UNIVERSIDADE FEDERAL DO PARÁ
INSTITUTO DE CIÊNCIAS EXATAS E NATURAIS
PROGRAMA DE PÓS-GRADUAÇÃO EM CIÊNCIA DA COMPUTAÇÃO

MAXWEL MACEDO DIAS

MINERAÇÃO DE DADOS EDUCACIONAIS: RELATO DE EXPERIÊNCIA NO AMBIENTE VIRTUAL LABSQL

Belém
2014

MAXWEL MACEDO DIAS

MINERAÇÃO DE DADOS EDUCACIONAIS: RELATO DE EXPERIÊNCIA NO AMBIENTE VIRTUAL LABSQL

Dissertação de Mestrado apresentada ao Programa de Pós-Graduação em Ciência da Computação da Universidade Federal do Pará como requisito para obtenção do título de Mestre em Ciência da Computação
Área de concentração: Sistemas de Computação
Orientador Prof. Dr. Eloi Luiz Favero.

Belém
2014

Dados Internacionais de Catalogação-na-Publicação (CIP)
Sistema de Bibliotecas da UFPA

Dias, Maxwel Macedo, 1985-
Mineração de dados educacionais: relato de experiência no ambiente virtual labsql / Maxwel Macedo Dias. - 2014.

Orientador: Eloi Luiz Favero.
Dissertação (Mestrado) - Universidade Federal do Pará, Instituto de Ciências Exatas e Naturais, Programa de Pós-Graduação em Ciência da Computação, Belém, 2014.

1. Mineração de dados-Educação. 2. Ambiente virtual de aprendizagem. 3. Tecnologia educacional. 4. Educação-Efeito das inovações tecnológicas. I. Título.

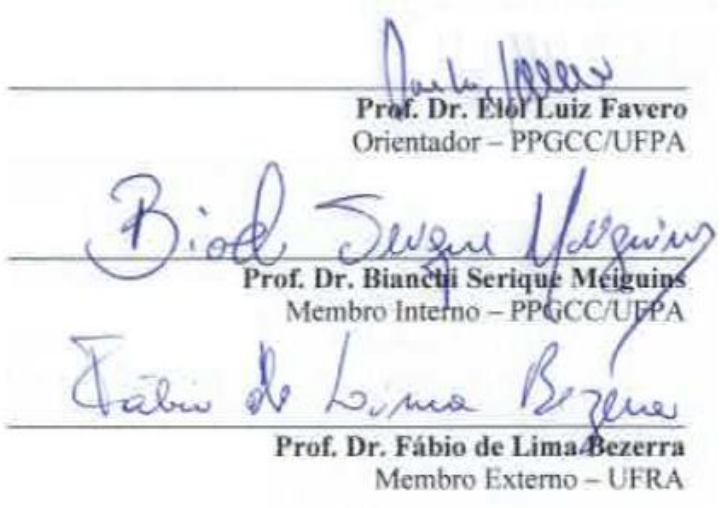
CDD 22. ed. 005.74071

UNIVERSIDADE FEDERAL DO PARÁ
INSTITUTO DE CIÊNCIAS EXATAS E NATURAIS
PROGRAMA DE PÓS-GRADUAÇÃO EM CIÊNCIA DA COMPUTAÇÃO

MAXWEL MACEDO DIAS

MINERAÇÃO DE DADOS EDUCACIONAIS: RELATO DE EXPERIÊNCIA NO AMBIENTE VIRTUAL LABSQL

Dissertação de Mestrado apresentada ao Programa de Pós-Graduação em Ciência da Computação da Universidade Federal do Pará como requisito para obtenção do título de Mestre em Ciência da Computação, defendida e aprovada em 12/09/2014, pela banca examinadora constituída pelos seguintes membros:



Prof. Dr. Eloi Luiz Favero
Orientador – PPGCC/UFPA

Prof. Dr. Bianchi Serique Meiguins
Membro Interno – PPGCC/UFPA

Prof. Dr. Fábio de Lima Bezerra
Membro Externo – UFRA

Visto:

Prof. Dr. Nelson Cruz Sampaio Neto
Coordenador do PPGCC/UFPA

À minha família, pra quem dedico toda minha vida.

AGRADECIMENTOS

Agradeço a Deus por tudo o que tem feito por mim até hoje. Por ter iluminado os meus caminhos e me ajudado a vencer grandes obstáculos.

À minha mãe Elizabeth, e ao meu pai Jairo. Aos quais devo toda a minha gratidão enquanto pessoa.

Aos meus tios Jackson e Nazaré, a minha irmã Marília, a meu cunhado Nivaldo e as minhas primas Juliana e Luciana que me deram total apoio durante essa jornada.

À minha querida Ana Carla que sempre me apoiou e soube me compreender acima de tudo.

Ao Prof. Eloi Luiz Favero pelos ensinamentos, pelo importante e fundamental apoio e orientação.

Ao Prof. Adriano Del Pino Lino pelo apoio e incentivo.

Aos meus grandes e prestativos amigos Luiz Alberto, Suelene de Jesus, Luciléia Rosa e Tácio Vinícius, que estiveram ao meu lado durante essa jornada. Meu muito obrigado a todos.

Gostaria de citar o nome das várias pessoas que permaneceram comigo e me ajudaram a dar prosseguimento a esta etapa da minha vida, mas fica aqui para aqueles que eu não citei o meu sincero e profundo agradecimento.

RESUMO

Uma das tecnologias digitais mais utilizadas nas atuais práticas de educação *online* é o Ambiente Virtual de Aprendizagem (AVA). Os educadores utilizam estes ambientes para disponibilizar informações *online*, mas possuem pouco suporte para avaliar o aprendizado dos educandos a distância, de forma que, a ausência da percepção do educador quanto ao estado de compreensão de seus educandos pode levar ao insucesso de um curso *online*.

A maioria dos AVAs armazenam grandes volumes de dados provenientes do histórico dos acessos aos recursos do sistema pelos educandos, suas avaliações, dentre outros. Nos últimos anos a Mineração de Dados Educacionais vem sendo utilizada para explorar os dados provenientes de ambientes educacionais, bem como entender melhor os educandos e o seu processo de ensino e aprendizagem.

O objetivo deste trabalho é avaliar o aprendizado *online* a partir dos dados provenientes do ambiente virtual LabSQL, utilizado na Universidade Federal do Pará, por meio da Mineração de Dados Educacionais, com a aplicação de técnicas como Árvore de Decisão, Redes Bayesianas, Regras de Associação e Análise de Agrupamento.

Os resultados obtidos mostraram-se eficientes para apoiar os educadores na avaliação das aprendizagens online, pois permitem analisar o perfil dos educandos em relação à utilização dessa tecnologia e ao processo de ensino-aprendizagem no ambiente LabSQL. Além disso, as regras geradas a partir da mineração de dados indicam como o educando pode aprimorar a aprendizagem utilizando melhor o ambiente.

PALAVRAS-CHAVES: Mineração de dados - Educação, Ambiente virtual de aprendizagem, Tecnologia educacional, Educação - Efeito das inovações tecnológicas.

ABSTRACT

One of the digital technologies used in current practices of online education is the Virtual Learning Environment (VLE). Educators use these environments to provide online information, but have little support to assess the learning of learners at a distance, so that the lack of perception of the educator as to the state of understanding of their students can lead to failure of an online course.

Most VLEs store large volumes of data from the history of accesses to system resources made by students, their assessments, among others. In recent years the Educational Data Mining has been used to explore the data from educational settings, as well as better understand the students and their teaching and learning.

The objective of this work is to assess online learning by the use of Educational Data Mining on the LabSQL virtual environment used in the Federal University of Pará, through the application of techniques called Decision Tree, Bayesian Network, Association Rules and Cluster Analysis.

The results obtained proved to be efficient to support educators in the assessment of online learning because they allow the analysis of student's profile regarding the use of this technology and the teaching-learning environment in LabSQL. Moreover, the rules generated from data mining indicate how the student can improve learning by better using the environment.

KEYWORDS: Data Mining - education, Virtual learning environment, Educational technology, Education - Effect of technological innovations.

LISTA DE ILUSTRAÇÕES

LISTA DE QUADROS

LISTA DE SIGLAS

AVA	Ambiente Virtual de Aprendizagem
ARFF	*Attribute-Relation File Format*
DCL	*Data Control Language*
DDL	*Data Definition Language*
EaD	Educação a Distância
MDE	Mineração de Dados Educacionais
KDD	*Knowledge Discovery in Databases*
LabSQL	Laboratório de Ensino e Aprendizagem de SQL
MEC	Ministério da Educação
PDF	*Portable Document Format*
SQL	*Structured Query Language*
TIC	Tecnologia de Informação e Comunicação
TPC	Tabelas de Probabilidade Condicional
UAB	Universidade Aberta Brasil
UFPA	Universidade Federal do Pará

SUMÁRIO

CAPÍTULO 1
INTRODUÇÃO

O Capítulo 1 apresenta uma visão geral da dissertação e os objetivos da pesquisa.

1.1 INTRODUÇÃO

A educação na modalidade *online* é uma realidade cada vez mais reconhecida e globalizada. As ofertas da educação *online* têm sido ampliadas devido à rápida evolução das Tecnologias de Informação e Comunicação (TIC), assim como das possibilidades oferecidas pelas mesmas. Para Santos (2010), a educação *online* é o conjunto de ações de ensino e aprendizagem mediados por interfaces digitais que potencializam práticas comunicacionais interativas e hipertextuais.

No Brasil há um crescimento exponencial da oferta de cursos *online*. Inicialmente, a Portaria do MEC n.° 2.253, de 18 de outubro de 2001, conhecida como "Portaria dos 20%", veio garantir às instituições de ensino superior a opção de oferecer até 20% de suas disciplinas regulares na modalidade à distância, que transita dos suportes tradicionais para a Internet. Pouco tempo depois, vieram a Portaria n.°4.059/2004 e o Decreto n.°5.622/2005 que ampliam muito mais os horizontes para a modalidade educacional a distância (o impresso via correio, rádio e a televisão) e para modalidade *online* (o computador e a Internet) (SILVA, 2010). Mais recentemente, destaca-se o investimento do governo brasileiro no Programa Universidade Aberta do Brasil (UAB)[1] que oferece cursos de formação universitária com prioridade para modalidade *online*, com a finalidade de expandir e interiorizar a oferta de cursos e programas de educação superior no País.

Segundo Santos (2010) uma das tecnologias digitais mais utilizadas nas atuais práticas de educação *online* são os Ambientes Virtuais de Aprendizagem, que constituem um conjunto de recursos para favorecer a aprendizagem, ou mesmo, um espaço virtual para a colaboração no qual se pode propor aos educandos um conjunto de atividades ou propostas de aprendizagem (SANCHO, 2010). Estes ambientes virtuais têm o propósito de apoiar classes de usuários a partir da Internet, sendo útil para usuários que não residem perto de instituições de ensino ou não dispõem de horários regulares para estudar. Além disso, eles também são uma importante ferramenta complementar para os cursos presenciais.

Os educadores deste novo processo de aprendizagem utilizam estes ambientes e ferramentas para disponibilizar informações *online*, porém possuem pouco suporte para

[1] http://uab.capes.gov.br

avaliar e discriminar os diferentes comportamentos das ações dos educandos sobre o AVA durante a realização dos cursos (ZAIANE e LUO, 2001).

Conforme Masetto (2000), um dos grandes problemas da educação *online* está na dificuldade de avaliar o aprendizado dos educandos à distância, mediado pelas tecnologias digitais. Esta dificuldade se justifica, entre outros fatores, pela falta de contato presencial entre educadores e educandos, visto que a ausência da percepção do educador, quanto ao estado de compreensão de seus educandos, pode levar ao insucesso de um curso *online*.

A maioria dos AVAs possui sistemas de registro automático dos percursos dos usuários, armazenando grandes volumes de dados gerados pela interação de educadores e educandos. Geralmente tais dados são provenientes do histórico dos acessos aos recursos do sistema pelos educandos, suas avaliações, comunicação (*chat* e *e-mail*) entre educandos e entre educandos e educadores, tempo de utilização do sistema, dentre outros. A existência destes registros assume grande importância no processo de monitoração do percurso dos educandos e pode ser crucial para a identificação antecipada, por parte do educador, de casos de potencial desmotivação e potencial abandono, revelados por um baixo nível de consultas dos materiais disponíveis, poucas entradas no sistema e poucas participações nos espaços de discussão. A identificação precoce destas situações permite que o educador possa agir de imediato junto do educando no sentido de resolver eventuais problemas e estimulá-lo ao envolvimento e à participação no curso (GOMES, 2010).

Recentemente, muitos pesquisadores da área de Informática na Educação (em particular, Inteligência Computacional Aplicada à Educação) têm mostrado interesse em utilizar mineração de dados para investigar perguntas científicas na área de educação (BAKER *et. al*., 2011). Conhecida como Mineração de Dados Educacionais (MDE), esta é uma disciplina emergente, preocupada com o desenvolvimento de métodos para explorar os tipos de dados provenientes de ambientes educacionais, bem como entender melhor os educandos e o seu processo de ensino e aprendizagem (INTERNATIONAL EDUCATIONAL DATA MINING SOCIETY, 2011).

Atualmente a MDE vem se estabelecendo como uma forte e consolidada linha de pesquisa que possui grande potencial para melhorar a qualidade do ensino (BAKER *et. al*., 2011). Embora ela seja uma área de pesquisa relativamente recente, há um número importante de contribuições publicadas (ROMERO e VENTURA, 2007).

Cortez e Silva (2008) abordam o desempenho dos educandos no ensino secundário utilizando técnicas de mineração de dados, como Árvore de Decisão e Redes Neurais Artificiais. Como um resultado direto desta pesquisa, ferramentas de previsão mais eficientes podem ser desenvolvidas, melhorando a qualidade da educação e reforçando o gerenciamento de recursos escolares. Wang e Meinel (2007) descobrem o interesse de aprendizagem dos educandos a partir de dados extraídos de um AVA, usando algumas técnicas de mineração de dados, como regras de associação. Os resultados obtidos ajudaram os educadores a avaliar os seus educandos de forma mais precisa e ajustar melhor suas programações de ensino. Observou-se, por exemplo, que o tempo médio de um educando para assistir uma palestra é de aproximadamente 10 minutos, enquanto a duração normal de uma palestra é de cerca de 90 minutos. Isso sugere que o educador deve segmentar toda a videoconferência em pedaços pequenos e organizá-los em uma rede semântica pesquisável, o que pode ajudar os educandos a encontrar o conhecimento adequado e relacionado durante a aprendizagem. Chen *et. al.*(2007) utilizou a técnica de mineração de dados, denominada análise de agrupamento para estudar o comportamento de aprendizagem dos educandos, tais como desempenhos em tarefas e testes. Eles examinaram também os registros de aprendizagem *online*, a fim de proporcionar aos professores um meio para observar os educandos durante o processo.

Em resumo, o estado da arte apresenta a mineração de dados como alternativa útil para a descoberta de conhecimento não trivial a partir dos dados dos AVAs. Adicionalmente, observa-se uma variedade de aplicabilidade de suas técnicas e tarefas para avaliar a aprendizagem *online*. Entretanto, não existe uma técnica específica de mineração de dados a ser aplicada em qualquer conjunto de dados educacionais, pois cada AVA possui particularidades que implicam diretamente na escolha de uma técnica. Associado a isso, existem técnicas de mineração de dados mais apropriadas para realizar determinadas tarefas relacionadas ao estudo da aprendizagem *online*. Por isso, existe a necessidade de mais investigação em relação às possibilidades de aplicação das técnicas de mineração de dados para analisar o processo de ensino-aprendizagem em AVAs. A comunidade de MDE vem crescendo rapidamente, contudo, no Brasil ainda são poucos os trabalhos publicados nesta área de pesquisa (BAKER *et. al.*, 2011).

Este trabalho propõe avaliar o aprendizado *online* a partir da Mineração de Dados Educacionais do AVA denominado LabSQL[2] utilizado na educação *online* da Universidade Federal do Pará - UFPA, para prever o desempenho dos educandos e gerar informações relevantes sobre o perfil dos educandos em relação à utilização dessa tecnologia e ao processo de ensino-aprendizagem, que possam apoiar os educadores na avaliação da aprendizagem *online*.

Para a obtenção desses resultados, são utilizadas as técnicas de mineração de dados denominadas Regras de Associação (*Apriori*), Análise de Agrupamento (*K-means*), Árvore de Decisão e Redes Bayesianas, pois elas podem, respectivamente, verificar associações entre características distintas dos educandos, agrupar um conjunto de educandos segundo suas características, prever o desempenho obtido pelos educandos a partir do seu perfil e contabilizar as relações de dependência entre as ações envolvidas no processo de aprendizagem, conforme observado em estudo preliminar (DIAS *et. al.*, 2008; DIAS *et. al.*, 2011).

Os resultados obtidos mostraram-se eficientes para apoiar os educadores na avaliação das aprendizagens *online*, pois permitem analisar o perfil dos educandos em relação à utilização dessa tecnologia e ao processo de ensino-aprendizagem no ambiente LabSQL. Além disso, as regras geradas a partir da mineração de dados indicam como o educando pode aprimorar a aprendizagem utilizando melhor o ambiente.

A principal contribuição deste trabalho foi apoiar o desenvolvimento da área de MDE no Brasil, analisando dados provenientes de um ambiente virtual de aprendizagem especializado no ensino da Linguagem de Consulta Estruturada SQL (*Structured Query Language*), visando a avaliação da aprendizagem *online* para fornecer auxílio aos educadores e proporcionar um aprendizado mais personalizado e de melhor qualidade.

1.2 RELEVÂNCIA DO TRABALHO

A crescente quantidade de educandos em cursos na modalidade *online* cria oportunidades excelentes para a pesquisa na área de MDE e pode, futuramente, beneficiar de forma significativa o processo de ensino e aprendizagem no Brasil. Para que a Educação *Online* e a MDE tenham impacto na sociedade brasileira é necessário

[2] O LabSQL é um ambiente virtual de ensino e aprendizagem da linguagem de manipulação de banco de dados SQL, utilizado nas disciplinas de banco de dados nos cursos de graduação e pós-graduação. Essa ferramenta é apresentada com mais detalhes na seção 2.2.1 (*Ambiente Virtual de Aprendizagem LabSQL).*

que os pesquisadores e educadores comecem a utilizar os dados obtidos em ambientes virtuais de aprendizagem de forma estruturada e com objetivos bem definidos (BAKER *et. al*., 2011).

Neste contexto, este trabalho apresenta grande relevância devido à baixa quantidade de estudos sobre o tema no Brasil. Além disso, o desenvolvimento de pesquisas no país que aplicam a MDE para avaliar a aprendizagem *online* em AVAs como o LabSQL, assume grande importância devido a expansão de projetos de EaD, a exemplo do programa UAB, e pela inserção maciça de TIC na educação, independentemente da modalidade ou nível de ensino.

Os resultados obtidos a partir da aplicação das técnicas de MDE no ambiente LabSQL possibilitam avaliar os processos de ensino-aprendizagem, e oferecem ao educador a possibilidade de detectar possíveis problemas na aprendizagem dos educandos, que por sua vez, possuem maiores possibilidades para aprendizagens mais efetivas. Além disso, o desenvolvimento deste trabalho abre oportunidades de pesquisa para outros AVAs, beneficiando o processo de avaliação e permitindo a evolução dos mesmos.

1.3 OBJETIVOS DA DISSERTAÇÃO

O objetivo geral desta dissertação é avaliar o aprendizado *online* a partir dos dados provenientes do AVA denominado LabSQL, utilizado na educação *online* da Universidade Federal do Pará, por meio da mineração de dados educacionais.

Os objetivos específicos desta pesquisa são:

- Apresentar os principais conceitos e aplicações relacionadas à área de MDE encontradas na literatura;
- Prever o desempenho dos educandos em exercícios e avaliações de aprendizado, com base na análise das atividades realizadas durante a utilização do LabSQL, o que permitirá o planejamento e a reação apropriada a algum problema relacionado à utilização dessa tecnologia e ao processo de ensino-aprendizagem;
- Identificar os educandos com desempenhos similares que podem receber conteúdo personalizado e participar de atividades mais focadas em suas necessidades;

- Verificar se os acertos nos exercícios e o desenvolvimento de trabalhos em equipe influenciam os acertos nas avaliações;
- Gerar informações relevantes sobre o perfil dos educandos que indiquem quais os recursos disponíveis no LabSQL (SQL-Livre, Material de Apoio, Exemplos de SQL) e práticas (quantidade de tentativas de resolução, quantidade de acessos) que afetam a aprendizagem, para manter o desempenho das características que estão acima da média e melhorar aquelas abaixo da média.

1.4 ORGANIZAÇÃO DO TEXTO

Este trabalho constitui-se, além desta introdução, de mais quatro capítulos distribuídos como segue.

O Capítulo 2 apresenta alguns conceitos de educação *online* e descreve os ambientes virtuais de aprendizagem, mais especificamente, o AVA LabSQL. Este capítulo aborda a avaliação das aprendizagens *online* e algumas de suas problemáticas.

O Capítulo 3 apresenta os conceitos e métodos da MDE, fundamentais para o entendimento desta pesquisa, como predição, agrupamento e mineração de regras de associação, entre outros.

O Capítulo 4 descreve a mineração de dados educacionais do ambiente virtual LabSQL com a aplicação das técnicas de mineração de dados: Árvore de decisão, Redes Bayesianas, Regras de Associação e Análise de Agrupamento. Além disso, neste capítulo é realizada a análise dos resultados obtidos para apoiar os educadores no acompanhamento das aprendizagens *online*, utilizando as regras geradas com pela MDE.

E, por fim, o Capítulo 5 apresenta a conclusão e as contribuições deste trabalho, a indicação de trabalhos futuros e as considerações finais desta dissertação.

CAPÍTULO 2

EDUCAÇÃO *ONLINE*

O Capítulo 2 tem como objetivo apresentar alguns conceitos que serão úteis para a compreensão da área da educação *online*, bem como apresentar em detalhes o AVA LabSQL.

2.1 CONCEITOS DE EDUCAÇÃO *ONLINE*

A educação na modalidade *online* vem crescendo junto com a web e com o desenvolvimento das tecnologias digitais. Segundo Santos (2010), educação *online* é o conjunto de ações de ensino e aprendizagem mediados por interfaces digitais que potencializam práticas comunicacionais interativas e hipertextuais.

Ainda que a expressão consolidada seja "Educação a Distância" ou "EaD", fazer educação *online* não é o mesmo que efetuar a conhecida modalidade via suporte analógicos unidirecionais, como o impresso, o rádio e a televisão. A modalidade "a distância" é operada por meios de transmissão em sua natureza, já a modalidade "*online*" utiliza os recursos favoráveis à interatividade cada vez mais presentes e em sintonia com a evolução da *web* na direção dos ambientes de comunicação e colaboração (SILVA, 2010).

Enquanto a modalidade "a distância", via meios unidirecionais separa emissão e recepção no tempo e no espaço, a modalidade *online* exige metodologia própria porque o suporte digital *online* conecta professores e educandos nos tempos síncrono e assíncrono, dispensa o espaço físico, favorece a convergência de mídias e contempla bidirecionalidade, multidirecionalidade, estar-junto "virtual" em rede e colaboração (SANTOS, 2010).

As tecnologias digitais mais utilizadas nas atuais práticas de educação *online* são os ambientes virtuais de aprendizagem, as teleconferências e as videoconferências.

2.2 AMBIENTE VIRTUAL DE APRENDIZAGEM

AVA é um sistema que fornece suporte a qualquer tipo de atividade realizada pelo educando, isto é, um conjunto de ferramentas que são utilizadas em diferentes situações do processo de aprendizagem (MARTINS e CAMPESTRINI, 2004). Os AVAs agregam uma das características fundamentais da Internet: a convergência de mídias, ou seja, a capacidade de hibridizar e permutar várias mídias em um mesmo ambiente.

Os ambientes virtuais de aprendizagem envolvem não só um conjunto de interfaces para socialização de informação e de conteúdos de ensino aprendizagem, mas também, sobre tudo, as interfaces de comunicação síncronas e assíncronas. As interfaces de conteúdos são os dispositivos que permitem produzir, disponibilizar e compartilhar

conteúdos digitalizados em diversas linguagens: texto, som e imagem. E as interfaces de comunicação são aquelas que contemplam troca de mensagens entre os interlocutores do grupo ou da comunidade de aprendizagem. Estas podem ser síncronas, quando contemplam a comunicação em tempo real (exemplos: *chats*, webconferências, entre outras), ou assíncronas, quando permitem a comunicação em tempos diferentes (exemplos: fóruns, listas de discussão, *blogs* e *wikis*, entre outras).

Além da autoaprendizagem, as interfaces dos AVAs permitem a interatividade e a aprendizagem colaborativa. O educando aprende com o material didático e na dialogia com outros sujeitos envolvidos (educadores, tutores e outros educandos) por meio de processos de comunicação síncronos e assíncronos (SANTOS, 2010).

2.2.1 AMBIENTE VIRTUAL DE APRENDIZAGEM LABSQL

O LabSQL é um ambiente interativo para auxiliar os educandos no aprendizado da linguagem de consulta estruturada SQL (*Structured Query Language*) e pode ser utilizado como ferramenta de apoio ao mediador para realizar automaticamente as avaliações nas atividades de laboratório (LINO *et. al*, 2007). Para os educandos envolvidos em cursos de educação *online*, esta possibilidade pode ajudá-los a aferirem o grau das suas aprendizagens, ajudando-os a regular o seu percurso de estudo e aprendizagem. Na avaliação do ambiente, Lino (2007) verificou que 96,67% dos educandos consideram que o LabSQL aumenta o aprendizado em relação ao processo de ensino-aprendizagem tradicional.

No ambiente LabSQL existem três tipos de exercícios: objetivos de múltipla escolha (ou V/F), subjetivos e exercícios de programação SQL. No momento em que o educando interage com o sistema e envia sua consulta SQL, o sistema executa e avalia a complexidade desta consulta em relação à consulta do mediador. Dessa forma, o educando recebe um retorno automático, contendo: o resultado da consulta, permitindo avaliar se a resposta está correta ou não; a avaliação automática da resposta do educando, levando-se em consideração o resultado da execução e o grau de complexidade comparado com a resposta do mediador; o número de tentativas e a avaliação global da avaliação ou exercício.

Além disso, o LabSQL disponibiliza o recurso SQL-Livre, interpretador de SQL embutido no ambiente, que fornece ao educando três opções: um ambiente de treinamento de comandos SQL, no qual os resultados dos comandos SQL são

visualizados; os históricos dos comandos executados; e os exemplos do conteúdo previamente selecionado. Todos os comandos SQL são salvos, permitindo assim estudos futuros em relação ao desempenho dos educandos.

Na Figura 2-1, está representada a arquitetura geral do LabSQL. Nela, observa-se a interface de mediação, que é utilizada pelo educador para definir as avaliações e questões e algumas soluções associadas. As questões podem ser disponibilizadas apenas para treinamento ou para a avaliação formal dos educandos e a seleção das mesmas pode ser feita previamente pelo educador ou a partir de um sorteio entre as questões armazenas da base de avaliações e questões, que é feito isoladamente para cada educando. No segundo caso, cada educando terá uma alta probabilidade de ter uma lista de questões bastante distinta dos demais educandos.

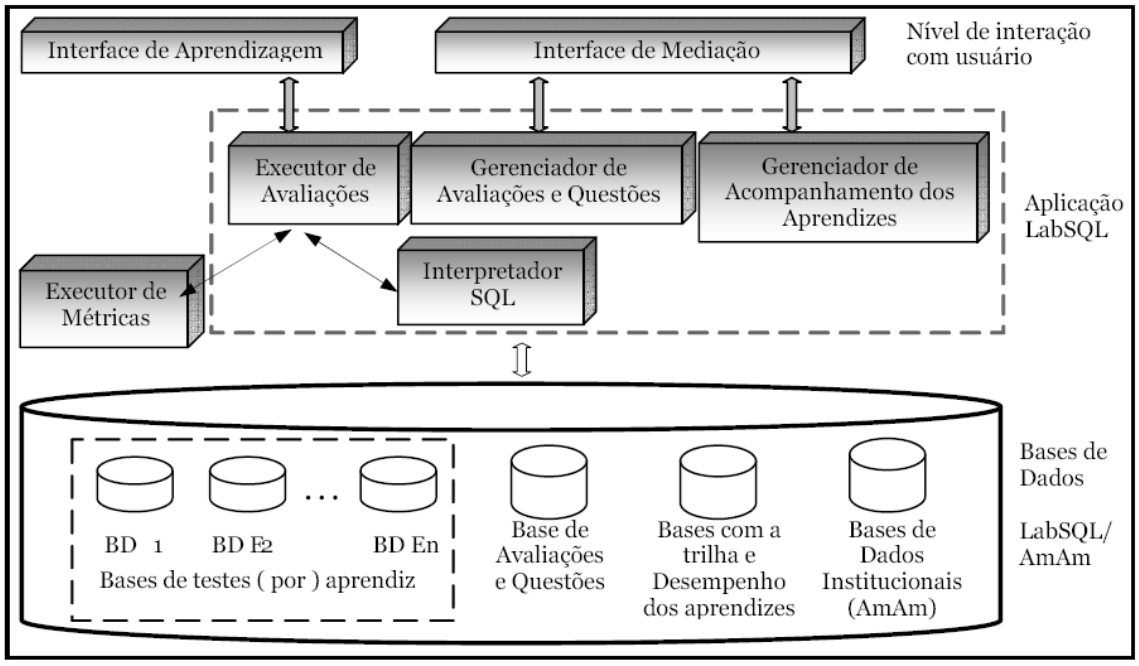


Figura 2-1 - Visão Geral da Arquitetura do LabSQL, LINO *et. al.*, (2007).

A interface de aprendizagem é utilizada pelos educandos para resolver as questões selecionadas anteriormente. Ao enviar uma questão, a requisição passa pelo executor de avaliações, que por sua vez aciona o interpretador SQL. O interpretador SQL retorna o resultado da consulta feita pelo educando e a compara com a base de testes daquele educando. Caso as consultas retornem os mesmos resultados, a consulta do educando é avaliada automaticamente pelo executor de métricas. Todos os erros e acertos são registrados nas bases com a trilha e desempenho dos aprendizes. Na base de dados institucionais do AVA persistem informações referentes aos cursos, educando e educadores.

Além do *feedback* para o aprendiz, é gerado um relatório detalhado para o mediador, contendo as informações de cada aprendiz e da turma em geral; permite visualizar a avaliação de cada questão resolvida por aprendiz e identificar os aprendizes com dificuldade de concluir os exercícios. Por exemplo, o ambiente mostra os educandos que já tentaram mais de 10 vezes. A partir dessa interface, o mediador pode enviar comentários associados às questões de cada aprendiz.

No relatório de avaliação, o mediador tem uma visão geral do andamento da turma em relação às avaliações cadastradas (listas de exercícios e avaliações). Este relatório tem como objetivo visualizar um *ranking* dos aprendizes por turma; facilitar o planejamento do tempo necessário para os aprendizes concluírem os exercícios e identificar grupos de aprendizes mais (ou menos) adiantados para propor exercícios em equipe.

O conteúdo do LabSQL é apresentado em 5 módulos, onde o grau de dificuldade aumenta do primeiro para o quinto. Porém, o ambiente promove bastante flexibilidade em relação à sequência de apresentação do conteúdo, pois os educandos não são obrigados a segui-lo em ordem pré-estabelecida. O Módulo I introduz os conceitos básicos de bancos de dados e da linguagem SQL; o Módulo II introduz o comando *select* e os operadores aritméticos e lógicos utilizados na linguagem; o Módulo III apresenta os conceitos da Linguagem de Definição de Dados (DDL- *Data Definition Language*) e da Linguagem de Controle de Dados (DCL- *Data Control Language*); o Módulo IV apresenta as funções de agregação; o Módulo V apresenta o conceito de subconsultas. Os módulos estão dispostos na interface de aprendizagem em formato de árvore, como mostrado na Figura 2-2.

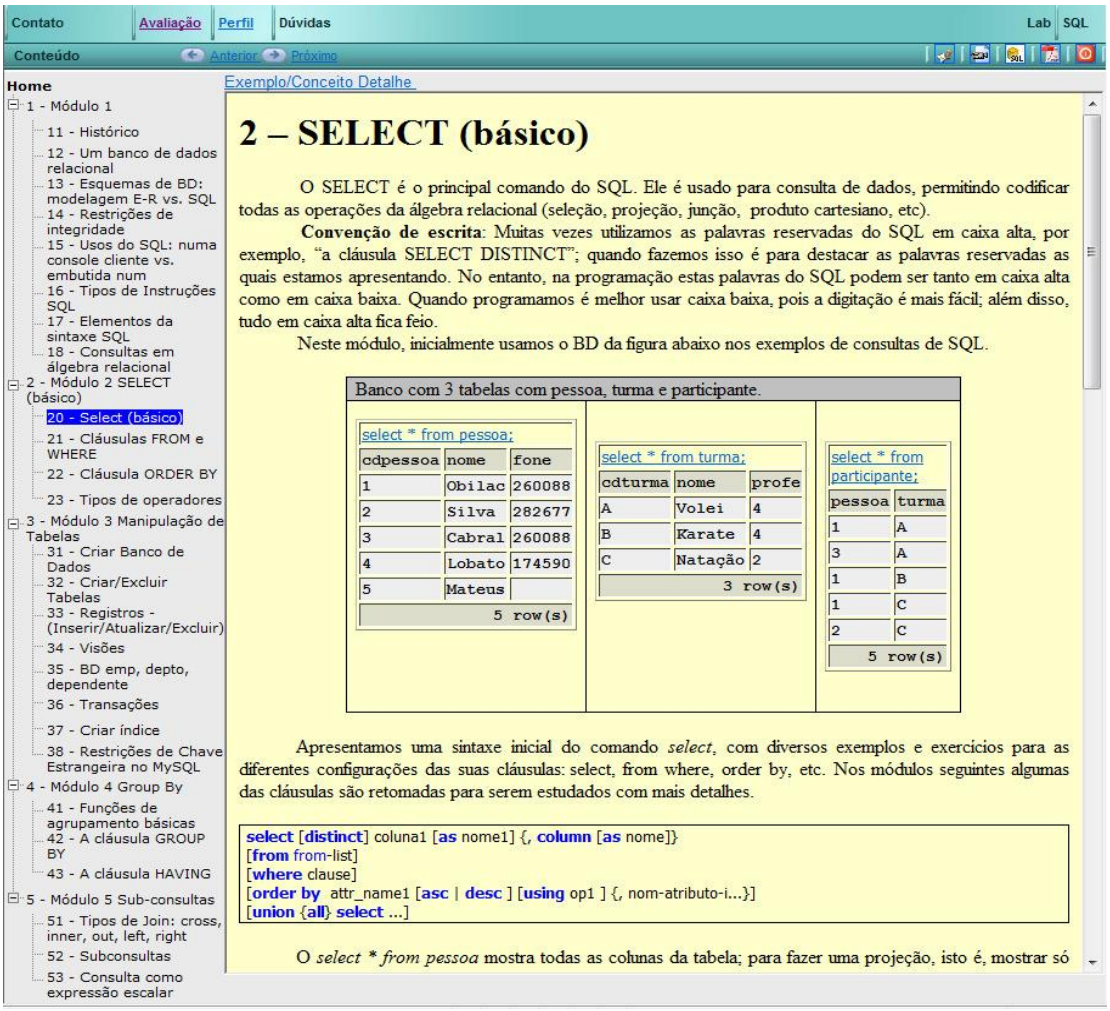


Figura 2-2 - Organização dos Módulos no LabSQL (LINO, 2007).

O LabSQL apresenta um grande número de recursos para os educandos e educadores cadastrados no sistema, possibilitando uma interação intensa entre os usuários e o sistema. Dentre os recursos presentes no ambiente, destacam-se:

i. **Fórum:** permite uma comunicação entre todos os participantes do LabSQL;

ii. **Analisar resultado**: exibe o desempenho do aprendiz quanto à realização das listas de exercícios, participação, frequência e aproveitamento nas provas;

iii. **Administrar Questão**: relatório que contém questões cadastradas por equipe e disponível no formato PDF (*Portable Document Format*) para impressão;

iv. **Administrar Usuário**: esse recurso permite ao aprendiz editar seus dados cadastrais;

v. **Material de Apoio**: materiais disponíveis para os aprendizes que são inseridos pelo professor. Atualmente contém um arquivo compactado com várias apresentações de Banco de Dados e *links* para o conteúdo no formato PDF e o vídeo de introdução ao sistema;

vi. **Relatório de Desempenho/Acessos**: o aprendiz tem a opção de visualizar os seguintes relatórios gráficos: acesso por usuário e usuário *online*, apresenta a mesma funcionalidade da área do professor; desempenho do educando, identifica seu progresso na avaliação a partir do gráfico de *Gantt*[3] interativo;

vii. **Avaliação**: é apresentada ao aprendiz quando existe uma prova ou lista de exercício;

viii. **Exercício**: fica disponível quando o professor associa questões a uma determinada sessão;

ix. **Agenda**: fornece um espaço para o aprendiz realizar qualquer anotação, funciona como um caderno no qual o aprendiz tem a liberdade de escrever, reescrever ou apagar uma informação.

2.3 AVALIAÇÃO DA APRENDIZAGEM *ONLINE*

Segundo Hoffmann (1998), avaliar vai muito além de verificar o desempenho do educando. Para o autor, o processo avaliativo é um método investigativo que pressupõe que o educador esteja cada vez mais alerta e se debruce compreensivamente sobre todas as manifestações do educando. Na essência uma avaliação tem a função de acompanhar a evolução dos educandos, detectando possíveis problemas durante o processo de ensino e procurando suprir suas deficiências.

"No paradigma educacional centrado nas aprendizagens significativas" (apoiado na pedagogia diferenciada e da autonomia), a avaliação é concebida como processo de coleta de informações, sistematização e interpretação das informações, julgamento de valor do objeto avaliado a partir das informações tratadas e decifradas, e, por fim, tomada de decisão (como intervir para promover o desenvolvimento das aprendizagens significativas) (SILVA, 2012). Neste contexto, a avaliação se materializa numa variedade de instrumentos avaliativos que tem uma função estratégica na coleta de um

[3] Gráfico de Gantt é uma ferramenta simples, inventada em 1917 por Henry L. Gantt (1861-1919), que representa o tempo a partir de barras horizontais.

maior número de variedade de informações sobre os percursos das aprendizagens e das histórias de vida dos educandos e das intervenções e das posturas dos educadores.

A avaliação deve ser contínua e considerar o envolvimento efetivo dos educandos nas diversas atividades propostas e desenvolvidas, tendo por suporte os diversos serviços disponíveis *online*. A participação em sessões de *chat*, o envio de contribuições para os fóruns de discussão, o compartilhamento de recursos (*sites*, *links*) com colegas, o número de vezes que o educando acessou a biblioteca e a sala de aula virtual, entre outras atividades, devem ser elementos a considerar nos processos de avaliação (GOMES, 2010).

Gomes (2010) considera também, que a disponibilização das informações dos percursos dos educandos nos ambientes *online*[4], aos próprios educandos é uma possibilidade que pode ser explorada com perspectivas pedagógicas, pois permite aos educandos tomarem mais facilmente consciência das atividades que realizam e, desse modo, facilitar a autorregulação dos mesmos, em termo do seu envolvimento nas atividades do curso. Além disso, a existência destes registros assume com grande importância na monitoração do percurso dos educandos e pode ser crucial para a identificação antecipada, por parte do educador, de casos de potencial desmotivação e potencial abandono, revelados por um baixo nível de consultas dos materiais disponíveis, poucas entradas no sistema e poucas participações nos espaços de discussão.

Destaca-se ainda que, do mesmo modo que os AVAs efetuam diversos tipos de registros de atividades dos educandos, também fazem o mesmo em relação às atividades dos educadores do curso. Desta forma, e à semelhança do que acontece com os educandos, a consulta destes registros pode ser um elemento de autoavaliação e heteroavaliação do envolvimento e do desempenho do educador na dinamização e na avaliação do curso.

2.4 PROBLEMÁTICAS DA AVALIAÇÃO DAS APRENDIZAGENS *ONLINE*

A problemática da avaliação das aprendizagens é recorrente no contexto educacional e são ainda maiores no contexto das novas práticas de educação *online*, pois

[4] A maioria dos Ambientes Virtual de Aprendizagem possui sistemas de registro automático dos percursos dos educandos no que se refere às entradas e às permanências, aos materiais consultados, às contribuições colocadas em fóruns, à participação em sessões de *chat* e à realização de atividades propostas.

quando feita a distância, a avaliação é mais complexa, por não ser possível ter o *feedback* das interações face a face, que possibilita uma avaliação informal do aprendiz, dando indícios da compreensão e interesse deste (ROCHA *et. al.*, 2006, apud GOMES, 2010). Na ausência da interação e do contato visual típico da educação presencial, o educador tem menos elementos para avaliar os educandos no decorrer do curso.

Segundo Gomes (2010), as problemáticas em torno da avaliação são, em termos globais, comuns aos modelos de educação presencial e a distância. Contudo, os novos contextos de educação a distância em ambientes *online* apresentam também um conjunto de questões totalmente específico e particularmente relevante. Frequentemente, as questões formuladas são: Como verificar a identidade dos educandos que pretendemos avaliar *online*? Como avaliar os processos de aprendizagem e não apenas os produtos? Como "conhecer" os educandos, as suas motivações, os interesses, dificuldades, quando com eles não interagimos diretamente? Como associar à avaliação uma componente de *feedback* relevante e temporalmente oportuno?

A implementação de práticas de avaliação contínua, envolvendo uma diversificação de instrumentos e de atividades de avaliação, que podem considerar aspectos como o grau e tipo de participação dos educandos em fóruns de discussão, a análise de níveis de consulta dos recursos disponibilizados e o desenvolvimento de portfolios digitais, podem ajudar o educador a construir o perfil de envolvimento e desempenho de cada um dos participantes em um curso *online*.

Conforme abordado por Gomes (2010), a maioria dos AVAs possui um conjunto de funcionalidades tendentes a facilitar as tarefas de avaliação. A avaliação do desenvolvimento dos portfolios dos educandos, por exemplo, permite também ao educador aumentar o grau de conhecimento em relação aos educandos, mesmo sem a possibilidade do contato presencial.

Os testes de múltipla escolha, testes de preenchimento de espaços lacunares ou outros tipos similares de provas, eventualmente com correção automática pelos AVAs, permitem que os educandos possam recorrer aos mesmos com frequência, podendo constituir situações de avaliação formativa em momentos determinados pelo próprio educando. Esta possibilidade pode ajuda-los a aferirem o grau das suas aprendizagens, ajudando-os a regular o seu percurso de estudo e aprendizagem.

Os fóruns são um dos elementos que permitem promover espaços de discussão coletiva e colaborativa do conhecimento. A participação dos educandos nestes fóruns é um dos elementos essenciais na promoção de atividades de discussão e de construção coletiva do conhecimento e, quando assim é, esse aspecto deve ser ponderado no processo de avaliação e classificação dos mesmos.

Os mapas conceituais podem ser utilizados como uma forma alternativa de avaliação. Uma das possibilidades é a sua utilização como forma de avaliação, nomeadamente por meio de elaboração de versões sucessivas de mapas conceituais sobre determinada temática que permitam verificar os progressos feitos pelos educandos (ROCHA *et. al*, 2005) (CHAVES. *et. al*, 2011).

Dessa forma, por meio da avaliação dos processos de aprendizagem, do conhecimento das motivações, dos interesses e das dificuldades de cada estudante e da interação frequente com cada um deles, que, mesmo em um contexto *online*, se pode estabelecer uma relação de conhecimento e construir um "perfil" de cada participante de um curso *online*.

A análise da problemática da avaliação das aprendizagens *online* e os conceitos de educação *online*, apresentados neste capítulo, são importantes para compreender a proposta desta pesquisa de avaliar o aprendizado *online*. Além disso, conhecer os recursos do ambiente virtual LabSQL apresentados é fundamental para compreender o conjunto de dados utilizado no processo de mineração de dados educacionais do LabSQL, no Capítulo 4.

CAPÍTULO 3

MINERAÇÃO DE DADOS EDUCACIONAIS

O Capítulo 3 tem como objetivo apresentar alguns conceitos que serão úteis para a compreensão da mineração de dados aplicados na área da educação.

3.1 CONCEITOS DE MDE

A mineração de dados consiste em extrair ou "minerar" conhecimento a partir de grandes quantidades de dados. Em parte da literatura relacionada, a mineração de dados é também tratada como sinônimo para outro termo, a descoberta de conhecimento em bases de dados (KDD, do inglês *Knowledge Discovery in Databases*). Outros autores consideram a mineração de dados uma etapa do processo de KDD, o qual é composto pelas etapas de seleção de dados, pré-processamento e limpeza, transformação, mineração de dados e interpretação, conforme apresentado por Fayyad *et. al*.(1996).

O objetivo geral do processo de KDD é extrair conhecimento de um conjunto de dados existentes e transformá-lo em uma estrutura mais compreensível de ser analisada pelo ser humano (SOUMEN *et. al.,* 2006). Por exemplo, é possível minerar dados de educandos para verificar a relação entre uma abordagem pedagógica e o aprendizado do educando. Por meio desta informação o professor poderia compreender se sua abordagem realmente está ajudando o educando e desenvolver novos métodos de ensino mais eficazes (BAKER *et. al.,* 2011).

A Mineração de Dados tem sido aplicada em diversas áreas do conhecimento, como por exemplo, finanças, bioinformática e ações contra terrorismo. Recentemente, com a expansão dos cursos a distância e também daqueles com suporte computacional, muitos pesquisadores da área de Informática na Educação (em particular, Inteligência Computacional Aplicada à Educação) têm mostrado interesse em utilizar mineração de dados para investigar perguntas científicas na área de educação, como por exemplo: quais são os fatores que afetam a aprendizagem? Ou como desenvolver sistemas educacionais mais eficazes? Dentro deste contexto, surgiu uma nova área de pesquisa conhecida como Mineração de Dados Educacionais (do inglês, *Educational Data Mining*, ou EDM) (BAKER *et. al.,* 2011).

A MDE é definida como a área de pesquisa que tem como principal foco o desenvolvimento de métodos para explorar conjuntos de dados coletados em ambientes educacionais. Assim, é possível compreender de forma mais eficaz e adequada os educandos, como eles aprendem, o papel do contexto na qual a aprendizagem ocorre, além de outros fatores que influenciam a aprendizagem. Por exemplo, é possível identificar em que situação um tipo de abordagem instrucional (aprendizagem individual ou colaborativa, por exemplo) proporciona melhores benefícios educacionais

ao educando. Também é possível verificar se o educando está desmotivado ou confuso e, assim, personalizar o ambiente e os métodos de ensino para oferecer melhores condições de aprendizagem (BAKER *et. al.*, 2011; PAIVA *et. al.*, 2012).

A comunidade de MDE vem crescendo rapidamente com a criação de conferências e o aumento do número de artigos e livros publicados, porém, no Brasil essa área de pesquisa ainda é embrionária (BAKER *et. al.*, 2011). Um dos trabalhos pioneiros no uso de mineração de dados na educação foi publicado por Brandão *et. al.* (2006), que analisou dados do programa nacional de informática na educação. Outro trabalho pioneiro no Brasil que analisou dados da avaliação de educandos é apresentado por Pimentel e Omar (2006).

3.2 MÉTODOS PARA MDE

De acordo com Baker (2010), muitas vezes os métodos utilizados em MDE precisam ser modificados, por causa da necessidade de considerar a hierarquia (em diversos níveis) da informação. Além disso, existe uma dependência estatística nos tipos de dados encontrados ao coletar informações em ambientes educacionais, que dificulta a adaptação dos dados aos algoritmos pré-existentes durante a etapa de pré-processamentos dos dados. A natureza destes dados é mais diversa do que a observada nos dados tradicionalmente utilizados. Por causa disso, diversos algoritmos e ferramentas utilizadas na área de mineração de dados não podem ser aplicadas para analisar dados educacionais sem modificação. Devido a esta lacuna na área de mineração de dados, muitos pesquisadores que publicam na área de MDE utilizam modelos desenvolvidos na área de psicometria (BARNES *et. al.*, 2005), (DESMARAIS e PU, 2005) e (PAVLIK *et. al.*, 2008).

Existem várias linhas de pesquisa na área de MDE, muitas delas derivadas diretamente da área de mineração de dados. Baker (2010) apresenta uma taxonomia das principais subáreas de pesquisa em MDE:

- Predição: Classificação, Regressão, Estimação de Densidade;
- Agrupamento;
- Mineração de relações: Mineração de Regras de associação, Mineração de Correlações, Mineração de Padrões Sequenciais, Mineração de Causas;
- Destilação de dados para facilitar decisões humanas;
- Descobertas com modelos.

As três primeiras categorias dessa taxonomia são de interesse tanto da área de MDE quanto da área de mineração de dados em geral. As subcategorias de Predição: Classificação, Regressão e Estimação de Densidade estão diretamente relacionadas as categorias dos métodos de mineração de dados apresentados por Moore (2005).

3.2.1 Área de Predição

Na área de predição, a meta é desenvolver modelos que deduzam aspectos específicos dos dados, conhecidos como variáveis preditivas, a partir da análise e fusão dos diversos aspectos encontrados nos dados, chamados de variáveis preditoras.

A predição necessita que parte dos dados seja manualmente codificada para viabilizar a correta identificação de uma ou mais variáveis preditoras previamente conhecidas. Como indicado na taxonomia, existem três tipos de predição: classificação, regressão e estimação de densidade. Em classificação, a variável preditora deve ser discreta, podendo ser binária ou categórica. Quando a variável preditora é um número real, diz-se que a predição é uma regressão, os algoritmos de regressão mais populares incluem regressão linear, redes neurais, e máquinas de vetores de suporte. A estimação de densidade é raramente utilizada na MDE devido à falta de independência estatística dos dados.

3.2.2 Área de Agrupamento

Na área de agrupamento, o objetivo principal é encontrar grupos e/ou categorias com características comuns ou padrões semelhantes, classificando os dados em diferentes grupos e/ou categorias. Estes grupos e categorias não são conhecidos inicialmente. Por meio de técnicas de agrupamento os grupos/categorias são automaticamente identificados a partir da manipulação das características dos dados. É possível criar esses grupos/ categorias utilizando diferentes unidades de análise, por exemplo, é possível achar grupos de escolas (para investigar as diferenças e similaridades entre escolas), ou achar grupos de educandos (para investigar as diferenças e similaridades entre educandos), ou até grupos de atos (para investigar padrões de comportamento) (AMERSHI e CONATI, 2009).

3.2.3 Área de Mineração de Relações

Em mineração de relações, a meta é descobrir possíveis relações entre variáveis em bancos de dados. Esta tarefa pode envolver a tentativa de aprender quais variáveis são mais fortemente associadas com uma variável específica, previamente conhecida e

importante, ou pode envolver as relações entre quaisquer variáveis presentes nos dados. Para identificar essas relações, existem quatro tipos de mineração: regras de associação, correlações, sequências e causas.

Na **mineração de regras de associação**, procura-se gerar/identificar regras do tipo se-então (*if-then*) que permitam associar o valor observado de uma variável ao valor de uma outra variável. Ou seja, caso uma condição seja verdadeira (por exemplo, variável Y possui valor 1) e uma regra associe essa condição ao valor de uma outra variável X, então podemos inferir o valor desta variável X. Por exemplo, ao analisar um conjunto de dados seria possível identificar uma regra que faz a associação entre a variável "objetivo do educando", uma variável binária que pode ter os valores alcançado ou não alcançado, e uma outra variável binária "pedir ajuda ao professor" que pode ter os valores sim ou não. Neste contexto, se o educando tem como objetivo aprender geometria, mas está com dificuldade, isto é, a variável meta do educando tem valor não alcançado, então é provável que ele peça ajuda do professor, ou seja, a variável pedir ajuda ao professor tem valor positivo.

Em **mineração de correlações**, a meta é achar correlações lineares (positivas ou negativas) entre variáveis. Por exemplo, ao analisar um conjunto de dados, seria possível identificar a existência de uma correlação positiva entre uma variável que indica a quantidade de tempo que um educando passa externalizando comportamentos que não estão relacionados as tarefas passadas pelo professor (por exemplo, conversas paralelas, brincadeiras e outras perturbações que ocorrem em sala de aula) e a nota que este educando recebe na próxima prova.

Em **mineração de sequências**, o objetivo principal é achar a associação temporal entre eventos e o impacto destes eventos no valor de uma variável. Neste caso, é possível determinar qual trajetória de atos e ações de um educando pode, eventualmente, levar a uma aprendizagem efetiva. Dessa forma, é possível criar um conjunto de atividades instrucionais que podem melhorar a qualidade do ensino fazendo com que os educandos externalizem ações que vão ajudá-los a construir seu conhecimento e desenvolver as habilidades necessárias para trabalhar com o conteúdo apresentado pelo professor.

Em **mineração de causas**, desenvolve-se algoritmos para verificar se um evento causa outro evento por meio da análise dos padrões de covariância (um sistema que faz isso é TETRAD) (SCHEINES *et. al.,* 1994). Por exemplo, se considerarmos o exemplo

anterior onde um educando externaliza comportamentos inadequados que não contribuem para resolver a tarefa dada pelo professor, o educando, em muitos casos, recebe uma nota ruim na prova final. Nesta situação, o comportamento do educando pode ser a causa dele não aprender e, assim, resulta em uma performance ruim na prova. Contudo, pode ser que o educando externalize tal comportamento inadequado devido a dificuldade em aprender, e portanto, a causa da performance ruim na prova não é o comportamento em sí, mas sim a dificuldade de aprendizagem do educando. Analisando o padrão de covariância, a mineração de causa pode auxiliar na inferência de qual evento foi a causa do outro.

3.2.4 Área de Destilação de Dados para facilitar decisões humanas

Na área de destilação de dados para facilitar decisões humanas, são realizadas pesquisas que tem como objetivo apresentar dados complexos de forma a facilitar sua compreensão e expor suas características mais importantes.

Os métodos dessa subárea da MDE facilitam a visualização da informação contida nos dados educacionais coletados por softwares educacionais (HERSHKOVITZ *et. al.,* 2008) e (KAY, *et. al.,* 2006). Estes métodos "purificam" os dados para auxiliar as pessoas na identificação de padrões. Em diversas ocasiões, esses padrões são previamente conhecidos, mas são difíceis de serem visualizados e/ou descritos formalmente. Por exemplo, uma visualização clássica em MDE é a curva de aprendizagem. Essa curva indica o nível de aprendizagem de um educando (ou de um conjunto de educandos) ao longo do tempo. Ela é apresentada num plano cartesiano conforme mostra a Figura 3-1. Nesta curva relaciona-se o número de oportunidades que o educando praticou um componente de conhecimento (apresentado no eixo x) e a sua performance (porcentagem de valores corretos, apresentada no eixo y). Uma curva que desce rapidamente no início do gráfico e depois gradativamente diminui sua inclinação indica que o modelo de conhecimento é bem especificado. Ou seja, o modelo representa corretamente quais as relações entre os componentes de conhecimento e as atividades realizadas pelos educandos.

Essas atividades oferecem a oportunidade de praticar os componentes de conhecimento relacionados e ao decorrer deste processo o educando aprende a medida que suas habilidades e conhecimentos são testados. Caso a curva de aprendizagem possua diversos pontos fora dos locais esperados, ou seja, porcentagem de erros muito acima ou muito abaixo do esperado, dado o número de oportunidades, isso indica que o

modelo utilizado não está bem refinado e provavelmente mais de um componente de conhecimento está sendo tratado no mesmo problema (COBERT, *et. al.*, 1995).

No caso da Figura 3-1, a curva em vermelho representa dados de educandos e a curva tracejada em azul representa a curva esperada calculada utilizando algoritmos de predição implementados na plataforma Datashop (KOEDINGER, *et. al.,* 2010). Observe que apesar de alguns pontos estarem um pouco acima ou abaixo do esperado, a curva em vermelho desce gradativamente seguindo a curva esperada. Ou seja, o modelo possui algumas falhas, mas que os educandos aprenderam ao longo do curso as habilidades (componentes de conhecimento) desejadas. Observe que esse modelo pode ser utilizado para identificar a evolução da aprendizagem de qualquer educando. Dependendo das interações entre o educando e as atividades, é possível verificar o quanto o educando aprendeu ou estimar o quanto ele irá aprender após um conjunto de atividades realizadas.

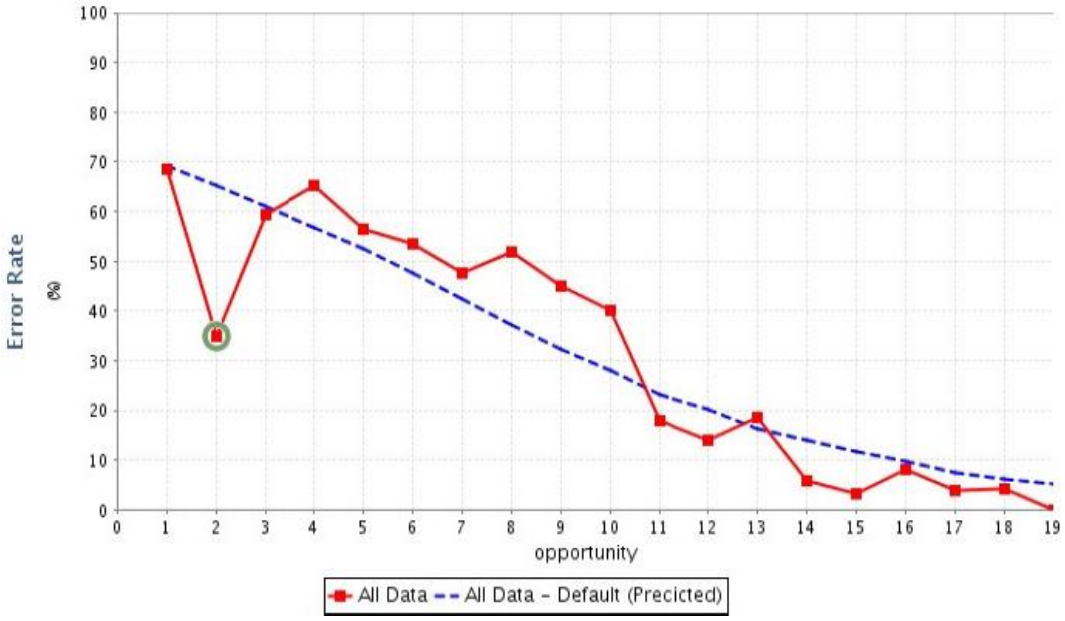


Figura 3-1 - Curva de aprendizagem utilizada na plataforma Datashop. A curva em vermelho representa os dados obtidos pelos educandos e a curva tracejada em azul modelo desenvolvido.

O uso da destilação de dados também é muito útil pra categorizar as ações dos educandos. A partir desta categorização é possível auxiliar o desenvolvimento de um modelo de predição mais robusto (BAKER *et. al.,* 2009).

3.2.5 Descoberta com Modelos

Na área descoberta com modelo, um modelo de um fenômeno é desenvolvido por meio de previsão, *clustering*, ou em alguns casos, engenharia do conhecimento, em

que o modelo é desenvolvido usando o raciocínio humano ao invés de métodos automatizados. Este modelo é então utilizado como um componente em outra análise, tal como a previsão ou mineração de relações.

No caso de previsão, as previsões do modelo criado são utilizadas como variáveis preditoras na previsão de uma nova variável. Por exemplo, análises de construções complexas, tais como o sistema de jogo dentro de aprendizagem *online* geralmente dependia da avaliação da probabilidade de que o educando conhece o componente do conhecimento atual que está sendo aprendido (BAKER *et. al.*, 2008; WALONOSKI e HEFFERNAN, 2006). Essas avaliações de conhecimento do educando têm, por sua vez, dependência de modelos dos componentes do conhecimento em um domínio, geralmente expressa como um mapeamento entre os exercícios dentro do software de aprendizagem e componentes do conhecimento.

No caso de mineração de relação, as relações entre as previsões do modelo criado e variáveis adicionais são estudadas. Isto pode permitir um pesquisador estudar a relação entre uma construção complexa e latente e uma grande variedade de construções observáveis.

3.3 TÉCNICAS DE MINERAÇÃO DE DADOS

Nos últimos anos a MDE vem sendo utilizada para a obtenção de diversos objetivos pedagógicos. A escolha da tarefa ou subárea de pesquisa de MDE depende dos objetivos definidos para mineração, do que pretende-se buscar nos dados, que tipo de regularidades ou categoria de padrões que tem-se interesse em encontrar. Nesse contexto, Romero (2011) propõe a formação de categorias para dispor diversos objetivos pedagógicos, as principais são:

- Comunicação com os *stakeholders*: provê auxilio aos educadores para avaliar as atividades e participação dos educandos. Os métodos geralmente utilizados são: mineração de processos, geração de relatórios, visualização de dados, e a análise estatística dos dados (FERREIRA, 2012);
- Realizar melhorias e manutenções em cursos: provê aos educadores estratégias que auxiliem para melhoria dos cursos. Os métodos de Mineração de Dados geralmente utilizados são: Associação, cluster e classificação;
- Gerar recomendações: provê uma recomendação de conteúdo no momento adequando vivenciado pelo educando. Os métodos de Mineração de Dados

geralmente utilizados são: associação, sequenciação, cluster e classificação (ABEL, 2010);

- Prever resultados de atividades/provas ou de avaliações de aprendizado: prevê resultados de testes e de outras avaliações educacionais, com base na análise das atividades realizadas pelos educandos. Os métodos de Mineração de Dados utilizados são: Associação, cluster e classificação (FERREIRA, 2012);
- Criar modelos de educandos: o objetivo é estudar determinadas características dos educandos. Os métodos utilizados são: Análise estatística, Redes Bayesianas, Modelos Psicométricos e Aprendizado por Reforço;

No contexto desse trabalho, os métodos de MDE mais adequadas e viáveis para prever o desempenho dos educandos e obter conhecimento relevante para o entendimento do perfil dos educandos na utilização do ambiente LabSQL são a associação, cluster e classificação, pois elas podem, respectivamente, verificar associações entre características distintas dos educandos, agrupar um conjunto de educandos segundo suas características, prever o desempenho obtido pelos educandos a partir do seu perfil e contabilizar as relações de dependência entre as ações envolvidas no processo de aprendizagem, conforme observado em estudo preliminar (DIAS *et. al.*, 2008; DIAS *et. al.*, 2011).

As técnicas utilizadas para realizar a tarefa de classificação são: Árvore de Decisão e Redes Bayesianas. Para realizar a tarefa de associação, destaca-se método de Regras de Associação (*Apriori*) e para realizar a tarefa de cluster, destaca-se a Análise de Agrupamento (*K-means*). A escolha destes métodos são justificadas e detalhadas nas seções 3.3.1, 3.3.2, 3.3.3 e 3.3.4, respectivamente.

3.3.1 Árvore de Decisão

A Árvore de Decisão é uma técnica de classificação simples, porém muito usada. Ela possui uma estrutura hierárquica que traduz uma progressão da análise de dados no sentido de desempenhar uma tarefa de previsão/classificação. Formalmente, uma Árvore de Decisão é um grafo acíclico direcionado em que cada nó ou é um nó de decisão, com dois ou mais sucessores, ou um nó folha corresponde a uma classe. Um nó de decisão contém um teste condicional baseado nos valores do atributo (FACELI *et. al.*, 2011).

3.3.1.1 Representação de uma Árvore de Decisão

A Figura 3-2 representa uma Árvore de Decisão e a divisão correspondente no espaço definido pelos atributos x_1 e x_2. Cada nó da árvore corresponde a uma região nesse espaço. Para cada um dos possíveis valores de atributos, tem-se um ramo para outra árvore de decisão (sub-árvore). Ela é formada por um conjunto de regras de classificação em que cada percurso da AD, desde um nó raiz até um nó folha, é convertido em uma regra, onde a classe do nó folha corresponde à classe prevista pelo consequente (parte "Então" da regra) e as condições ao longo do caminho correspondem às condições do antecedente (parte "Se" da regra).

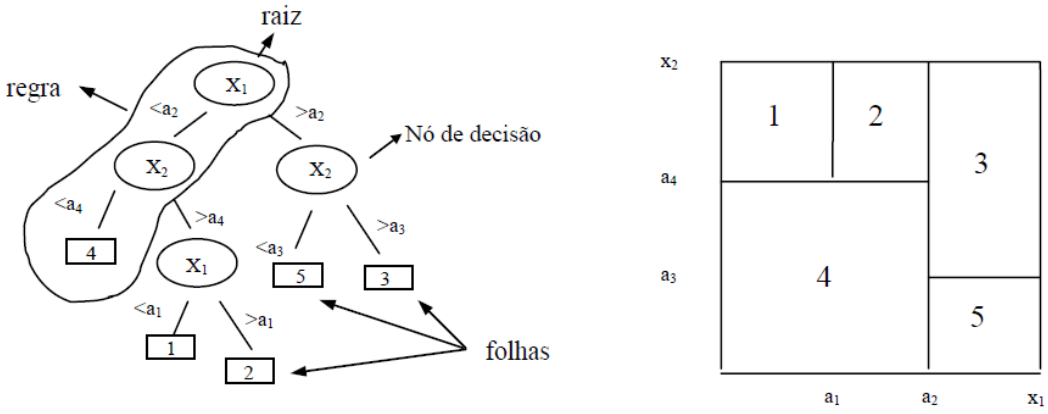


Figura 3-2 - Representação de Uma Árvore de Decisão (FACELI et. al., 2011).

De acordo com Fayyad *et. al.* (1996), as regras de classificação que resultam da transformação de árvores de decisão podem ter as seguintes vantagens:

i. São uma forma de representação do conhecimento amplamente utilizadas em sistemas especialistas;

ii. Em geral são de fácil interpretação pelo ser humano;

iii. Geralmente melhoram a precisão preditiva pela eliminação das ramificações que expressam peculiaridades do conjunto de treinamento que são pouco generalizáveis para os dados do teste.

É importante que as regras sejam acompanhadas de medidas relativas à sua precisão (ou confiança) e a sua cobertura. A precisão informa o quanto a regra é correta, ou seja, qual a porcentagem de casos que, se o antecedente é verdadeiro, então o consequente é verdadeiro. Uma alta precisão indica uma regra com uma forte dependência entre o antecedente e o consequente da regra.

Uma árvore de decisão usa a estratégia dividir para conquistar para resolver um problema de decisão. Um problema complexo é dividido em problemas mais simples, aos quais recursivamente é aplicada a mesma estratégia. As soluções dos subproblemas podem ser combinadas, na forma de árvore, para produzir uma solução do problema complexo (FACELI *et. al.*, 2011). Essa é a ideia básica por trás de algoritmos baseados em AD, tais como: ID3 (QUILAN, 1986), C4.5 (QUILAN, 1986) e CART (ÁRVORES DE DECISÃO, 2007). No entanto, o algoritmo para a construção da árvore em si pode variar, além de outros detalhes como a forma de realizar a decisão do melhor caminho em um nó ou até mesmo fazer o tratamento de atributos contínuos.

Muitos algoritmos de indução de Árvore de Decisão existentes, incluindo ID3, C4.5 e CART funcionam selecionando recursivamente o melhor atributo para dividir os dados e expandir os nós folha da árvore até que um critério de parada seja satisfeito.

3.3.1.2 Métricas para Selecionar a Melhor Divisão

O critério utilizado para selecionar a melhor divisão é o da utilidade do atributo para a classificação. Aplica-se, por este critério, um determinado ganho de informação a cada atributo. O atributo escolhido como atributo teste para o corrente nó é aquele que possui o maior ganho de informação. A partir desta aplicação, inicia-se um novo processo de partição. As métricas desenvolvidas para selecionar a melhor divisão são muitas vezes baseadas no grau de impureza dos nós filhos. Exemplos de métricas de impureza incluem Entropia, Gini e Erro de classificação. Nos casos em que a árvore é usada para classificação, os critérios de partição mais conhecidos são baseados na entropia (TAN *et. al.*, 2009).

3.3.1.3 Entropia

A Entropia é o cálculo do ganho de informação baseado em uma medida utilizada na teoria da informação. A entropia caracteriza a pureza ou impureza dos dados: em um conjunto de dados, é uma medida da falta de homogeneidade dos dados de entrada em relação a sua classificação. Por exemplo, a entropia é máxima (igual a 1) quando o conjunto de dados é heterogêneo (OVERVIEW, 2005).

A construção de uma árvore de decisão tem três objetivos: diminuir a entropia (a aleatoriedade da variável objetivo), ser consistente com o conjunto de dados e possuir o menor número de nós.

3.3.1.4 Podagem

Após construir a AD, um passo de poda da árvore pode ser executado para reduzir o tamanho da AD. Árvores de decisão que sejam grandes demais são susceptíveis ao excesso de ajustes (*overfitting*) aos dados de treinamento. A poda ajuda a retirar as ramificações da árvore inicial de uma forma que melhore a capacidade de generalização da AD. Existem duas possibilidades de podagem em árvore de decisão: parar com o crescimento da árvore mais cedo (pré-poda) ou crescer uma árvore completa e, em seguida, podar a árvore (pós-poda) (CARVALHO, 1999). Verifica-se que a pós-poda é mais lenta, porém mais confiável que a pré-poda (QUILAN, 1986).

Para entender o mecanismo de podagem, precisa-se antes entender o conceito de taxa de estimativa de erro, a qual pode ser obtida da seguinte forma: se *N* exemplos são cobertos por determinado nó folha e *E* dentre estes *N* são classificados de forma incorreta, então a taxa de estimativa de erro dessa folha é *E*/*N* (BERSON e SMITH, 1997).

3.3.2 Redes Bayesianas

Os problemas que envolvem as tarefas de predição e classificação, especialmente aqueles em que as informações disponíveis são incompletas ou imprecisas, têm encontrado solução a partir da utilização de algoritmos baseados no Teorema de Bayes, os métodos probabilísticos Bayesianos.

3.3.2.1 A Estatística Bayesiana

A noção fundamental da Estatística Bayesiana é a Probabilidade Condicional, definida por $P(H|E)$ no qual H é a hipótese e E é a evidência. Para computar a probabilidade de uma hipótese H, é necessário levar em consideração o valor da evidência E. Quando não existir evidências, tem-se a probabilidade incondicional $P(H)$ (RUSSEL, 2004).

O cálculo é feito a partir da Equação 3.1, dada por:

$$P(H \mid E) = \frac{P(H \cap E)}{P(E)}, \qquad (3.1)$$

onde o numerador é a probabilidade de H *e* E ocorrerem simultaneamente e o denominador é a probabilidade de ocorrer E.

A formulação do teorema de Bayes envolve estas probabilidades. A Equação 3.2 apresenta o teorema formulado por Bayes,

$$P(H \mid E) = \frac{P(E \mid H)P(H)}{P(E)}. \qquad (3.2)$$

3.3.2.2 Classificadores Bayesianos

A estatística bayesiana pode ser usada para classificação de uma forma relativamente simples, sendo chamada de classificadores Bayesianos, cujo objetivo é a descrição e identificação de classes e também a previsão de classes de objetos que não foram classificados.

O classificador bayesiano mais simples é conhecido como Naïve Bayes e considera a hipótese de que todas as variáveis são independentes. Ele apresenta bom desempenho em vários domínios e é considerado robusto à presença de ruídos e atributos irrelevantes, no entanto, a suposição de independência total entre os atributos pode torna-se um pouco rígida, tornando o classificador incapaz de tratar problemas de classificação reais onde as variáveis geralmente apresentam interdependência (FACELI *et. al.*, 2011; TAN *et. al.*, 2009).

Devido à dificuldade característica do Naïve Bayes, existem abordagens mais flexíveis para a modelagem de probabilidades condicionais, uma delas chamada de Redes de Crenças Bayesianas ou simplesmente Redes Bayesianas (TAN *et. al.*, 2009).

As Redes Bayesianas são modelos probabilísticos representados por grafos acíclicos e direcionados, mostrando as relações de causalidade entre as variáveis de um problema (RUSSEL, 2004). As Redes Bayesianas possuem uma parte gráfica que é qualitativa, e uma parte quantitativa que são as tabelas com a distribuição de probabilidades das variáveis.

Formalmente, seja x = {x1, x2,...,xn} um conjunto de variáveis aleatórias para um dado domínio. Uma rede Bayesiana sobre x é uma tupla (S, ΘS) em que o primeiro componente, a estrutura da rede S, é um grafo acíclico direcionado cujos nós representam as variáveis aleatórias e as arestas representam dependências diretas entre variáveis, influência ou correlação. O conjunto de variáveis aleatórias que influenciam uma variável x_i é usualmente designado por Pais de x_i. A segunda componente Θ_S é o conjunto de tabelas de probabilidade condicional. A ausência de aresta entre dois vértices corresponde à premissa da independência condicional, que significa que uma

variável é independente de qualquer outra que não é sua descendente no grafo, dada a observação dos pais. Essa premissa é útil para reduzir o número de variáveis necessárias para definir uma distribuição de probabilidade, assim, as redes representam um resumo das possíveis circunstâncias envolvidas em um domínio, devido à complexidade em descobrir probabilidades de muitas variáveis, ou a irrelevância de fatores.

A construção do modelo envolve duas etapas: a primeira é a construção do grafo e a segunda é a avaliação dos valores de probabilidades nas tabelas associadas a cada nodo, chamadas de Tabelas de Probabilidade Condicional (TPC); essas etapas podem ser realizadas utilizando o conhecimento de especialistas do domínio, utilizando bases de dados, ou pela junção das duas abordagens (TAN *et. al*., 2009).

A Equação 3.3 demonstra como são feitos os cálculos das probabilidades para cada variável.

$$P(U) = P(x_1, x_2, ..., x_n) = P(U) = \prod_{i=1}^{n} P(x_i \mid pa(xi)), \qquad (3.3)$$

onde P(U) é a probabilidade conjunta para a rede e $P(x_i \mid pa(x_i))$, são as probabilidades condicionais de A em relação aos seus pais.

Para realizar a tarefa de classificação uma das variáveis é selecionada como atributo alvo, e todas as outras variáveis são atributos de entrada. O conjunto de variáveis que influenciam o atributo alvo é designado por Markov Blanquet: que é constituído pelas variáveis pais da variável alvo, pelos filhos da variável alvo e pelos pais dos filhos da variável alvo.

Assim uma Rede Bayesiana pode ser utilizada como um classificador que, dado um exemplo x, fornece a distribuição de probabilidade a posteriori P(y | x) do nó classe y ϵ Y. É possível calcular a probabilidade a posteriori P(y | x, S) para cada classe y ϵ Y marginalizando a distribuição de probabilidade conjunta P(y, x | S) e então retornar a classe.

Com a utilização de uma ferramenta de análise de Redes Bayesianas é possível definir hipóteses ou fazer inferências sobre um determinado atributo, tendo respostas sobre as influências dele de acordo com as ligações existentes entre os outros atributos. As inferências podem ser do tipo Diagnóstico (dos efeitos para as causas, ex: Dado Efeito, P(Causa|Efeito)) ou do tipo Causas (das causas para os efeitos, ex: Dado Causa, P(Efeito| Causa)).

Por exemplo, considerando o grafo representado pela Figura 3-3, pode-se observar que a ocorrência dos eventos *Ladrão* e *Terremoto* são prováveis causas direta da ocorrência de *Alarme*, assim como, a ocorrência do evento *Alarme* é uma provável causa das ocorrências de *João_liga* e *Maria_liga*. Os nós *Ladrão* e *Terremoto* são pais de *Alarme*, que por sua vez é pai de *Maria_liga* e *João_liga*. Como *Ladrão* e *Terremoto* não têm pai, são considerados nós raízes da rede.

O problema mostrado a seguir e representado na Figura 3-3 é exemplificado por Russel (2004), para um melhor entendimento sobre os conceitos de uma Rede Bayesiana.

Existe um alarme contra ladrões em casa. Este alarme é muito confiável na detecção de ladrões, entretanto, ele também pode disparar caso ocorra um terremoto. Há dois vizinhos, João e Maria, os quais prometeram telefonar para o trabalho do dono da casa caso o alarme dispare. João sempre liga quando ouve o alarme, entretanto, algumas vezes confunde o alarme com o telefone e também liga nestes casos. Maria, por outro lado, gosta de ouvir música alta e às vezes não escuta o alarme.

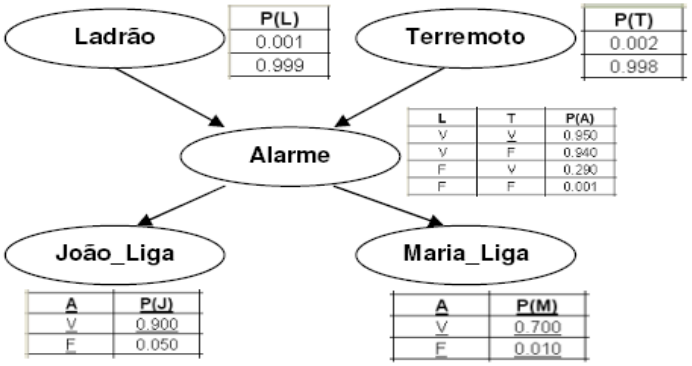


Figura 3-3 Rede Bayesiana com tabelas de Probabilidade de cada Variável (Adaptado de RUSSEL, 2004).

O modelo gráfico proposto na Figura 3-3 trata-se de uma simplificação do problema, pois alguns fatos, como Maria ouvindo música alta e João escutando o barulho do telefone, estão implícitos. Estabelecida a topologia da rede, é necessário construir a TPC para cada variável. Para isso, é necessária a identificação de todas as combinações de possíveis valores das variáveis pais e, os valores que a variável em questão pode assumir, como mostrado nas TPC's de cada nó do grafo.

Para Luna (2004) e Tan *et. al.* (2009) existem muitos pontos positivos de se utilizar Redes Bayesianas, dentre suas principais características destacam-se:

- Permitem expressar as assertivas de independência de forma visual facilitando a percepção;
- Tornam o processo de inferência eficiente computacionalmente;
- Permitem analisar grandes quantidades de dados;
- Podem ser utilizadas em vários domínios.
- São apropriadas para tratar situações com valores de atributos incompletos por meio de soma ou integração de probabilidades pelos valores possíveis do atributo;
- São robustas para tratar *overfitting* devido à combinação probabilística dos dados com conhecimento anterior sobre a situação.

3.3.3 Associação

Regras de associação consistem na descoberta de relacionamentos existentes entre variáveis. Esses relacionamentos são descobertos efetuando-se múltiplos passos iterativos sobre a base de dados. A cada iteração é levado em consideração o conjunto de itens (*itemset*) gerados no passo anterior, chamado de conjunto de itens candidatos. A partir de conjuntos frequentes, é possível derivar regras de associação. Uma regra de associação é uma expressão de implicação no formato de regra $X \rightarrow Y$ (**se** X **então** Y), onde X (antecedente) e Y (consequente) são conjuntos disjuntos de itens ($X \cap Y = \emptyset$). A inferência feita por uma regra de associação não implica necessariamente em causalidade. Em vez disso, sugere um forte relacionamento de co-ocorrência entre os itens no antecedente e o consequente da regras (TAN *et. al.*, 2009).

Para cada Regra de Associação é computado um fator de suporte e um fator de confiança. O suporte determina a frequência na qual uma regra é aplicável a um determinado conjunto de dados. Ele é definido como a razão do número de registros que satisfaçam tanto X quanto Y sobre o número total de registros, isto é, Suporte = $|X \cap Y|/N$, onde N corresponde ao número total de registros.

A confiança determina a frequência na qual os itens em Y aparecem em transações que contenham X. Ela é definida como a razão do número de registros que satisfaçam tanto X quanto Y sobre o número de registros que satisfazem X, isto é, Confiança = $|X \cap Y| \ / \ |X|$. Para uma determinada regra $X \rightarrow Y$, quanto maior a

confiança, maior a probabilidade de que Y esteja presente em transações que contenha X.

Uma estratégia comum adotada por muitos algoritmos de mineração de regras de associação é decompor o problema em duas subtarefas principais:

1) **Geração de Conjuntos de Itens Frequentes**, cujo objetivo é encontrar todos os conjuntos de itens que satisfaçam o limite de suporte mínimo.

2) **Geração de Regras**, cujo objetivo é extrair todas as regras de alta confiança dos conjuntos de itens frequentes no passo anterior.

Pivato (2006) mostra um exemplo que ilustra esses conceitos. Considere o conjunto A de itens comprados em uma loja e o conjunto de transações T, representando as compras ocorridas, com suporte e confiança mínima de 50% como mostrado na Figura 3-4.

A = {bermuda, calça, camiseta, sandália, tênis}
T = {{calça, camiseta, tênis}, {camiseta, tênis}, {bermuda, tênis}, {calça, sandália}},

Suporte Mínimo = 50% (duas transações)
Confiança Mínima = 50%

Figura 3-4 - Dados de Entrada para o Processo de Extração de Regras de Associação.

A representação dos itens comprados, referente a cada transação é mostrada no Quadro 3-1. Por exemplo, na transação "1" foi comprado calça, camiseta e tênis.

Quadro 3-1 - Representação das transações e itens comprados.

Transações	Itens Comprados
1	calça, camiseta, tênis
2	camiseta, tênis
3	bermuda, tênis
4	calça, sandália

Tendo as transações disponíveis, extraem-se os *itemsets* a partir de combinações dos elementos do conjunto de itens A e de *itemsets* já encontrados. A formação de *itemsets* varia de acordo com os algoritmos utilizados. Nesse exemplo, para facilitar a compreensão do conceito de regras de associação e *itemsets*, são mostradas todas as combinações possíveis na Figura 3-5.

I = {{bermuda}, {calça}, {camiseta}, {sandália}, {tênis}, {bermuda, calça}, {bermuda, camiseta}, {bermuda, sandália}, {bermuda, tênis}, {calça, camiseta}, {calça, sandália}, {calça, tênis}, {camiseta, sandália}, {camiseta, tênis}, {sandália, tênis}, {bermuda, calça, camiseta}, {bermuda, calça, sandália}, {bermuda, calça, tênis}, {bermuda, camiseta, sandália}, {bermuda, camiseta, tênis}, {bermuda, sandália, tênis}, {calça, camiseta, sandália}, {calça, camiseta, tênis}, {calça, sandália, tênis}, {camiseta, sandália, tênis}, {bermuda, calça, camiseta, sandália}, {bermuda, calça, camiseta, tênis}, {bermuda, calça, sandália, tênis}, {bermuda, camiseta, sandália, tênis}, {calça, camiseta, sandália, tênis}, {bermuda, calça, camiseta, sandália, tênis}};

Figura 3-5 - Combinações Possíveis dos Itens Comprados.

O suporte é calculado para cada *itemset*, mas somente são considerados os *itemsets* com suporte acima do suporte mínimo, como mostrado no Quadro 3-2. Considerando o suporte mínimo de 50%, pode-se observar que existe a regra {camiseta}⇒{tênis}. A confiança dessa regra é calculada como sendo o suporte de ocorrer o *itemset* {camiseta, tênis} dividido pelo suporte de ocorrer somente o *itemset* que forma o antecedente da regra, ou seja, {camiseta}. Nesse caso, a confiança é 100%. Também é possível encontrar a regra {tênis}⇒{camiseta}, com suporte igual a 50%, mas com confiança igual a 66,6%, uma vez que o *itemset* {tênis} ocorreu em outras transações que não ocorreu o *itemset* {camiseta}. As regras compostas somente por um item, como {tênis}⇒verdadeiro, indicam que o item foi adquirido independentemente da compra de outro item, no caso de tênis comprado em 75% das transações.

Quadro 3-2 - *Itemsets* Frequentes e seus Respectivos Suportes.

Conjunto de atributos	Suporte
{tênis}	75%
{calça}	50%
{camiseta}	50%
{camiseta, tênis}	50%

Existem algumas variações do algoritmo básico de geração de Regras de Associação. A grande maioria é uma extensão do algoritmo Apriori, proposto por Agrawal e Srikant (1994). Ele foi o primeiro algoritmo para mineração de *itemsets* e regras de associação.

A parte de geração de conjuntos de itens frequentes do Algoritmo Apriori se baseia no principio de que, se um conjunto de itens é frequente, então todos os seus subconjuntos também devem ser frequentes. De forma inversa, se um conjunto de itens for infrequente, então todos os seus superconjuntos devem ser infrequentes também, e dessa forma, todas as regras candidatas podem ser podadas imediatamente sem termos

que calcular seus valores de confiança. Esta abordagem pode reduzir efetivamente o número de conjuntos candidatos.

O algoritmo *Apriori* gera regras a partir dos conjuntos de termos frequentes encontrados e que obedeçam ao critério de confiança mínima. O algoritmo seleciona as regras cuja confiança é superior à confiança mínima definida pelo usuário. Inicialmente, todas as regras de confiança alta que tenham apenas um item no consequente da regra são extraídas. Essas regras são então usadas para gerar novas regras candidatas. Por exemplo, se {acd} → {b} e {abd} → {c} forem regras de confiança alta, então a regra condidata {ad} → {bc} é gerada pela fusão dos consequentes de ambas as regras (TAN *et. al.*, 2009).

3.3.4 Agrupamento

Técnicas de agrupamento são utilizadas para dividir dados em grupos (*clusters*) de acordo com sua significância e/ou utilidade. Segundo Tan *et. al.* (2009), esse tipo de análise vem sendo utilizada em diversas áreas, como a de reconhecimento de padrões, recuperação de informações e mineração de dados.

Tratando-se de significado ou compreensão, os dados são agrupados de acordo com características similares, visando a descoberta automática de classes. Tratando-se da utilidade, a análise de grupos visa encontrar os objetos de dados que são os mais representativos do seu grupo, chamados de protótipos de grupos, para serem usadas por outras técnicas de análise ou processamento de dados.

A análise de agrupamentos pode ser definida como uma técnica que agrupa um conjunto de itens, indivíduos ou objetos, sendo que os objetos incluídos em um mesmo agrupamento são os mais similares entre si e menos similares em relação a outros que estão em outros agrupamentos (GIUDICI, 2003). Essa análise é feita apenas com informações contidas nos próprios dados, assim, a definição dos grupos é imprecisa e a melhor definição depende da natureza dos dados e dos resultados buscados (TAN *et. al.*, 2009).

Os agrupamentos são realizados por meio de uma distância de similaridade (dissimilaridade). Dessa maneira, a pessoa que realiza a análise deve possuir conhecimento suficiente sobre o problema, visando distinguir grupos úteis, necessários à realização de consultas. A análise de agrupamentos é uma metodologia objetiva para

quantificar uma característica estrutural de um conjunto de observações (HAIR *et. al.*, 1998), e apresenta três desafios básicos:

- *Medir a similaridade entre os itens*. Torna-se necessário a adoção de um parâmetro de qualificação dos itens;
- *Formar os agrupamentos*. Deve-se determinar quais variáveis fazem parte da geração de determinados agrupamentos;
- *Definir o número de grupos*. Existem basicamente duas abordagens, uma que define o número de agrupamentos desejados, e outra que usa um critério, tal como um raio de abrangência do agrupamento.

Segundo Tan *et. al.* (2009) existem vários tipos de agrupamentos como: a) **bem separados**, onde objetos de um mesmo grupo são mais próximos que objetos de grupos diferentes; b) **baseados em protótipos**, onde os objetos são mais semelhantes ao protótipo que define o grupo do que ao protótipo de qualquer outro grupo; c) **baseados em densidade**, o grupo é uma região densa cercada por uma região de baixa densidade; d) **propriedades compartilhadas**, onde há um conjunto de objetos que partilham alguma propriedade.

Nesse trabalho será utilizada a técnica K-means, baseada em protótipos.

3.3.4.1 K-means

Segundo Tan *et. al.* (2009) e Smaragdakis *et. al.*(2004) o algoritmo K-means é uma técnica de agrupamento baseada em protótipos, que busca identificar um número K de grupos especificado pelo usuário, e representados por centróides, que são em geral formados pela média de um grupo de pontos em um espaço *n* dimensional. Ela é uma técnica antiga e amplamente utilizada em várias áreas.

O procedimento do K-means é descrito pela execução de algumas etapas: 1) Seleção dos centróides iniciais para K grupos definidos pelo usuário; 2) Cálculo da distância de similaridade (medida de proximidade) de todos os elementos (pontos) com base em uma função de proximidade, como a distância Euclidiana (Equação 3.4); 3) Cada elemento é atribuído ao centroíde mais próximo, e cada coleção de elementos atribuídos a um centroíde é um grupo; 4) O centróide de cada grupo é então atualizado baseado em uma função objetiva que mede a qualidade de um agrupamento e depende das proximidades dos pontos entre si ou dos centróides do grupo. Os passos de

atribuíção e atualização são repetidos até que não ocorram mais alterações nas posições (WANG e SU, 2011).

Existem diferentes funções de proximidade (Manhattan, Euclidiana e Coseno), centróides (Mediana e Média) e funções objetivas (Minimizar a soma da distância, Minimizar a soma da distância quadrada e Maximizar a soma da semelhança de coseno) que podem ser usadas no algoritmo K-means básico. A métrica Euclidiana é a medida de distância mais popular, e uma das mais utilizadas em análise de agrupamento (FACELI *et. al*, 2011).

Alguns autores como Lozano (1999) e Tan *et. al.* (2009) apontam problemas nos resultados do algoritmo devido a posição inicial dos centroides, em função da ocorrência de mínimos locais. Assim, algumas recomendações são dadas para a escolha dos centróides iniciais como: Não fazer a seleção de forma totalmente aleatória, uma boa escolha é colocá-los tanto quanto possível para longe um do outro e fazer testes das diferentes posições e selecionar com base em uma função objetivo que, por exemplo, minimiza a soma do erro quadrado de cada ponto com seu centróide mais próximo (Equação 3.5)

$$distância = \sqrt{\sum_{j=1}^{n} \left(x_i^{(j)} - c_j \right)^2}. \qquad (3.4)$$

$$obj = \sum_{j=1}^{k} \sum_{i=1}^{n} distância^2. \qquad (3.5)$$

onde *distância* é uma média da distância entre um ponto escolhido de dados $x_i^{(j)}$ e o centróide c_j do grupo. Ou seja, é um indicador da distância entre os *n* pontos de dados dos seus respectivos centróides de fragmentação.

Nas Figuras 3-6 (a) e 3-6 (b) são apresentadas as etapas e o modo de atualização para formação dos grupos em relação ao centróide, respectivamente.

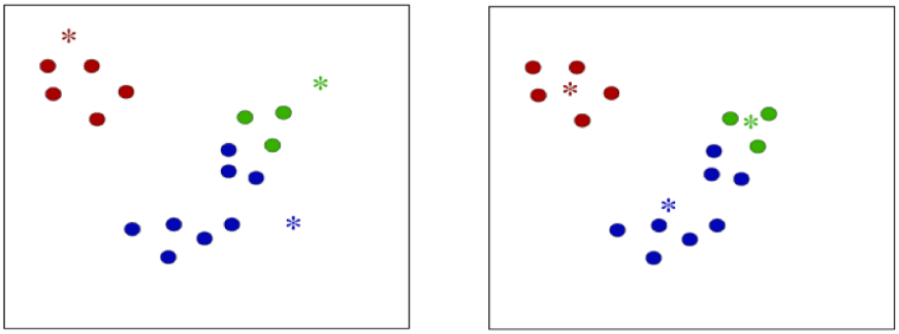


Figura 3-6 (a) Atribuição ao representante mais próximo (b) Reapresentação da Atribuição

Também é importante citar as limitações do algoritmo K-means. As figuras 3-7 (a) e 3-7 (b) mostram as limitações no agrupamento com diferentes densidade, onde a associação falha quando existe um centróide com baixa densidade.

Outro problema é durante o cálculo de similaridade e atribuição, podendo não ter nem um nó associado, desta forma gerando um cluster vazio, devendo substituir seu respectivo centróide, caso a substituição não ocorra é necessário a eliminação do cluster (TAN *et. al.*, 2009).

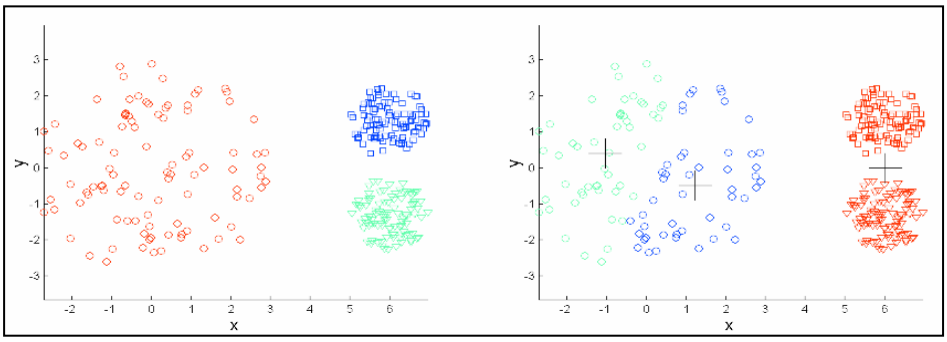


Figura 3-7 (a) Formação original ***(b) K-means (3 Clusters)***

3.4 CONSIDERAÇÕES FINAIS

Estudar e apresentar os principais conceitos e aplicações relacionadas à área de Mineração de Dados Educacionais encontradas na literatura é um dos objetivos específicos deste trabalho, necessário para a realização desta pesquisa. Além disso, conhecer as técnicas de mineração de dados apresentadas neste capítulo é fundamental para compreender como elas são aplicadas no processo de mineração de dados educacionais do LabSQL, no Capítulo 4.

CAPÍTULO 4

MINERAÇÃO DE DADOS EDUCACIONAIS DO AVA LABSQL

O Capítulo 4 apresenta a mineração de dados educacionais realizada com base nos registros das atividades dos educandos no ambiente virtual LabSQL, utilizado na Educação Online da Universidade Federal do Pará – UFPA. As técnicas de Mineração de Dados utilizadas foram: Árvore de decisão, Redes Bayesianas, Regras de Associação e Agrupamento.

4.1 INTRODUÇÃO

No contexto da educação *online*, as problemáticas da avaliação das aprendizagens são ainda maiores e mais complexas, pois na ausência da interação e do contato visual típico da educação presencial, o educador tem menos elementos para avaliar os educandos no decorrer do curso. Uma baixa percepção do educador quanto ao estado de compreensão de seus educandos pode levar ao insucesso de qualquer curso.

Diante das problemáticas da avaliação das aprendizagens *online* e conforme os objetivos propostos no início deste trabalho, este capítulo apresenta a mineração de dados educacionais do ambiente LabSQL para prever o desempenho dos educandos e gerar informações relevantes sobre o perfil dos educandos em relação à utilização dessa tecnologia e ao processo de ensino-aprendizagem, que possam apoiar os educadores na avaliação da aprendizagem *online*.

Na seção 4.2, são descritas as etapas de seleção, pré-processamento e transformação dos dados realizados durante o processo de mineração de dados educacionais do LabSQL, conforme o processo de KDD. Em seguida, a Seção 4.3 apresenta a aplicação da técnica de Árvore de Decisão, a Seção 4.4 apresenta a aplicação de Redes Bayesianas, a Seção 4.5 apresenta a aplicação de Regras de Associação. E por fim, a Seção 4.6 apresenta a aplicação de Análise de Agrupamento.

4.2 SELEÇÃO, PRÉ-PROCESSAMENTO E TRANSFORMAÇÃO DOS DADOS

Os dados obtidos a partir do ambiente LabSQL referem-se a 25 turmas em um modelo de ensino-aprendizagem semipresencial, contendo, em média, 26 educandos cada, durante 10 semestres letivos, da Universidade Federal do Pará, no período de 2008 a 2012. No total, o sistema foi utilizado por 667 educandos. O quadro 4-1 apresenta o quantitativo de turmas por curso.

Quadro 4-1 Quantitativo de turmas por curso.

nível	**curso**	**qtd**
pós-graduação	especialização em Banco de dados	5
graduação	Ciência da computação	10
	Sistemas de informação	10
total		**25**

O banco de dados do LabSQL possui cerca de 20 tabelas com informações associadas aos seus usuários. Dentre os atributos selecionados destacam-se: sexo do usuário (masculino ou feminino); nome do curso (Ciência da Computação, Sistema de Informação ou Especialização em Banco de Dados); nome do tipo de curso (graduação ou especialização); turma (vinte e cinco valores); coordenador da turma (dois valores); trabalhou em equipe (sim ou não); usou agenda de anotações do sistema (sim ou não); usou material de apoio (sim ou não), usou o recurso do SQL-Livre (sim ou não); notas obtidas em exercícios e avaliações.

Em seguida foram realizados alguns tratamentos nos dados obtidos para uma melhor aplicação das técnicas de mineração de dados. Dentre as atividades realizadas nesta etapa destacam-se: a retirada de registros de usuários de testes cadastrados no ambiente LabSQL e o preenchimento manual de dados em branco, como o sexo do usuário, inferido a partir do seu nome. Criaram-se ainda novos atributos a partir de outros (atributos derivados), visando avaliar o desempenho dos educandos, por exemplo, para avaliar se o educando está abaixo ou acima da média de pontos ou de acessos, ou se o educando usou ou não determinado recurso do ambiente LabSQL, como o SQL-Livre, o material de apoio, os exemplo de SQL e o trabalho em equipe.

Além disso, houve a definição de quais atributos são relevantes, baseando-se em conversas e entrevistas com os educadores que utilizam o ambiente. Nesse sentido, destacam-se os atributos relacionados à frequência (quantidade) de acesso; aos acertos nos exercícios, aos acertos nas avaliações (provas), ao número de submissões (SQL-Livre e exercícios), aos trabalhos em equipe e a utilização dos recursos disponíveis do ambiente, como o SQL-Livre e o material de apoio. Esses atributos são considerados relevantes, pois são comumente utilizados pelos professores para avaliar o desempenho e atribuir as notas finais dos educandos na disciplina.

Foram trabalhados aproximadamente 30 atributos. O Quadro 4-2 apresenta a descrição dos principais atributos utilizados na etapa de mineração dos dados que refletem o contexto da turma.

Para uma maior compreensão acerca do domínio dos dados, o Apêndice A apresenta uma análise descritiva dos dados coletados no LabSQL por meio de tabelas e gráficos.

Quadro 4-2 Descrição dos principais atributos utilizados na etapa de mineração dos dados.

Atributo	**Descrição (Domínio)**
acima_média_pontos_alternativas_avaliação	Ficou acima da média de pontos em questões de múltipla escolha nas avaliações (sim ou não)
acima_média_pontos_alternativas_exercícios	Ficou acima da média de pontos em questões de múltipla escolha nos exercícios (sim ou não)
acima_média_pontos_discursivas_avaliação	Ficou acima da média de pontos em questões discursivas nas avaliações (sim ou não)
acima_média_pontos_discursivas_exercícios	Ficou acima da média de pontos em questões discursivas nos exercícios (sim ou não)
acima_média_pontos_SQL_avaliação	Ficou acima da média de pontos em questões de programação SQL nas avaliações (sim ou não)
acima_média_pontos_SQL_exercícios	Ficou acima da média de pontos em questões de programação SQL nos exercícios (sim ou não)
acima_média_de_tentativas_SQL_avaliação	Ficou acima da média de tentativas de programação SQL nas avaliações (sim ou não)
acima_média_ de_tentativas_SQL_exercícios	Ficou acima da média de tentativas de programação SQL nos exercícios (sim ou não)
acima_média_total_de_tentativas_SQL	Ficou acima da média de tentativas de programação SQL nos exercícios e avaliações (sim ou não)
acima_média_total_pontos_alternativas	Ficou acima da média de pontos em questões de múltipla escolha nos exercícios e avaliações (sim ou não)
acima_média_total_pontos_discursivas	Ficou acima da média de pontos em questões discursivas nos exercícios e avaliações (sim ou não)
acima_média_total_pontos_SQL	Ficou acima da média de pontos em questões de programação SQL nos exercícios e avaliações (sim ou não)
acima_qtd_acessos	Ficou acima da média de acessos ao ambiente (sim ou não)
acima_qtd_acessos_material_de_apoio	Ficou acima da média de acesso ao material de apoio disponível no ambiente (sim ou não)
acima_qtd_acessos_exemplo_SQL	Ficou acima da média de acesso aos exemplos de SQL disponíveis no ambiente (sim ou não)
acima_qtd_acessos_SQL_Livre	Ficou acima da média de acesso ao SQL-Livre (sim ou não)
coord_turma	coordenador da turma (2 valores)
nome_tipo_curso	Código do tipo de curso (graduação ou especialização)
sexo_usuário	Gênero do usuário (masculino ou feminino)
trabalhou_em_equipe	Trabalhou em equipe (sim ou não)

4.3 APLICAÇÃO DE ÁRVORE DE DECISÃO

Para aplicação da Árvore de decisão, após a conversão do banco de dados do LabSQL para o formato ARFF (*Attribute-Relation File Format*), utilizou-se a ferramenta WEKA (*Waikato Environment for Knowledge Analysis*) (WEKA, 2014), versão 3.6.5, que executou a tarefa de classificação, utilizando o algoritmo J48 em validação cruzada *10-Fold-Cross-Validation*, a fim testar a acurácia do modelo no mesmo conjunto de dados utilizado para construir o modelo. A acurácia é uma métrica

que avalia os modelos de classificação a partir da porcentagem de predições corretas que o modelo executou sobre o total de predições realizadas. Ela é importante, pois permite avaliar um classificador para determinar o quanto ele será eficiente para predizer dados futuros, ou seja, qual a sua capacidade de generalização.

No total, foram geradas cerca de 20 árvores de decisão para os atributos discretos do banco de dados, sendo geradas cerca de 400 regras. Após a fase de mineração de dados, foram selecionadas sete árvores de decisão, levando-se em consideração a acurácia dos modelos de classificação e a relevância dos atributos para a análise do desempenho dos educandos do LabSQL. Dentre os atributos metas dos modelos de classificação selecionados, destacam-se:

• *acima_media_pontos_alternativas_exercícios*, visando classificar os educandos que estão, ou não, acima da média de pontos em questões de múltipla escolha nos exercícios;

• *acima_qtd_acessos*, visando classificar os educandos que estão, ou não, acima da média de acesso ao ambiente;

• *acima_qtd_acessos_SQL_Livre*, visando classificar os educandos que estão, ou não, acima da média de acesso ao SQL-Livre;

• *acima_média_de_tentativas_sql_avaliação*, visando classificar os educandos que estão, ou não, acima da média de tentativas de programação SQL nas avaliações;

• *acima_média_pontos_sql_exercícios*, visando classificar os educandos que estão, ou não, acima da média de pontos em questões de programação SQL nos exercícios.

4.3.1 **Resultados da Árvore de Decisão**

A média de acurácia dos modelos de classificação selecionados foi de aproximadamente, 80,24%. A partir das árvores de decisão geradas foram extraídas cerca de 80 regras, dentre as quais, foram selecionadas as mais relevantes. Um dos critérios que podem ser utilizados para medir a qualidade das regras geradas por um sistema de aprendizado é a precisão. A precisão é o grau de confiabilidade das regras, geralmente representada a partir da taxa de estimativa de erro.

A Figura 4-1 apresenta a árvore de decisão na forma de regras de classificação para análise dos educandos que estão, ou não, acima da média de pontos em questões de

programação SQL nas avaliações. Nela, observa-se que, aproximadamente, 91% dos educandos que estão abaixo da média de pontos obtidos nas questões de programação SQL nos exercícios (1), e estão abaixo da média de tentativas de resolução das questões de programação SQL nas avaliações (3), estão abaixo da média de pontos nas questões de programação SQL nas avaliações. Destaca-se ainda, que o desempenho nos exercícios de programação SQL, identificado por (1) e (2), é o atributo mais representativo para classificar o desempenho dos educandos nas provas de programação SQL, por ser o nó raiz da árvore de decisão.

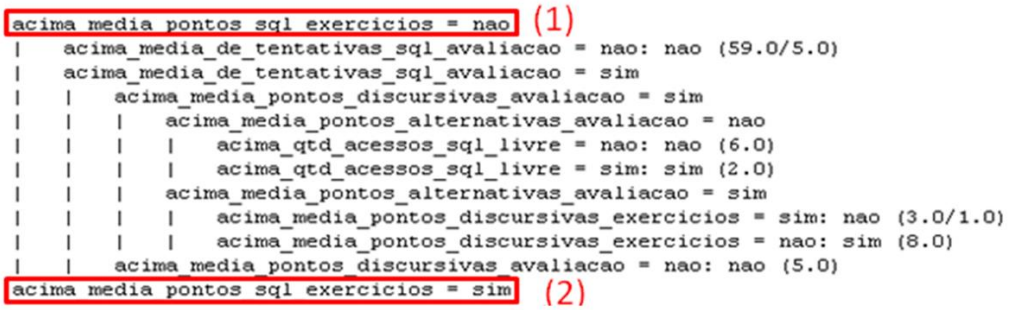


Figura 4-1 Análise dos educandos que estão, ou não, acima da média de pontos em questões de programação SQL nas avaliações.

A partir da árvore de decisão gerada para exibir os educandos que estão, ou não, acima da média de tentativas de resolução das questões de programação SQL nas avaliações, destaca-se a regra acima_média_de_tentativas_sql_exercícios = não: não (201/26). Nela, observa-se que, aproximadamente, 87% dos educandos que estão abaixo da média de tentativas de resolução das questões de programação SQL nos exercícios, também estão abaixo da média de tentativas nas avaliações.

4.4 APLICAÇÃO DE REDES BAYESIANAS

Para aplicação das Redes Bayesianas, após a conversão do banco de dados do LabSQL para o formato separado por tabulação, utilizou-se a ferramenta Bayesware Discoverer (BAYESWARE LIMITED, 2011), na versão que pode ser livremente usada para fins de pesquisa e por instituições acadêmicas. Após a geração das redes bayesianas nesta ferramenta, buscou-se executar inferências nas redes para descobrir informações e padrões que podem ser úteis para gestores do domínio da aplicação. Foram realizadas diferentes análises para as redes bayesianas geradas.

4.4.1 **Resultados de Redes Bayesianas**

A ferramenta *Bayesware Discoverer* construiu as redes a partir dos atributos do banco de dados criado, exibindo as tabelas de probabilidade condicional ou incondicional de cada nó (atributo). Dentre as ligações observadas, a Figura 4-2 apresenta a Rede Bayesiana utilizada para análise dos educandos que estão, ou não, acima da média de pontos em questões de programação SQL nos exercícios e avaliações, representados pela variável preditora acima_média_total_pontos_SQL (1). Nela, observa-se que o atributo acima_média_total_pontos_SQL (1) tem dependência direta dos atributos acima_média_total_de_tentativas_SQL (2) e do atributo acima_média_total_pontos_alternativas (3). Nesse sentido, a média de acertos em questões de SQL nos exercícios e avaliações, depende da média de tentativas de resolução das questões de programação SQL, e da média de pontos em questões de múltipla escolha nos exercícios e avaliações.

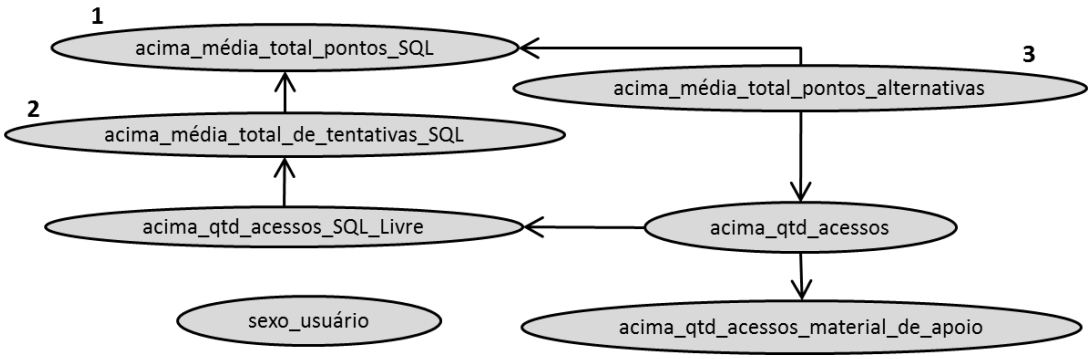


Figura 4-2 Análise dos educandos que estão, ou não, acima da média de pontos em questões de programação SQL nos exercícios e avaliações - atributo acima_média_total_pontos_SQL (1).

Após realizar algumas inferências, conforme observado na Figura 4-3 foi possível observar que ao colocar o atributo acima_média_total_de_tentativas_SQL (2), com 100% para o valor “sim” e o atributo acima_média_total_pontos_alternativas (3), com 100% para o valor “sim”, o atributo acima_média_total_pontos_SQL (1), aumentou de 0,722 (72,2% de probabilidade a priori) para 0,875 (87,5% de probabilidade a posteriori) em “sim”. Observou-se que a acurácia foi de aproximadamente, 77,43%, ou seja, foram classificados corretamente 516 instancias, das 667. Dessa forma, se o usuário está acima da média de tentativas de programação SQL, e se ele está acima da média de pontos em questões de múltipla escolha, então o usuário tem 15,3% de chance a mais de estar acima da média de pontos em questões de programação SQL.

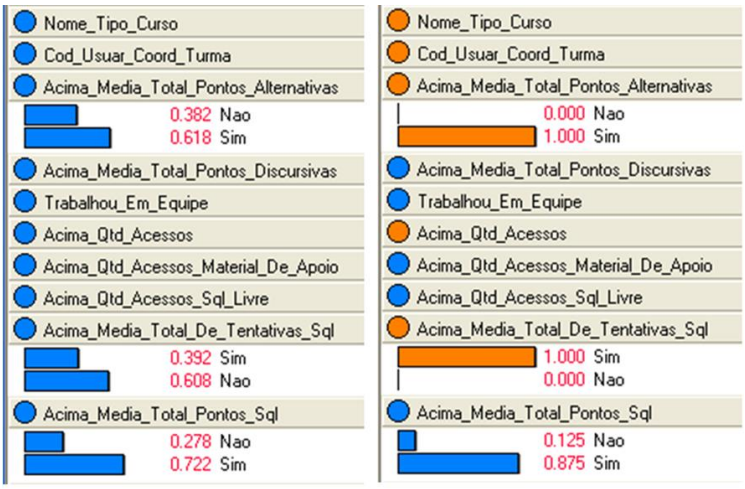


Figura 4-3 Análise dos educandos que estão, ou não, acima da média de pontos em questões de programação SQL nos exercícios e avaliações.

A Figura 4-4 apresenta a rede bayesiana para análise dos educandos que estão, ou não, acima da média de pontos em questões de programação SQL nas avaliações. Nela, observa-se que o atributo meta acima_média_pontos_SQL_avaliação (1) tem dependência direta dos atributos acima_média_pontos_SQL_exercícios (2) e trabalhou_em_equipe (3). Portanto, observa-se que o desempenho nas avaliações de SQL depende da prática de SQL nos exercícios e do desenvolvimento de trabalhos em equipe que favoreçam a aprendizagem colaborativa.

Além disso, destaca-se que o atributo acima_média_pontos_SQL_exercícios (2) depende do atributo acima_qtd_acesso_SQL_Livre (4). Dessa forma, entende-se que o desempenho dos educandos nos exercícios de SQL depende do acesso ao recurso do Interpretador de SQL (SQL-Livre) disponível no ambiente LabSQL. Portanto, deve-se incentivar a utilização deste recurso durante o processo de aprendizagem para melhorar o desempenho dos educandos nos exercícios de SQL e, consequentemente, nas avaliações.

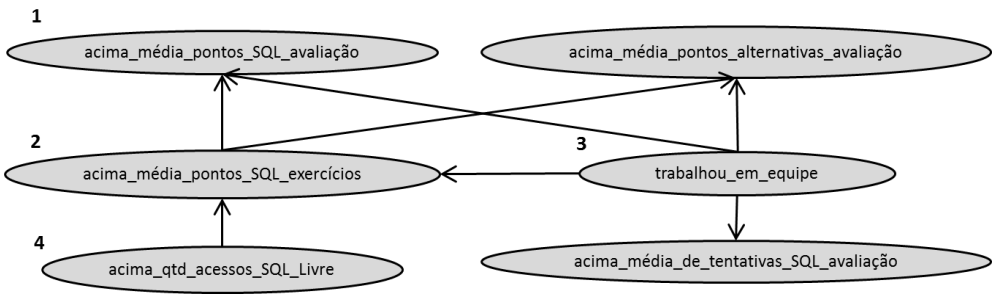


Figura 4-4 Análise dos educandos que estão, ou não, acima da média de pontos em questões de programação SQL nas avaliações.

4.5 APLICAÇÃO DE REGRAS DE ASSOCIAÇÃO

Para aplicação de Regras de associação, utilizou-se a ferramenta WEKA por implementar diversos algoritmos de Associação, dentre eles, o algoritmos *Apriori*, utilizado neste trabalho. Antes de aplicar o algoritmo *Apriori* no WEKA para obtenção de Regras de Associação, é necessário configurar alguns parâmetros, tais como: nome do arquivo de treinamento, o número máximo de regras a serem descobertas pelo algoritmo, a confiança mínima das regras, e o limite superior e inferior para suporte mínimo das regras descobertas.

Para a maioria dos valores de parâmetros definidos neste estudo foi utilizado o valor padrão sugerido pelo WEKA, exceto o número de regras, o valor do suporte mínimo e o valor da confiança mínima. Esses valores foram alterados, pois o valor padrão do WEKA é baixo para o número máximo de regras (10) e muito alto para o suporte (0,1) e para a confiança mínima (0,9). A confiança mínima utilizada foi igual a 0,80 com valor do suporte mínimo igual 0,075.

Para explicar como é realizada a leitura de uma Regra de Associação gerada pelo Weka, analisa-se a regra do Quadro 4-3 e sua descrição em seguida:

Quadro 4-3 Regra de Associação Gerada pelo WEKA.

acima_média_de_tentativas_SQL_exercício = ***sim*** **E** *conceito_SQL_exercício =* ***exc 98***   **ENTÃO** *conceito_SQL_avaliação =* ***exc 79*** *conf:(**0,81**).*

Nela, observa-se que, nas questões de programação SQL, dentre os 98 educandos que obtiveram conceito *excelente* e estão acima da média de tentativas de resolução, nos exercícios, 79 educandos obtiveram também o conceito *excelente* nas avaliações, ou seja, 81% (0,81). A confiança é calculada dividindo 79/98 que é igual a 0,81. Enquanto que o suporte é calculado dividindo 79/667 que é igual a 0,12, onde o valor 667 é o total de instâncias utilizadas neste estudo.

4.5.1 **Resultados de Regras de Associação**

O Quadro 4-4 apresenta a relação das principais regras de associação geradas a partir da mineração de dados educacionais. Essas regras foram selecionadas, pois são consideradas importantes para analisar o perfil dos educandos em relação ao conceito obtido nas avaliações de programação SQL.

Quadro 4-4 Relação das Principais Regras de Associação Selecionadas.

Identificação	**Regra de Associação Selecionada**
Regra 1	*acima_qtd_acessos_exemplo_SQL=**sim*** **E** *acima_média_de_tentativas_SQL_exercício=**sim*** **E** *conceito_SQL_exercício=**exc** 63* **ENTÃO** *conceito_SQL_avaliação=**exc** 53 conf:(0.84)*
Regra 2	*acima_média_ de_tentativas_SQL_avaliação=**não*** **E** *conceito_alternativa_avaliação=**exc*** **E** *conceito_SQL_exercício=**exc** 88* **ENTÃO** *conceito_SQL_avaliação=**exc** 72 conf:(0.82)*
Regra 3	*acima_qtd_acessos=**sim*** **E** *acima_média_ de_tentativas_SQL_avaliação=**não*** **E** *conceito_SQL_exercício=**exc** 72* **ENTÃO** *conceito_SQL_avaliação=**exc** 58 conf:(0.81)*
Regra 4	*acima_qtd_acessos=**sim*** **E** *acima_média_ de_tentativas_SQL_exercício=**sim*** **E** *conceito_SQL_exercício=**exc** 71* **ENTÃO** *conceito_SQL_avaliação=**exc** 57 conf:(0.80)*
Regra 5	*acima_qtd_acessos=**sim*** **E** *acima_média_atraso_inscrição=**não*** **E** *conceito_alternativa_avaliação=**exc*** **E** *conceito_SQL_exercício=**exc** 72* **ENTÃO** *conceito_SQL_avaliação=**exc** 61 conf:(0.85)*
Regra 6	*acima_qtd_acessos=**sim*** **E** *acima_qtd_acessos_exemplo_SQL=**sim*** **E** *acima_média_ de_tentativas_SQL_avaliação=**não*** **E** *conceito_alternativa_avaliação=**exc** 67* **ENTÃO** *conceito_SQL_avaliação=**exc** 57 conf:(0.85)*
Regra 7	*acima_qtd_acessos=**sim*** **E** *acima_qtd_acessos_SQL_Livre=**sim*** **E** *conceito_alternativa_avaliação=**exc*** **E** *conceito_SQL_exercício=**exc** 65* **ENTÃO** *conceito_SQL_avaliação=**exc** 54 conf:(0.83)*
Regra 8	*acima_qtd_acessos=**sim*** **E** *acima_média_atraso_inscrição=**não*** **E** *conceito_SQL_avaliação=**exc*** **E** *conceito_SQL_exercício=**exc** 65* **ENTÃO** *conceito_alternativa_avaliação=**exc** 61 conf:(0.94)*
Regra 9	*acima_qtd_acessos_exemplo_SQL=**não*** **E** *acima_qtd_acessos=**não*** **E** *acima_qtd_acessos_SQL_Livre=**não*** **E** *acima_média_ de_tentativas_SQL_avaliação=**não*** **E** *acima_média_ de_tentativas_SQL_exercício=**não*** **E** *conceito_alternativa_exercício=**ins** 85* **ENTÃO** *conceito_SQL_avaliação=**ins** 70 conf:(0.82)*
Regra 10	*acima_média_ de_tentativas_SQL_avaliação=**não*** **E** *acima_média_ de_tentativas_SQL_exercício=**não*** **E** *conceito_alternativa_avaliação=**ins*** **E** *conceito_SQL_exercício=**ins** 79* **ENTÃO** *conceito_SQL_avaliação=**ins** 79 conf:(1)*

A seguir são descritas cada uma das dez regras selecionadas, apresentadas no Quadro 4-3:

Regra 1: Com confiança de 84%, verifica-se que, o educando obtém conceito *excelente* nas avaliações de programação SQL quando ele está acima da média de

acesso aos exemplos e, nos exercícios de programação SQL, está acima da média de tentativas de resolução e possui conceito *excelente*.

Regra 2: Com confiança de 82%, verifica-se que, o educando obtém conceito *excelente* nas avaliações de programação SQL quando ele está abaixo da média de tentativas de resolução nas avaliações e possui conceito *excelente* nos exercícios de programação SQL e nas avaliações do tipo alternativa.

Regra 3: Com confiança de 81%, verifica-se que, o educando obtém conceito *excelente* nas avaliações de programação SQL quando ele está acima da média de acessos ao LabSQL e nas questões de programação SQL, está abaixo da média de tentativas de resolução nas avaliações e também possui conceito *excelente* nos exercícios.

Regra 4: Com confiança de 80%, verifica-se que, o educando obtém conceito *excelente* nas avaliações de programação SQL quando ele está acima da média de acesso ao LabSQL e nos exercícios de programação SQL, está acima da média de tentativas de resolução e também possui conceito *excelente*.

Regra 5: Com confiança de 85%, verifica-se que, o educando obtém conceito *excelente* nas avaliações de programação SQL quando ele está acima da média de acesso ao LabSQL e abaixo da média de atraso de inscrição, além disso, também possui conceito *excelente* nas avaliações do tipo Alternativa e nos exercícios de programação SQL.

Regra 6: Com confiança de 85%, verifica-se que, o educando obtém conceito *excelente* nas avaliações de programação SQL quando ele está acima da média de acessos ao ambiente LabSQL e aos exemplos de SQL e abaixo da média de tentativas de resolução das questões de programação SQL nas avaliações, além disso, também possui conceito *excelente* nas avaliações do tipo Alternativa.

Regra 7: Com confiança de 83%, verifica-se que, o educando obtém conceito *excelente* nas avaliações de programação SQL quando ele está acima da média de acessos ao ambiente LabSQL e ao recurso do SQL-Livre, além disso, também possui conceito *excelente* nas avaliações do tipo Alternativa e nos exercício de programação SQL.

Regra 8: Com confiança de 94%, verifica-se que, o educando obtém conceito *excelente* nas avaliações do tipo alternativa quando ele está acima da média de acesso ao

ambiente LabSQL e abaixo da média de atraso de inscrições na turma, além disso, também possui conceito *excelente* nas avaliações e exercícios de programação SQL.

Regra 9: Com confiança de 82%, verifica-se que, o educando obtém conceito *insuficiente* nas avaliações de programação SQL quando ele está abaixo da média de acesso aos exemplos, ao ambiente LabSQL e ao recurso do SQL-Livre e abaixo da média de tentativas de resolução nos exercícios e avaliações, além disso, também possui conceito *insuficiente* nos exercícios do tipo alternativa.

Regra 10: Com confiança de 100%, verifica-se que, o educando obtém conceito *insuficiente* nas avaliações de programação SQL quando ele está abaixo da média de tentativas de resolução nos exercícios e avaliações e também possui conceito *insuficiente* nas avaliações do tipo alternativa e nos exercícios de programação SQL.

A análise de regras de associação permite identificar alguns aspectos relevantes sobre o processo de ensino e aprendizagem e a interação do educando no ambiente LabSQL, são eles:

1) Nas regras 1 e 4, observa-se que os educandos possuem melhores condições para obter um conceito *excelente* nas avaliações de Programação SQL, quando possuem um número elevado de acessos ao LabSQL, ou mais especificamente aos exemplos de SQL, e nos exercícios, possuem conceito *excelente* com um número elevado de tentativas de resolução.

2) Nas regras 2, 3 e 6, observa-se que geralmente os educandos com conceito *excelente* nas avaliações de programação SQL realizam um número relativamente baixo de tentativas de resolução nas avaliações.

3) Nas regras 5 e 8, observa-se que geralmente os educandos com conceito *excelente* nas avaliações de programação SQL realizam um número elevado de acessos ao LabSQL e não possuem atraso de inscrição na turma dentro do ambiente. Portanto, desde o início, deve-se motivar a participação dos educandos na utilização do ambiente para desenvolver melhor seu aprendizado, e criar outras possibilidades para aproximar os educandos que estão atrasados em relação ao restante da turma.

4) Na regra 7, observa-se que geralmente o número elevado de acessos ao LabSQL, ou mais especificamente ao recurso do SQL-Livre, garantem aos educandos melhores condições para obter um conceito *excelente* nas avaliações de Programação SQL.

5) Nas regras 9 e 10, observa-se que geralmente os educandos com conceito *insuficiente* nas avaliações de programação SQL possuem um número baixo de acessos ao LabSQL e aos recursos do SQL-Livre e aos exemplos de SQL. Além disso, possuem um número baixo de tentativas de resolução das questões de programação SQL nos exercícios e avaliações.

4.6 APLICAÇÃO DE ANÁLISE DE AGRUPAMENTO

Durante o treinamento foram gerados agrupamentos de 4 até 10 clusters (grupos). Identificou-se que o agrupamento com 7 clusters resultou em centróides mais determinantes para formação de grupos homogêneos.

A Figura 4-5 apresenta as características dos 7 grupos obtidos com o algoritmo de agrupamento (clusterização) *K-means* na ferramenta WEKA, gerados com base nos 667 registros transformados a partir dos dados das atividades dos educandos.

Atributos / **Grupo** / Total de instâncias (667)	**1** 55 (8%)	**2** 42 (6%)	**3** 101 (15%)	**4** 106 (16%)	**5** 64 (10%)	**6** 208 (31%)	**7** 91 (14%)
nome_tipo_curso	especializacao	graduacao	graduacao	graduacao	graduacao	graduacao	graduacao
acima_qtd_acessos_exemplo_SQL	nao	nao	nao	nao	nao	sim	nao
acima_qtd_acessos	sim	nao	nao	nao	sim	sim	nao
acima_qtd_acessos_SQL_Livre	nao	nao	nao	nao	sim	sim	nao
acima_média_pontos_alternativa_avaliação	sim	nao	nao	nao	nao	sim	sim
acima_média_pontos_alternativa_exercício	sim	nao	nao	sim	nao	sim	nao
acima_média_pontos_SQL_avaliação	nao	nao	nao	sim	nao	sim	nao
acima_média_pontos_SQL_exercício	nao	nao	nao	nao	sim	sim	nao
acima_média_de_tentativas_SQL_avaliação	sim	nao	nao	sim	sim	nao	nao
acima_média_de_tentativas_SQL_exercício	sim	nao	nao	nao	nao	sim	nao
conceito_alternativa_avaliação	exc	bom	ins	bom	bom	exc	exc
conceito_alternativa_exercício	bom	ins	ins	bom	reg	bom	reg
conceito_SQL_avaliação	bom	reg	ins	exc	bom	exc	bom
conceito_SQL_exercício	reg	ins	ins	bom	bom	exc	ins
média_de_tentativas_SQL_avaliação	4,06	2,21	0,28	2,94	3,48	2,45	2,03
média_de_tentativas_SQL_exercício	1,92	0,57	0,28	1,95	2,57	3,19	0,71
nota_alternativa_avaliação	9,51	7,75	0,57	8,34	8,15	9,62	9,65
nota_alternativa_exercício	7,85	5,14	2,45	7,85	6,35	7,88	5,71
nota_SQL_avaliação	7,83	7,07	0,58	9,58	7,93	9,52	7,91
nota_SQL_exercício	6,36	3,24	1,39	7,35	7,84	9,03	4,82
qtd_acessos_exemplo_SQL	70,24	36,00	8,13	51,91	92,91	87,37	30,57
qtd_acessos	53,04	29,43	10,42	34,62	52,58	54,94	30,58
qtd_acessos_material_de_apoio	11,02	5,52	2,89	5,60	10,03	7,56	5,07
qtd_acessos_SQL_Livre	34,65	13,74	6,44	14,49	26,91	26,16	12,25
qtd_questões_alternativa_avaliação	12,49	10,86	0,95	9,66	12,41	10,38	11,37
qtd_questões_alternativa_exercício	30,62	17,07	6,27	28,86	30,75	34,26	20,75
qtd_questões_SQL_avaliação	13,24	11,29	1,13	13,75	16,25	14,43	11,54
qtd_questões_SQL_exercício	22,69	8,67	2,46	25,45	30,84	37,31	13,37

Figura 4-5 Grupos de Educandos Gerados

Analisando os clusters da Figura 4-5 é possível realizar algumas interpretações:

Grupo 1: Esse cluster agrupa os educandos de cursos do nível de *especialização*. Ele apresenta uma evolução no desempenho do educando em questões de programação SQL, observada pelo progresso no conceito de *regular* nos exercícios para *bom* nas avaliações, associado ao aumento do número de tentativas de resolução das questões de programação SQL, de 1,92 em exercícios para 4,06 nas avaliações. Analogamente, ocorre uma evolução no conceito de *bom* em exercícios para *excelente* nas avaliações, nas questões de alternativa. Essa evolução de conceito nas questões de alternativa está associado a um número elevado de exercícios realizados (30 questões), em relação ao número máximo de exercícios respondidos do tipo alternativa no ambiente LabSQL (39 questões).

Grupo 2: Este cluster representa os educandos de nível de *graduação*. Ele apresenta uma evolução no desempenho do educando em questões de programação SQL, observada pelo progresso no conceito de *insuficiente* nos exercícios para *regular* nas avaliações, associado ao aumento do número de tentativas de resolução das questões de programação SQL, de 0,57 em exercícios para 2,21 nas avaliações. Analogamente, ocorre uma evolução no conceito de *insuficiente* em exercícios para *bom* nas avaliações, nas questões de alternativa.

Grupo 3: Esse cluster representa os educandos de nível de *graduação* com conceito *insuficiente* em todos os tipos de questões em exercícios e avaliações. Neste grupo, a média de acesso ao sistema é baixa (10 acessos) em relação à média das turmas (36 acessos). Além disso, o número de questões realizadas em exercícios e avaliações é baixo, apenas 17% do total de questões. As notas de questões de programação SQL e de alternativas nas avalições são respectivamente 0,58 e 0,57.

Grupo 4: Esse cluster agrupa os educandos de cursos do nível de *graduação*. Ele apresenta uma evolução no desempenho do educando em questões de programação SQL, observada pelo progresso no conceito de *bom* nos exercícios para *excelente* nas avaliações, associado ao aumento do número de tentativas de resolução das questões de programação SQL, de 1,95 em exercícios para 2,94 nas avaliações. Nas questões de alternativa, o conceito *bom* se mantém nos exercícios e nas avaliações.

Grupo 5: Esse cluster agrupa os educandos de cursos do nível de *graduação*. Nas questões de programação SQL, o conceito *bom* se mantém nos exercícios e nas avaliações. Nas questões de alternativa, ocorre uma evolução no conceito de *regular* em exercícios para *bom* nas avaliações. Essa evolução de conceito nas questões de

alternativa está associado a um número elevado de exercícios realizados (30 questões), em relação ao número máximo de exercícios respondidos do tipo alternativa no ambiente LabSQL (39 questões), semelhante aos grupos **1** e **6** nessa característica. Adicionalmente, observa-se que este grupo apresenta o maior número de acesso aos exemplos de SQL (92,91) e o maior número de questões de programação SQL resolvidas nas avaliações (16,25).

Grupo 6: Esse cluster agrupa os educandos de cursos do nível de *graduação*. O conceito *excelente* se mantém nas avaliações e exercícios em questões de programação SQL. Ocorre uma evolução no conceito de *bom* em exercícios para *excelente* nas avaliações, nas questões de alternativa. Essa evolução de conceito nas questões de alternativa está associado a um número elevado de exercícios realizados (34 questões), em relação ao número máximo de exercícios respondidos do tipo de alternativa no ambiente LabSQL (39 questões), semelhante aos grupos **1** e **5** nessa característica. Adicionalmente, observa-se que este grupo apresenta um desempenho geral acima da média em relação aos demais grupos. Além disso, o destaque está para o maior número de acessos ao sistema e maior número de questões resolvidas.

Grupo 7: Este cluster representa os educandos de nível de *graduação*. Ele apresenta uma evolução no desempenho do educando em questões de programação SQL, observada pelo progresso no conceito de *insuficiente* nos exercícios para *bom* nas avaliações, associado ao aumento do número de tentativas de resolução das questões de programação SQL, de 0,71 em exercícios para 2,03 nas avaliações. Analogamente, nas questões de alternativa, ocorre uma evolução no conceito de *regular* em exercícios para *excelente* nas avaliações.

4.6.1 **Resultados de Análise de Agrupamento**

A análise do agrupamento permite identificar alguns aspectos relevantes sobre o processo de ensino e aprendizagem e a interação do educando no ambiente LabSQL, são eles:

1) O número elevado de tentativas de resolução de questões de programação SQL em exercícios, associado ao número elevado de acessos aos exemplos de SQL e ao recurso do SQL-Livre, corrobora na obtenção de melhor conceito nos exercícios e nas avaliações (rendimento geral).

2) Em exercícios, o número elevado de tentativas de resolução de questões de programação SQL corrobora na obtenção de melhor conceito. Porém, isso ocorre de forma inversa nas tentativas de resolução de questões de programação SQL em avaliações, conforme observado nos grupos **1** e **5**, que têm um número elevado de tentativas em avaliações (4,06 e 3,48) e não obtiveram conceito *excelente*. Diferente dos grupos **4** e **6**, que têm um número menor de tentativas (2,94 e 2,45) e obtiveram o conceito *excelente* nas avaliações de programação SQL. Pode-se concluir que os educandos que realizam várias tentativas em exercícios conseguem solucionar com mais facilidade as questões nas avaliações, portanto não precisam realizar um número elevado de tentativas para solucionar as questões na avaliação. Por outro lado, constata-se que os educandos que executaram poucas tentativas nos exercícios não conseguem evoluir o conceito para *excelente*, mesmo que executem um número elevado de tentativas nas avaliações.

3) Os grupos **2** e **7** possuem uma característica comum, a evolução no conceito de programação SQL, de *insuficiente* para *regular* e *insuficiente* para *bom*, respectivamente. Destaca-se que existe uma pequena diferença entre as características desses grupos que permitem que o educando ao invés de evoluir do conceito de *insuficiente* para *regular*, possa evoluir diretamente para o conceito *bom*. Para isso, o educando que for identificado com as características do grupo **2** deve interagir com o LabSQL para aumentar o número de tentativas de resolução nos exercícios de programação SQL e fazer um número adicional de resoluções de exercícios em questões de: programação SQL (5 questões) e de alternativa (3 questões). Adicionalmente, existe a possibilidade desse educando atingir ao invés do conceito *bom* nas questões de alternativa nas avaliações, conseguir o conceito *excelente*.

4) Os grupos **5** e **6** de uma forma geral, apresentam um desempenho superior comparados com os demais grupos, obtendo os conceitos *bom* e *excelente* nas avaliações, respectivamente. Destaca-se que existe uma pequena diferença entre as características desses grupos que permitem que o educando ao invés de obter do conceito *bom*, passe a obter o conceito *excelente*. Para isso, o educando que for identificado com as características do grupo **5** deve interagir com o LabSQL para aumentar o número de tentativas de resolução nos exercícios de programas SQL e fazer um número adicional de exercícios de programação SQL (7 questões) e de alternativa (4

questões). Adicionalmente, existe a possibilidade deste educando atingir ao invés do conceito *bom* nas questões de alternativa nas avaliações, conseguir o conceito *excelente*.

Dessa forma, a aplicação das técnicas de Mineração de Dados: Árvore de Decisão, Redes Bayesianas, Regras de Associação e Análise de Agrupamento, apresentadas no Capítulo 3, mostraram-se eficientes para alcançar o objetivo desta pesquisa, pois possibilitaram: prever o desempenho dos educando, identificar os educandos com desempenhos similares, verificar a influência dos exercícios e dos trabalhos em equipe nas avaliações, além de gerar informações relevantes sobre o perfil dos educandos.

Além disso, os resultados obtidos a partir da Mineração de Dados Educacionais realizada com base nos registros de atividades dos educandos no ambiente virtual LabSQL, apresentado no Capítulo 2, mostram a grande importância que a existência destes registros provenientes de ambientes virtuais assume para a avaliação das aprendizagens *online*, por possibilitar entender melhor os educandos e o seu processo de ensino e aprendizagem.

CAPÍTULO 5

CONCLUSÃO

O Capítulo 5 apresenta as considerações finais do trabalho, faz uma síntese dos resultados obtidos a partir da mineração de dados educacionais do ambiente virtual LabSQL, além de propor trabalhos futuros.

5.1 CONCLUSÕES

Diversificar os períodos, as fontes e as ferramentas de avaliação é medida importante na educação *online*, pois auxilia o educador a construir o perfil dos educandos por meio do cruzamento de informações. Neste estudo do ambiente virtual LabSQL, as técnicas de mineração de dados utilizadas mostraram-se eficientes para alcançar o objetivo de apoiar os educadores na avaliação das aprendizagens *online*, pois permitiram analisar o perfil dos educandos, em relação à utilização dessa tecnologia e ao processo de ensino-aprendizagem.

Assim, para cada objetivo específico desta pesquisa, relacionado à aplicação das técnicas de Mineração de Dados Educacionais do ambiente virtual LabSQL, destacam-se os resultados a seguir.

O primeiro objetivo, que consiste em prever o desempenho dos educandos em exercícios e avaliações de aprendizado, foi alcançado a partir da aplicação de Árvore de Decisão. Como resultado dessa aplicação, foi possível prever que o educando possui 91% de chance de ter um desempenho abaixo da média nas avaliações, se estiver abaixo da média na pontuação dos exercícios e no número de tentativas de resolução das avaliações. Observou-se também, que o educando possui 87% de chances de estar abaixo da média de tentativas de resolução das questões nas avaliações, se estiver abaixo da média de tentativas de resolução nos exercícios.

Dessa forma, as árvores de decisão permitiram perceber padrões referentes ao processo de aprendizagem relacionado ao comportamento dos educandos, levando em consideração as regras que são mais relevantes para indicar como o educando pode aprimorar a aprendizagem utilizando melhor o ambiente.

O segundo objetivo visa identificar os educandos com desempenhos similares para receberem direcionamento personalizado. A aplicação de Análise de Agrupamento possibilitou identificar grupos de educandos que necessitam de um número adicional de exercícios, realizando um número maior de tentativas de resolução para melhorar o conceito nas avaliações. Portanto, os educandos que realizam um número elevado de tentativas de resolução em exercícios conseguem solucionar com mais facilidade as questões nas avaliações. Por outro lado, constatou-se que os educandos que executam poucas tentativas nos exercícios não conseguem evoluir o conceito para *excelente*, mesmo que executem um número elevado de tentativas nas avaliações.

Com relação ao terceiro objetivo, que pretende verificar a influência de trabalhos em equipe e de exercícios nas avaliações, os resultados foram obtidos a partir da aplicação das Redes Bayesianas. Verificou-se que o desempenho dos educandos nas avaliações de programação SQL depende da prática de exercícios e do desenvolvimento de trabalhos em equipe que favoreçam a aprendizagem colaborativa. Além disso, as Redes Bayesianas permitiram contabilizar relações de dependência entre as ações envolvidas no processo de aprendizagem e o desempenho obtido pelos educandos.

Finalmente, com relação ao quarto objetivo, que consiste em gerar informações relevantes sobre o perfil dos educandos, os resultados foram obtidos a partir da aplicação de Regras de Associação. Constatou-se que o número elevado de tentativas de resolução dos exercícios de SQL e o número elevado de acessos ao ambiente LabSQL, principalmente o acesso aos recursos do SQL-Livre e aos exemplos de SQL, garantem aos educandos melhores condições para obter um conceito excelente nas avaliações de programação SQL. Observou-se também que, geralmente os educandos com conceito excelente nas avaliações de programação SQL realizam um número elevado de acessos ao LabSQL e não possuem atraso de inscrição na turma dentro do ambiente. Portanto, desde o início, deve-se motivar a participação dos educandos na utilização do ambiente e na descoberta das funcionalidades para desenvolver melhor seu aprendizado, e criar outras possibilidades para auxiliar e aproximar os educandos que estão atrasados em relação ao restante da turma.

Dessa forma, os resultados obtidos a partir da aplicação das técnicas de MDE no ambiente LabSQL oferecem ao educador a possibilidade de detectar possíveis problemas na aprendizagem dos educandos, que por sua vez, possuem maiores possibilidades para aprendizagens mais efetivas.

Assim, por meio da avaliação dos processos de aprendizagem com o uso de tecnologias de Mineração de Dados pode-se estabelecer melhor o perfil e o desempenho de cada educando, beneficiando o processo de avaliação e permitindo a evolução dos AVAs.

5.2 TRABALHOS FUTUROS

Como trabalhos futuros o projeto tem quatro principais metas:

(a) Criar um agente que usa o perfil do educando e as regras extraídas para compor um *feedback* que subsidia o educando a obter maior aprendizagem no curso;

(b) Implementar um módulo dentro do LabSQL para gerar um *feedback* para os educandos a partir das regras geradas pela MDE;

(c) Realizar a mineração de dados dos logs de navegação dos educandos dentro do LabSQL;

(d) Aplicar técnicas de visualização de informação e outras técnicas de Mineração de Dados, como Redes Neurais e Algoritmos Genéticos, sobre a base de dados do ambiente LabSQL.

(e) Dado os grupos de educandos, identificar o fluxo de navegação utilizados no AVA LabSQL.

5.3 PUBLICAÇÕES

O trabalho desenvolvido na linha de pesquisa dessa dissertação gerou algumas publicações, tais como:

i) Trabalhos Completos Publicados em Anais de Congressos:

- DIAS, M. M.; FAVERO, E. L.; LINO, A. D. P.. **Mineração de Dados na Avaliação de Aprendizagem Online: um Estudo de Caso no Ambiente Virtual LabSQL**. In: XXXVII Conferência Latino-Americana de Informática - CLEI, 2011, Quito - Equador. Anais da XXXVII Conferencia Latinoamericana de Informática - CLEI 2011, 2011.

- DIAS, M. M.; SILVA FILHO, L. A.; LINO, A. D. P.; FAVERO, E. L.; RAMOS, E. M. L. S.. **Aplicação de Técnicas de Mineração de Dados no Processo de Aprendizagem na Educação a Distância**. In: XIX Simpósio Brasileiro de Informática na Educação (SBIE 2008), 2008, Fortaleza - CE. Anais do XIX Simpósio Brasileiro de Informática na Educação (SBIE 2008). Porto Alegre - RS: Sociedade Brasileira de Computação, 2008. p. 105-114.

ii) Resumos publicados em anais de congressos:

- DIAS, M. M.; SILVA FILHO, L. A.. **Acompanhamento do Aprendizado na Educação a Distância a partir da Aplicação de Técnicas de Mineração de Dados**. In: CNMAC 2008 - 31º Congresso Nacional de Matemática Aplicada e Computacional, 2008, Belém. CNMAC 2008, 2008.

- DIAS, M. M; FAVERO, E. L.; **Mineração de Dados Educacionais para Acompanhamento da Aprendizagem Online: um Estudo de Caso no Ambiente Virtual LabSQL**. IV Seminário Regional de Política e Administração da Educação da Região Norte, 2012, Santarém, PA - Brasil.

REFERÊNCIAS

AGRAWAL, R.; SRIKANT, R. **Fast algorithms for mining association rules**. In: J. B. Bocca, M. Jarke, and C. Zaniolo (Eds.), Proccedings 20th International Conference on Very Large Data Bases, VLDB, Santiago, Chile, p. 487–499. Morgan Kaufmann. 1994.

AMERSHI, S., CONATI, C. COMBINING **Unsupervised and Supervised Classification to Build User Models for Exploratory Learning Environments**. Journal of Educational Data Mining, 1(1): 18-71. 2009.

BAKER, R. S. J. D., ISOTANI, S., de CARVALHO, A. **Mineração de Dados Educacionais: Oportunidades para o Brasil**. In Revista Brasileira de Informática na Educação. v. 19, n. 02, 2011.

BAKER, R.S.J.D. **Data Mining for Education**. McGaw, B., PETERSON, P., BAKER, E. (Eds.) International Encyclopedia of Education (3rd edition). Oxford, UK: Elsevier. 2010.

BAKER, R.S.J.D., CORBETT, A.T., ROLL, I., KOEDINGER, K.R. **Developing a Generalizable Detector of When Students Game the System**. User Modeling and User-Adapted Interaction, 18 (3): 287- 314. 2008.

BAKER, R.S.J.D., de CARVALHO, A.M.J.A., RASPAT, J., ALEVEN, V., CORBETT, A.T., KOEDINGER, K.R. **Educational Software Features that Encourage and Discourage Gaming the System.** In Proceedings of the International Conference on Artificial Intelligence in Education, 475- 482. 2009.

BARNES, T., BITZER, D., VOUK, M. **Experimental Analysis of the Q-Matrix Method** in Knowledge Discovery. Lecture Notes in Computer Science 3488: Foundations of Intelligent Systems. 603-611. 2005.

BAYESWARE LIMITED. "**Bayesware Discoverer**". Janeiro (2011), http://www.bayesware.com.

BERSON, A. e SMITH, S.J. **Data Warehousing, Data Mining and OLAP**. EUA. Mac-Graw-Hill, 1997.

BRANDÃO, M. F. R., RAMOS, C. R. S., TRÓCCOLI, B. T. **Análise de agrupamento de escolas e Núcleos de Tecnologia Educacional**: Mineração na base de dados de avaliação do Programa Nacional de Informática na Educação, 366-374, 2006.

CHAVES, R.O. ; MIRANDA, T.; TAVARES, E. M.; Oliveira S. R. B.; FAVERO, E. L.. **DESIGMPS: Um jogo de apoio ao ensino de modelos de qualidade de processos de software baseado em mapas conceituais**. In: 22º Simpósio Brasileiro de Informática na Educação e 17º Workshop de Informática na Escola, 2011, Aracaju. anais 22º Simpósio Brasileiro de Informática na Educação e 17º Workshop de Informática na Escola, 2011.

CHEN, C. M., Li, C.; CHAN, T.Y.; JONG, B.S. e LIN, T.W. **Diagnosis of Students' *Online* Learning Portfolios**. 37th ASEE/IEEE Frontiers in Education Conference, Milwaukee, WI, 2007.

CORTEZ P. e SILVA A. **Using Data Mining To Predict Secondary School Student Performance** 5th FUture BUsiness TEChnology Conference. Porto, Portugal, 2008.

DESMARAIS, M.C., PU, X. **A Bayesian Student Model without Hidden Nodes and Its Comparison with Item Response Theory**. International Journal of Artificial Intelligence in Education 15: 291-323, 2005.

DIAS, M. M.; SILVA FILHO, L. A.; LINO, A. D. P.; FAVERO, E. L. e RAMOS, E. M. L. S. **Aplicação de Técnicas de Mineração de Dados no Processo de Aprendizagem na Educação a Distância.** In: XIX Simpósio Brasileiro de Informática na Educação, Fortaleza - CE (2008), p. 105-114.

DIAS, MAXWEL, L. A., LINO, A. D. P., ELÓI L. FAVERO. **Mineração de Dados na Avaliação de Aprendizagem *Online*: um Estudo de Caso no Ambiente Virtual LabSQL** In: XXXVII Conferencia Latinoamericana de Informática - CLEI, 2011, Quito Equador. Conferencia LatinoAmericana de Informática CLEI , 2011. v.1. p.1 – 10.

FACELI, K.; LORENA, A. C.; GAMA, J.; CARVALHO, A. C. P. L. F. **Inteligência Artificial: uma Abordagem de Aprendizado de Máquina**. Rio de Janeiro: LTC, 2011.

FAYYAD, U. M.; PIATETSKY-SHAPIRO, G.; SMYTH, P. **From data Mining to knowledge Discovery**: An overview. In: Advances in Knowledge Discovery and Data Mining, AAAI Press / The MIT Press, MIT, Cambridge, Massachusetts, and London, England, 1996, p.1-34.

GOMES, M. J. **Problemáticas da Avaliação em Educação *Online***, In: Educação *Online*: cenário, formação e questões didático-metodológicas. Organizado por Marco Silva, Lucila Pesce e Antônio Zuin. Rio de janeiro: Wak. (2010), p. 309-336.

GIUDICI, P. **Applied Data Mining : Statistical Methods for Business and Industry**. John Wiley & Sons Ltd, The Atrium, Southern Gate, Chichester, West Sussex PO19 8SQ, England. 2003.

HAIR, J. F.; ANDERSON, R. E.; TATHAN, R. L.; BLACK, W. C.. **Multivariate Data Analysis**. New Jersey: Prentice Hall, 5th Ed., 1998.

HERSHKOVITZ, A., NACHMIAS, R. **Developing a Log-Based Motivation Measuring Tool**. In Proceedings of the International Conference on Educational Data Mining, 226-233. 2008.

HOFFMANN, JUSSARA. **Avaliação: Mito e Desafio. Uma Perspectiva Construtivista**. Rio grande do Sul: Mediação, 1998.

International Educational Data Mining Society. julho (2011). http://www.educationaldatamining.org.

KAY, J., MAISONNEUVE, N., YACEF, K. REIMANN, P. **The Big Five and Visualisations of Team Work Activity**. In Proceedings of Intelligent Tutoring Systems (ITS06). 197-206. 2006.

KOEDINGER, K.R., BAKER, R.S.J.d., CUNNINGHAM, K., SKOGSHOLM, A., LEBER, B., STAMPER, J. **A Data Repository for the EDM commuity**: The PSLC DataShop. ROMERO, C., VENTURA, S., PECHENIZKIY, M., BAKER, R.S.J.d. (Eds.) Handbook of Educational Data Mining. Boca Raton, FL: CRC Press. 2010.

LINO, A. D. P. **LabSQL - Laboratório de Ensino de SQL**. Dissertação de Mestrado. Programa de Pós-Graduação em Engenharia Elétrica - Universidade Federal do Pará, 2007.

LINO, A. D. P.; SILVA, A. do S.; FAVERO, E. L.; BRITO, S. R.; HARB, M. P. A. A. **Avaliação Automática de Consultas SQL em Ambiente Virtual de Ensino-Aprendizagem**. II Conferência Ibérica de Sistemas e Tecnologias de Informação, Porto, 2007.

LUNA, J. E. O. **Algoritmos EM para Aprendizagem de Redes Bayesianas a partir de Dados Incompletos**. Dissertação (Mestrado em Ciência da Computação) - Departamento de Computação e Estatística, Universidade Federal de Mato Grosso do Sul - UFMS, Campo Grande, 2004.

MARTINS, J. G. e CAMPESTRINI, B. B. **Ambiente Virtual de Aprendizagem Favorecendo o Processo Ensino-Aprendizagem em Disciplinas na Modalidade de Educação *Online* no Ensino Superior**. In: XI Congresso Intenacional de Educação a Distância. Salvador-BA, 2004.

MASETTO, M. T. **Mediação Pedagógica e o Uso da Tecnologia**. In: Moran, J. M., Masetto, M. T., Behrens, M. A. Novas Tecnologias e Mediação Pedagógica. Campinas: Editora Papirus, 2000.

MOORE, A. **Statistical Data Mining Tutorials.** (2005) Available *online* at : <http://www.autonlab.org/tutorials/> Acesso em 05 maio de 2012.

OVERVIEW DATA MINING: Curso de Inteligência Tecnológica - IME, Rio de Janeiro, 2005. 6 p.

PAIVA, R. ; BITTENCOURT, I. I. ; PACHECO, H. ; SILVA, A. P. ; JAQUES, P. ; ISOTANI, S. . **Mineração de Dados e a Gestão Inteligente da Aprendizagem: Desafios e Direcionamentos**. In: Workshop de Desafios da Computação Aplicada à Educação, 2012, Curitiba. Anais do Congresso da Sociedade Brasileira de Computação, 2012. p. 1-10.

PAVLIK, P., CEN, H., WU, L. and KOEDINGER, K. **Using Item-type Performance Covariance to Improve the Skill Model of an Existing Tutor.** In Proceedings of the International Conference on Educational Data Mining, 77-86. 2008.

PIMENTEL, E.P., OMAR, N. **Descobrindo Conhecimentos em Dados de Avaliação Aprendizagem com Técnicas de Mineração de Dado.** Workshop sobre Informática na Escola. Anais do Congresso da Sociedade Brasileira de Computação, 147- 155, 2006.

PIVATO, M. **Mineração de Regras de Associação em Dados Georreferenciados**. Dissertação de Mestrado. ICMC-USP. 2006.

ROCHA, F. E. L. da; COSTA JUNIOR, J. V.; FAVERO, E. L.. **Como usar ontologias na avaliação de aprendizagem significativa mediado por mapas conceituais**. Revista Brasileira de Informática na Educação, sp, v. 13, n.2, p. 53-64, 2005.

ROMERO, C. e VENTURA, S. **Educacional data mining: A survey from 1995 to 2005**. Expert Systems with Applications. v. 33. (2007), p.135-146.

ROMERO, C.; VENTURA, S.; PECHENIZKIY; BAKER, R.. **Handbook of Educational Data Mining**. Florida: CRC Press, 2011.

RUSSELL, S. J., NORVIG, P. (2004), **Inteligência Artificial**, 2ª Edição, Editora Elsevier, Rio de Janeiro - RJ.

SANCHO, J. M. **Para promover o debate sobre os ambientes virtuais de ensino e aprendizagem**, In: Educação *Online*: cenário, formação e questões didático-metodológicas. Organizado por Marco Silva, Lucila Pesce e Antônio Zuin. Rio de janeiro: Wak. (2010), p. 95-106.

SANTOS, E. **Educação *online* além da EAD: um fenômeno da cibercultura**, In: Educação *Online*: cenário, formação e questões didático-metodológicas. Organizado por Marco Silva, Lucila Pesce e Antônio Zuin. Rio de janeiro: Wak. (2010), p. 29-48.

SCHEINES, R., SPRITES, P., GLYMOUR, C., MEEK, C. **Tetrad II: Tools for Discovery**. Lawrence Erlbaum Associates: Hillsdale, NJ. 1994.

SILVA, J. F. **Avaliação do Ensino e da Aprendizagem numa Perspectiva Formativa Reguladora**. In: SILVA, J. F.; HOFFMANN, J.; ESTEBAN, M. T. Práticas Avaliativas e Aprendizagens Significativas: em Diferentes Áreas do Currículo. Editora Mediação, Porto Alegre, 2012.

SILVA, M. **Desenho didático: contribuições para a pesquisa sobre formação de professores para a docência *online***, In: Educação *Online*: cenário, formação e questões didático-metodológicas. Organizado por Marco Silva, Lucila Pesce e Antônio Zuin. Rio de janeiro: Wak. (2010), p. 215-231

SMARAGDAKIS, G., I. M., BESTAVROS, A. **SEP: A Stable Election Protocol for clustered heterogeneous wireless sensor networks**, in: Second International Workshop on Sensor and Actor Network Protocols and Applications (SANPA), 2004.

SOUMEN C., MARTIN E., USAMA F., JOHANNES G., JIAWEI H., SHINICHI M., GREGORY P.S., WEI W. **Data Mining Curriculum: A Proposal (Version 1.0),** Intensive Working Group of ACM SIGKDD, Curriculum Committee, April 30, 2006

TAN, P.; STEINBACH, M.; KUMAR, V. **Introdução ao Data Mining Mineração de dados**. Rio de Janeiro: Ed. Ciência Moderna, 2009.

WANG, L.; MEINEL, C. **Mining the Student's Learning Interest in Browsing Web-Streaming Lectures**. IEEE - Symposium on Computational Intelligence and Data Mining, 2007.

WANG, J.; SU, X. **An improved K-means clustering algorithm**, in 3rd International Conference on Communication Software and Networks (ICCSN), Xi'an, 2011.

WEKA. **Data Mining Software in Java**. Janeiro (2014). http://www.cs.waikato.ac.nz/ml/weka.

ZAIANE, O. e LUO, J. **Towards Evaluating Learners Behaviour in a Web-Based Distance Learning Environment**. In: International Conference on Learning Technologies, Madison, 2001.

Apêndice A – ANÁLISE DESCRITIVA DOS DADOS

Gênero

A Tabela 1 apresenta o percentual de usuários que utilizaram o LabSQL, no período de 2008 a 2012, por gênero. Nela, verifica-se que a maioria dos usuários é do gênero masculino, com 78,70%. A Figura 1 apresenta graficamente este percentual.

Tabela 1 Percentual de Usuários que Utilizaram o LabSQL, no Período de 2008 a 2012, por Gênero.

Gênero	Percentual
Masculino	78,70
Feminino	21,30
Total	100,00

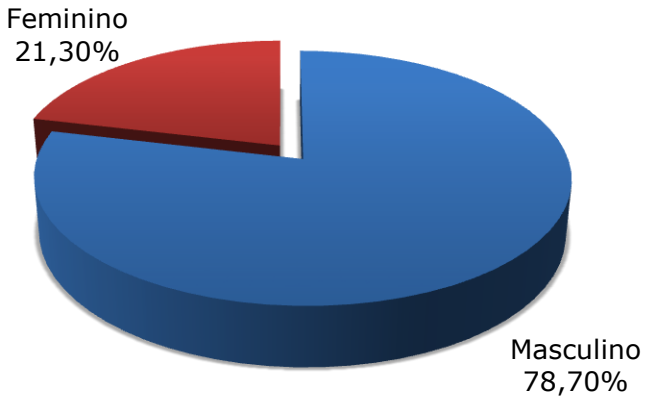


Figura 1 Percentual de Usuários que Utilizaram o LabSQL, no Período de 2008 a 2012, por Gênero.

Tipo de Curso

A Tabela 2 apresenta o percentual de usuários que utilizaram o LabSQL, no período de 2008 a 2012, por tipo de curso. Nela, verifica-se que a maioria dos usuários é educando de graduação, com 74,47%. A Figura 2 apresenta graficamente este percentual.

Tabela 2 Percentual de Usuários que Utilizaram o LabSQL, no Período de 2008 a 2012, por Tipo de Curso.

Tipo de curso	Percentual
Graduação	74,47
Especialização	25,53
Total	100,00

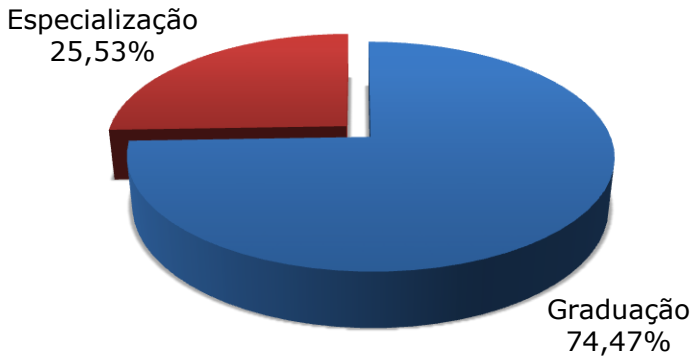


Figura 2 Percentual de Usuários que Utilizaram o LabSQL, no Período de 2008 a 2012, por Tipo de Curso.

Curso

A Tabela 3 apresenta o percentual de usuários que utilizaram o LabSQL, no período de 2008 a 2012, por curso. Nela, verifica-se que a maioria dos usuários de graduação são educandos do curso de Ciência da Computação, com 50,67%, e a maioria dos educandos de especialização são do curso de Banco de Dados, com 64,88%.

Tabela 3 Percentual de Usuários que Utilizaram o LabSQL, no Período de 2008 a 2012, por Tipo de Curso.

Curso	Graduação	Especialização
Ciência da Computação	50,67	-
Engenharia da Computação	2,51	-
Sistemas de Informação	46,82	
Especialização em Aplicações WEB	-	35,12
Especialização em BD	-	64,88
Total	100,00	100,00

Turma

A Tabela 4 apresenta a quantidade e o percentual de usuários que utilizaram o LabSQL, no período de 2008 a 2012, por turma. Nela, observa-se que a média de usuários por turma é de 26 educandos, aproximadamente.

Tabela 4 Quantidade de Usuários que Utilizaram o LabSQL, no Período de 2008 a 2012, por Turma.

Turma	Quantidade
T01	28
T02	26
T03	21
T04	23
T05	34
T06	35
T07	33
T08	31
T09	27
T10	22
T11	27
T12	23
T13	27
T14	21
T15	25
T16	28
T17	27
T18	29
T19	29
T20	28
T21	38
T22	23
T23	18
T24	26
T25	18
Total	667
Média	26,68

Conceito em Questões de Programação SQL

A Tabela 5 apresenta o percentual dos conceitos em questões de programação SQL obtidos pelos usuários que utilizaram o LabSQL, no período de 2008 a 2012. Nela, verifica-se que a maioria dos usuários obteve conceito Excelente, com 63,80%. A Figura 3 apresenta graficamente este percentual.

Tabela 5 Percentual dos Conceitos em Questões de Programação SQL Obtidos pelos Usuários que Utilizaram o LabSQL, no Período de 2008 a 2012.

Conceito	Percentual
Insuficiente	21,02
Regular	2,49
Bom	12,69
Excelente	63,80
Total	100,00

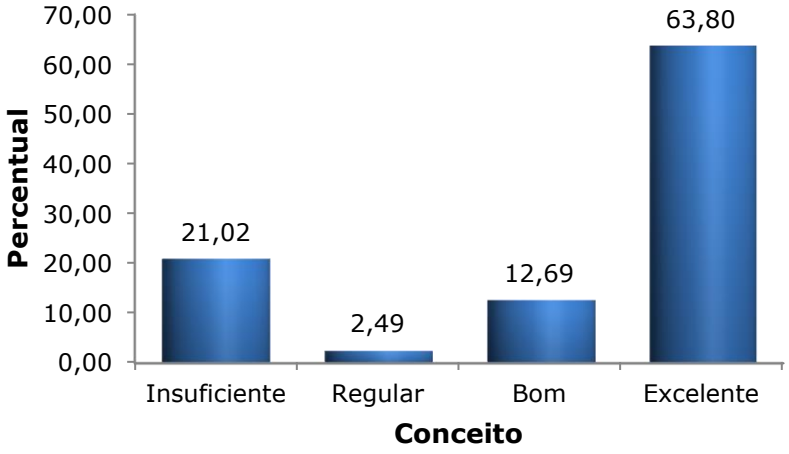


Figura 3 Percentual dos Conceitos em Questões de Programação SQL Obtidos pelos Usuários que Utilizaram o LabSQL, no Período de 2008 a 2012.

Acesso ao LabSQL - Acima ou Abaixo da Média da Turma

A Tabela 6 apresenta o percentual de usuários que estão acima ou abaixo da média de acesso ao LabSQL, no período de 2008 a 2012. Nela, verifica-se que a maioria dos usuários está abaixo da média de acesso ao LabSQL em relação a sua turma, com 55,60%. A Figura 4 apresenta graficamente este percentual.

Tabela 6 Percentual de Usuários que Estão Acima ou Abaixo da Média de Acesso ao LabSQL, no Período de 2008 a 2012.

Login	Percentual
Não	55,60
Sim	44,40
Total	100,00

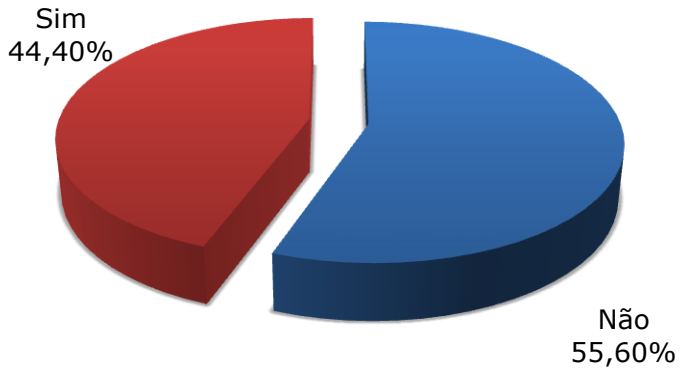


Figura 4 Percentual de Usuários que Estão Acima ou Abaixo da Média de Acesso ao LabSQL, no Período de 2008 a 2012.

Acesso ao Material Apoio - Acima ou Abaixo da Média da Turma

A Tabela 7 apresenta o percentual de usuários que estão acima ou abaixo da média de acesso ao material de apoio disponível no LabSQL, no período de 2008 a 2012. Nela, verifica-se que a maioria dos usuários está abaixo da média de acesso ao material de apoio em relação a sua turma, com 69,78%. A Figura 5 apresenta graficamente este percentual.

Tabela 7 Percentual de Usuários que Estão Acima ou Abaixo da Média de Acesso ao Material de Apoio do LabSQL, no Período de 2008 a 2012.

Material de Apoio	Percentual
Não	69,78
Sim	30,22
Total	100,00

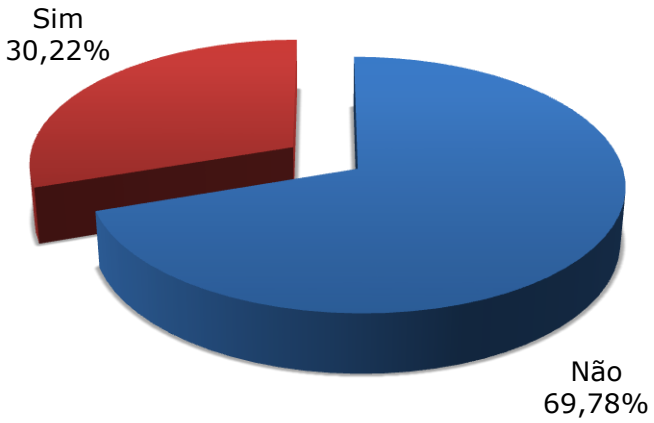


Figura 5 Percentual de Usuários que Estão Acima ou Abaixo da Média de Acesso ao Material de Apoio do LabSQL, no Período de 2008 a 2012.

Acesso ao SQL-Livre - Acima ou Abaixo da Média da Turma

A Tabela 8 apresenta o percentual de usuários que estão acima ou abaixo da média de acesso ao recurso do SQL-Livre do LabSQL, no período de 2008 a 2012. Nela, verifica-se que a maioria dos usuários está abaixo da média de acesso ao SQL-Livre em relação a sua turma, com 63,81%. A Figura 6 apresenta graficamente este percentual.

Tabela 8 Percentual de Usuários que Estão Acima ou Abaixo da Média de Acesso ao Recurso do SQL-Livre do LabSQL, no Período de 2008 a 2012.

Acesso ao SQL-Livre	Percentual
Sim	36,19
Não	63,81
Total	100,00

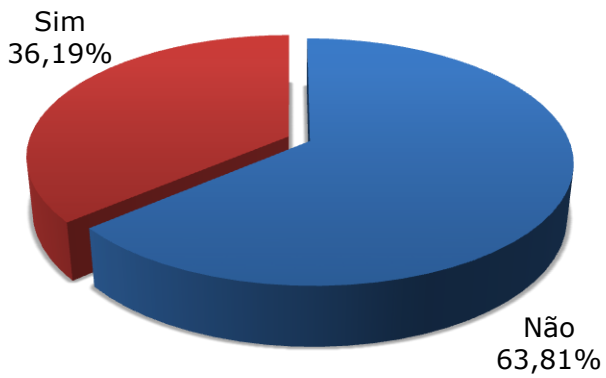


Figura 6 Percentual de Usuários que Estão Acima ou Abaixo da Média de Acesso ao Recurso do SQL-Livre do LabSQL, no Período de 2008 a 2012.

Atraso de Inscrição

A Tabela 9 apresenta o percentual de usuários que estão acima ou abaixo da média de atraso de inscrição no LabSQL, no período de 2008 a 2012. Nela, verifica-se que a maioria dos usuários está abaixo da média do atraso de inscrição em relação a sua turma, com 66,67%. A Figura 7 apresenta graficamente este percentual.

Tabela 9 Percentual de Usuários que Estão Acima ou Abaixo da Média de Atraso de Inscrição no LabSQL, no Período de 2008 a 2012.

Atraso de Inscrição	Percentual
Sim	33,33
Não	66,67
Total	100,00

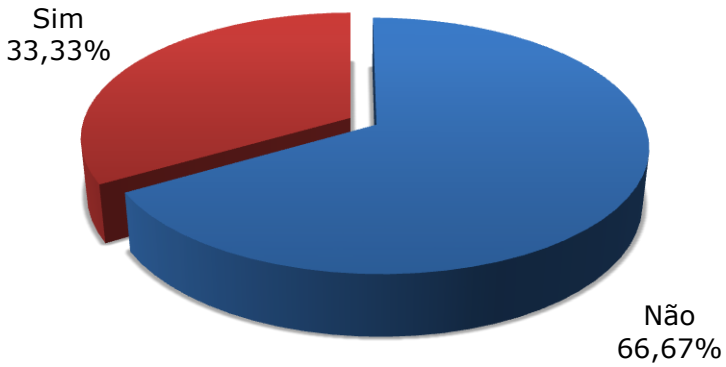


Figura 7 Percentual de Usuários que Estão Acima ou Abaixo da Média de Atraso de Inscrição no LabSQL, no Período de 2008 a 2012.

Questões de Alternativa (Múltipla Escolha) - Acima ou Abaixo da Média de Pontos da Turma

A Tabela 10 apresenta o percentual de usuários que estão acima ou abaixo da média de pontos em questões de múltipla escolha do LabSQL, no período de 2008 a 2012. Nela, verifica-se que a maioria dos usuários está abaixo da média de pontos em questões de múltipla escolha em relação a sua turma, com 53,48%. A Figura 8 apresenta graficamente este percentual.

Tabela 10 Percentual de Usuários que Estão Acima ou Abaixo da Média de Pontos em Questões de Múltipla Escolha do LabSQL, no Período de 2008 a 2012.

Múltipla escolha	Percentual
Sim	46,52
Não	53,48
Total	100,00

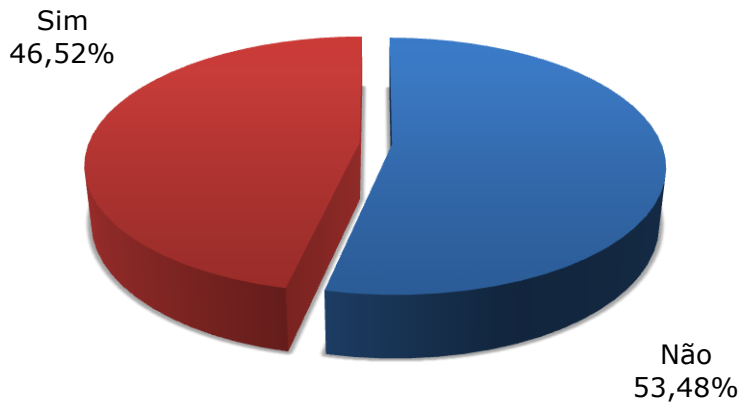


Figura 8 Percentual de Usuários que Estão Acima ou Abaixo da Média de Pontos em Questões de Múltipla Escolha do LabSQL, no Período de 2008 a 2012.

Questões de Programação SQL - Acima ou Abaixo da Média de Pontos da Turma

A Tabela 11 apresenta o percentual de usuários que estão acima ou abaixo da média de pontos em questão de programação SQL do LabSQL, no período de 2008 a 2012. Nela, verifica-se que a maioria dos usuários está abaixo da média de pontos em questões de programação SQL em relação a sua turma, com 54,98%. A Figura 9 apresenta graficamente este percentual.

Tabela 11 Percentual de Usuários que Estão Acima ou Abaixo da Média de Pontos em Questão de Programação SQL do LabSQL, no Período de 2008 a 2012.

SQL	Percentual
Sim	45,02
Não	54,98
Total	100,00

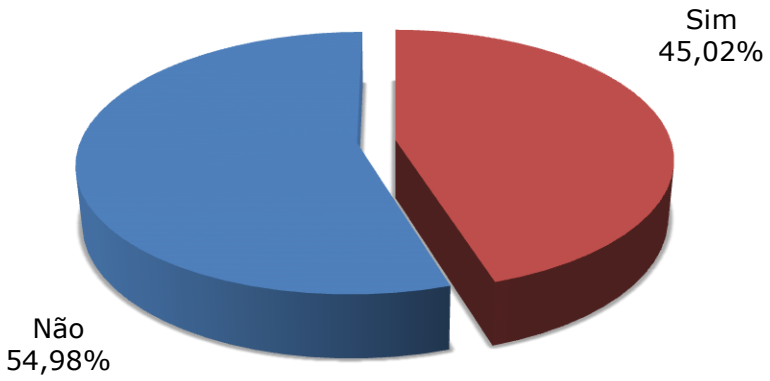


Figura 9 Percentual de Usuários que Estão Acima ou Abaixo da Média de Pontos em Questão de Programação SQL do LabSQL, no Período de 2008 a 2012.

Média de Pontos em Questões de Programação SQL

A Tabela 12 apresenta a média de pontos em questões de programação SQL no LabSQL, no período de 2008 a 2012, por turma. Nela, verifica-se que a turma T11 obteve a maior média de pontos em questões de programação SQL, com 9,16, seguido da turma T3 e T21 que obtiveram 9,09 e 8,80, respectivamente. Observa-se que a média geral das 25 turmas é de 7,46 pontos.

Tabela 12 Média de Pontos em Questões de Programação SQL do LabSQL, no Período de 2008 a 2012, por Turma.

Turma	Média
T01	7,30
T02	6,78
T03	9,09
T04	6,75
T05	7,70
T06	8,09
T07	7,76
T08	7,95
T09	5,59
T10	6,75
T11	9,16
T12	6,18
T13	8,03
T14	5,20
T15	8,52
T16	7,71
T17	8,30
T18	7,09
T19	7,48
T20	6,33
T21	8,80
T22	8,42
T23	6,60
T24	8,06
T25	6,75
Média geral	7,46

Tentativas de Resolução de Questões de Programação SQL

A Tabela 13 apresenta o percentual de usuários que estão acima ou abaixo da média de tentativas de resolução de questões de programação SQL do LabSQL, no período de 2008 a 2012. Nela, verifica-se que a maioria dos usuários está abaixo da média de tentativas de resolução de questões de programação SQL em relação a sua turma, com 65,80%. A Figura 10 apresenta graficamente este percentual.

Tabela 13 Percentual de Usuários que Estão Acima ou Abaixo da Média de Tentativas de Resolução de Questões de Programação SQL do LabSQL, no Período de 2008 a 2012.

Tentativas SQL	Percentual
Sim	34,20
Não	65,80
Total	100,00

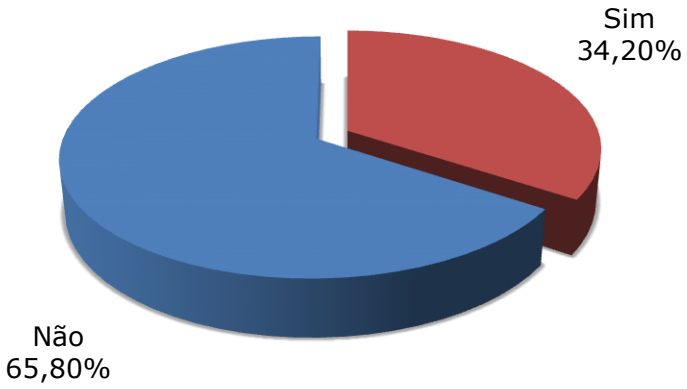


Figura 10 Percentual de Usuários que Estão Acima ou Abaixo da Média de Tentativas de Resolução de Questões de Programação SQL do LabSQL, no Período de 2008 a 2012.

Média de Tentativas de Resolução de Questões de Programação SQL

A Tabela 14 apresenta a média de tentativas de resolução das questões de programação SQL do LabSQL, no período de 2008 a 2012, por turma. Nela, verifica-se que a maior média de tentativas de resolução das questões de programação do LabSQL é da turma T09, com 4,94 tentativas. Observa-se que a média geral das 25 turmas é de 3,31 tentativas.

Tabela 14 Média de Tentativas de Resolução das Questões de Programação do LabSQL, no Período de 2008 a 2012, por Turma.

Turma	Média
T01	2,20
T02	2,37
T03	4,37
T04	2,20
T05	4,89
T06	4,09
T07	3,55
T08	3,83
T09	4,94
T10	0,17
T11	3,48
T12	3,30
T13	2,78
T14	1,87
T15	3,98
T16	4,64
T17	4,44
T18	2,37
T19	3,88
T20	4,05
T21	2,26
T22	4,30
T23	2,97
T24	3,27
T25	2,67
Média geral	3,31

Acesso aos Exemplos de SQL - Acima ou Abaixo da Média de Acessos da Turma

A Tabela 15 apresenta o percentual de usuários que estão acima ou abaixo da média de acesso aos exemplos de SQL disponíveis no LabSQL, no período de 2008 a 2012. Nela, verifica-se que a maioria dos usuários está abaixo da média de acesso aos exemplos SQL em relação a sua turma, com 65,17%. A Figura 11 apresenta graficamente este percentual.

Tabela 15 Percentual de Usuários que Estão Acima ou Abaixo da Média de Acesso aos Exemplos de SQL disponíveis no LabSQL, no Período de 2008 a 2012.

Exemplo	Percentual
Não	65,17
Sim	34,83
Total	100,00

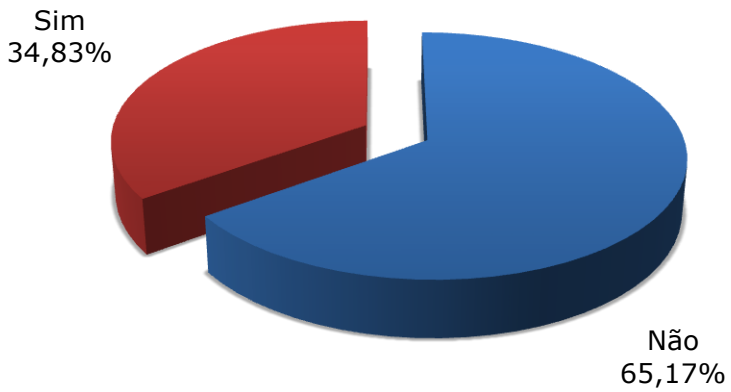


Figura 11 Percentual de Usuários que Estão Acima ou Abaixo da Média de Acesso aos Exemplos de SQL disponíveis no LabSQL, no Período de 2008 a 2012.

Média de Acesso aos Exemplos de SQL

A Tabela 16 apresenta a média de acesso dos usuários aos exemplos de SQL disponíveis no LabSQL, no período de 2008 a 2012, por turma. Nela, verifica-se que a turma que apresentou a maior média de acesso aos exemplos de SQL é a turma T16, com média de 103 acessos. Observa-se que a média geral das 25 turmas é de 52,52 acessos aos Exemplos de SQL.

Tabela 16 Média de Acesso dos Usuários aos Exemplos de SQL Disponíveis no LabSQL, no Período de 2008 a 2012, por Turma.

Turma	Média
T01	58
T02	32
T03	66
T04	42
T05	71
T06	64
T07	31
T08	89
T09	21
T10	50
T11	81
T12	42
T13	40
T14	14
T15	67
T16	103
T17	44
T18	64
T19	62
T20	59
T21	39
T22	86
T23	43
T24	33
T25	12
Média geral	52,52

Média de Acesso ao LabSQL

A Tabela 17 apresenta a média de acesso dos usuários ao LabSQL, no período de 2008 a 2012, por turma. Nela, verifica-se que a turma T11 obteve a maior média de acesso ao LabSQL, com 69,44, seguido da turma T12 e T22 que realizaram em média 62,48 e 56,35 acessos, respectivamente. Observa-se que a média geral das 25 turmas é de 36,25 acessos ao LabSQL.

Tabela 17 Média de Acesso dos Usuários ao LabSQL, no Período de 2008 a 2012, por Turma.

Turma	Média
T01	43,67
T02	40,24
T03	50,67
T04	48,93
T05	30,00
T06	38,97
T07	22,91
T08	50,74
T09	13,47
T10	8,00
T11	69,44
T12	62,48
T13	45,31
T14	9,09
T15	54,92
T16	32,09
T17	26,30
T18	41,44
T19	35,92
T20	38,22
T21	27,79
T22	56,35
T23	21,82
T24	28,56
T25	9,00
Média geral	36,25

Média de Acesso ao Material de Apoio

A Tabela 18 apresenta a média de acesso dos usuários ao material de apoio do LabSQL, no período de 2008 a 2012, por turma. Nela, verifica-se que a turma T08 obteve a maior média de acesso ao material de apoio, com 16,42 acessos, seguido da turma T22 e T19 que realizaram em média 12,57 e 11,50 acessos, respectivamente. Observa-se que a média geral das 25 turmas é de 6,17 acessos ao material de apoio.

Tabela 18 Média de Acesso dos Usuários ao Material de Apoio do LabSQL, no Período de 2008 a 2012, por Turma.

Turma	Média
T01	4,18
T02	5,76
T03	5,24
T04	10,67
T05	4,84
T06	10,26
T07	5,22
T08	16,42
T09	2,22
T10	2,13
T11	3,89
T12	3,64
T13	6,69
T14	2,73
T15	6,20
T16	8,40
T17	6,82
T18	7,78
T19	11,50
T20	5,84
T21	5,04
T22	12,57
T23	1,35
T24	1,76
T25	3,00
Média geral	6,17

Média de Acesso ao SQL-Livre

A Tabela 19 apresenta a média de acesso dos usuários ao SQL-Livre do LabSQL, no período de 2008 a 2012, por turma. Nela, verifica-se que a turma T19 obteve a maior média de acesso ao SQL-Livre, com 35,08 acessos, seguido da turma T11, T18 e T01 que realizaram em média 31,00, 27,75 e 27,55 acessos, respectivamente. Observa-se que a média geral das 25 turmas é de 17,58 acessos ao recurso do SQL-Livre.

Tabela 19 Média de Acesso dos Usuários ao SQL-Livre do LabSQL, no Período de 2008 a 2012, por Turma.

Turma	Média
T01	27,55
T02	18,65
T03	15,81
T04	20,41
T05	17,61
T06	16,77
T07	8,03
T08	16,97
T09	5,06
T10	0,13
T11	31,00
T12	19,90
T13	20,38
T14	8,18
T15	24,08
T16	19,91
T17	12,52
T18	27,75
T19	35,08
T20	16,81
T21	12,71
T22	25,61
T23	21,59
T24	14,82
T25	2,22
Média geral	17,58