos citados na Figura 7.15.

ID	*CC(%)*	*CI(%)*	*TP*	*FP*	*TN*	*FN*	*TPR*	*FPR*	*P*	*SR*
III.a	61(88,4)	8(11,6)	34	1	27	7	A=0,829 R=0,964	A=0,036 R=0,171	A=0,971 R=0,794	0,88

A Figura 7.16, mostra a árvore (AD_7) gerada a partir da metodologia apresentada no diagrama da Figura 7.15, usando os dados de 6, dos 7 conjuntos de dados para induzir o classificador (do CAP1_2009_E1 até CAP1_2014_E5). O modelo foi avaliado pelo conjunto de dados CAP1_2015_E6, o resultado está na Tabela 7.30.

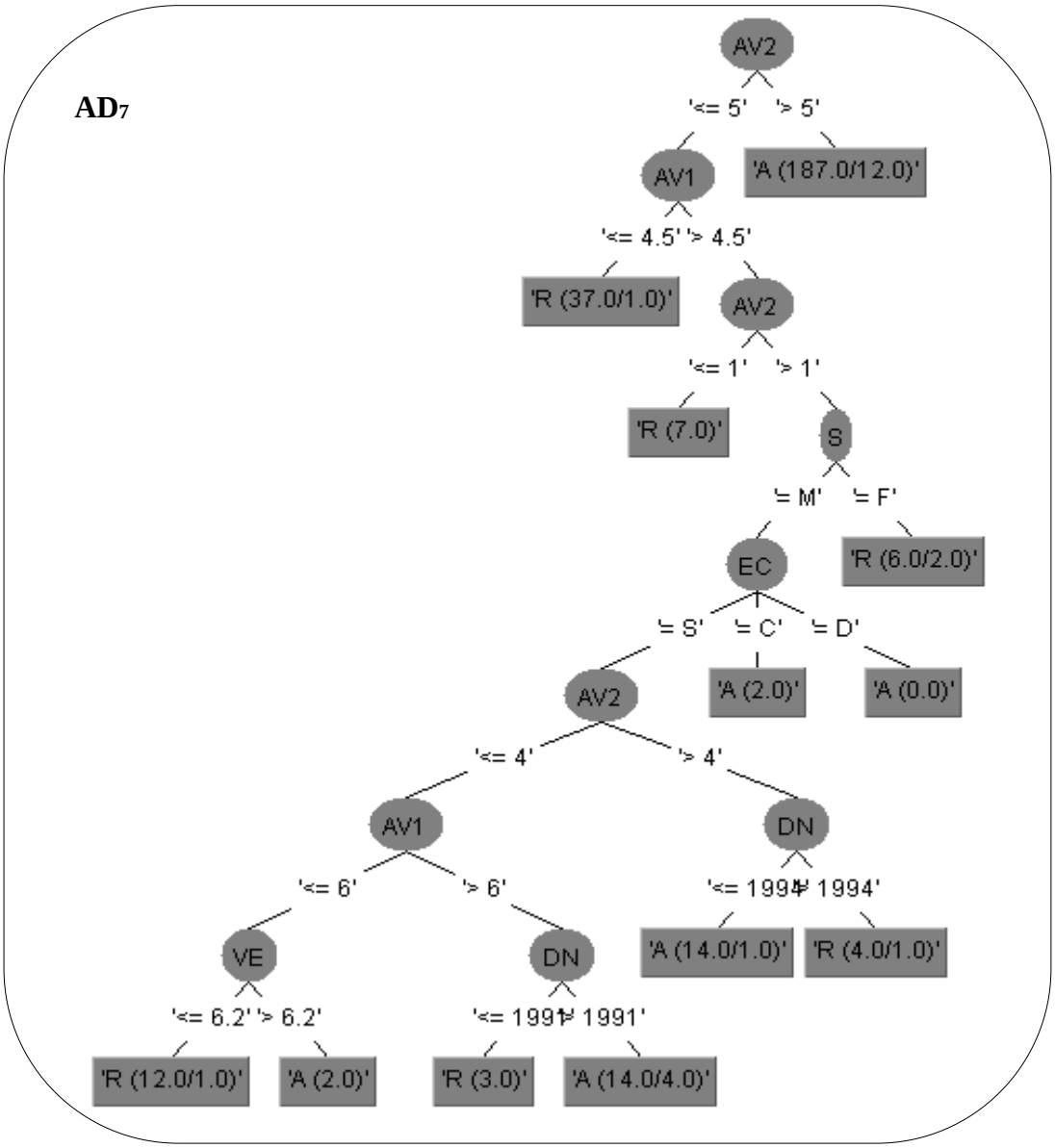

Figura 7.16 Classificador (árvore AD_7) gerado a partir da metodologia apresentada para a situação III, na Figura 7.15. Esta árvore foi induzida a partir dos dados dos 6 conjuntos (de CAP1_2009_E1 até CAP1_2014_E5) anteriores a CAP1_2015_E6.

7.2.2.4 Resumo e Análise dos Resultados do Segundo Experimento

7.2.2.4.1 Segundo Experimento (Situação I)

A Tabela 7.28 contém os resultados da Situação I do Segundo Experimento, que utiliza o conjunto de atributos {DN, EC, S, GI, VE, AV_1, G} para os diferentes conjuntos de dados.

A informação anotada no ID, I.a da Tabela7.28, corresponde aos resultados gerados pela classificação do conjunto de dados CAP1_2010_E2, que obteve o maior valor de

correta classificação (92%) se comparado aos 6 outros conjuntos. Das 41 instâncias presentes no conjunto (CAP1_2010_E2) 38 foram corretamente classificadas e apenas três foram classificadas incorretamente. O classificador induzido pelo conjunto do ano anterior (CAP1_2009_E1) aparece na Figura 7.12 como AD_1, possui apenas um nó (o atributo o AV_1) e apresenta a seguinte divisão: Se AV1 > 1,5 então "A", senão "R". A Tabela 3.5 mostra os métodos de avaliação para se alcançar o resultado final, o conjunto CAP1_2009_E1 utiliza o seguinte método: AF = ((AV_1*0,27) + (AV_2*0,31) + (AV_3*0,36) + ((LE_1+LE_2+LE_3+LE_4) / 4*0,06)); dos atributos que são usados na fórmula para a Avaliação Final, somente o atributo AV_1 comparece no conjunto deste experimento, sendo assim, está é a razão para que o algoritmo tenha escolhido apenas este atributo para induzir o classificador.

O segundo melhor resultado foi alcançado pela classificação do conjunto CAP1_2015_E6, com 89% de acurácia, como mostra a Tabela 7.28, no ID, I.f. O classificador utilizado nesta classificação foi o AD_6 que aparece na Figura 7.12, induzido pelo conjunto CAP1_2014_E5. Este classificador também usa apenas um atributo o AV_1 na seguinte condição: Se AV_1 > 5,5 então "A", senão "R".

O terceiro melhor resultado foi obtido pelo classificador AD_5, o resultado está na tabela 7.28, no ID I.e. O classificador foi induzido pelo conjunto CAP1_2013_E5 e avaliado pelo conjunto CAP1_2014_E5, obtendo 83% de correta classificação, o terceiro melhor resultado. O classificador escolheu apenas o atributo AV1 que foi subdividido em: Se AV1 > 6,0 então "A", senão "R".

Dos seis classificadores apresentados na Figura 7.12, os três citados acima com as melhores classificações usaram apenas o atributo AV_1, para obter os resultados. Os demais classificadores (AD_2, AD_3 e AD_4) da Figura 7.12 serão comentados a seguir.

O classificador AD_2, da Figura 7.12, induzido pelo conjunto CAP1_2010_E2 e avaliado pelo conjunto do próximo ano (CAP1_2011_E3) classificou corretamente 62% das instâncias (é o 5º melhor classificador em 6, neste experimento). Este classificador usa dois atributos para formar a árvore, o AV_1 e o VE. O AV_1 se divide assim: Se AV1 > 7,5 então "A", senão Se VE > 2,7 então "R", senão "A". Verificados os dados do conjunto CAP1_2010_E2, que induziu este classificador é possível observar que para o atributo VE das 4 instâncias reprovadas 50% obteve 1,0 ponto na nota do vestibular e 50% > 1,0 ponto, das 37 instâncias aprovadas 48% obtiveram valor igual a 1,0 ponto (18 instâncias), 18% valor entre 2,0 e 2,9 e 32% obtiveram valor maior que 3,0. Sobre o atributo VE

(vestibular) o estudante pode fazer o vestibular nacional da instituição composto questões da avaliação mais redação, usar a nota do ENEM, ou fazer o vestibular continuado composto apenas pela redação. E sobre o preenchimento das vagas, como citado no capítulo 5, na seção 5.3.1 "O ingresso se dá pela classificação de notas da maior para a menor e pelo preenchimento da quantidade de vagas disponíveis no curso". Neste conjunto de dados ocorreu que grande parte dos alunos aprovados na disciplina tiraram nota muito baixa no vestibular.

O classificador AD_3 usa dois dos atributos em sua árvore que são: AV_1 e EC. Que se dividem assim: Se $AV_1 > 6,5$ então "A" senão Se EC = S então "R", Se EC = C então "A", Se EC = D então "R". O conjunto de dados que induziu este classificador é o CAP1_2011_E3. O atributo EC é dividido em C(asado), S(olteiro) e D(ivorciado), das 43 instâncias do conjunto: uma é igual a D, que possui 100% das instâncias R(eprovadas); cinco são iguais a C, que possui 100% das instâncias A(provadas) e as trinta e sete instâncias restantes são iguais a S dessas 70% são A(provadas) o que corresponde a 26 instâncias e 30% são R(eprovadas) igual a 11 instâncias.

O último classificador AD_4 induzido por CAP1_2012_E4 e avaliado por CAP1_2013_E5, obteve 60% de acurácia como mostra a Tabela 7.28, no ID I.d, o menor desempenho deste experimento. Este classificador usou três atributos em sua árvore, AV_1, VE e DN.

Os seis classificadores da Situação I, alcançaram bons índices de classificação com variação entre 60 e 92% de acurácia, o que significa que o conjunto do ano anterior com os 6 atributos usados nesta situação pode ajudar a prever o desempenho do ano seguinte com boa margem de acerto.

7.2.2.4.2 Segundo Experimento (Situação II)

A Situação II seguiu a mesma metodologia da Situação I, apenas adicionando ao conjunto de atributos o atributo AV_2. O classificador AD1, induzido pelo conjunto CAP1_2009_E1 e avaliado pelo conjunto CAP1_2010_E2, obteve o melhor desempenho com 92% de acurácia na classificação das instâncias. O classificador AD1 da Figura 7.14, usou apenas o atributo da primeira avaliação AV_1 da seguinte maneira: Se $AV_1 > 1.5$ então "A", senão "R". Como este classificador foi induzido pelo conjunto do ano de 2009, se os dados forem analisados é possível perceber que as notas de AV_1 foram notas altas, são 19 instâncias com nota 10, outras 16 instâncias com notas entre 9,0 e 9,9, sendo assim

as notas abaixo de 1,5 provavelmente em sua maioria são de instâncias que foram reprovadas ao final da disciplina.

O segundo melhor desempenho com 91% das instâncias corretamente classificadas foi alcançado pelo classificador AD_6 presente na Figura 7.14, induzido pelo conjunto CAP1_2014_E5 e avaliado pelo conjunto CAP1_2015_E6. O resultado está anotado na Tabela 7.29 no ID II.f. Este classificador usou os dois atributos de avaliação presentes no conjunto: AV_1 e AV_2, da seguinte maneira: Se $AV_2 > 3,5$ então "A", senão Se $AV_1 > 6,5$ então "A", senão "R", para montar a árvore. Como mostra a Tabela 3.5 na fórmula da avaliação final (AF = ((AV_1*2) + (AV_2*2) + (AV_3*1))/5 * 0,9 + PI) do conjunto CAP1_2014_E5 que induziu o classificador, as avaliações AV_1 e AV_2 tem peso igual e ambas são usadas na árvore com sucesso, para o auxílio da predição do desempenho do próximo ano.

O classificador AD_2, induzido pelo conjunto CAP1_2010_E2, e avaliado pelo conjunto do ano seguinte, classificou 83% das instâncias corretamente, o terceiro melhor classificador deste experimento. É possível visualizar este resultado na Tabela 7.29, no ID II.b, e o classificador na Figura 7.14, é a árvore AD_2. Este classificador usou apenas o atributo da segunda avaliação (AV_2) e se comporta da seguinte maneira: Se $AV_2 > 5$ então "A", senão "R".

Da Tabela 7.29, o ID II.e traz o resultado da classificação do conjunto CAP1_2014_E5, com 75% de correta classificação. O classificador usado foi induzido pelo conjunto do ano imediatamente anterior (CAP1_2013_E5) que está representado na Figura 7.14, como a árvore AD_5. A árvore é composta apenas pelo atributo da avaliação 2 (AV_2) da seguinte maneira: Se $AV_2 > 4,5$ então "A", senão "R".

Na Tabela 7.29, no ID II.d, é possível ver o resultado da classificação do conjunto de dados CAP1_2013_E5, que alcançou 63% na acurácia. Este resultado foi possível usando o classificador AD_4, induzido pelo conjunto CAP1_2012_E4 presente na Figura 7.14. O classificador AD_4, escolheu os dois atributos de avaliação (AV_1 e AV_2) para compor a sua árvore. Se as informações da Tabela 3.5 forem observadas a respeito da fórmula de avaliação aplicada ao conjunto de dados do ano de 2012 (AF = ((AV_1*30)+(AV_2*30)+(AV_3*30)+(AV_4*10))/100), é possível perceber que apesar de serem as duas únicas avaliações presentes no nosso conjunto de dados, as duas avaliações

usadas na árvore AD_4 estão entre as três avaliações com maior peso. A árvore trabalha da seguinte maneira: Se AV1 > 4,5 então "A", senão Se AV2 > 5,5 então "A", senão "R".

O classificador AD_3, da Figura 7.14 escolheu apenas o atributo da segunda avaliação (AV_2) para classificar as instâncias. Este classificador foi induzido pelo conjunto de dados CAP1_2011_E3 e foi avaliado pelo conjunto CAP1_2012_E4, onde alcançou 49% de correta classificação, sendo o classificador com o menor índice de acurácia deste experimento.

7.2.2.4.3 Segundo Experimento (Situação III)

A terceira situação do segundo experimento, usou um conjunto robusto de instâncias (se comparado aos demais experimentos). Dos sete conjuntos de dados, seis foram usados para formar este grande conjunto com 288 instâncias, e um conjunto de 69 instâncias foi utilizado para avaliar o classificador. Foram escolhidos o maior número de atributos comuns aos conjuntos de dados usados nesta pesquisa que contempla informações pessoais, nota do vestibular e duas instâncias de avaliação {DN, EC, S, GI, VE, AV_1, AV_2, G}.

O classificador induzido pelo grande conjunto de instâncias está na Figura 7.16, como AD_7. O atributo raiz escolhido foi a segunda avaliação AV_2, que volta a aparecer em mais dois nós da árvore. Dos sete atributos (excluída a classe = G), foram usados seis atributos na árvore do classificador AD_7 {AV2, AV1, S, EC, DN, VE}. Apenas o atributo GI não foi usado para filtrar a classe à qual a instância poderia pertencer. Os atributos AV_1, DN, aparecem duas vezes subdividindo a árvore. Os atributos S, EC e VE aparecem apenas uma vez na árvore. A árvore AD7 tem maior complexidade e tamanho (Tamanho = 22, destes 12 são folhas) se comparada às demais árvores geradas para as situações I e II. O conjunto CAP1_2015_E6 usado para avaliar o classificador AD_7 alcançou 88% na acurácia da classificação. Entretanto a árvore AD6 da Figura 7.14, induzida apenas pelo conjunto CAP1_2014_E5, que também foi avaliada pelo conjunto de dados CAP1_2015_E6, com complexidade inferior e menor tamanho (Tamanho = 5, destes 3 são folhas) conseguiu classificar porcentagem pouco maior de instâncias (91%) corretamente. O que mostra que a complexidade do classificador e o tamanho não são indicadores de bom desempenho e também indica que o conjunto de informações do ano anterior é capaz de predizer com certa eficiência o desempenho do ano atual.

7.2.3 Terceiro Experimento

Os experimentos descritos nesta seção exploram os resultados obtidos com os processos de seleção de atributos relevantes, apresentados e discutidos no Capítulo 6, tendo em vista, também, a possibilidade de antever o desempenho acadêmico final, durante o andamento da disciplina. Inicialmente, os resultados relativos à seleção de atributos relevantes e os resultados dos ranqueamentos, para cada conjunto de dados, foi considerado, com vistas a arbitrar um conjunto mínimo de atributos com potencial para a indução de um modelo robusto.

A metodologia adotada para a realização do *Terceiro Experimento* com relação a atributos foi:

(1) caracterizar como relevante o atributo que comparecesse em pelo menos 7 dos 12 conjuntos obtidos (por conjunto de dados – Seção 6.2), como mostrado na Tabela 6.36;

(2) considerando os experimentos de ranqueamento, escolher os atributos comuns entre os sete primeiros ranqueados pelos três métodos considerados (*Chi*, *GainR* e *InfoGain*), como mostra a Tabela 6.37;

(3) considerando os dois conjuntos de atributos construídos (1) e (2) (reunidos na Tabela 7.31), determinar sua interseção e, no conjunto resultante, escolher aqueles atributos que, temporalmente, tenham seus respectivos valores determinados o quanto antes, durante o andamento da disciplina (como mostra a Tabela 7.32).

Tabela 7.31 Conjuntos de atributos relevantes extraídos via seleção de subconjuntos e via ranqueamento.

Conjunto de dados	Subconjunto de atributos relevantes (S)	Atributos comuns dentre os sete primeiros, nos 3 ranqueamentos (R)
CAP1_2009_E1	S,AV_1,AV_2,AV_3	AV_1,AV_2,AV_3,FC,S,GI,EC
CAP1_2010_E2	S,AV_1,AV_2,AV_3,LD,FC	AV_2,FC,LD,AV_1,AV_3,GI,S
CAP1_2011_E3	EC,S,AV_1,AV_2,AS_1	AV_2,AV_1,AS_1,FC,EC,S,GI
CAP1_2012_E4	EC,AV_1,AV_2,AV_3,AV_4,LD,FC	AV_1,AV_3,AV_2,LD,AV_4,EC
CAP1_2013_E5	AV_1,AV_2,AV_3,PI	AV_2,AV_3,AV_1,PI,LD,S,EC
CAP1_2014_E5	S,PI,AV_1,AV_2,FC	AV_2,AV_1,FC,PI,LD,S,EC
CAP1_2015_E6	DN, AV_1,AV_2,PI,LP,	AV_2,AV_1,LP,PI,DN,LD

Das três tabelas a seguir cada uma apresenta os resultados obtidos com o processo de aprendizado automático (4-validação cruzada) usando os algoritmos: NB (Tabela 7.33), NN (Tabela 7.34) e J48(AD) (Tabela 7.35), respectivamente. Os arquivos de dados usados no aprendizado são os sete arquivos de dados da Tabela 7.32, mostrados na

primeira coluna da tabela, cada um deles descrito apenas com aqueles atributos identificados na coluna Redução (temporalidade) associada à linha em questão.

Tabela 7.32 Conjuntos de atributos que representam a implementação do passo (3) da heurística empregada para escolha do conjunto de atributos para o Experimento 3.

ID	Conjunto de dados	S ∩ R	Redução (temporalidade)
CAP1	CAP1_2009_E1	S,AV_1,AV_2,AV_3	S,AV_1
CAP2	CAP1_2010_E2	S,AV_1,AV_2,AV_3,LD,FC	S,AV_1,AV_2
CAP3	CAP1_2011_E3	EC,S,AV_1,AV_2,AS_1	S,EC,AV_1,AV_2
CAP4	CAP1_2012_E4	EC,AV_1,AV_2,AV_3,AV_4,LD	EC,AV_1,AV_2
CAP5	CAP1_2013_E5	AV_1,AV_2,AV_3,PI	AV_1,AV_2
CAP6	CAP1_2014_E5	S,PI,AV_1,AV_2,FC	S,PI,AV_1
CAP7	CAP1_2015_E6	DN,AV_1,AV_2,PI,LP	DN,AV_1

Tabela 7.33 Resultados do processo de 4-validação cruzada usando o NB, para os conjuntos de dados da Tabela 7.32, descritos pelos atributos mostrados na coluna Redução (temporalidade).

ID	*CC(%)*	*CI(%)*	*TP*	*FP*	*TN*	*FN*	*TPR*	*FPR*	*P*	*SR*
CAP1	48(96,0)	2(4,0)	39	1	9	1	A=0,975 R=0,900	A=0,100 R=0,025	A=0,975 R=0,900	0,96
CAP2	40(97,6)	1(2,4)	37	1	3	0	A=1,000 R=0,750	A=0,250 R=0,000	A=0,974 R=1,000	0,97
CAP3	39(90,7)	4(9,3)	28	1	11	3	A=0,903 R=0,917	A=0,083 R=0,097	A=0,966 R=0,786	0,90
CAP4	54(94,7)	3(5,3)	43	2	11	1	A=0,977 R=0,846	A=0,154 R=0,023	A=0,956 R=0,917	0,94
CAP5	50(83,3)	10(16,7)	27	5	23	5	A=0,844 R=0,821	A=0,179 R=0,156	A=0,844 R=0,821	0,83
CAP6	34(91,9)	3(8,1)	22	2	12	1	A=0,957 R=0,857	A=0,143 R=0,043	A=0,917 R=0,923	0,91
CAP7	63(91,3)	6(8,7)	38	3	25	3	A=0,927 R=0,893	A=0,107 R=0,073	A=0,927 R=0,893	0,91

Tabela 7.34 Resultados do processo de 4-validação cruzada usando o NN, para os conjuntos de dados da Tabela 7.32, descritos pelos atributos mostrados na coluna Redução (temporalidade).

ID	*CC(%)*	*CI(%)*	*TP*	*FP*	*TN*	*FN*	*TPR*	*FPR*	*P*	*SR*
CAP1	48(96,0)	2(4,0)	39	1	9	1	A=0,975 R=0,900	A=0,100 R=0,025	A=0,975 R=0,900	0,96
CAP2	38(92,7)	3,(7,3)	37	3	1	0	A=1,000 R=0,250	A=0,750 R=0,000	A=0,925 R=1,000	0,92
CAP3	39(90,7)	4(9,3)	30	3	9	1	A=0,968 R=0,750	A=0,250 R=0,032	A=0,909 R=0,900	0,90
CAP4	52(91,2)	5(8,8)	42	3	10	2	A=0,955 R=0,769	A=0,231 R=0,045	A=0,933 R=0,833	0,91
CAP5	47(78,3)	13(21,7)	25	6	22	7	A=0,781 R=0,783	A=0,214 R=0,219	A=0,806 R=0,759	0,78
CAP6	30(81,0)	7(19,0)	19	3	11	4	A=0,826 R=0,786	A=0,214 R=0,174	A=0,864 R=0,733	0,81
CAP7	63(91,3)	6(8,7)	39	4	24	2	A=0,951 R=0,857	A=0,143 R=0,049	A=0,907 R=0,923	0,91

Tabela 7.35 Resultados do processo de 4-validação cruzada usando o J48 (AD), para os conjuntos de dados da Tabela 7.32 descritos pelos atributos mostrados na coluna Redução (temporalidade).

ID	*CC(%)*	*CI(%)*	*TP*	*FP*	*TN*	*FN*	*TPR*	*FPR*	*P*	*SR*
CAP1	46(92,0)	4(8,0)	39	3	7	1	A=0,975 R=0,700	A=0,300 R=0,025	A=0,929 R=0,875	0,92
CAP2	39(95,1)	2(4,9)	36	1	3	1	A=0,973 R=0,750	A=0,250 R=0,027	A=0,973 R=0,750	0,95
CAP3	37(86,0)	6(14,0)	27	2	10	4	A=0,871 R=0,833	A=0,167 R=0,129	A=0,931 R=0,714	0,86
CAP4	52(91,2)	5(8,8)	41	2	11	3	A=0,932 R=0,846	A=0,154 R=0,068	A=0,953 R=0,786	0,91
CAP5	43(71,7)	17(28,4)	22	7	21	10	A=0,688 R=0,750	A=0,250 R=0,313	A=0,759 R=0,677	0,71
CAP6	28(75,7)	9(24,3)	21	7	7	2	A=0,913 R=0,500	A=0,500 R=0,087	A=0,750 R=0,778	0,75
CAP7	65(94,2)	4(5,8)	40	3	25	1	A=0,976 R=0,893	A=0,107 R=0,024	A=0,930 R=0,962	0,94

Os sete conjuntos de dados utilizados no *Terceiro Experimento* variam entre si com relação ao número de instâncias de dados e aos atributos usados na descrição de cada instância. A Tabela 7.36 apresenta um resumo das classificações realizadas nas tabelas 7.33, 7.34 e 7.35, com apenas a informação da acurácia que cada conjunto alcançou quando avaliado em determinado classificador. Na Tabela 7.36, o asterisco (*) indica o maior valor de SR obtido por um (ou mais) dos três algoritmos, associado a cada um dos sete conjuntos considerados.

Tabela 7.36 Taxa de classificação correta da coluna SR (*Success Rate*), das tabelas 7.33, 7.34 e 7.35. $^{(*)}$ caracteriza o melhor resultado.

ID	*NB*	*NN*	*AD*
CAP1	$0,96^{(*)}$	$0,96^{(*)}$	0,92
CAP2	$0,97^{(*)}$	0,92	0,95
CAP3	$0,90^{(*)}$	$0,90^{(*)}$	0,86
CAP4	$0,94^{(*)}$	0,91	0,91
CAP5	$0,83^{(*)}$	0,78	0,71
CAP6	$0,91^{(*)}$	0,81	0,75
CAP7	0,91	0,91	$0,94^{(*)}$

O resultado de SR mais próximo de 100%, foi aquele obtido pelo NB, usando o conjunto CAP2, com 97% de acurácia na classificação. A Tabela 7.37 representa uma compilação das colunas SR participantes das tabelas 7.33, 7.34 e 7.35, com os resultados obtidos quando do uso do conjunto original (CO) e do conjunto reduzido (CR). Os resultados do conjunto original foram extraídos da Seção 7.2.1 *Primeiro Experimento*, onde os sete conjuntos de dados estão divididos em subconjuntos (ver Subseção 7.2.1.1,

onde a divisão do primeiro conjunto, CAP1_2009_E1, é detalhada). Cada subseção das sete primeiras subseções da Seção 7.2.1 (isto é, subseções de 7.2.1.1 a 7.2.1.7) trata de um dos conjuntos de dados. Em cada uma dessas subseções, o conjunto original (CO, isto é, descrito por todos os atributos) é tratado no último experimento descrito na respectiva subseção). Os resultados do conjunto reduzido de dados foram extraídos da Tabela 7.36, apresentada nesta subseção (i.e., 7.2.3 Terceiro Experimento).

Tabela 7.37 Taxa de correta classificação da coluna SR (*Success Rate*), do conjunto original (CO) e do conjunto reduzido (CR) (*) caracteriza o melhor resultado de cada classificador.

ID	*NB*		*NN*		*AD*	
	CO	*CR*	*CO*	*CR*	*CO*	*CR*
CAP1_2009_E1	0,98(*)	0,96	1,00(*)	0,96	0,92(*)	0,92(*)
CAP1_2010_E2	0,95	0,97(*)	0,90	0,92(*)	0,90	0,95(*)
CAP1_2011_E3	0,83	0,90(*)	0,86	0,90(*)	0,90(*)	0,86
CAP1_2012_E4	0,82	0,94(*)	0,89	0,91(*)	0,85	0,91(*)
CAP1_2013_E5	0,88(*)	0,83	0,83(*)	0,78	0,78(*)	0,71
CAP1_2014_E5	0,86	0,91(*)	0,72	0,81(*)	0,75(*)	0,75(*)
CAP1_2015_E6	0,92(*)	0,91	0,91(*)	0,91(*)	0,88	0,94(*)

Se comparados os resultados de desempenho de cada classificador descritos pelos valores nas duas colunas de cada classificador, a começar pelo NB o conjunto reduzido obteve melhor desempenho em quatro dos sete conjuntos utilizados (conjuntos dos anos 2010, 2011, 2012 e 2014); nos três restantes (2009, 2013 e 2015) o desempenho varia entre 1 a 5% quando comparados os resultados de COs em relação ao CRs. Para o classificador NN o CR, em 4 situações, teve melhor resultado (nos anos de 2010, 2011, 2012 e 2014) e o CO em duas situações (2009 e 2013), os dados de 2015 não apresentaram diferença nos resultados usando ambos os conjuntos (CO ou CR). Para o classificador AD (i.e. J48) o CR apresentou melhores resultados em três dos experimentos, aqueles relativos aos anos de 2010, 2012, e 2015 e o uso do CO em duas situações (2011 e 2013). Nos dados referentes a 2009 e 2014, usando CO ou CR não fez diferença porque os resultados apresentados foram os mesmos. De certa forma os resultados apresentados no *Terceiro Experimento* evidenciam que com uma seleção prévia de atributos relevantes, a acurácia do classificador induzido, quando menor (comparada àquela do classificador gerado usando todos os atributos), é por um fator estatisticamente irrelevante e, em vários casos, pode mesmo melhorar o desempenho do classificador, como evidencia os valores da Tabela 7.37.

Capítulo 8
Conclusões e Trabalhos Futuros

8.1 Introdução

A predição do desempenho acadêmico em ambientes universitários, dentro de um período de tempo pré-determinado, permite que o docente responsável pela disciplina, adeque a condução das técnicas de ensino-aprendizagem de acordo com as necessidades apresentadas pelo desempenho dos estudantes. Esta pesquisa teve por objetivo principal investigar o uso de técnicas de Aprendizado de Máquina e Mineração de Dados, com o foco na predição do desempenho acadêmico de alunos em uma disciplina específica. Para a realização desta pesquisa foi utilizado o processo de aprendizado automático (4-validação cruzada) usando os algoritmos: Naïve Bayes (NB) [Domingos & Pazzani 1997], J48 (versão Weka do C4.5 [Quinlan 1993]) e o 1-Nearest Neighbor (1Bk) [Cover & Hart 1967]), disponíveis no ambiente Weka (versão 3.6.12) utilizada em todos os experimentos descritos neste documento.

São apresentadas duas seções a seguir. A Seção 8.2, resume os principais pontos investigados na pesquisa descrita neste trabalho, e também informa as principais contribuições alcançadas, e a Seção 8.3 sugere continuação para a pesquisa descrita neste documento.

8.2 Principais Pontos Investigados e Contribuições desta Pesquisa

Para alcançar o objetivo principal desta pesquisa, de investigar o uso de técnicas de AM na predição do desempenho acadêmico, foi necessário usar os conjuntos de dados, como entrada em cada algoritmo, observando-se a acurácia (taxa de acerto) obtida pelos classificadores resultantes. Os experimentos foram realizados usando 4-validação cruzada.

A pesquisa descrita nesta dissertação foi desenvolvida a partir de dados históricos sobre a disciplina Construção de Algoritmos e Programação 1 (CAP1), do curso

presencial de Bacharelado em Ciência da Computação, em que a cada semestre pode ocorrer variação na fórmula avaliativa usada na disciplina. Nesta pesquisa foram utilizados sete conjuntos de dados de CAP1, cada um deles contendo dados referentes a um particular ano, em um período de sete anos (2009-2015). Em cada ano foi usada uma fórmula avaliativa distinta. Foram desenvolvidos seis esquemas de avaliação para classificar cada conjunto de atributos, e.g, as fórmulas usadas nos anos de 2013 e 2014 compartilham o mesmo esquema de avaliação, porque ambos possuem os mesmos atributos, apesar de o cálculo usado ser diferente em cada ano.

Sobre o primeiro experimento descrito no Capítulo 7, na subseção 7.2.1, diferente do imaginado inicialmente com os atributos das informações pessoais IP por não possuírem alto poder preditivo em relação à classe, houve alta porcentagem de predição de desempenho somente com estes atributos (IP), como mostra a Tabela 8.1. Acredita-se que esses resultados tenham sido alcançados devido a distribuição heterogênea dos valores dos atributos em relação à classe. Os melhores resultados alcançados neste experimento são derivados dos classificadores gerados pela AD, como mostra a média da classificação em 72%.

Tabela 8.1 Resumo da taxa de classificação correta da coluna SR (*Success Rate*) apenas da primeira instância do conjunto (informações de IP) das 20 tabelas da Seção 7.2.1 (7.2 até 7.22) $^{(*)}$ caracteriza o melhor resultado.

ID	*NB*	*NN*	*AD*
CAP1	0,76	0,80$^{(*)}$	0,80$^{(*)}$
CAP2	0,90$^{(*)}$	0,85	0,87
CAP3	0,72$^{(*)}$	0,72$^{(*)}$	0,72$^{(*)}$
CAP4	0,73	0,66	0,77$^{(*)}$
CAP5	0,50	0,53$^{(*)}$	0,53$^{(*)}$
CAP6	0,72$^{(*)}$	0,48	0,72$^{(*)}$
CAP7	0,66$^{(*)}$	0,52	0,63
Média	0,71	0,65	0,72

O segundo experimento, dividido em três situações descrito no Capítulo 7, subseção 7.2.2, usa somente o conjunto de atributos comuns a todos os conjuntos (DN, EC, S, GI, VE, AV_1, AV_2, G), e apenas o classificador J48, gerador de AD, para induzir e predizer o desempenho. A primeira situação (subseção 7.2.2.1) usou o conjunto comum de atributos menos o atributo correspondente à segunda avaliação (DN, EC, S, GI, VE,

AV_1, AV_2, G), para induzir a árvore com as instâncias do ano anterior e nesta árvore classificar as instâncias do ano seguinte. Os resultados da Tabela 7.28, mostram uma variação entre 60% e 92% nos resultados dos seis experimentos. A segunda situação (subseção 7.2.2.2) se difere da primeira apenas por agregar o atributo AV_2 aos conjuntos de dados. Os resultados descritos na Tabela 7.29, evidenciam uma variação entre 49% e 92%. Estas duas situações mostraram que ora os dados de um ano podem prever com bom desempenho o ano seguinte e ora os conjuntos de dados podem não obter tão bom desempenho na predição. A terceira situação do segundo experimento (descrita na subseção 7.2.2.3) induz o conceito do classificador com um conjunto de dados robusto, formado pela junção dos 6 arquivos de dados (2009-2014) para induzir a classificação do conjunto de 2015. Os resultados mostraram um desempenho de 88% de classificação correta. Este resultado se aproxima do melhor resultado encontrado nas duas situações anteriores; das 69 instâncias apenas 8 foram classificadas incorretamente.

O terceiro experimento descrito na Seção 7.2.3, utilizou o conjunto original (CO) de atributos, que contempla todos os atributos de cada conjunto de dados (ver subseção 7.2.1) e o conjunto reduzido (CR) de atributos obtido a partir do 'ranqueamento' de atributos (ver Seção 6.3) e da redução por temporalidade (ver Tabela 7.2). A Tabela 7.37, compara os resultados obtidos com o CO e o CR, em cada classificador. Em todos os classificadores o conjunto reduzido obteve melhor desempenho, mostrando que é possível antever o desempenho do estudante ainda em períodos iniciais da disciplina, se selecionados os atributos disponíveis inicialmente com maior relevância em relação à classe. Em seis das sete situações o classificador NB apresentou melhores resultados corroborando o apontado na pesquisa de [Kotsiantis *et al.* 2004].

Foram encontradas dificuldades relacionadas aos dados, como o desbalanceamento dos valores dos atributos em relação à classe. Por se tratarem de dados reais, foram encontrados valores de atributos desatualizados ou ausentes. O Sistema de Gestão Acadêmica do Centro Universitário de onde os dados foram exportados, não possui documentação disponível, portanto o entendimento dos dados e do funcionamento da disciplina foi realizado a partir da manipulação dos dados e transformação para que servissem de entrada para os algoritmos.

Esta pesquisa contribui para a auxiliar na diminuição da permanência de estudantes além do tempo necessário para a conclusão de disciplinas e tende a ajudar a diminuir os índices de reprovação ou abandono. Os resultados obtidos nos experimentos descritos nesta dissertação, podem fornecer suporte para que gestores acadêmicos e professores possam monitorar o progresso de alunos a fim de identificar aqueles mais propensos à reprovação, para que estes sejam auxiliados a alcançar o desempenho esperado para aprovação.

8.3 Trabalhos Futuros

Como trabalhos futuros, é sugerido investigar o uso de diferentes tipos de algoritmos na predição de desempenho de estudantes, e.g, aqueles que induzem o conceito como um conjunto de regras de decisão ou induzem a representação do conceito como uma rede neural, entre outros. E aplicar técnicas de AM e MD em outros cursos ou disciplinas, que abordem o estudo de diferentes áreas de conhecimento.

Referências & Bibliografia

[Adomavicius & Tuzhilin 2005] Adomavicius, G.; Tuzhilin, A. (2005) Toward the next generation of recommender systems: A survey of the state-of-the-art and possible extensions, In: IEEE Transactions on Knowledge and Data Engineering, vol. 17, n.6, pp. 734-749.

[Agrawal *et al.* 1993] Agrawal, R.; Imielinski, T.; Swami, A. N. (1993) Mining associations rules between sets of items in large databases, In: Proceedings of the 1993 ACM-SIGMOD International Conference on Management of Data, pp. 207-216.

[Agrawal & Srikant 1994] Agrawal, R.; Srikant, R. (1994) Fast algorithms for mining association rules, In: Proceedings of the 20th International Conference on Very Large Databases, pp.478-499.

[Agrawal & Srikant 1995] Agrawal, R.; Srikant, R. (1995) Mining sequential patterns, In: Proceedings of the 11th International Conference on Data Engineering, pp. 3-14.

[Aha *et al.* 1991] Aha, D. W., Kibler, D. & Albert, M. K. (1991) Instance-based learning algorithms, In: Machine Learning, v. 6, pp. 37-66.

[Aha 1992] Aha, D. W. (1992) Tolerating noisy, irrelevant and novel attributes in instance-based learning algorithms, In: International J. Man-Machine Studies, v. 36, pp. 267-287.

[Angiulli 2007] Angiulli, F. (2007) Fast nearest neighbor condensation for large data sets classification, In: IEEE Transactions on Knowledge and Data Engineering, v. 19, no. 11, pp. 1450-1464.

[Baker & Yacef 2010] Baker, R.; Yacef, K. (2010). The state of educational data mining in 2009: A review and future visions, In: Journal of Educational Data Mining, pp. 3−17.

[Beale & Jackson 1994] Beale, R & Jackson, T. (1994) Neural computing: an introduction, In: Institute of Physics Publishing.

[Blum & Langley 1997] Blum, A.; Langley, P. (1997) Selection of relevant features and examples in machine learning. In: Artificial Intelligence 97, pp. 245-271.

[Blum & Mitchell 1998] Blum, A.; Mitchell, T. (1998) Combining labeled and unlabeled data with co-training, In: Proceedings of the Eleventh Annual Conference on Computational Learning Theory. pp. 92-100.

[Cohen 1995] Cohen, W. W. (1995) Fast effective rule induction, In: Proc. of The 12th International Conference on Machine Learning (ICML 1995), pp. 115-123.

[Cover & Hart 1967] Cover, T. M.; Hart, P. E. (1967) Nearest neighbor pattern classification, In: IEEE Transactions on Information Theory, v. IT-13, no. 1, pp. 21-27.

[Dash & Liu 1997] Dash, M.; Liu, H. (1997) Feature selection for classification. In: Intelligent Data Analysis, pp. 131-156.

[Domingos & Pazzani 1996] Domingos, P.; Pazzani, M. (1996) Simple Bayesian classifiers do not assume independence, In: Proc. AAA-I, pp. 1386.

[Domingos & Pazzani 1997] Domingos, P.; Pazzani, M. (1997) On the optimality of the simple Bayesian classifier under zero-one loss, In: Machine Learning, v. 29, pp. 103-130.

[Fayyad, Piatetsky-Shapiro & Smyth 1996] Fayyad, U.; Piatetsky-Shapiro, G.; Smyth, P. (1996) From data mining to knowledge discovery in databases, In: American Association for Artificial Intelligence, pp. 37-54.

[Forman & Cohen 2014] Forman, G.; Cohen, I. (2004) Learning from little: comparison of classifiers given little training, In: Proc. of The Eight European Conference on Principles and Practices of Knowledge Discovery in Databases, pp. 161-172.

[Friedman *et al.* 1997] Friedman, N.; Geiger, D.; Goldszmidt, M. (1997) Bayesian network classifiers, In: Machine Learning, v. 29, pp. 131-163.

[García *et al.* 2011] García, E.; Romero, C.; Ventura, S.; de Castro, C. (2011) A collaborative educational association rule mining tool, In: Internet and Higher Education, v. 14, pp. 77-88.

[Gates 1972] Gates, G. W. (1972) The reduced nearest neighbor rule, In: IEEE Transactions on Information Theory 18, pp. 431-433.

[Gottardo *et al.* 2012a] Gottardo, E.; Kaestner, C.; Noronha, R. V. (2012) Avaliação de desempenho de estudantes em cursos de educação a distância utilizando mineração de dados, In: XXXII Congresso da Sociedade Brasileira de Computação (CSBC 2012).

[Gottardo *et al.* 2012b] Gottardo, E.; Kaestner, C.; Noronha, R. V. (2012) Previsão de desempenho de estudantes em cursos EaD utilizando mineração de dados: Uma estratégia baseada em séries temporais, In: 23º Simpósio Brasileiro de Informática na Educação (SBIE 2012).

[Gotardo *et al.* 2013] Gotardo, R.; Cereda, P. R. M; Hruschka Jr., E. R. (2013) Predição do desempenho do aluno usando sistemas de recomendação e acoplamento de classificadores, In: II Congresso Brasileiro de Informática na Educação (CBIE 2013), pp. 657-666.

[Grzymala-Busse & Hu 2000] Grzymala-Busse, J. W.; Hu, M. (2000) A comparison of several approaches to missing atribute values in data mining. In: Proceeding of the Second International Conference on Rough Sets and Current Trends in Computing, pp. 340-347.

[Hart 1968] Hart, P. E. (1968) The condensed nearest neighbor rule, In: IEEE Transactions on Information Theory 14, pp. 515-516.

[Haruechaiyasak 2008] Haruechaiyasak, C. (2008) A tutorial on Naïve Bayes classification, RT-Choochart82255418560, Suanpalm/TH, Bangkok, 6 pg.

[Hruschka & Nicoletti 2013] Hruschka Jr, E. R; Nicoletti, M. C. (2013) Roles played by Bayesian networks in machine learning: an empirical investigation, In: Emerging Paradigms in Machine Learning, pp. 75-116.

[Hssina *et al.* 2014] Hssina, B.; Merbouha, A.; Ezzikouri, H.; Erritali, M. (2014) A comparative study of decision tree ID3 and C4.5, In: International Journal of Advanced Computer Science and Applications, Special Issue on Advances in Vehicular Ad Hoc Networking and Applications, pp. 13-19.

[Kohavi & John 1997] Kohavi, R.; John, G. H. (1997) Wrappers for feature subset selection. In: Artificial Intelligence 97, pp 273-324.

[Kotsiantis *et al.* 2003] Kotsiantis, S. B.; Pierrakeas, C. J.; Zaharakis, I. D.; Pintelas, P. E. (2003) Efficiency of machine learning techniques in predicting student's performance in distance learning systems, Recent Advances in Mechanics and Related Fields, University of Patras, Greece, (Technical Report).

[Kotsiantis *et al.* 2004] Kotsiantis, S.; Pierrakeas, C.; Pintelas. P. (2004) Predicting students' performance in distance learning using machine learning techniques, In: Taylor & Francis Inc. Applied Artificial Intelligence, v. 18, n. 5, pp. 411-426.

[Koutina & Kermanidis 2011] Koutina, M.; Kermanidis, K. L. (2011) Predicting Postgraduate Students' Performance Using Machine Learning Techniques, L. Iliadis et al. (Eds.): EANN/AIAI 2011, Part II, IFIP AICT 364, pp. 159–168.

[Liu & Motoda 2012] Liu, H.; Motoda, H. (2012) Feature selection for knowledge discovery and data mining. In: Springer Science & Busines Media, v. 454.

[Liu & Setiono 1996] Liu, H.; Setiono, R. (1996) A probabilistic approach to feature selection-a filter solution. In: Machine Learning: Proceedings of the Thirteenth International Conference on Machine Learning. Morgan Kaufmann.

[Long 1997] Long, S. J. (1997) Regression models for categorical and limited dependent variables, In: Advanced Quantitative Techniques in the Social Sciences, v. 7. CA: Sage Publications Inc..

[Mitchell 1997] Mitchell, T. M. (1997) Machine learning. McGraw-Hill Science/Engineering/Math, New York.

[Murphy 2006] Murphy, K. P. (2006) Naïve Bayes classifiers, RT-CS340-Fall06, UBC/CA, Vancouver, 8 pg.

[Ng 1990] Ng, K.; Abramson, B. (1990) Uncertainty management in expert systems, IEEE Expert, Abr. 1990, pp. 29-47.

[Platt 1998] Plat J. C. (1998) Sequential minimal optimization: a fast algorithm for training support vector machines, MSR-TR-98-14, Microsoft Research, 21 pg.

[Quinlan 1986] Quinlan, J. R. (1986) Induction of decision trees, In: Machine Learning, v. 1, pp. 81-106.

[Quinlan 1991] Quinlan, J. R. (1991) Determinate literals in inductive logic programming. In: Proceedings of the Eighth International Workshop on Machine Learning, Morgan Kaufmann, New York.

[Quinlan 1993] Quinlan, J. R. (1993) C4.5 Programs for Machine Learning. Morgan Kaufmann, San Francisco.

[Quinlan 1996] Quinlan, J. R. (1996) Improved use of continuous attributes in C4.5, In: Journal of Artificial Intelligence Research, v. 4, pp. 77-90.

[Ricci *et al.* 2010] Ricci, F.; Rokach, L.; Shapira, B.; Kantor, P. B. (2010) Recommender systems handbook, In: Springer, pp. 4-7.

[Rosenblatt 1958] Rosenblatt, F. (1958) The Perceptron: a probabilistic model for information storage and organization in the brain, Cornell Aeronautical Laboratory, Psychological Review, v. 65, pp. 386-408.

[Rumelhart *et al.*1986] Rumelhart, D. E.; Hinton, G. E.; Williams, R. J. (1986) Learning internal representations by error propagation. In: Parallel Distributed Processing, v. 1 e 2.

[Russel & Norvig 1995] Russel, S.; Norvig, P. (1995) Artificial Intelligence: A Modern Approach, Prentice Hall Series in Artificial Intelligence.

[Santos & Nicoletti 1996] Santos, F. O.; Nicoletti, M. C. (1996) O uso do modelo de Bayes em sistemas baseados em conhecimento, RT-DC 003/96, UFSCar/DC, São Carlos, 19 pg.

[Schafer 1999] Schafer, J. L. (1999) Multiple imputation: A primer. In: Statistical Methods in Medical Research, v. 8, pp. 3-15.

[Stoffalette 2015] Stoffalette João, R. (2015) Mineração de padrões sequenciais e geração de regras de associação envolvendo temporalidade, Dissertação de Mestrado, DC-UFSCar, 2015.

[Wasikowski & Chen 2010] Wasikowski, M.; Chen, X (2010) Combining the small sample class imbalance problem using feature selection, IEEE Computer Society v. 22, pp. 1388-1400.

[Weka 2012] *WEKA Manual*, Version 3-6-8, New Zealand: University of Waikato.

[Witten & Frank 2005] Witten, I. H.; Frank, E. (2005) Data Mining – Practical Machine Learning Tools and Techniques, 2nd ed., USA: Elsevier.

[Witten, Frank & Hall 2011] Witten, I. H.; Frank, E.; Hall, M. A. (2011) Data Mining – Practical Machine learning Tools and Techniques. 3rd ed., USA: Elsevier.

[Yang & Honavar 1998] Yang, J.; Honavar, V. (1998) Feature subset selection using a genetic algorithm. In: Feature extraction, construction and selection. Springer, pp 117-136.