

Capítulo

2

Tomada de Decisão Baseada em Dados (DDDM) e Aplicações em Informática em Educação

Luciano Silva

Faculdade de Computação e Informática, Universidade Presbiteriana Mackenzie
luciano.silva@mackenzie.br

Abstract

The Data-Driven Decision Making (DDDM) is a data governance approach which establishes a flow of data acquisition, analysis and reporting to support decision-making. DDDM has already been applied with great success in the corporate environment and recently began to awaken interest in the Education. Several decisions about teaching and learning process improvements can be made based on census data analysis, test results, satisfaction surveys and portfolios of students and classes. With the development of teaching and learning environments as well as the Educational Data Mining, large volumes of data can be stored and analyzed with a view to implementing a DDDM flow. Within this context, the aim of this work is to present the fundamentals of DDDM, its main tools and especially their possibilities and potential for teachers and managers to make sound decisions for the teaching and learning process improvement.

Resumo

A Tomada de Decisão Baseada em Dados (DDDM – Data-Driven Decision Making) é uma abordagem de governança de dados que estabelece um fluxo de aquisição, análise e comunicação de dados para suportar processos decisórios. DDDM já tem sido aplicada com bastante sucesso no ambiente corporativo e, recentemente, começou a despertar interesse no segmento educacional. Diversas decisões de melhorias de processos de ensino e aprendizagem podem ser tomadas com base em análises de dados censitários, resultados de testes, pesquisas de satisfação e portfólios de alunos e turmas. Com a evolução dos ambientes de ensino e aprendizagem, assim como das técnicas de Mineração de Dados Educacionais (MDE), grandes volumes de dados podem ser armazenados e analisados visando à implantação de um fluxo de DDDM. Dentro deste contexto, o objetivo deste texto é apresentar os fundamentos da DDDM, suas principais ferramentas e, principalmente, suas possibilidades e potencialidades para que professores e gestores possam tomar decisões seguras para processos de melhoria de ensino e aprendizagem.

1. Introdução

Grandes volumes de dados têm sido gerados nos mais variados segmentos educacionais: ambientes virtuais de aprendizagem, resultados de testes, portfólios de acompanhamento de alunos e turmas, pesquisas censitárias, pesquisas de CPAs (Centros Próprios de Avaliação), dentre outros. Estas bases de dados constituem um ativo educacional valioso para que professores e gestores possam analisar ambientes educacionais e tomar decisões sobre os mais variados aspectos dos processos de ensino e aprendizagem como, por exemplo, avaliar se determinada metodologia de ensino tem sido efetiva ou não durante um certo período.

Neste contexto, já houve um avanço bastante significativo com a Mineração de Dados Educacionais, que permite descobrir “conhecimento” nestas bases de dados educacionais. O próximo passo natural após a aplicação de MDE é um processo de tomada de decisão. Uma das estratégias para isto é a DDDM (*Data-Driven Decision Making*) – Tomada de Decisão Baseada em Dados, que estabelece um fluxo de trabalho bem determinado para que processos decisórios sobre dados educacionais possam ser tomados. DDDM já tem sido aplicada com bastante sucesso em outros segmentos corporativos e, recentemente, despertou o interesse das comunidades de pesquisa em Informática em Educação.

O desenvolvimento de novas técnicas, processos e metodologias que visem a apoiar DDDM é de grande interesse tanto do ponto de vista computacional quando educacional, além de representar um desafio para avanços em Informática em Educação. No contexto computacional, como organizar bases de dados educacionais, desenvolver e aplicar novas técnicas de mineração de dados, e apresentar resultados de forma concisa e interativa com usuários, representam alguns dos possíveis desafios para se obter melhores fluxos de DDDM. Já no contexto educacional, como analisar o “conhecimento” descoberto pelos processos de mineração e tomar decisões a partir disto, configuram-se como os principais desafios na adoção de DDDM.

Assim, para apresentar à comunidade de Informática em Educação as possibilidades e potencialidades da DDDM, este minicurso se propõe aos seguintes objetivos:

- apresentar os conceitos fundamentais de DDDM, desde os processos de coleta e análise de dados até a consolidação de dados e posterior tomada de decisão
- discutir cenários de aplicação de DDDM em diversos segmentos educacionais
- apresentar técnicas, metodologias e ferramentas computacionais que dão suporte à implantação de fluxos de DDDM

Este capítulo de apoio ao minicurso está organizado da seguinte forma:

- a Seção 4.2 apresenta os fundamentos do processo de tomada de decisão baseada em dados;
- a Seção 4.3 mostra os princípios de aquisição de dados para processos decisórios sobre dados;
- a Seção 4.4 detalha os mecanismos de mineração de dados, com ênfase em dados educacionais;

- a Seção 4.5 mostra algumas estratégias para consolidação de dados como *dashboards* e *heatmaps*;
- a Seção 4.6 mostra alguns exemplos de tomadas de decisão baseadas em dados educacionais; e,
- finalmente, a Seção 4.7 oferece recomendações finais.

Ao final do texto, encontra-se uma seleção de referências bibliográficas importantes para os interessados em se aprofundar no assunto. O autor espera que, com este texto, os participantes conheçam os fundamentos de DDDM, suas principais aplicações e o arcabouço computacional necessário para a sua implantação. Além disto, espera-se que este contato com a DDDM possa suscitar novas pesquisas nos grupos de pesquisa em Informática em Educação no país.

2. Fundamentos de Tomada de Decisão Baseada em Dados

2.1 Processo Decisório

Decisões são escolhas que as pessoas fazem para enfrentar problemas e aproveitar oportunidades (Datnow, 2014). Tomar decisões para enfrentar problemas e aproveitar oportunidades é um ingrediente importante do trabalho de administrar. Muito do que os gerentes fazem é resolver problemas e enfrentar outros tipos de situações que exigem escolhas. Tomar decisões é identificar e selecionar um curso de ação para lidar com um problema específico ou extrair vantagens em uma oportunidade.

Os problemas e as situações variam muito em termos de natureza, urgência, impacto sobre as organizações e outros fatores. Por isso, as decisões podem ser classificadas de diferentes maneiras. A classificação das decisões de acordo com as situações a que se aplicam permite aos gerentes definir quanta energia e tempo dedicar a cada uma delas (BERRY e LINOFF, 2004). A classificação do processo de tomada de decisão é mostrado na Tabela 1:

Tabela 1: Classificação dos tipos de decisão.

Grau de Familiaridade	<ul style="list-style-type: none">• Decisões Programadas• Decisões Não programadas
Quanto a Natureza	<ul style="list-style-type: none">• Decisões Estratégicas• Decisões Administrativas ou Gerenciais• Decisões Operacionais
Grau de Participação	<ul style="list-style-type: none">• Decisões Individuais• Decisões Coletivas
Quanto ao Tipo de resultado	<ul style="list-style-type: none">• Decisões Satisfatórias• Decisões Maximizadas• Decisões Otimizadas

As decisões programadas aplicam-se a problemas repetitivos, enquanto que as decisões não-programadas aplicam-se a problemas que não são familiares.

As decisões estratégicas escolhem objetivos organizacionais e os meios para realizá-los. As decisões administrativas ou gerenciais colocam decisões estratégicas em prática, enquanto que as decisões operacionais definem meios e recursos para a execução de atividades. Normalmente são tomadas no nível operacional, mas também podem ser tomadas no nível da gerência.

As definições individuais ou coletivas são aquelas que se referem ao grau de participação das pessoas. Certas decisões são individuais, outras são tomadas por grupos ou por meio de consultas a grupos.

Para uma decisão satisfatória, qualquer alternativa serve, pode ser motivada por falta de tempo, informação ou outros recursos. Para uma decisão maximizada, procura-se o melhor resultado possível, ou seja, aquele que produz as maiores consequências positivas e reduz ao mínimo as consequências negativas e, nas decisões otimizadas, equilibram vantagens e desvantagens de diversas alternativas. Às vezes o tomador de decisões procura uma solução média, que atenda a um número de critérios e objetivos.

2.2 Estrutura Geral do Fluxo DDDM

Decisões podem ser tomadas com base em dados, sendo o processo decisório denominado Tomada de Decisão Baseada em Dados (*DDDM*). Um fluxo típico de DDDM envolve (Bambryck, 2010):

1. Aquisição de dados
2. Mineração de Dados
3. Consolidação e comunicação de dados
4. Tomada de decisão

A seguir, serão definidas de forma geral as atividades de cada uma destas fases. Cada uma delas será, posteriormente, detalhada numa seção específica.

2.3 Aquisição de Dados

O processo de aquisição de dados consiste nas atividades de obtenção de dados visando à aplicação de técnicas de Mineração de Dados. Existem diversas estratégias para se realizar a aquisição de dados para DDDM:

- formulários *online* e *offline*
- consultas a bases de dados vinculados ao contexto de tomada de decisão
- monitoração em tempo real do contexto

É comum os dados estarem organizados em planilhas ou bases de dados para efeitos de utilização em um processo de DDDM.

2.4 Mineração de Dados

O processo de Mineração de Dados engloba as atividades de análise de dados, objetivando-se à chamada “descoberta de conhecimento” nos dados com processos de agrupamento e classificação, mineração de relação e análise baseada em modelos.

2.5 Consolidação e Comunicação de Dados

Na consolidação e comunicação de dados, o objetivo principal é a utilização de mecanismos de visualização de dados (tabelas, gráficos, dentre outros), de tal forma a destilar possíveis volumes de dados em artefatos que possam facilitar o processo decisório. Como a área de Visualização de Dados está bastante avançada, esta fase possui diversas alternativas de artefatos com *dashboards* e *heatmaps*.

2.6 Tomada de Decisão

Trata-se da fase que culmina o processo de DDDM. Existem diversas técnicas e modelos de tomada de decisão: determinísticos, probabilísticos, *fuzzy*, dentre outros. Nesta fase, o objetivo principal é definir uma ação com base nos dados brutos que estão sendo analisados, assim como no conhecimento descoberto na fase de Mineração de Dados e nos artefatos de visualização.

3. Aquisição de Dados para DDDM

Na aquisição de dados para DDDM, as atividades principais referem-se aos métodos de obtenção e recuperação de informações para a fase de mineração de dados (Corrigan e Vincent, 2011). Neste contexto, as bases de dados configuram-se como elementos essenciais para o armazenamento de grandes volumes de dados.

3.1 Bases de Dados

Há duas alternativas essenciais para armazenamento de dados para os processos de aquisição em DDDM, conforme mostrado na Figura 1:

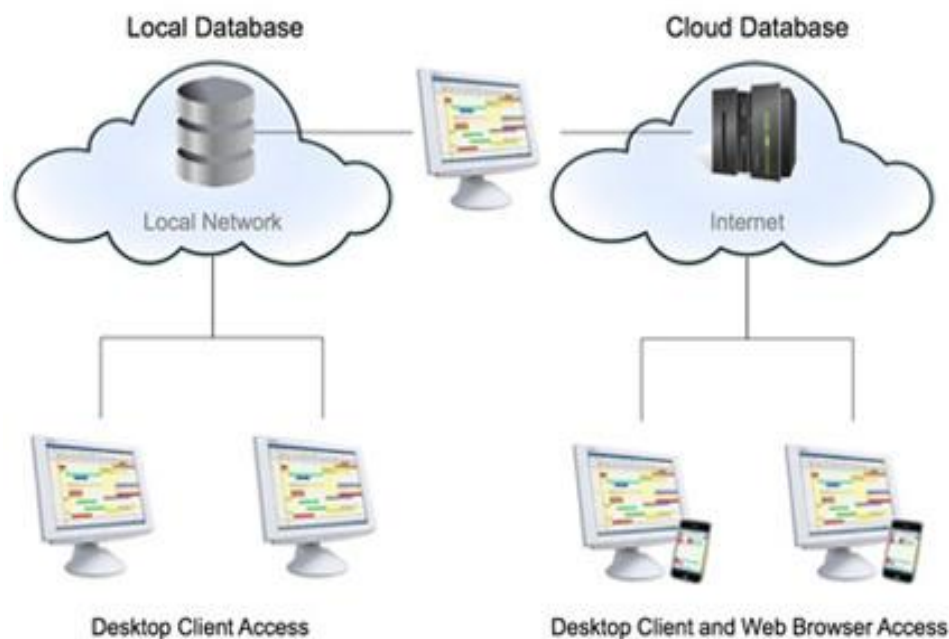


Figura 1. Bases locais e em nuvens para armazenamento de dados em DDDM.

No armazenamento local (*Local Database*), a base de dados está fisicamente na mesma rede local onde se localiza o processo de análise DDDM. Isto permite, por exemplo, maior velocidade de acesso no processamento de grandes volumes de dados. Porém, o acesso a pontos remotos pode apresentar algumas dificuldades.

No armazenamento em nuvem (*Cloud Database*), não há necessidade de se saber onde a base de dados está fisicamente (diz-se, neste caso, que a base está na nuvem). O acesso é feito por meio de um portal e diversas plataformas podem acessar estas bases

como, por exemplo, estações de trabalho (desktops), dispositivos móveis e navegadores. Apesar desta diversidade de acesso, a qualidade de banda de comunicação da rede de acesso é o principal fator limitante do desempenho deste tipo de armazenamento.

3.2 Métodos de Aquisição de Dados

Para o processo de aquisição de dados, há diversas estratégias. Uma das estratégias mais comuns é o uso de formulários, onde são solicitados dados específicos para um determinado grupo que se queira tomar alguma decisão.

A Figura 2 ilustra um exemplo destes formulários:



Campos para Cadastro do Aluno

* Nome

* RG * CPF * Data Nascimento * Sexo

Nacionalidade Naturalidade Estado Civil

* Endereço Residencial (Rua, Avenida)

Número Complemento Bairro

CEP Cidade Estado

Telefone Residencial Celular E-mail

Código do Responsável

Figura 2. Exemplo de formulário para obtenção de dados de alunos.

Neste formulário estão sendo solicitados dados pessoais para a montagem de cadastro de alunos. Informações importantes como sexo, data de nascimento, estado civil e cidade podem direcionar processos decisórios como, por exemplo, alocação de docentes por faixa etária.

Outro método possível para obtenção de dados são as próprias consultas em bases pré-existentes, conforme mostrado na Figura 3:

Consulta Geral

Codigo do Aluno: 29
Contratante:
Aluno: ANTONIO DIAS NETO
Telefone: 1295887455

Matricula: 1020
Status do Aluno: Estudando
Turma: 1222/1
Tipo de curso: REDES NT4

Contrato cadastrado em: 14/04/03
Vendedor: 1
Cadastrado Por:
Nada consta no SPC
Empresa:
Consulta:
Sair

Parcela	Vencimento	Valor	Situação
1 /12	25/03/03	R\$620,00	Vencido
10 /12	05/12/03	R\$620,00	A Vencer
11 /12	05/01/04	R\$620,00	A Vencer
12 /12	05/02/04	R\$620,00	Pago
2 /12	05/04/03	R\$620,00	Vencido
3 /12	05/05/03	R\$620,00	Vencido
4 /12	05/06/03	R\$620,00	A Vencer
5 /12	05/07/03	R\$620,00	A Vencer
6 /12	05/08/03	R\$620,00	A Vencer
7 /12	05/09/03	R\$620,00	A Vencer
8 /12	05/10/03	R\$620,00	A Vencer
9 /12	05/11/03	R\$620,00	A Vencer

Valor da Mensalidade: R\$620,00
1 /12 Parcelas de: R\$620,00
Valor Total do curso: R\$7.440,00
Valor total Pago pelo Aluno: R\$620,00
Valor total a Vencer: R\$4.960,00
Número de Parcelas Vencidas: 3

Figura 3: Consulta a uma base de alunos para geração de grupos de dados.

As consultas realizadas diretamente nas bases de dados permitem expor segmentos específicos de dados para a fase de mineração de dados. No exemplo particular da Figura 3, o interesse maior está vinculado no acompanhamento de parcelas pagas, vencidas ou a vencer de um determinado aluno.

Uma forma mais complexa de obtenção de dados consiste no processamento de *streams*, normalmente capturadas em fluxos contínuos de terminados processos. Na Figura 4, tem-se um cenário mostrando o uso desta forma de captura de dados:

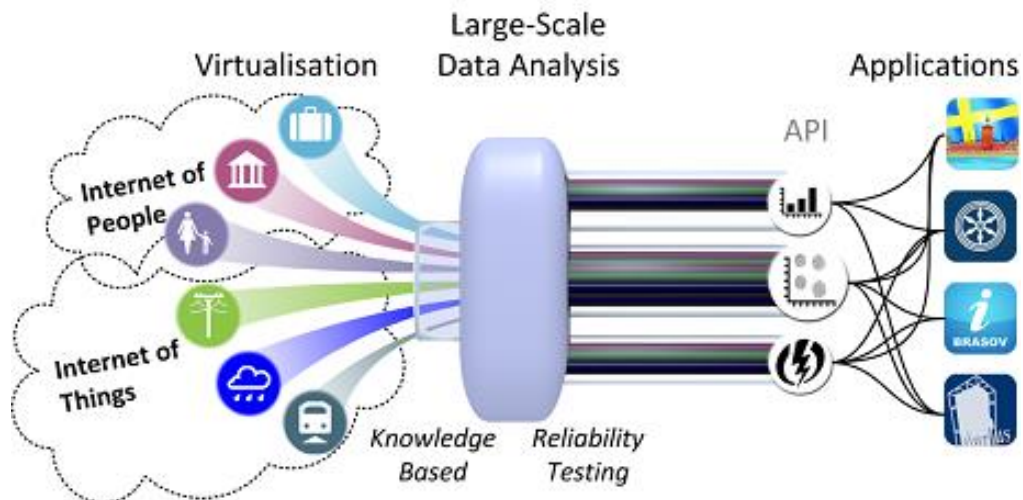


Figura 4. Cenário de processamento de streams.

No esquema de obtenção de dados por *streaming*, vários pontos geradores de dados (*Internet of People, Internet of Things*) geram sequências contínuas de dados (*streams*) que alimentam diretamente um sistema de análise (*Data Analysis*), normalmente sem a necessidade de um armazenamento intermediário. Geralmente, também não há necessidade de armazenamento do resultado desta análise e os resultados podem ser direcionados diretamente para os processos decisórios. Este esquema, em particular, poderia ser utilizado para monitorar as atividades em tempo real de estudantes realizando algum tipo de tarefa.

4. Mineração de Dados para DDDM

4.1 Fundamentos de Mineração de Dados

Mineração de Dados é uma área de pesquisa multidisciplinar, envolvendo basicamente Banco de Dados, Estatística e Aprendizado de Máquina. A MD é parte principal de um processo que tem como entrada uma Base de Dados e como saída um Conhecimento (Fayyad et al., 1996). Ela é dividida em tarefas como predição, clusterização e associação que devem ser escolhidas de acordo com análises exploratórias inicialmente feitas sobre os dados (Han et al., 2006; Tan et al., 2009; Faceli et al., 2011). A MD tem sido amplamente utilizada em diferentes áreas, principalmente medicina, indústria, marketing, agronegócios, educação, entre outras (Berry e Linoff, 2004; Dunstone e Yager, 2008; Stone et al., 2008; Silva et al., 2008; Romero e Ventura, 2010).

Mineração de Dados é o termo usado para uma tarefa aplicada em um grande volume de dados com o objetivo de descobrir padrões relevantes em um contexto previamente definido para apoio em tomada de decisão. Portanto, a descoberta de padrões constitui-se de um processo que se inicia pela escolha dos dados que documentam de alguma maneira a pergunta que o especialista deseja responder. Os dados são integrados e pré-processados para que sejam entregues estruturados, higienizados, selecionados e padronizados à tarefa de Mineração de Dados. Na tarefa de mineração aplica-se alguma técnica inteligente capaz de encontrar soluções que auxiliam o especialista na descoberta de uma resposta. O resultado desta tarefa deve ser pós-processado para que se apresentem análises qualitativa e/ou quantitativa dos elementos encontrados e, quando possível, apresentados de maneira que possa ser interpretada para facilitar a tomada de decisão. Na Figura 5, descreve-se todo este processo, o qual pode ser chamado de Descoberta de Conhecimento em Bases de Dados ou KDD (*Knowledge Discovery in Database*) (Fayyad et al., 1996).

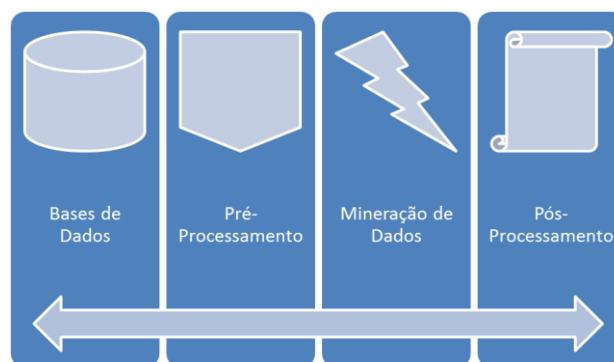


Figura 5. Processo de Descoberta de Conhecimento em Bases de Dados.

Note, nesta figura, de processo do KDD que há uma seta bidirecional na integração de todas as etapas. Isto significa que o processo pode ser iterativo, ou seja, seguindo cada passo por vez, ou interativo, o processo pode ser executado em passos arbitrários ou mesmo repetidos caso haja necessidades. A seguir, serão discutidas com mais detalhes cada uma das etapas do KDD.

4.2 Bases de Dados para Mineração de Dados

O processo KDD envolve a descoberta de conhecimento de um conceito específico intrinsecamente armazenado em bases de dados. Portanto, é importante envolver neste processo um especialista que conhece o problema e que possa auxiliar na escolha de bases de dados. Isto significa que podem ocorrer casos onde será necessário utilizar mais de uma base de dados. E ainda, esta base poderá representar dados organizados em uma tabela, os quais são chamados de estruturados, ou podem estar em documentos descritos na forma de texto ou multimídia (imagem, vídeo e som) os quais são chamados de não estruturados.

Nesta etapa do processo os dados devem ser integrados e assim constituir uma base de dados única. Se a pergunta que se deseja responder envolve dados não estruturados, então esses devem passar por um processo anterior e, conseqüentemente serem representados de maneira estruturada. No caso de mais de uma base de dados estruturada, a integração pode gerar uma série de problemas que serão tratados na etapa seguinte de pré-processamento.

Uma base de dados é constituída por uma tabela, na qual geralmente os exemplares ficam em linhas e os atributos em coluna. Não há regras que impeça a organização ser feita invertendo linhas por colunas. O exemplar é um registro da base de dados que ainda é chamado na literatura como objeto, padrão, amostra, etc. O atributo, por outro lado, descreve o exemplar e é chamado na literatura por características ou variáveis. Os atributos podem ser definidos como regulares ou especiais. Atributos regulares são medidas feitas sobre a pergunta em análise, e atributos especiais são rótulos que se atribui ao exemplar. Por exemplo, em um problema de MDE onde se deseja aplicar Mineração de Dados sobre os dados de alunos de um curso e suas notas em todos os semestres para encontrar um modelo que seja capaz de inferir se o aluno se formará no prazo. Cada aluno é um exemplar desta base e a nota atribuída ao aluno em cada disciplina é um atributo regular. A situação do aluno no final do curso, por exemplo, formado no prazo ou fora do prazo é um exemplo típico de atributo especial.

O valor armazenado nos atributos pode ser numérico ou categórico. O valor numérico pode ser contínuo ou discreto, e o valor categórico pode ser nominal ou ordinal. Estas nomenclaturas e definições são importantes para a escolha da tarefa de Mineração de Dados e os respectivos algoritmos no problema em questão. Na seção seguinte, daremos continuidade ao processo KDD, discutindo o pré-processamento.

4.3 Pré-Processamento

Como discutido antes, o processo de KDD pode ser iterativo ou interativo. Portanto, a necessidade de envolver a etapa de pré-processamento deve ser avaliada a cada

problema. Muitos dos casos tratados no pré-processamento são oriundos da integração de base de dados. Exemplos típicos destes casos são valores ausentes, ruidosos, inconsistentes e redundantes. Ainda há de considerar também que, em outras situações com ou sem integração de bases de dados, o aumento no número de exemplares e atributos da base de dados podem interferir na execução dos algoritmos usados nas tarefas de mineração de dados e, por essa razão, devem ser tratados.

Como as razões para a necessidade de pré-processamento variam muito de acordo com a área de conhecimento e considerando-se que o foco deste trabalho é base de dados educacionais, contextualizaremos as causas neste domínio de interesse. Embora as causas sejam variantes ao domínio de aplicação, as soluções, por outro lado, são robustas. Isso significa que elas podem ser aplicadas não somente à MDE.

Adicionalmente às situações acima expostas, a sequência de diagnósticos de problemas nas bases de dados não deve ser entendida como ótima solução. Pode haver casos em que o tratamento pode ser feito em uma ordem contrária a aqui apresentada.

Para um diagnóstico de possíveis problemas é comum inicialmente fazer uma análise descritiva dos dados. Nesta análise, medições são feitas sobre os atributos dos dados, como média, desvio padrão, quartis, valor mínimo e máximo. Estas medidas auxiliam no encaminhamento da solução de pré-processamento a ser adotada e, também, em caso de valores ausentes já será possível verificar a sua existência e, conseqüentemente, a sua solução.

Valores ausentes (*missing values*) são atributos que não têm valores preenchidos. As razões são diversas, por exemplo, em uma integração de dados de alunos de uma faculdade envolvendo diferentes cursos, pode haver disciplinas incomuns e, portanto haverá o surgimento deste tipo de valores que precisam ser tratados. O tratamento pode ser feito pela simples remoção do atributo (em caso de grande incidência como poderia ser para o caso do exemplo apresentado) ou do exemplar (em caso de poucas ocorrências). Ou, ainda, o valor pode ser substituído por uma constante calculada pela média, mediana, valor máximo ou mínimo. Por fim, pode-se aqui praticar a interatividade do KDD e utilizar alguma tarefa de Mineração de Dados para fazer a estimação do melhor valor para o atributo.

Outro tratamento que pode ser diagnosticado na análise descritiva são os valores ruidosos ou que estão fora do padrão (*outliers*). Este tipo de situação ocorre quando surge algum exemplar com valor de atributo que foge de um padrão. Isto pode ser diagnosticado com a construção de um *box plot*. Esse gráfico tem o valor dos 3 primeiros quartis e os valores mínimo e máximo e, portanto, pode-se verificar visualmente a variação dos valores de cada atributo e diagnosticar a presença de *outliers*. Alternativamente, o desvio padrão também pode ser usado como uma informação qualitativa de valores fora de um padrão. Por exemplo, se em uma prova unificada da faculdade que historicamente tem média 5 e, esporadicamente um grupo de alunos aparecem com nota 10, isso pode significar um ruído. A razão do surgimento pode ser uma fraude (no exemplo, o grupo de alunos pode ter acesso antecipado às questões da prova) ou pode ser uma mudança de comportamento no padrão da base de dados ou *outlier* (no exemplo, os alunos são mais dedicados que em anos anteriores). O tratamento para este tipo de ocorrência pode ser feito por diferentes abordagens (Han et al., 2006). Tipicamente, ela pode ser resolvida com a separação dos valores em faixas pré-definidas,

técnica conhecida como *binning* (Tan et al., 2009). Para o exemplo de contexto, o *binning* seria apresentar notas dos alunos não por números, mas sim por letras. Neste caso, cada letra representa uma faixa de valores e assim evitam-se além do *outliers* efeitos de subjetividade em uma correção da prova.

Por fim, uma situação que surge tipicamente quando se faz integração de dados é a inconsistência de valores. A inconsistência ocorre quando há falta de um critério bem definido entre os valores dos atributos ou dos exemplares. Exemplo que ilustra o critério para os atributos é quando em uma análise de dados de notas de alunos e situação final, um aluno teve nota 5 e está aprovado e, na mesma base, um outro aluno tem a mesma nota 5, mas está reprovado. Isto pode ocorrer por diversas razões, como a mudança de critério de aprovação de um curso. Outra situação, agora com os valores de exemplares, considere que uma disciplina tenha aula teoria e prática. Na aula teórica, o professor define uma escala de nota de 0 a 10 e, por outro lado, na aula prática outro professor define uma escala de 0 a 1. Na integração destas notas haverá inconsistência entre as escalas dos valores. Para o primeiro exemplo, a solução poderia ser a remoção dos exemplares ou o ajuste dos valores, desde que se conheça alguma informação adicional sobre os dados fornecida por um especialista. No segundo caso, a informação de valores mínimos e máximos da análise descritiva pode ajudar na solução chamada de normalização.

A normalização de valores consiste em uma técnica para deixar os valores dos atributos em uma mesma escala. No exemplo anterior, alunos com notas de 0 a 1 e de 0 até 10, a normalização faz com que se tenha uma escala única. Abordagem de solução comum é calcular o valor máximo de um atributo para dividi-lo aos demais exemplares com mesmo atributo. A normalização faz parte de um tratamento chamado transformação de valores que ainda compreende a mudança de tipos categóricos para numéricos. Ou seja, caso queira fazer uma análise mais profunda dos alunos considerando informações de escolaridade dos pais e gênero do aluno. Neste caso, a escolaridade deveria ser um atributo categórico ordinal com valores típicos segundo grau, graduação, pós-graduação que seriam transformados para 1, 2 e 3, por exemplo. No caso do gênero, atributo nominal, como são apenas dois valores, eles poderiam ser transformados para binário 0 e 1. No entanto, deve-se ter cuidado para não transformar um atributo nominal em ordinal no processo de transformação, isto é, o valor não pode ter ideia de ordem. Para mais de dois tipos de valores nominais, a transformação pode ser feita com uso de sequências de 0s e 1s, por exemplo, 01, 10, 11 e assim conseqüentemente.

Finalmente, como última investigação feita no pré-processamento é a seleção de exemplares ou atributos. As causas que levam a se fazer este tipo de análise são muitas, tais como, integração de bases, falta de definição clara de atributos que representam um problema, grande disponibilidade de dados, e outras. A seleção consiste basicamente em escolher o melhor conjunto de dados que representam a base original com a mesma capacidade analítica. A seleção de exemplares é um processo bastante simples, basta verificar redundâncias e eliminar o exemplar que está a mais na base de dados. A seleção de atributos, por outro lado, é preciso ter um pouco mais cuidado. O que se faz basicamente é verificar o quanto dois atributos estão correlacionados e, conseqüentemente, escolher um deles para análise. Portanto, a correlação, por exemplo, de Pearson, é uma técnica típica para seleção de atributos (Tan et al., 2009).

5. Tarefas de Mineração de Dados

A tarefa de Mineração de Dados é onde se extrai algum tipo de conhecimento da base de dados. Isso será feito a partir da escolha de uma tarefa e, conseqüentemente, de um algoritmo, os quais dependem basicamente dos atributos disponíveis. O conhecimento descoberto dependerá das tarefas que estão em uso, categorizadas como: Modelo Preditivo, Análise de Agrupamento e Regras de Associação.

5.1 Modelo Preditivo

A predição pode ser definida como uma tentativa de se descobrir o que acontecerá em um momento futuro (Han et al., 2006). Em Mineração de Dados, essa tentativa é feita a partir de um modelo construído usando a base de dados como referência. Por esta razão, a base usada na construção do modelo é chamada de treinamento (base de treinamento) e, que para o modelo preditivo, é necessário que ela tenha o atributo especial com a classe ou valor que se deseja prever. O processo de construção do modelo recebe o nome de aprendizado e, por ter o atributo especial como referência, designamos como sendo supervisionado. Portanto, o aprendizado supervisionado é o processo de construção de um modelo preditivo onde se conhece, para finalidades de ajustes de parâmetros dos algoritmos, o que se deseja prever (Han et al., 2006; Tan et al., 2009; Faceli et al., 2011).

Em linhas gerais, o conhecimento descoberto no modelo preditivo é a função que permite mapear um conjunto de atributos de um exemplar que não faz parte da base de treinamento em um dos valores de atributo especial. Isto significa que após o treinamento, o modelo preditivo é capaz de inferir novas situações que não estavam armazenadas na base de dados. Dependendo do tipo do atributo especial, chamamos o modelo preditivo de classificação ou previsão.

Classificação de dados é quando o atributo especial da base de treinamento é formando por uma categoria ou, como também chamado na literatura, por um rótulo (em inglês, *label*) ou alvo (em inglês, *target*). No contexto de educação, um processo de classificação poderia ser feito para prever se o aluno merece ou não uma bolsa de estudos, se ele está apto para participar de um intercâmbio, se ele se formará no prazo ou não, etc. Apesar dos exemplos serem com apenas duas classes, a classificação ainda pode ser feita para múltiplas classes.

Previsão de séries é o nome dado quando o atributo especial é um valor contínuo. No contexto de educação, a previsão poderia ser feita para inferir a nota do aluno em uma disciplina, a nota média do aluno no final do curso, a quantidade de faltas no final de um semestre e etc.

Na literatura encontram-se muitos algoritmos de classificação e previsão de séries. Alguns deles ainda podem ser usados para ambos os tipos de atributos especiais ou então adaptados para operarem com diferentes tipos de atributos. De maneira geral, os algoritmos são divididos basicamente em Aprendizado de Máquina e Inteligência Computacional. Exemplos típicos de algoritmos de Aprendizado de Máquina são: k Vizinhos mais Próximos ou kNN (*k Nearest Neighbor*), Máquinas de Vetores de Suporte ou SVM (*Support Vector Machine*), Árvores de Decisão (*Decision Tree*), Bayes, e muitos outros. Os algoritmos de Inteligência Computacional são aqueles inspirados na

biologia, principalmente no humano (Han et al., 2006; Tan et al., 2009; Faceli et al., 2011; Witten et al., 2001). Exemplo típico são as Redes Neurais Artificiais com algoritmos do tipo Perceptron de Múltiplas Camadas ou MLP (*Multi Layer Perceptron*), Funções de Base Radial (Radial Bases Function), dentre outros (Haykin, 2009).

5.2 Análise de Agrupamentos

A descoberta de grupos (*clustering*) é um modelo que procura encontrar exemplares ou padrões com atributos semelhantes na base de dados. Portanto, a segmentação da base em grupos é feita a partir de medidas de similaridade. Em problemas que envolvem agrupamento, não se encontra disponível o atributo especial classe ou rótulo. Por esta razão, dizemos que o aprendizado realizado pelos algoritmos de agrupamento é não-supervisionado (Jain e Dubes, 1988; Han et al., 2006; Tan et al., 2009; Jain, 2010; Faceli et al., 2011; Witten et al., 2001).

O aprendizado não-supervisionado é de grande desafio, pois não se conhece a saída que se deseja alcançar, o que significa não conhecer o número de grupos da base de dados. E ainda, os exemplares estão distribuídos em um espaço de dimensão elevada com diferentes formatos e separação. Estes aspectos demonstraram a grande dificuldade para lidar em problemas de agrupamento de dados.

Em um contexto geral, o que o agrupamento faz é a descoberta de perfil. Mais especificamente em educação, a descoberta poderia ser útil para descobrir estilos de aprendizado dos alunos, disciplinas de interesses comuns e muitas outras aplicações.

Assim como na análise preditiva, os algoritmos de agrupamento também podem ser divididos em Aprendizado de Máquina e Inteligência Computacional. Algoritmos típicos de Aprendizado de Máquina são o Agrupamento Hierárquico, k-Médias (k-Means), Agrupamento Espacial baseado em Densidade ou DBScan (*Density Based Spatial Clustering of Applications with Noise*). Por outro lado, como abordagem baseada em Redes Neurais temos o Mapa Auto-Organizável ou SOM (*Self-Organizing Map*) (Jain e Dubes, 1988; Han et al., 2006; Tan et al., 2009; Haykin, 2009; Jain, 2010; Faceli et al., 2011; Witten et al., 2001).

5.3 Regras de Associação

A associação de dados é uma abordagem de Mineração de Dados bastante diferente das demais apresentadas. A particularidade desta tarefa é que a análise é feita a partir dos atributos regulares, os quais são chamados de itens e, a partir deles gera-se a regra de associação. Os exemplares também recebem outro nome em associação de dados, são chamados de conjunto de itens. Os algoritmos não fazem parte de nenhum tipo de aprendizado, apenas metodologias para geração de regras (Agrawal et al., 1993; Han et al., 2006; Tan et al., 2009; Jain, 2010; Faceli et al., 2011; Witten et al., 2001; Aggarwal e Reddy, 2013).

Uma regra de associação é um conhecimento que apresenta itens que ocorrem associadamente. No contexto deste trabalho, uma associação seria, por exemplo, evidenciar que aluno aprovado na disciplina A também é aprovado na disciplina B, ou

que aluno aprovado em A e B também é aprovado na disciplina C. Ou então, aluno que acerta uma questão X na prova também acerta a questão Y e assim por diante.

A associação de dados é feita em dois momentos distintos: no primeiro verifica-se a interseção de itens e no segundo descobre-se a regra de associação. Portanto, em casos típicos os valores dos itens são binários (numéricos ou categóricos) e o que se faz é percorrer o conjunto de itens para verificar as possíveis combinações a partir de análise de interseção. Nesta fase, valores limites são definidos para a filtragem de combinações que não ocorrem minimamente. Na literatura, os dois algoritmos mais utilizados são o A Priori e FP-Growth (*Frequent Patterns Growth*) (Agrawal et al., 1993; Faceli et al., 2011). O resultado de ambos algoritmos é o mesmo, conjunto de itens que acontecem minimamente em um dado conjunto. Por outro lado, na geração de regras verifica-se a probabilidade de cada item acontecer dado que um outro ocorreu. E neste caso faz-se a verificação cruzada, ou seja, no exemplo anterior verifica-se a regra SE A ENTÃO B, e também, SE B ENTÃO A.

5.4 Pós-Processamento

Nesta última etapa do processo de descoberta de conhecimento é onde fazemos análises dos resultados. As análises podem ser quantitativas ou qualitativas dependendo da tarefa que foi escolhida.

O modelo preditivo é construído por um conjunto de dados rotulados. Estes dados podem ser separados em dois conjuntos, sendo geralmente chamados de treinamento e teste. Usa-se o conjunto de treinamento para gerar o modelo, e o conjunto de teste, por outro lado, para fazer análises quantitativas. Ou seja, apresenta-se um exemplar do conjunto de teste ao modelo, verifica-se o resultado estimado, e compara-se o resultado desejado (que está na base de dados) para quantificar desempenho. Este exemplo é uma medida típica chamada de acerto, que pode ser transformada em taxa quando multiplicamos por 100 e dividido pelo número de padrões na base de dados (Han et al., 2006; Tan et al., 2009; Faceli et al., 2011; Witten et al., 2001).

Na análise de agrupamento, por outro lado, não há valor desejado. Entretanto, mesmo assim, a quantificação pode ser feita com medidas que avaliam a coesão de um grupo e isolamento entre grupos. Mas ainda, sobretudo, a segmentação feita por um agrupamento permite também análises qualitativas. Isto deve-se ao fato que é necessário entender o perfil dos dados presentes em cada grupo formado para se definir um conceito ou categoria (Jain e Dubes, 1988; Han et al., 2006; Tan et al., 2009; Jain, 2010; Faceli et al., 2011; Witten et al., 2001).

Em regras de associação, por sua vez, o resultado é a própria regra. No entanto, a avaliação das proposições da regra é feita no momento de sua construção. Sendo assim, elas são apresentadas já com medidas que quantificam a abrangência e confiança que se pode assumir em cada regra. Mesmo assim, ainda há análises para se fazer nestas regras, que avaliam a correlação das regras geradas (Han et al., 2006).

5.5 Mineração de Dados Educacionais

Na Educação, a Mineração de Dados Educacionais ou MDE (*Educational Data Mining*) é uma área de pesquisa interdisciplinar que lida com o desenvolvimento de métodos para explorar dados oriundos de contextos educacionais (Romero e Ventura, 2010; Paiva et al., 2012). Os tipos de estudos desta área são classificados, segundo Romero e Ventura (2010) em: educação *offline* para análises em dados de desempenho do aluno, comportamento, currículo, etc., ou seja, gerados em ambientes de sala de aula; aprendizado eletrônico (*e-learning*) e Sistema de Gestão da Aprendizagem ou LMS para análise de dados armazenados em sistemas LMS no formato de log e bases de dados; e Sistemas Tutores Inteligentes ou ITS (*Intelligent Tutoring System*) e Sistemas Hipermídias Adaptativos Educacionais ou AEHS (*Adaptive Educational Hypermedia System*) os quais aplicados sobre dados de sistemas que se adaptam a cada estudante em particular, aos cursos oferecidos que estão em forma de log, aos modelos de usuários, dentre outros.

Na literatura existem diversos trabalhos que usam diferentes técnicas de Mineração de Dados no contexto educacional. Singh e Kumar (2012), por exemplo, utilizaram a técnica de mineração de dados chamada árvore de decisão para gerar conhecimento aos gestores da instituição para avaliar o desempenho de seus alunos. Dejaeger et al. (2011), por outro lado, utilizaram a técnica de mineração de dados chamada Clusterização de Dados para identificar os principais fatores de satisfação dos alunos em duas instituições de ensino e conseqüentemente para a construção de modelos para apoiar os gestores no processo de tomada de decisão estratégica.

Há duas grandes categorias de técnicas em MDE:

- **MDE Clássica:** envolve processos de predição, agrupamento e descoberta de relações
- **MDE Não-Clássica:** envolve descoberta com modelos e destilação de dados para julgamento humano.

A predição é um processo que depende de uma janela de tempo. Existem, essencialmente três janelas de tempo importantes em MDE:

- Exatamente agora ou agora: o estudante está usando o ambiente neste momento?
- Futuro próximo: um estudante que usou o ambiente há pouco tempo desenvolveu habilidades para a próxima unidade?
- Futuro: qual será a nota do estudante no próximo ENEM/ENADE?

Para o processo de predição, é necessário um conjunto de atributos de entrada e, dependendo do tipo de saída desejado (categorizada ou valorada), pode-se utilizar classificadores ou regressores. Um classificador permite uma saída categorizada e determina quais componentes e suas combinações podem predizer um determinado rótulo.

A Figura 5 exibe um exemplo de classificador:

KnowledgeComp	pknow	time	totalactions	right
ENTERINGGIVEN	0.704	9	1	WRONG
ENTERINGGIVEN	0.502	10	2	RIGHT
USEDIFFNUM	0.049	6	1	WRONG
ENTERINGGIVEN	0.967	7	3	RIGHT
REMOVECOEFF	0.792	16	1	WRONG
REMOVECOEFF	0.792	13	2	RIGHT
USEDIFFNUM	0.073	5	2	RIGHT

Figura 5. Exemplo para classificador.

Na Figura 5, há um conjunto de componentes de conhecimento (*KnowledgeComp*) que foi adquirido por um determinado aprendiz com certa probabilidade (*pknow*). A partir deste contexto, estes aprendizes foram submetidos a um conjunto de questões, que foram respondidas num determinado tempo (*time*), com um determinado número de ações (*totalactions*) e produzindo resultados corretos (RIGHT) ou incorretos (WRONG). Os resultados correspondem, exatamente, às categorias (RIGHT ou WRONG). Com o processo de predição por classificação, pode-se inferir o quanto a probabilidade de aquisição de conhecimento pode influenciar numa categoria RIGHT ou WRONG.

Já num regressor, conforme exemplo mostrado na Figura 6, ao invés de se trabalhar com categorias, utiliza-se valores numéricos:

KnowledgeComp	pknow	time	totalactions	numhints
ENTERINGGIVEN	0.704	9	1	0
ENTERINGGIVEN	0.502	10	2	0
USEDIFFNUM	0.049	6	1	3
ENTERINGGIVEN	0.967	7	3	0
REMOVECOEFF	0.792	16	1	1
REMOVECOEFF	0.792	13	2	0
USEDIFFNUM	0.073	5	2	0

Figura 6. Exemplo para regressor.

A regressão para obtenção da variável *numhints* foi obtida através da seguinte regressão linear:

$$\text{numhints} = 0.12 * \text{pKnow} + 0.932 * \text{time} + 0.11 * \text{totalactions}$$

Existem diversos regressores que podem ser utilizados no contexto de MDE: regressores lineares (como o exemplo), quadráticos, logarítmicos e exponenciais.

O agrupamento em MDE, assim como em Mineração de Dados tradicional, é utilizado quando não se conhece propriedades acerca dos dados que estão sendo minerados.

Por exemplo, considere-se o exemplo mostrado na Figura 7:

KnowledgeComp	pknow	time	totalactions	right
ENTERINGGIVEN	0.704	9	1	WRONG
ENTERINGGIVEN	0.502	10	2	RIGHT
USEDIFFNUM	0.049	6	1	WRONG
ENTERINGGIVEN	0.967	7	3	RIGHT
REMOVECOEFF	0.792	16	1	WRONG
REMOVECOEFF	0.792	13	2	RIGHT
USEDIFFNUM	0.073	5	2	RIGHT

Figura 7. Exemplo para agrupamento.

No exemplo desta figura, vamos utilizar como característica comum o tempo gasto para terminar os testes: se o tempo for maior ou igual a 10, o aprendiz pertencerá ao grupo vermelho; caso contrário, pertencerá ao grupo azul. Observe que, neste processo de agrupamento, a definição de uma característica comum foi essencial para se construir os grupos.

Na mineração de relações, o objetivo é encontrar inter-relações entre os dados que estão sendo minerados. Por exemplo, considere-se o conjunto de dados mostrado na Figura 8:

KnowledgeComp	pknow	time	totalactions	right
ENTERINGGIVEN	0.704	9	1	WRONG
ENTERINGGIVEN	0.502	10	2	RIGHT
USEDIFFNUM	0.049	6	1	WRONG
ENTERINGGIVEN	0.967	7	3	RIGHT
REMOVECOEFF	0.792	16	1	WRONG
REMOVECOEFF	0.792	13	2	RIGHT
USEDIFFNUM	0.073	5	2	RIGHT

Figura 8. Exemplo para mineração de relações.

Neste exemplo, o objetivo será prospectar a probabilidade de aquisição de uma determinada componente de aprendizado (*pknow*) e o tempo gasto (*time*) para resolver questões acerca desta componente. Neste tipo de mineração, pode-se procurar por relações lineares, quadráticas, dentre outras. É muito comum o uso de técnicas estatísticas de ajustes de dados neste contexto.

Num processo mais complexo, a MDE pode ser baseada em modelos para se realizar uma análise direcionada dos dados. Um exemplo típico desta técnica é o Rastreamento Bayesiano do Conhecimento, utilizando o modelo de aprendizagem de Corbett e Anderson.

O esquema geral deste modelo é mostrado na Figura 9:

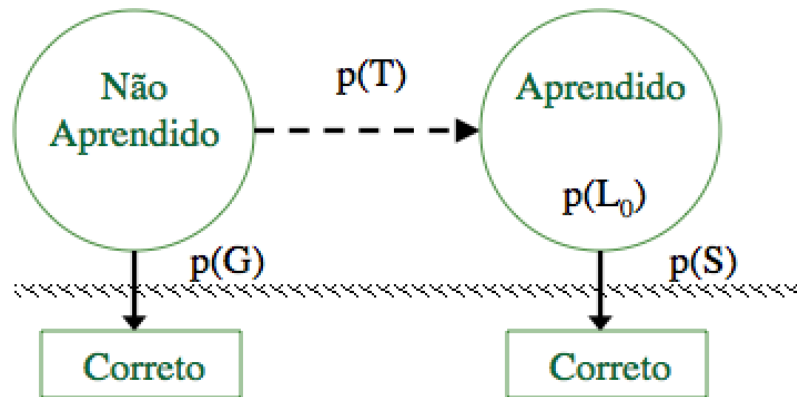


Figura 9. Esquema geral do Modelo de Corbett e Anderson.

Neste modelo, um determinado aprendiz pode estar em dois contextos: não aprendeu um determinado assunto ou aprendeu este assunto. Há uma probabilidade $p(T)$ de haver uma transição entre o estado de Não-Aprendido para o estado de Aprendido.

Em ambos os estados, trabalha-se com duas probabilidades:

- $p(G)$: probabilidade do aprendiz responder corretamente um grupo de questões dado que não aprendeu o conteúdo (por exemplo, o aluno “chutou” a resposta);
- $p(S)$: probabilidade do aprendiz responder corretamente um grupo de questões dado que aprendeu o conteúdo. Adicionalmente, há uma quarta probabilidade, $p(L_0)$, que especifica a probabilidade de um aprendiz possuir um conhecimento (um conhecimento prévio, por exemplo) no estado de aprendizado de um determinado conteúdo.

Com base neste modelo, pode-se rastrear o conhecimento numa cadeia de respostas corretas ($Correct_n$)/incorretas ($Incorrect_n$) em um determinado instante n , em função de conhecimentos anteriores (L_{n-1}) utilizando as fórmulas abaixo:

$$P(L_{n-1}|Correct_n) = \frac{P(L_{n-1}) * (1 - P(S))}{P(L_{n-1}) * (1 - P(S)) + (1 - P(L_{n-1})) * (P(G))}$$

$$P(L_{n-1}|Incorrect_n) = \frac{P(L_{n-1}) * P(S)}{P(L_{n-1}) * P(S) + (1 - P(L_{n-1})) * (1 - P(G))}$$

A primeira fórmula especifica qual a probabilidade do aprendiz ter adquirido um conhecimento anterior ($n-1$) dado que, no passo n , ele produziu uma resposta correta para uma determinada questão. De maneira análoga, interpreta-se a segunda fórmula. As

probabilidades $P(L_{n-1})$ presentes nas fórmulas poderiam ser as colunas p_{know} dos exemplos anteriores.

5.6 Consolidação e Comunicação de Resultados para DDDM

A consolidação e comunicação de dados configura-se como uma fase importante para DDDM, que permite sumarizar dados e informações através de diversas estratégias como tabelas e gráficos.

Especialmente nas tarefas de comunicação de resultados, esquemas de visualização de dados desempenham papel importante para que tomadores de decisão possam, além da observação dos próprios dados brutos e da descoberta de conhecimento proporcionada pela Mineração de Dados, utilizar diversos artefatos gráficos para complementar dados e informações para uma tomada de decisão mais segura (Mandinach e Jackson, 2012).

5.6.1 Dashboards

Um *dashboard* é uma apresentação gráfica que tem como objetivo ser fácil de ler (e entender), normalmente em uma única página e, preferencialmente, com recursos de atualização dinâmica. A Figura 10 mostra um exemplo de *dashboard*:

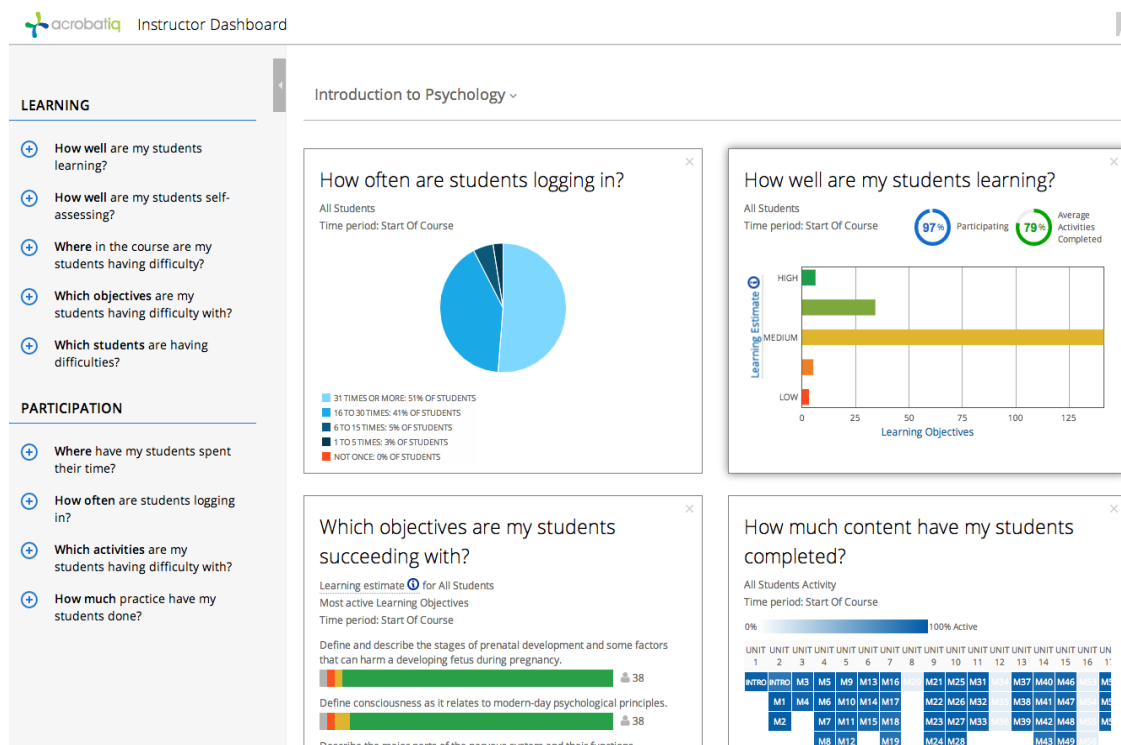


Figura 10. Exemplo de dashboard com dados de monitoração de estudantes.

No exemplo de *dashboard* acima, pode-se observar a frequência de entradas de estudantes em um determinado sistema online de ensino/aprendizagem, quais são as tarefas completadas por estes estudantes e suas taxas de sucesso, assim como sua evolução de aprendizado.

5.6.2 Heatmaps

Heatmaps são interfaces 3D, normalmente vinculadas a georeferências, onde se pode acumular dados sobre determinados tipos de eventos. Jogos educativos podem se beneficiar deste tipo de visualização, pois estes mapas permitem que as ocorrências de eventos possam retroalimentar os próprios cenários dos jogos de tal forma a ilustrar frequências, por exemplo.

A área de *Game Analytics*, ou Mineração de Dados em Jogos, permite que eventos vinculados a jogos educativos não só possam ser monitorados, assim como disponibiliza várias ferramentas para geração de *dashboards* e *heatmaps* associados a comportamentos de usuários dentro dos jogos. A Figura 11 exibe um exemplo de *heatmap* construído sobre o *game engine* Unity 3D:



Figura 11. Heatmap renderizado diretamente sobre o nível de um jogo.

Os *heatmaps* são renderizados como superfícies sobre o cenário do jogo. No exemplo acima, correspondente às regiões variando de verde a vermelho. Normalmente, quanto mais próximo do vermelho é a região, maior será a frequência de ocorrência de eventos nesta região. Indica-se, geralmente, com a cor verde ocorrências menores de eventos.

6. Tomada de Decisão em DDDM

Finalmente, tem-se a tomada de decisão em si como a última fase do processo de DDDM. Para que a tomada de decisão em DDDM possa ser efetiva, pode-se utilizar diversos elementos já discutidos anteriormente:

- os próprios dados brutos oriundos das atividades de aplicação de questionários, consultas às bases de dados e processamento de *streams*;
- o “conhecimento” descoberto com as atividades de Mineração de Dados; e
- com os artefatos gerados nas atividades de consolidação e comunicação de dados.

Existem diversas técnicas de tomada de decisão. Neste texto introdutório, optou-se por apresentar somente a técnica baseada em árvores de decisão, já bastante conhecidas de processos de classificação em Mineração de Dados e que podem ser adaptadas para contextos decisórios (Spillane, 2012).

6.1 Árvores de Decisão

Uma árvore de decisão é uma figura que mostra a sequência do processo decisório e o desdobramento das alternativas de cursos de ação e as decisões seguintes (Spillane, 2012). A Figura 12 mostra o esquema geral de uma árvore de decisão:

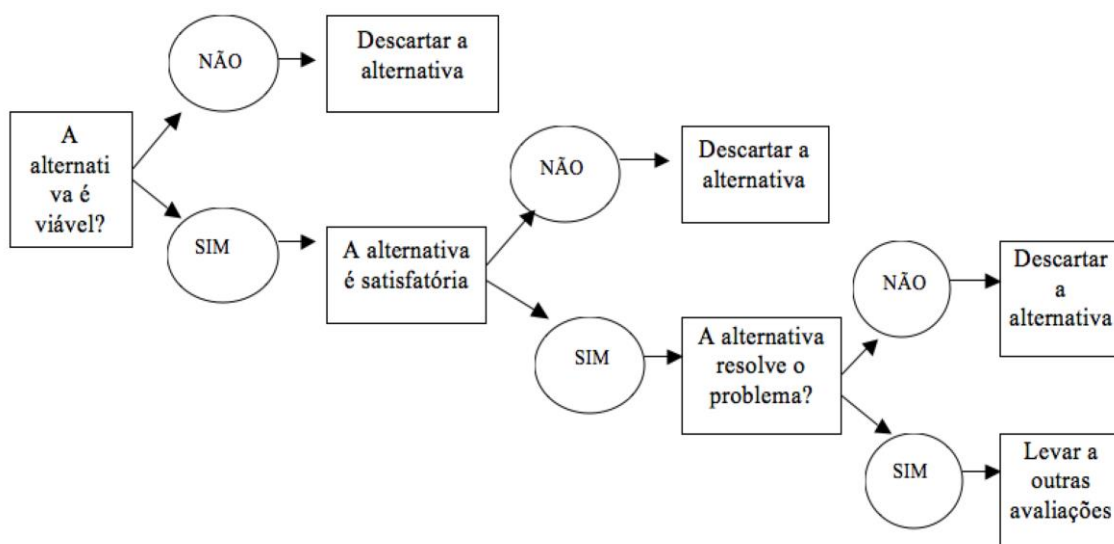


Figura 12. Esquema geral de um processo decisório baseado em árvore de decisão.

O primeiro ponto é verificar se a alternativa de decisão considerada é viável. Caso não seja, a alternativa é descartada. Caso seja viável, verifica-se se ela é satisfatória dentro do contexto de tomada de decisão. Ser satisfatório pode ser considerado um conceito *fuzzy* e pode ter rótulos qualitativos para escolha. Caso não seja satisfatória, a alternativa é descartada. Caso seja satisfatória, verifica-se se a alternativa resolve a

questão ou problema vinculado à decisão. Caso positivo, esta alternativa de decisão pode ser considerada e continua-se o processo decisório, caso necessário. Caso negativo, descarta-se a alternativa.

6.2 Decisão com Granularidade de Aluno

Como exemplo de tomada de decisão tomando-se o contexto de um aluno (granularidade de aluno) num ambiente de ensino à distância e como orientar suas ações dentro deste ambiente, construiu-se a seguinte árvore de decisão (Silva, 2002):

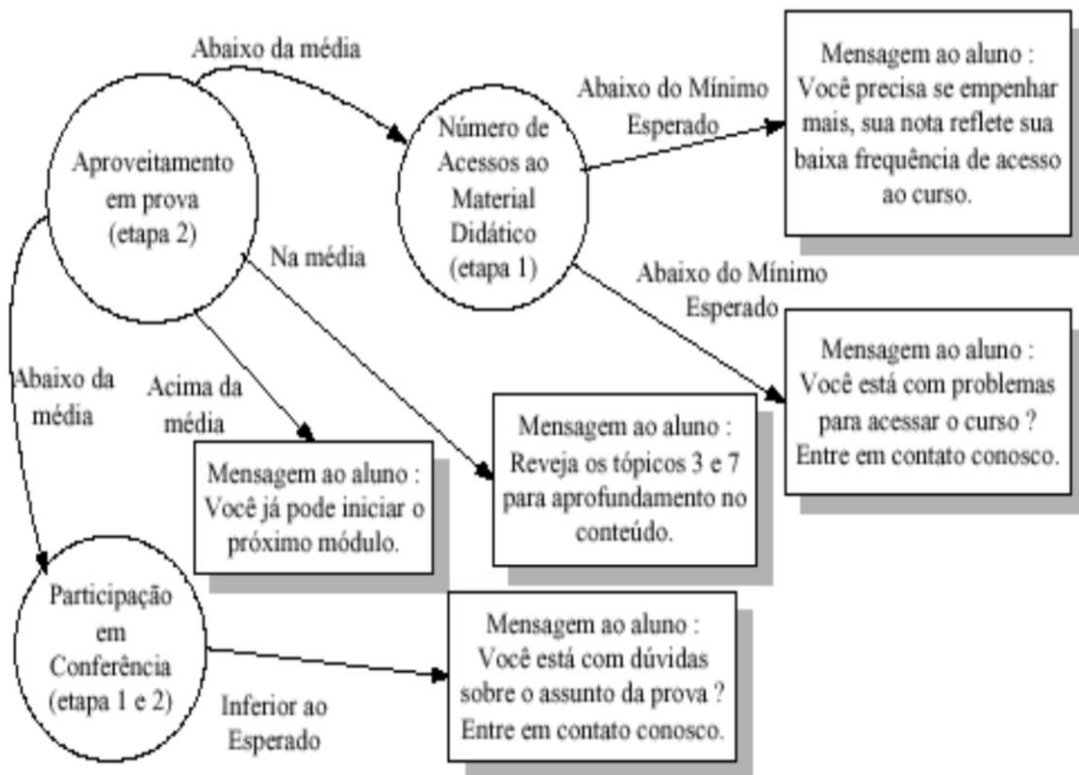


Figura 13. Exemplo de árvore de decisão com granularidade de aluno.

Neste exemplo, o objetivo é orientar o aluno dependendo do seu aproveitamento na prova da etapa 2. Para isto:

- obteve-se média geral da turma nesta prova
- obteve-se o número de acessos do aluno ao material didático da etapa 1
- obteve-se a participação do aluno nas etapas 1 e 2 do curso

Todos estes dados funcionaram como pontos decisórios em cada um dos níveis de decisão.

6.3 Decisões com Granularidade de Curso

Um exemplo típico de decisão com granularidade de curso é o próprio critério de avaliação para aprovação no curso, conforme mostrado na árvore de decisão da Figura 14, próxima página.

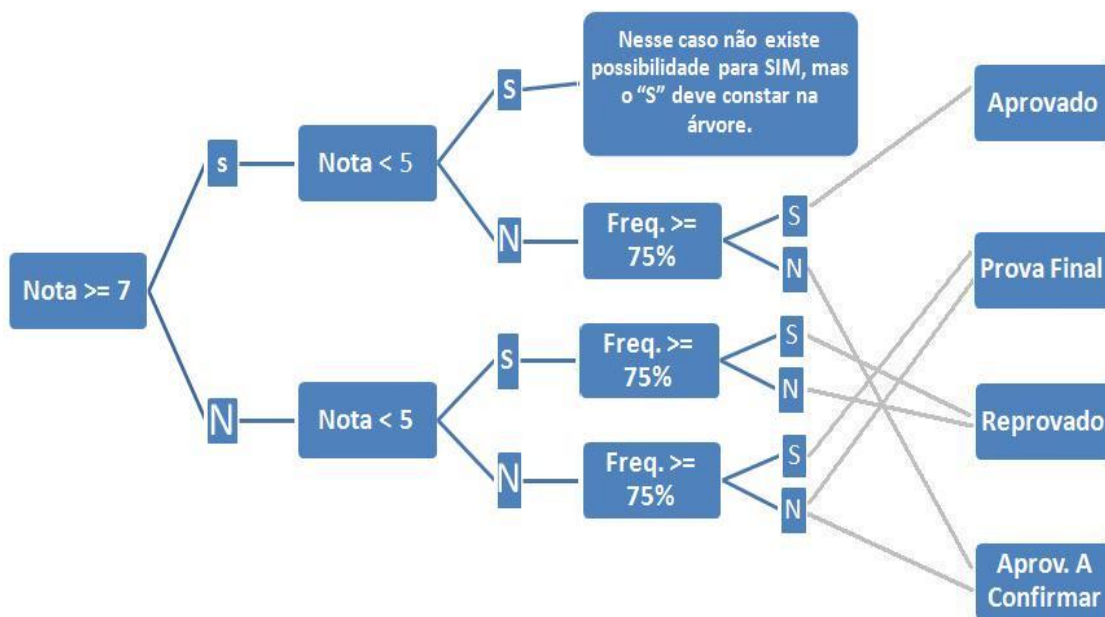


Figura 14. Árvore de decisão para aprovação em um determinado curso.

Neste exemplo, a decisão consiste em definir um critério de aprovação no curso, baseado na nota obtida no curso, assim com a avaliação frequência.

6.4 Decisões com Outras Granularidades

Outras granularidades podem ser consideradas como grupos, universidades, municípios, estados e países. Assim como ocorreu nas granularidades de alunos e cursos, pode-se

considerar pontos de decisão tanto quantitativos quanto qualitativos. Pode-se ter, também, probabilidades associadas aos pontos de decisão.

6.5 Comentários Finais

Processos decisórios correspondem a uma evolução natural dos mecanismos de Mineração de Dados. Em particular, há um campo muito vasto para pesquisa em DDDM para dados educacionais e o autor espera que este texto introdutório possa guiar o leitor nos primeiros passos dentro desta área.

Referências

- Aggarwal, C. C., Reddy, C. K. (2013) *Data Clustering: Algorithms and Applications*. USA: CRC Press.
- Agrawal, R., Imieliński, T., e Swami, A. (1993) Mining Association Rules Between Sets Of Items In Large Databases. In *ACM SIGMOD Record*(Vol. 22, No. 2, pp. 207-216). ACM.
- Bambrick-Santoyo, P. (2010) *Driven by Data: A Practical Guide to Improve Instruction*. New York: Jossey-Bass.
- Berry, M. J., Linoff, G. S. (2004). *Data Mining Techniques: For Marketing, Sales, And Customer Relationship Management*. New York: John Wiley e Sons.
- Chu, W.W. (2013) *Data Mining and Knowledge Discovery for Big Data: Methodologies, Challenge and Opportunities*. New York: Springer, 2013.
- Corrigan, M.W., Vincent, P.F. (2011) *Multi-Dimensional Education: A Common Sense Approach to Data-Driven Thinking*. New York: Corwin.
- Datnow, A. (2014) *Data-Driven Leadership*. New York: Jossey-Bass.
- Dejaeger, K., Goethals, F., Giangreco, A., Mola, L. and Baesens, B. (2011) Gaining Insight Into Student Satisfaction Using Comprehensible Data Mining Techniques. In: *European Journal of Operational Research*, Vol. 218(2), pages. 548-562.
- Dunstone, T., Yager, N. (2008) *Biometric System And Data Analysis: Design, evaluation, and data mining*. New York: Springer.
- Faceli, K., Lorena, A. C., Gama, J., Carvalho, A. C. P. L. F. (2011). *Inteligência Artificial: Uma Abordagem de Aprendizagem de Máquina*. Rio de Janeiro: Editora LTC.
- Fayyad, U., Piatetsky-Shapiro, G., Smyth, P. (1996) The KDD Process For Extracting Useful Knowledge From Volumes Of Data. In: *Communications of the ACM*, 39(11), 27-34.
- Han, J., Kamber, M. (2011) *Data Mining: Concepts and Techniques*. 3.ed. New York: Morgan Kauffman.
- Han, J., Kamber, M., PEI, J. (2006) *Data Mining: Concepts And Techniques*. New York: Morgan Kaufmann.
- Haykin, S. (2009) *Neural Networks And Learning Machines* (Vol. 3). Upper Saddle River: Pearson Education.
- Jain, A. K. (2010) Data Clustering: 50 Years Beyond K-Means. *Pattern Recognition Letters*, 31(8), 651-666.
- Jain, A. K., Dubes, R. C. (1988) *Algorithms For Clustering Data*. New Jersey: Prentice-Hall.

- Kallick, B., Colosimo, J. (2008) Using Curriculum Mapping and Assessment Data to Improve Learning. New York: Corwin.
- Mandinach, E.B., JACKSON, S.S. (2012) Transforming Teaching and Learning Through Data-Driven Decision Making. New York: Corwin.
- O’Neal, C. (2012) Data-Driven Decision Making: A Handbook for School Leaders: New York: International Society for Technology in Education.
- Paiva, R., Bittencourt I. I., Pacheco H., Silva A. P., Jacques P., Isotani S. (2012) Mineração De Dados E A Gestão Inteligente Da Aprendizagem: Desafios E Direcionamentos. In: XXXII Congresso da Sociedade Brasileira de Computação, 2012, Curitiba. Anais do DEsaflE! - I Workshop de Desafios da Computação Aplicada à Educação.
- Romero, C., Ventura, S. (2010) Educational Data Mining: A Review Of The State Of The Art. In: Systems, Man, and Cybernetics, Part C: Applications and Reviews, IEEE Transactions on, 40(6), 601-618.
- Silva, L. A., Rangayyan, R. M., Hernandez, E. D. M. (2008). Classification Of Breast Masses Using A Committee Machine Of Artificial Neural Networks. In: Journal of Electronic Imaging, 17(1), 013017-013017.
- Silva, L., Silva, L.A. (2014) Fundamentos de Mineração de Dados Educacionais. In: I Workshop de Mineração de Dados em Ambientes Virtuais de Aprendizagem, III Congresso Brasileiro de Informática na Educação, 1-23.
- Silva, D. R. (2002) Um Ambiente para Descoberta de Conhecimento com Suporte de Data Warehousing e sua Aplicação para Acompanhamento do Aluno em Educação a Distância. Dissertação de Mestrado. UFSCAR.
- Singh, S., Kumar, V. (2012) Classification of Student’s data Using Data Mining Techniques for Training & Placement Department in Technical Education. In: International Journal of Computer Science and Network, Vol. 1(4), pages.121-126.
- Spillane, J.P. (2012) Data in Practice: Conceptualizing the Data-Based Decision-Making Phenomena. American Journal of Education, 118(2), 113–141.
- Stone, Z., Zickler, T., Darrell, T. (2008) Autotagging Facebook: Social Network Context Improves Photo Annotation. In: Computer Vision and Pattern Recognition Workshops, 2008. CVPRW’08. IEEE Computer Society Conference, pp. 1-8.
- Tan, P. N., Steinbach, M., Umar, V. (2009) Introdução ao Data Mining: mineração de dados. Rio de Janeiro: Ciência Moderna.
- Witten, I. H., Frank, E., Mark, A. (2011) Data Mining: Practical Machine Learning Tools And Techniques. New York: Prentice Hall.