UNIVERSIDADE ESTADUAL DA PARAÍBA
CAMPUS I - CAMPINA GRANDE-PB
CENTRO DE CIÊNCIA E TECNOLOGIA
DEPARTAMENTO DE COMPUTAÇÃO
CURSO DE LICENCIATURA PLENA EM COMPUTAÇÃO

JOANNA LIGIA DE QUEIROZ MARQUES

Mineração de Dados Educacionais: um estudo de caso utilizando o Ambiente Virtual do SENAI

CAMPINA GRANDE – PB
2014

JOANNA LIGIA DE QUEIROZ MARQUES

Mineração de Dados Educacionais: um estudo de caso utilizando o Ambiente Virtual do SENAI

Monografia apresentada ao Curso de Licenciatura Plena em Computação da Universidade Estadual da Paraíba, em cumprimento à exigência para obtenção do grau de graduado.

Orientador: Prof. Dr. Djalma de Melo Carvalho Filho

CAMPINA GRANDE – PB
2014

M357m Marques, Joanna Ligia de Queiroz.
Mineração de dados educacionais [manuscrito] : um estudo de caso utilizando o ambiente virtual do SENAI / Joanna Ligia de Queiroz Marques. - 2014.
72 p. : il. color.

Digitado.
Trabalho de Conclusão de Curso (Graduação em Computação) - Universidade Estadual da Paraíba, Centro de Ciências e Tecnologia, 2014.
"Orientação: Prof. Dr. Djalma de Melo Carvalho Filho, Departamento de Computação".

1. Mineração de dados. 2. Evasão escolar. 3. Educação a distância. 4. Plataforma Moodle. I. Título.

21. ed. CDD 371.291

Mineração de Dados Educacionais: um estudo de caso utilizando o Ambiente Virtual do SENAI

Monografia apresentada ao Curso de Licenciatura Plena em Computação da Universidade Estadual da Paraíba, em cumprimento à exigência para obtenção do grau de graduado.

Aprovada em 16 / 04 /2014.

Djalma de Melo Carvalho Filho

Prof. Dr. Djalma de Melo Carvalho Filho / UEPB
Orientador

Kátia Elizabete Galdino

Prof.. Dra. Kátia Elizabete Galdino / UEPB
Examinador

Prof. Dr. Frederico Moreira Bublitz / UEPB
Examinador

DEDICATÓRIA

A minha mãe, Regina Maria de Queiroz Marques (in memoriam), incentivadora e responsável por tudo que sou. Impossível expressar esse amor com palavras.

AGRADECIMENTOS

Quero neste espaço deixar o meu reconhecimento a todos aqueles com quem pude contar nos momentos difíceis, e em que pensei não mais suportar. Se esse momento chegou, é graças ao apoio e afago de vocês.

Ao meu Pai Maior, que sempre esteve comigo, me iluminando, me dando forças, e me protegendo de todo mal. Ao meu pai, João Batista Marques, pelo cuidado ao me acompanhar durante todos estes anos e pelos ensinamentos e valores que me foram passados. Aos meus irmãos, Junia, Joyce e Vitor, pelo carinho, força e compreensão durante este período tão delicado de nossas vidas.

A minha vó e ao meu vô, pela sabedoria infinita que sempre me aconselharam. Aos tios, primos, e agregados que estiveram sempre por perto quando precisei.

Aos meus amigos Edna Maciel, Edcley Dantas, Felipe Thamay, Gustavo Nóbrega, Jefferson Felipe, Renato Mello *(in memoriam)* e Thalles Santos, pelo companheirismo, noites de estudo e risadas.

Ao meu orientador, Djalma Filho, pela paciência e pelo auxílio, além de todo o conhecimento que transmite pela sua garra, persistência e sabedoria.

Aos amigos do SENAI, que me acompanharam e me ajudaram neste período, participando do desenvolvimento deste estudo, compreendendo as minhas faltas, e ouvindo meus lamentos.

Em especial ao meu namorado, André Lima, ouvinte atento das minhas inquietações, entusiasta do meu trabalho, e o melhor presente que essa graduação me deu.

“E ainda que tivesse o dom de profecia, e conhecesse todos os mistérios e toda a ciência, e ainda que tivesse toda a fé, de maneira tal que transportasse os montes, e não tivesse amor, nada seria.”

1 Coríntios, capítulo 13, versículo2. *In*: Bíblia.

R E S U M O

O Brasil atravessa um momento de grande desafio: ampliar e aperfeiçoar o acesso a Educação. A modalidade de Educação a Distância surge como um recurso eficiente para esse propósito visto o baixo custo, alcance e facilidade tempo espacial de acesso. Contudo esse escopo ainda perpassa por despreparo e dificuldades, sendo a evasão escolar a maior delas. O objetivo desse Estudo de Caso é identificar padrões de acesso dos alunos que evadem cursos na modalidade de Educação a Distância, por meio de Mineração de Dados Educacionais, e gerar regras que caracterizem o perfil de acesso desses alunos. Esse estudo tornará possível prever desistências por meio da análise de características de acesso e características sociais do estudante no Ambiente Virtual, possibilitando sugerir soluções para o dado problema, e em tempo hábil para evitar a reprovação do aluno. Para a realização desse Estudo de Caso foram utilizados dados obtidos no Ambiente Virtual de Aprendizagem (AVA) do SENAI DR-PB, com as iterações ao longo de 2013. O benefício esperado com o resultado desse estudo de caso é contribuir com as ações do SENAI DR-PB para tratar do problema da evasão escolar nos cursos da modalidade de Educação a Distância ao investigar o cenário exposto.

PALAVRAS-CHAVE: Mineração de Dados Educacionais, Mineração de dados, Análise da Aprendizagem, Educação a Distância, Evasão Escolar, Moodle, Educação.

A B S T R A C T

Brazil is going through a time of great challenge: to expand and improve access to education. The modality of E-learning emerged as an effective resource for this function, because it is cheap, reach and powerful. However this goal intersect with incapacity and difficulties, and the greatest of these is truancy. The objective of this case study is to identify access patterns of student that give up courses in E-learning, and generate rules that characterize the access profile of these students. This study will make it possible to predict dropouts by analysis the access characteristics and social characteristics of the student in the Learning Platform, suggest solutions to this problem and timely to prevent evasion of the student. For this research was used data obtained during the iteration with the Learning Platform DR SENAI-PB in 2013. The hope with the outcome of this case study is to contribute to the action of DR SENAI-PB to solve the problem of truancy in E-learning courses so to investigate the above scenario.

KEYWORDS: Educational Data Mining, Data Mining, Learning Analytics, E-learning, school supply, Moodle, Education.

LISTA DE FIGURAS

LISTA DE TABELAS

LISTA DE ABREVIATURAS E SIGLAS

AVA	Ambiente Virtual de Aprendizagem
DM	*Data Mining*
DN	Departamento Nacional do SENAI
DR	Departamento Regional do SENAI
EAD	Educação a Distância
Edm	*Educational Data Mining*
IA	Inteligência Artificial
KDD	*Knowledge Discovery in Database*
Lms	*Learning Management System*
PN-EAD	Plano Nacional de Educação a Distância do SENAI
STI	Sistema Tutor Inteligente

SUMÁRIO

1 Introdução

No cenário Brasileiro, a Educação a Distância surge como um meio prático de ensino direcionado a parte da população que por algum motivo (alto custo, distância, trabalho) busca uma alternativa para o acesso ao ensino. Apesar de oferecer uma gama de possibilidades (democratização do ensino, flexibilidade na aprendizagem, exercício da autonomia, ensino colaborativo) essa modalidade trás consigo questões preocupantes como os altos índices de evasão que denunciam desafios e limites futuros.

O potencial dessa modalidade ainda é pouco explorado. Os Ambientes Virtuais de Aprendizagem, de maneira geral, oferecem recursos limitados para que o professor obtenha entendimento sobre as percepções, aprendizagens e realizações dos alunos no escopo do curso, o que na educação tradicional é realizado no contato direto com o aluno (*face-to-face*). Turmas com muitos alunos agrava ainda mais essa situação, tornando progressivamente mais difícil manter-se atento a todas as atividades realizadas pelos alunos no ambiente.

Um avanço significativo foi possível com o surgimento dos Ambientes Virtuais de Aprendizagem (AVA), o que possibilitou um melhor acompanhamento da interação entre os participantes do curso virtual, sendo composto por ferramentas que facilitam a autoria de conteúdos e a interação síncrona[1] e assíncrona[2] entre o aluno e o professor.

Atrelado a isso, *hardware* cada vez mais evoluído possibilitou o armazenamento maciço de dados na área de educação. O Ensino a Distância permitiu armazenar dados em *logs* sobre todas as ações de alunos, professores e técnicos educacionais em um contexto educacional, gerando uma quantidade cada vez maior de dados detalhados. Decisões anteriormente baseadas em suposições, ou colhidas ao longo de custosas pesquisas de campo, facilmente podem ser feitas com base em dados reais, permitindo observação sobre as necessidades do estudante e modificações na abordagem das interações, além de viabilizar a personalização do ensino.

[1] Em tempo real. O emissor envia uma mensagem e o receptor recebe quase que instantaneamente.
[2] Dispensa a participação simultânea das pessoas. O emissor envia uma mensagem ao receptor, o qual poderá ler e responder em outro momento.

Surge então o desafio de obter conhecimento por meio da análise desses dados. Questões importantes sobre a análise desses dados trouxe a tona uma forte e consolidada linha de pesquisa: a Mineração de Dados Educacionais. O objetivo primordial dessa linha de pesquisa é o desenvolvimento de métodos para explorar conjuntos de dados coletados em ambientes educacionais que ofereçam grande potencial para melhorar a qualidade do ensino, projetando abordagens mais eficazes para o processo de Ensino e Aprendizagem (COSTA *et. al*. 2012).

Todos esses avanços oferecem uma oportunidade para a melhoria do ensino a distância, por tornar possível automatizar grande parte do esforço na análise de densos dados disponíveis. Esse diagnóstico atualizado registra o comportamento do estudante segundo por segundo e possibilita ao educador manter, ainda que na distância física, uma constante presença social, cognitiva e docente, comportamento indispensável para o ensino online proposto por Garrison e Anderson (2003). O educador, além de promover a melhoria na qualidade do curso, também pode oferecer um melhor *feedback* ao aluno, ajudando-o a ampliar a qualidade de sua experiência com o material disponibilizado.

1.1 Relevância

Analisar relatórios de iterações nos Ambientes Virtuais de Aprendizagem (AVA) é uma atividade dispendiosa para o Educador, principalmente quando esses estão em formato rudimentar. A Mineração de Dados transforma os registros, evidenciando dados valiosos, possibilitando fazer previsões sobre os acessos dos alunos, traçar suas curvas de aprendizagem e intervir quanto a dificuldades. Essa área está se consolidando, e requer maiores detalhes sobre as particularidades de sua aplicação. O presente trabalho apresenta uma metodologia para minerar dados educacionais baseada no estudo de Fayyad, Piatetsky-Shapiro e Smyth (1996) sobre Mineração de Dados.

Um conjunto de etapas é elaborado para compor a metodologia utilizada, e em seguida colocado em prática com dados reais obtidos no Ambiente Virtual de Aprendizagem do Sistema Nacional de Aprendizagem Industrial (SENAI). O experimento visou mapear o perfil de acesso dos alunos que desistem dos cursos oferecidos na modalidade a distância. Espera-se que, em trabalhos futuros, o sistema identifique esses perfis, principalmente em 25% iniciais das aulas do curso,

e evidencie automaticamente os resultados encontrados para os responsáveis educacionais. Foram analisadas também as limitações da aplicação da metodologia ao caso estudado, explicitando os problemas oriundos de suas particularidades.

Como contribuições desta pesquisa lista-se uma metodologia expansível, e a proposta de uma solução para atenuar os altos índices de evasão em Ambientes Virtuais de Aprendizagem por meio de um sistema de alerta.

1.2 Objetivos

Objetivo geral:

O propósito desse estudo de caso é aplicar técnicas de Mineração de Dados Educacionais para identificar estudantes, em um Ambiente Virtual de Aprendizagem (AVA), com características que podem levar à evasão ou à reprovação em cursos *on-line* e propor soluções para o dado problema.

Objetivo Específico:

Utilizar os dados obtidos em um AVA para identificar padrões de acesso dos alunos em um dado curso a distância, para analisar a problemática da evasão escolar, bem como sugerir soluções.

Gerar regras que caracterizem o perfil de acesso dos alunos, tornando possível prever desistências ou reprovações por meio da análise das iterações no Ambiente Virtual e das características sociais do aluno.

1.3 Procedimentos Metodológicos

Realizou-se uma pesquisa bibliográfica de caráter investigativo no desenvolvimento desse estudo de caso, e uma pesquisa documental como experimento, utilizando os registros de acesso a um dado Ambiente Virtual de Aprendizagem. A instituição de ensino sujeito do experimento foi escolhida por ceder acesso aos dados para a investigação.

> *Para Caldas (1986, p. 15) a pesquisa bibliográfica constitui a "coleta e armazenagem de dados de entrada para a revisão, processando-se mediante levantamento das publicações existentes sobre o assunto ou problema em estudo, seleção, leitura e fichamento das informações relevantes".*

Vergara (2005) e Bogdan *et al.* (1994) conceituam estudo de caso como uma observação detalhada de um contexto ou indivíduo, de uma única fonte de documentos ou de um acontecimento específico.

No que se refere aos meios, o instrumento de avaliação foi a pesquisa de campo, por meio da análise de registros de dados e observação pessoal.

1.4 Estrutura do Trabalho

Nos próximos capítulos estão explanados os principais métodos para minerar dados educacionais, seus pontos fortes e fracos, e como usá-los para promover a descoberta científica, e conduzir intervenções relevantes e melhoria do sistema de ensino aplicado.

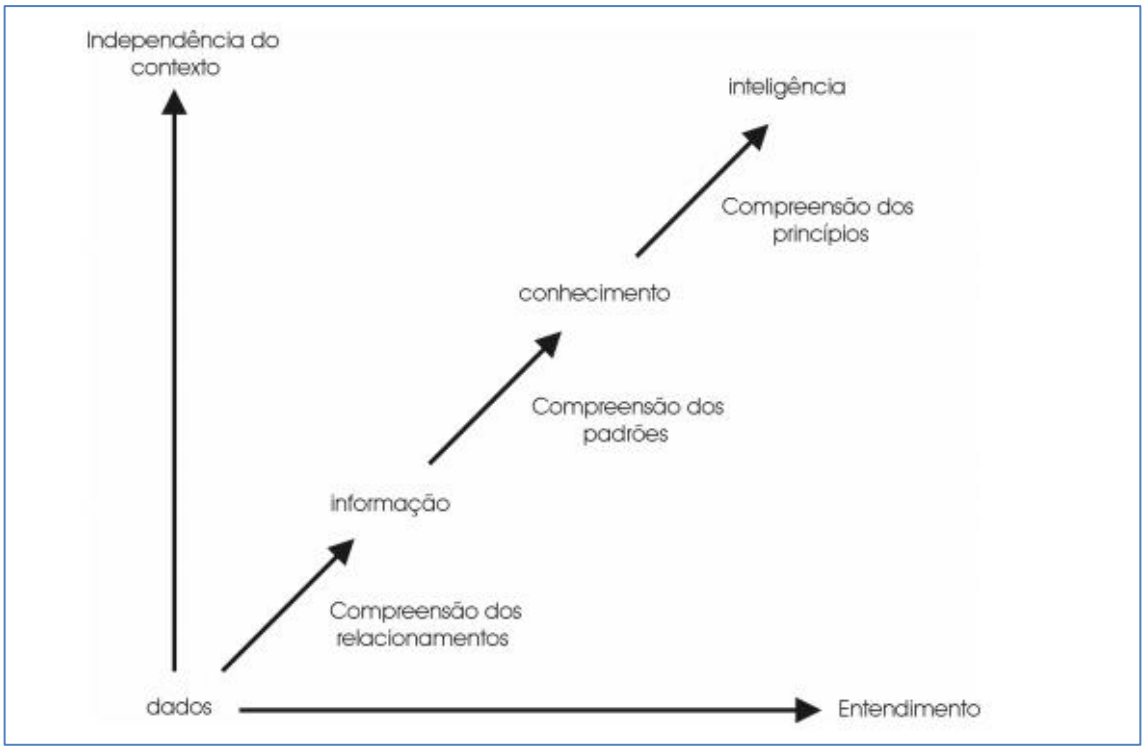
O restante desse trabalho está organizado da seguinte maneira: no capítulo 2 estão explicitados conceitos importantes para entender o funcionamento da pesquisa. No capítulo 3 define-se a área de Mineração de Dados Educacionais e suas particularidades. No capítulo 4 encontra-se a base teórica desse estudo, com os trabalhos relacionados disponíveis na literatura. No capítulo 5 apresenta-se a Metodologia para Minerar dados Educacionais. No capítulo 6 contextualiza-se o espaço de realização do Experimento e efetiva-se a pesquisa com a aplicação da metodologia considerada. Ainda no capítulo 6 medidas de confiabilidade são aplicadas ao modelo gerado, para identificar se a análise contida nesse estudo é confiável e aplicável. No capítulo 7 são descritas as considerações finais sobre o trabalho e a abertura existente para trabalhos futuros.

2 Definições importantes

2.1 Dado, Informação e Conhecimento

Com a revolução tecnológica dos últimos anos é comum ouvir-se que “estamos vivendo na era da informação”. No entanto, esse período ficaria mais bem conceituado como a era dos dados (HAN; KAMBER; PEI, 2011). O crescimento explosivo do volume de dados nos mais diversos ramos de aplicação (medicina, engenharia, telecomunicações) é resultado da informatização da sociedade atual e do desenvolvimento de rápidas e poderosas ferramentas de coleta e armazenamento de dados. Como procedente, ferramentas poderosas para descobrir automaticamente informações valiosas a partir de maciços dados e transformá-los em conhecimento organizado e versátil são extremamente necessárias. O que torna esse período legitimamente a “era dos dados”. Os conceitos de dados, informação e conhecimento estão interligados, porém o relacionamento entre eles é dado em função da capacidade de entendimento e da independência de contexto que cada um implica. Quando inteligência, a afirmação é a mais generalizável, podendo ser aplicada a diferentes contextos. Quando dado, os valores se opõem (Figura 1).

Figura 1: Relacionamento Entre Dados, Informação e Conhecimento.



Fonte: KOCK JR. *et al.* 1996, *apud* REZENDE *et al.,* 2003.

O dado é um elemento puro, sobre um determinado evento. Dados são fatos, números, texto ou qualquer mídia armazenada e processada pelo computador que, por si só, não oferece entendimento sobre o domínio da situação. A informação é o dado analisado e contextualizado, onde envolve a interpretação de um conjunto de dados, ou seja, a informação é constituída por padrões, associações ou relações que todos aqueles dados acumulados podem proporcionar. Analogamente, conhecimento refere-se à habilidade de criar um modelo mental que descreva o objeto e indique ações a implementar e as decisões a tomar, utilizado essencialmente para fornecer uma base de previsão com um determinado grau de certeza (REZENDE, 2003).

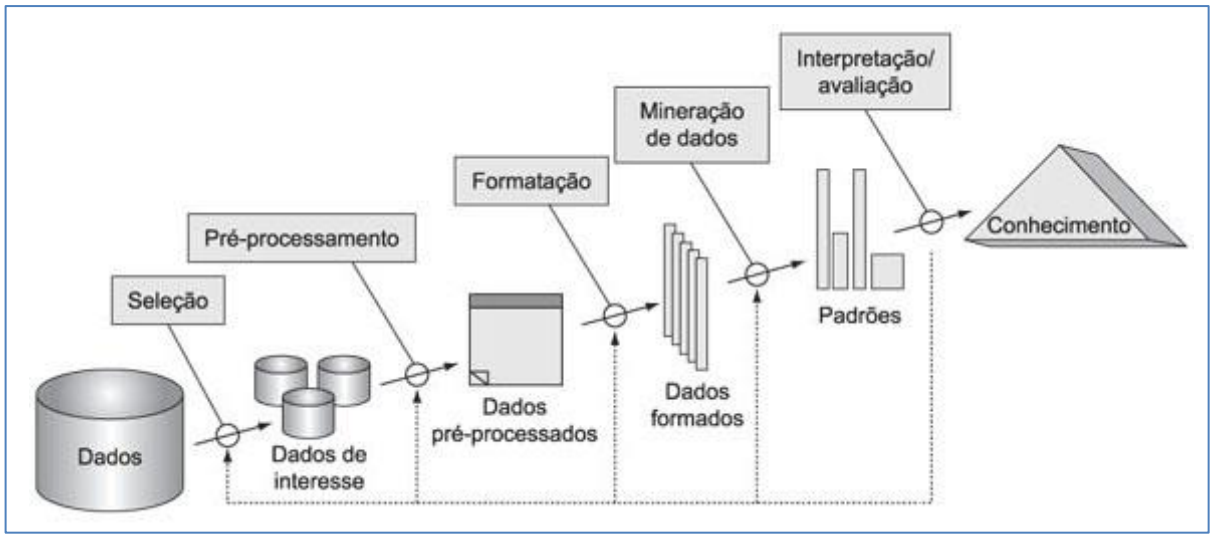
O desenvolvimento de Sistemas de Informação permitiu analisar dados e organizá-los como informação. Nos últimos anos apresenta-se como grande desafio o desenvolvimento de sistemas que sejam capazes de processar as informações para gerar o conhecimento de forma automática.

2.2 Mineração de dados

A Mineração de Dados (Data Mining, DM) pode ser vista como a etapa principal de um processo mais amplo conhecido como de descoberta do conhecimento em bases de dados (Knowledge Discovery in Databases, KDD) – sendo considerados sinônimos em alguns contextos. O conceito de KDD foi definido por Fayyad *et al.* (1996) como sendo: “... processo não trivial de identificar, em dados, padrões válidos, novos, potencialmente úteis e compreensíveis”. Considerando a relevância de cada componente: “Processo” refere-se a um conjunto de ações sequenciais para obtenção do conhecimento (ver 2.1) a partir de dados. “Padrão” alude unidades de informação que se repetem, ou regularidade discernível que se repete de maneira previsível. “Válido” diz respeito a um grau de certeza estatisticamente aceitável para os padrões descobertos, tornando-os verdadeiros sobre todas as interpretações. “Novo” visa que os padrões encontrados devem oferecer alguma informação não conhecida anteriormente sobre os dados. “Compreensível” quando se refere ao grau de entendimento do contexto da solução que esse padrão, ou a visualização dele proporciona aos que utilizarão o conhecimento.

O processo para Minerar Dados dispõe de diversos algoritmos que processam os dados em busca de padrões. Embora esses algoritmos sejam capazes de descobrir padrões “válidos e novos”, ainda são apoiados por decisões de analistas humanos para determinar padrões valiosos[3]. Os responsáveis por esse processo estão divididos na literatura em três classes: Especialista do domínio, o qual possui conhecimentos na aplicação em questão; Especialista em Mineração de Dados, experiente no processo de Extração de Conhecimento; e o usuário final, que utilizará o conhecimento extraído. Dessa maneira, não se pode esperar que a extração de conhecimento seja útil simplesmente submetendo um conjunto de dados a uma “caixa preta” (MANNILA, 1997).

Figura 2: Etapas fundamentais de uma mineração bem sucedida



Fonte: FAYYAD et. al. 1996, apud Camilo, 2009, tradução do autor.

Existem diversas abordagens para a divisão das etapas do processo de Extração de Conhecimento de Bases de Dados, proposta inicialmente por Fayyad *et al.* (1996), conforme visão geral na Figura 2, e que serviu de base para esse estudo, considerando-se nove etapas, com ressalva a alterações para o contexto aplicado.

O processo de descobrir conhecimento a partir de dados gera uma hierarquia, onde começa com instâncias elementares e volumosas, e termina em um ponto relativamente concentrado, mas muito valioso. Encontrar padrões requer “simplificar” os dados brutos, desconsiderando aquilo que é específico, e “privilegiar” aquilo que

[3] É necessário que a máquina possua uma Base de Conhecimento sobre o domínio para influenciar em seu processo decisório, por meio de ferramentas de Inteligência Artificial (I.A.) ou sistemas especialistas.

é genérico, perdendo um pouco dos dados para conservar apenas a essência da informação. A Figura 3 representa a tradicional pirâmide da informação.

Figura 3: Pirâmide da Informação

Fonte: NAVEGA, 2002.

O processo de descobrir conhecimento a partir de dados gera uma hierarquia, onde começa com instâncias volumosas e termina com um fragmento mais concentrado. Encontrar padrões requer simplificar os dados brutos, desconsiderando aquilo que é específico, e favorecer o que é genérico para conservar a essência da informação. A Figura 3 representa a tradicional pirâmide da informação.

2.3 Ambiente Virtual de Aprendizagem

Segundo Cole e Foster (2007) um Ambiente Virtual de Aprendizagem (AVA - *Learning Management System*, LMS) disponibiliza uma série de recursos, síncronos e assíncronos, que dão suporte ao processo de aprendizagem, permitindo seu planejamento, implementação e avaliação. Esta categoria de Ambiente Virtual de Aprendizagem ostenta um conjunto de funcionalidades delineadas para armazenar, distribuir e gerir conteúdos de aprendizagem, de forma progressiva e interativa, podendo, ainda, registar e relatar atividades do aprendiz, bem como o seu desempenho.

Algumas funcionalidades principais consistem em: Acesso protegido e gestão de perfis - Necessita de um *login* para acesso às funcionalidades ativas, de acordo com o perfil do usuário, existindo um sistema de gestão do perfil de cada usuário; Gestão do acesso a conteúdos - Os conteúdos (texto, áudio, vídeo, etc.) são configurados pelo autor e posteriormente geridos pelo *LMS*, indicando o progresso e

o desempenho do usuário; Comunicação Autor/Usuário - Comunicação “assíncrona” e “síncrona”, classificados em função dos tipos de participantes e em função do desenho pedagógico do curso; Controle de atividade - Registro das atividades de cada usuário (data do *login*, tempo de permanência, documentos acessados, seções visitadas, etc.); Gestão de utilizadores e gestão do processo – Potencialidade na automatização da gestão pedagógica, administrativa e organizacional.

Alguns AVA’s bem difundidos na academia: TelEduc, desenvolvido pela UNICAMP - Universidade Estadual de Campinas; AulaNet, desenvolvido pela PUC-Rio - Pontifícia Universidade Católica do Rio de Janeiro; WebCt - desenvolvido pela Universidade de British Columbia, Canadá; Moodle - desenvolvido pelo australiano Martin Dougiamas.

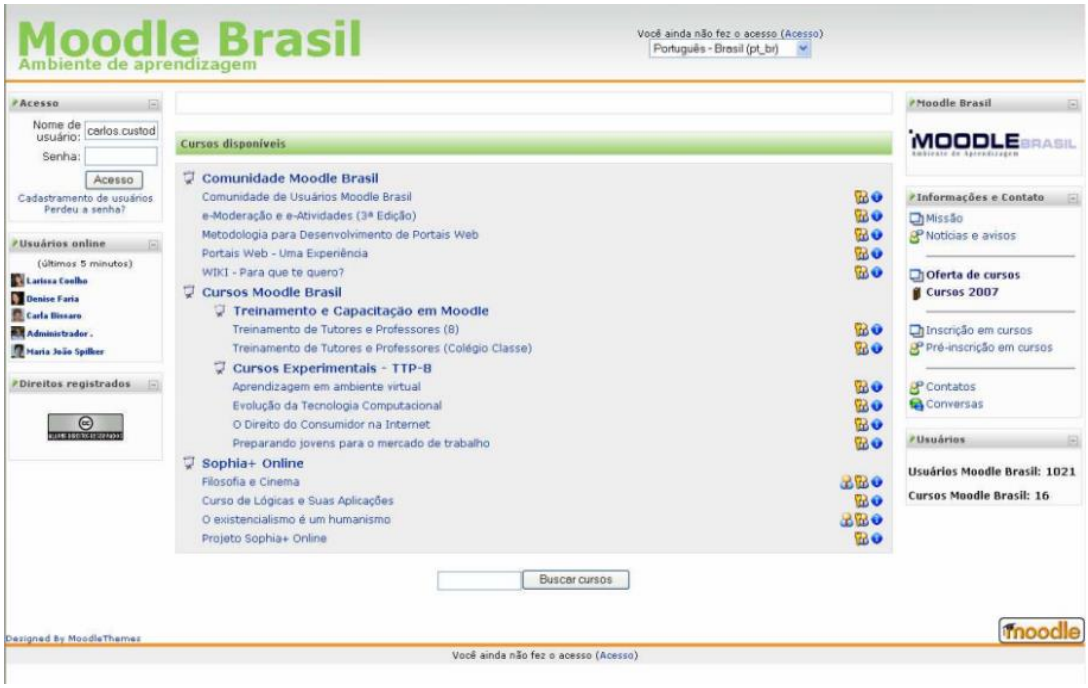
2.4 Moodle

O Moodle (Modular Object Oriented Dynamic Learning Environment) é um Sistema Gerenciador de Aprendizagem, distribuído livremente sob a licença GNU20 via Web, tendo como metodologia pedagógica o sócio construtivismo, no qual os participantes constroem conhecimentos de forma colaborativa, interagindo no ambiente (COLE E FOSTER, 2007). Há cerca de 20 diferentes tipos de atividades disponíveis (fóruns, glossários, wikis, tarefas, questionários, jogos SCORM, etc.) personalizáveis, além de uma série de outras ferramentas para atividades colaborativas (blogs, mensagens).

O Ambiente Moodle é adequado para atividades totalmente à distância, podendo ainda ser uma ferramenta para apoiar e complementar as atividades do ensino presencial. Considerada uma das ferramentas mais reconhecidas e largamente utilizadas no mundo para construção de ambientes virtuais de aprendizagem, o Ambiente, que deve ser instalado em um servidor *web*, pode ser acessado por qualquer Browser padrão que suporte a linguagem PHP. Seu desenvolvimento de forma modular permitiu uma evolução rápida de suas funcionalidades. O suporte do ambiente é realizado por uma comunidade internacional, garantindo o funcionamento e a personalização do ambiente para diversas necessidades.

Atualmente, o Moodle já foi traduzido para mais de setenta línguas, inclusive o português, tendo se destacado como um importante ambiente de ensino à distância, devido a sua flexibilidade, baixo custo e valor educativo (CUSTÓDIO, 2008). A ferramenta possui relatórios completos de acesso, porém em formato rudimentar, de todas as atividades realizadas no sistema pelos alunos. A tela inicial do Ambiente Virtual de Aprendizagem Moodle Brasil, desenvolvido com base no Moodle é mostrada na Figura 4.

Figura 4: Interface do Ambiente Virtual de Aprendizagem Moodle Brasil.



Fonte: MOODLE Brasil 2007, *apud* Custódio, 2008.

3 Mineração de dados Educacionais

A área de Mineração de dados Educacionais (Educational Data Mining, EDM) é uma área recente de pesquisa que tem como objetivo desenvolver ou adaptar métodos e algoritmos de Mineração de Dados (Data Mining, DM) existentes na literatura, para explorar dados originados em ambientes educacionais, considerando as especificidades inerentes aos dados produzidos pelos Ambientes Virtuais de Aprendizagem (AVA) e Sistemas Tutores Inteligentes (STI), tais como a não independência estatística nos tipos de dados encontrados ao coletar informações em ambientes educacionais e a necessidade de considerar a hierarquia da informação (Baker et. al., 2011a).

A origem dos métodos da EDM encontra-se nas áreas de aprendizagem de máquina, na psicometria e na estatística tradicional. Essa linha de pesquisa emergiu anos após a ascensão dos algoritmos de análise para *Big Data*[4], justificando o retardamento por não haver dados educacionais suficientes para o estudo nesse período. Nos dias atuais, com a ascensão da modalidade de Educação a Distância, essa barreira foi superada: muitos estudantes passaram a utilizar *software* educacional por intermédio da Web e na própria sala de aula. Os sistemas chegam a registrar processos de aprendizagem e comportamentos do aluno em uma escala de granulação fina, como a interação do estudante no *software* a cada segundo. Uma grande fonte para esses dados educacionais é Datashop[5], do Centro de Ciência e Aprendizagem de Pitsburgo.

A área de EDM está bem consolidada internacionalmente, mas, ainda em estágio inicial no Brasil (Baker *et. al.*, 2011a).

3.1 Origens da EDM

A comunidade de pesquisadores na área de Mineração de Dados Educacionais vem surgindo nos últimos anos. Em 2005 foi organizado o primeiro Workshop, *Educational Data Mining*, como parte do *20th National Conference on Artificial*

[4] Conceito que se refere ao grande armazenamento de dados, principalmente nos últimos anos, e as soluções digitais capazes de lidar com esse volume.

[5] Repositório público para dados de interação de *software* educacional, com mais de 30 milhões de ações estudantis, respostas e anotações do sistema.

Intelligence - AAAI 2005, em Pittsburgh, EUA. Em 2008 foi realizada a primeira conferência em EDM: *First International Conference on Educational Data Mining*, em Montreal, Canadá, que encontra-se em 2014 na sua sétima edição. Os pesquisadores oficializaram também a criação de um periódico que publicou o seu primeiro volume em 2009, o JEDM - *Journal of Educational Data Mining*. Já em 2011 constituiu-se a sociedade científica para EDM (*International Educational Data Mining Society*), conectando os pesquisadores da área.

3.2 Métodos, técnicas e algoritmos de Mineração de Dados Educacionais

Grande parte dos Métodos utilizados na área de EDM são provindas da área de mineração de dados (BAKER *et al*., 2010, tradução nossa) adaptados às necessidades de particularidades existentes em dados educacionais. Uma porção deles não é muito útil para dados extraídos do contexto educacional por não representa-los bem, como as Redes neurais que podem gerar over-fitting[6] para dados altamente baseados em contexto, e com poucas amostras. Os métodos, apresentados a seguir, estão conforme categorização na taxonomia proposta por Baker *et al*. (2010, tradução nossa):

- **Predição** *(prediction)* – induz um atributo presente nos dados;
    - **Classificação** *(classification)* - visa mapear, por meio do aprendizado de uma função, um dado atributo para uma de várias classes pré-definidas (atributo binário ou categórico);
    - **Regressão** *(regression)* - procura predizer um dado atributo numérico por meio de relações funcionais entre dadas variáveis;
- **Agrupamento** *(clustering)* – seu objetivo é identificar um conjunto finito de grupos ou categorias que descrevam os dados, não se especificando o alvo ou variável de predição.
- **Mineração de Relações** *(Relationship Mining)* – descobrem-se relações entre as variáveis em um conjunto de dados com muitas variáveis.
    - **Regras de Associação** (*Association Rule Mining*) - visa encontrar formas compactas de descrever subgrupos de dados, apresentando a

[6] Quando a hipótese é muito específica para o conjunto de dados utilizado, ou ajusta-se a ele.

forma: SE atributo X ENTÃO atributo Y. Se relaciona diretamente com a sumarização descrita por Fayyad *et al.* (1996).

- **Correlações** *(Correlation Mining)* – encontra-se relação conjunta, ou correlação linear de uma ou mais variáveis dentro de um contexto de análise.
    - **Padrões Sequenciais** *(Sequential Pattern Mining)* – procura-se por padrões que ocorrem em uma sequência temporal.
    - **Causas** *(Causal Mining)* – verifica-se a causa de um evento por outro, no decorrer da análise dos padrões de covariância.
- **Destilação de dados** *(Distillation of Data for Human Judgment)* – apresenta-se dados de alta complexidade, facilitando a compreensão dos usuários.
- **Descobrimento com Modelos** *(Discovery with Models)* – utiliza-se um modelo pré-existente (desenvolvido com métodos de predição, clusterização, etc.) como um componente em outra análise.

Para o presente estudo consideramos o método da Predição por meio da Classificação de exemplos, descritos em sequência.

3.2.1 Predição

Na predição, a meta é o desenvolvimento de um modelo para inferir um único aspecto dos dados, chamado de variável preditiva, a partir de uma combinação de outros aspectos dos dados, as variáveis preditoras. Esse método pode ser utilizado para fazer inferências sobre o presente, ou até mesmo para prever algo mais difícil, como os acontecimentos futuros.

No contexto educacional é possível investigar com esse método quais alunos estão frequentando as aulas e quais alunos correm risco de evadir o curso, assim como modelar os alunos que encontram-se entediados. Analogamente, o conhecimento adquirido a partir de suas ações no Ambiente Virtual e o desempenho do aluno diante de uma avaliação futura podem ser mensurados e deduzidos. Uma análise pertinente pode identificar a abrangência dos benefícios educacionais na utilização de uma determinada ferramenta no Ambiente Virtual.

Segundo Baker *et al*. (2010) e Romero *et al*. (2008), há vários benefícios relacionados à utilização da predição em EDM, citando a) Design Instrucional – identificar o momento em que o aluno fica entediado no Ambiente Virtual permite

melhorar o conteúdo, evitando consequentes evasões; b) Inteligência artificial – identificar que o aluno não alcançou as habilidades necessárias com o curso, admite oferecer ajuda, utilizando decisões automáticas pelo *software*; c) Acompanhamento a Distância – modelar as ações do estudante torna-se relevante para professores e outros interessados nessa capacidade de monitoramento, apoiando decisões com base em eventos observados no Ambiente Virtual, “fator significativo para a eficiência nas experiências de aprendizagem” (Moore, 1989).

Em EDM a predição é dividida frequentemente em duas taxonomias: Regressão, que prevê uma variável contínua – um número chamado de Regressor – e a Classificação, que prevê uma variável categórica – uma de várias classes pré-definidas.

As investigações podem ser direcionadas ao tipo de atributo que se deseja inferir, listando como possíveis regressores: o número de sugestões solicitadas pelo aluno em uma atividade proposta; o tempo gasto para responder a atividade; a porcentagem assistida de uma vídeo-aula; o desempenho em um teste de múltipla escolha. Da mesma maneira especificando como categorias da previsão: o acerto/erro do aluno em uma dada questão subjetiva; a desistência ou a conclusão do curso; um tipo de material didático recomendável para a aprendizagem; o perfil de interesse do aluno.

Modelo de Regressão

Para construir um modelo de regressão, é necessário obter um conjunto de dados onde já é conhecido o valor do Regressor. Esse conjunto de amostras, ou instâncias é chamado de **conjunto de treinamento**, utilizado para construir um modelo que, em seguida, irá prever o valor do atributo quando este não existir. A ideia básica da regressão é determinar quais características, em cada combinação, podem prever o atributo de valores numéricos.

São Pedro et. al. (2013) buscou prever a quantidade de dicas solicitadas pelo aluno no desenvolvimento de uma atividade (Tabela 2). Cada número de dicas, a variável preditiva, é uma amostra do conjunto de treinamento. Associado a cada variável preditiva encontra-se um vetor de atributos, outras variáveis além da variável preditiva. Na Tabela 2 é possível observar as prováveis variáveis preditoras:

Desempenho (medido pela probabilidade do aluno possuir uma habilidade), Tempo, TotalDeAções e Habilidade. Sugere-se usar essas amostras para tentar construir um modelo de regressão do atributo “NumeroDeDicas” e, em seguida, prever o valor do mesmo quando não houver o atributo.

Tabela 1: Representação de um Conjunto de Treinamento para a Regressão

Habilidade	Desempenho	Tempo	TotalDeAções	NumeroDeDicas
H01	0.704	9	1	0
H02	0.502	10	2	0
H03	0.049	6	1	3
H04	0.967	7	3	0
H05	0.792	16	1	1
H06	0.792	13	2	0
H07	0.073	5	2	0

Fonte: Sao Pedro, 2013, adaptado.

Modelo de Classificação

Analogamente, para construir um modelo de classificação, é necessário obter um conjunto de dados onde a variável preditiva seja conhecida, também chamado de **conjunto de treinamento**. O modelo em seguida será utilizado para prever o valor do atributo categórico. A ideia básica é determinar quais características pode prever à qual categoria, entre várias, o atributo preditivo pertence. Essa categoria pode ser binária – “o aluno foi aprovado ou reprovado?” – ou estar contida em um conjunto de valores categóricos – “qual a habilidade cognitiva[7] do aluno?”.

No estudo de São Pedro *et. al.* (2013), é possível utilizar esse método para prever se o estudante acertou ou errou uma questão subjetiva a partir da combinação de atributos já conhecidos (Tabela 3). A ideia básica é determinar com quais atributos, e qual combinação é capaz de prever a variável categórica “AcertoDoAluno”.

[7] Baseando-se na teoria das Inteligências Múltiplas proposta pelo psicólogo Howard Gardner em 1998, que descreveu sete tipos de habilidades nos seres humanos, posteriormente modificada para nove tipos.

Tabela 2: Representação de um Conjunto de Treinamento para a Classificação.

Habilidade	Desempenho	Tempo	TotalDeAções	AcertoDoAluno
H01	0.704	9	1	WRONG
H02	0.502	10	2	RIGHT
H03	0.049	6	1	WRONG
H04	0.967	7	3	RIGHT
H05	0.792	16	1	WRONG
H06	0.792	13	2	RIGHT
H07	0.073	5	2	RIGHT

Fonte: Sao Pedro 2013, adaptado.

Algoritmos característicos de cada método funcionam melhor para domínios e problemas específicos. No próximo tópico encontra-se a técnica utilizada nesse estudo de caso.

3.3 Arvores de decisão

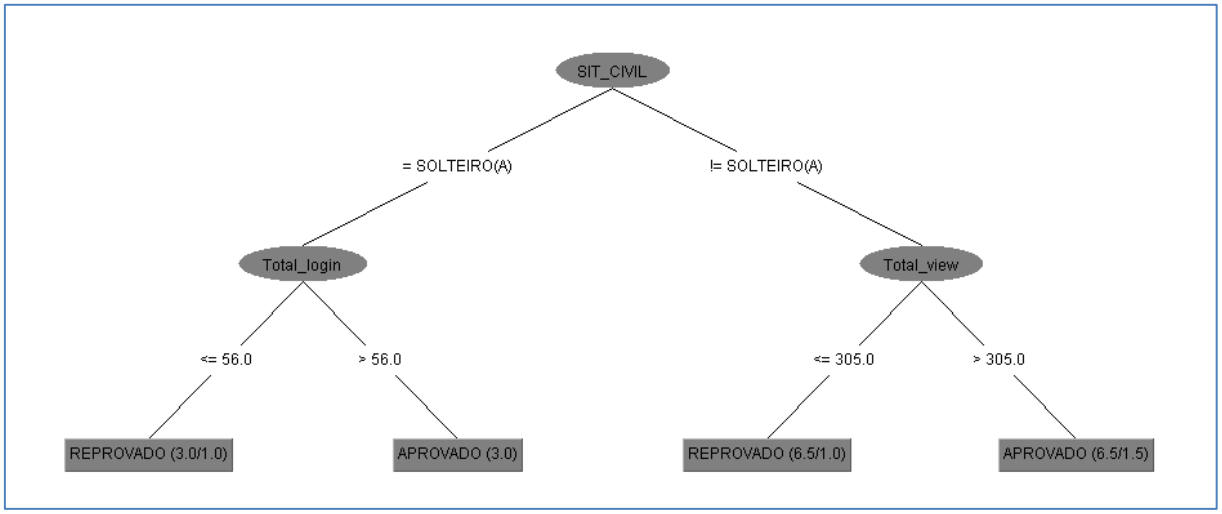
Árvores de decisão (*Decision Trees*) são ferramentas que podem ser utilizadas para tomar decisões e inferir valores categóricos. A ideia é um aprendizado indutivo: Cria-se uma hipótese baseada em instâncias particulares que gera conclusões gerais. É uma estrutura muito usada na implementação de sistemas especialistas e no desenvolvimento de agentes inteligentes na Inteligência Artificial. As árvores de decisão tomam como entrada uma situação descrita por um conjunto de atributos e retornam uma decisão, que é a categoria para o valor de entrada.

Árvores de decisão são simples representações de conhecimento e classificam exemplos em um número finito de classes. Sua estrutura é composta por: um nó de decisão que especifica um teste a ser realizado no valor de um atributo, com um galho para cada resposta possível do teste, levando a uma sub-árvore ou uma folha – os nós são rotulados com nomes de atributos; os arcos são rotulados com possíveis valores para o atributo preditivo; as folhas são rotuladas com diferentes categorias do atributo preditivo.

Objetos são classificados percorrendo um caminho da árvore - seguindo os arcos que contêm valores que correspondem a atributos no objeto. Em uma árvore de decisão a classificação de um caso se inicia pela raiz da árvore, e esta árvore é

percorrida até que se chegue a uma folha. Em cada nó de decisão será feito um teste que irá direcionar o caso para uma sub-árvore. Este processo irá guiar-se para uma folha. A classe do caso se pressupõe que seja a mesma que está armazenada nesta folha. Na Figura 5 é possível observar um exemplo.

Figura 5: Arvore de Decisão – Estudo de Caso AVA SENAI.



Fonte: *print screen* da aplicação RapidMiner®.

A árvore de decisão chega a sua decisão pela execução de uma sequência de testes. Cada nó interno da árvore corresponde a um teste do valor de uma das propriedades, e os ramos deste nó são identificados com os possíveis valores do teste. Cada nó folha da árvore especifica o valor de retorno se a folha for atingida.

Em um contexto educacional, cada nó da árvore de decisão pode representar os atributos pertencentes a um conjunto de alunos. Ao descer a árvore, conduzindo a busca a um dos filhos desse nó – descendo da raiz em direção as folhas da árvore – pode-se selecionar a configuração que melhor se ajusta às características de cada objeto, nesse caso o aluno, e prever a categoria tomando como critério o comportamento associado.

Como gerar uma árvore

A árvore de decisão é gerada a partir de exemplos do domínio (conjunto de treinamento, especificado no tópico anterior referente a classificação). Seu papel é escolher as regras mais importantes e descartar regras que menos se ajustam. Na Figura 6 encontra-se evidenciado parte do conjunto de treinamento objeto desse estudo de caso.

Figura 6: Conjunto de treinamento do estudo de caso AVA SENAI

Amostra	Atributos preditores									Alvo
id	turma	totallogin	total_postfo...	total_discu...	total_mens...	total_mens...	total_iterac...	total_visuali...	total_view	status_matr...
46	5	48	1	1	?	?	?	20	305	REPROVADO
48	5	99	6	1	?	25	263	182	315	APROVADO
49	5	44	1	?	?	19	?	24	275	REPROVADO
50	5	76	4	3	?	22	?	81	420	REPROVADO
51	5	?	1	?	?	21	?	30	416	APROVADO
52	5	16	?	?	?	13	67	4	106	REPROVADO
53	5	123	5	5	4	23	284	75	524	APROVADO
62	9	61	2	?	?	?	25	20	155	REPROVADO
63	9	127	5	?	?	43	107	154	558	APROVADO
66	8	102	2	1	?	?	205	125	618	APROVADO
73	8	31	?	?	?	15	72	17	95	REPROVADO
74	8	68	5	?	?	?	133	112	352	REPROVADO
75	8	56	1	?	?	19	125	31	221	APROVADO
76	8	107	1	?	?	18	42	50	236	APROVADO
77	8	35	1	1	?	18	?	?	116	REPROVADO
78	8	5	1	?	?	?	?	10	29	REPROVADO
79	8	60	2	1	?	17	75	51	203	APROVADO
80	8	75	1	?	?	14	206	74	351	REPROVADO
81	8	62	1	?	34	25	264	31	428	APROVADO

Fonte: *print screen* da aplicação RapidMiner®.

Algoritmo de aprendizado para gerar a Arvore de Decisão: Quando a função é chamada, submete-se o vetor contendo todas as variáveis preditoras elegíveis, a variável preditiva (atributo alvo) e o conjunto de exemplos (conjunto de treinamento). O algoritmo escolhe a melhor variável preditora para repartir o conjunto de treinamento e criar o nó de decisão correspondente. Cada nó possui um conjunto com os exemplos restantes e um histograma[8] dos exemplos de acordo com o atributo alvo. A recursão do algoritmo pára quando uma das três condições for verdadeira: todos os exemplos têm o mesmo atributo alvo; não existem mais atributos; não existem mais exemplos. Na Figura 7 demonstra-se um pseudocódigo do algoritmo de extração de regras.

[8] Distribuição de Frequências dos atributos restantes e a frequência em que os valores do atributo estão presentes no conjunto de dados.

Figura 7: Algoritmo de aprendizado para Arvores de decisão.

```
Node criaArvore (exemplos, atributoAlvo, atributos)
{
    se todos exemplos tem o mesmo valor de atributoAlvo
       retorna a folha com o valor
    senão se o conjunto de atributos é vazio
       retorna a folha com o valor de atributoAlvo mais comum entre os exemplos
    senão
    {
      A = melhor atributo entre atributos com um variações v1, v2, .. vk
      Particione os exemplos segundo seus valores para A em conjuntos S1, S2, ...Sk
      Crie um nó de decisão N com atributo A
      Para i=1 até K
          Conecte um no B para o nó N com teste vi
          Se Si tem elementos (não vazio)
             Conecte ramo B a criaArvore(Si, atributoAlvo, atributos - {A})
          Senao
             Conecte B para a folha do nó com atributoAlvo mais comum
      Retorna no N
   }
}
```

Fonte: POZZER, 2006.

3.4 Algoritmo J48 / C4.5

Nesse estudo adotamos o algoritmo J48, que se baseia no algoritmo de árvores de decisão C4.5 (QUINLAN, 1993), formando a arvore mais adequada sobre o conjunto de dados, ao podar as regras que melhoram a sua precisão. Os algoritmos de árvores de decisão são conhecidos pelo seu poder de expressividade, encadeando um conjunto de testes, os quais atuam diretamente no ganho de informação a respeito dos dados. Essa característica possibilita transformar árvores em regras de classificação.

O algoritmo J48 permite a criação de modelos de decisão em árvore em que cada nó da árvore avalia a existência ou significância de cada atributo individual. As árvores de decisão são construídas do topo para a base, mediante escolha do atributo mais apropriado para cada situação. Uma vez escolhido o atributo, os dados de treino são divididos em subgrupos, correspondendo aos diferentes valores dos atributos e o processo é repetido para cada subgrupo até que uma grande parte dos atributos em cada subgrupo pertença a uma única classe. A indução por árvore de decisão é um algoritmo que habitualmente aprende um conjunto de regras com elevada qualidade.

3.5 Ferramentas para Minerar dados

Diversas ferramentas para Minerar Dados encontram-se disponíveis na literatura. Citam-se algumas delas: RapidMiner®, SAS OnDemand, Weka, Microsoft Excel. Para esse estudo de caso utilizamos a ferramenta RapidMiner® versão 5.3.

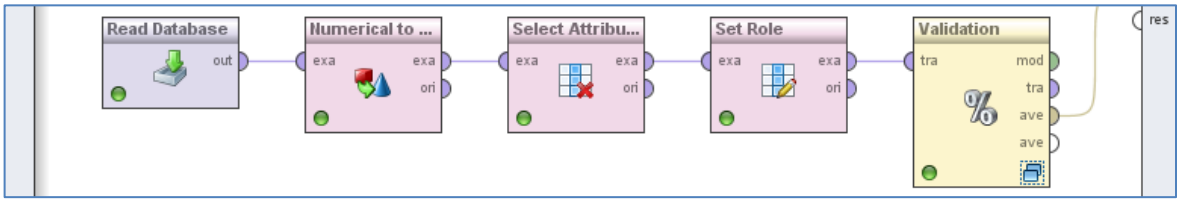
RapidMiner®

O RapidMiner® é uma ferramenta de Mineração de Dados, Mineração de Texto, análise de dados e inteligência de negócio; utilizado nas áreas de investigação, educação, projetos experimentais e em aplicações industriais. Por ser desenvolvida em Java, permite a sua utilização versátil em qualquer sistema operativo e ambiente de trabalho. O projeto RapidMiner® começou em 2001 por Ralf Klinkenberg, Ingo Mierswa e Simon Fischer na Unidade de Inteligência Artificial da Universidade de Dortmund (Alemanha). Em 2006, Ingo Mierswa e Ralf Klinkenberg fundaram a empresa Rapid-I. Esta ferramenta encontra-se disponível em duas versões: Community Edition – gratuita, mas limitada em termos de funcionalidades e recursos; Enterprise Editon – é a versão profissional do software que, além de todas as vantagens da versão Community contêm soluções empresariais específicas para utilizadores profissionais. Possui igualmente capacidades avançadas de criação de relatórios e serviços específicos de garantia e assistência.

Na implementação do algoritmo J48 desse estudo foi utilizada a versão Community Edition 5.3.008 com a instalação do pacote de expansão da ferramenta Weka. A ferramenta foi eleita no estudo por oferecer um melhor apoio a validação cruzada, embora exija do usuário um grande esforço para obter a validação desejada. Em particular, ferramentas importantes de mineração de dados, como a Weka, não oferecem apoio para esse tipo de validação entre os dados no nível do aluno ou da classe. A validação cruzada permite verificar a corretude de um modelo gerado a partir de dados de treinamento, e da análise em dados de teste, oferecendo uma estimativa de como o modelo irá se comportar ao analisar um conjunto novo de dados. Validação cruzada ao nível de aluno ou classe é fundamental em dados educacionais, pois existe uma grande quantidade de dados para ser minerada, e as conclusões obtidas precisam garantir que o modelo encontrado possa ser utilizado para inferir o comportamento ou a aprendizagem de novos alunos e/ou classe (Baker *et. al.* 2011b, tradução nossa).

Para a produção da árvore de decisão, foram carregados na ferramenta os conjuntos de dados por intermédio de uma conexão com o Banco de Dados MySQL, utilizado como gerenciador da base de dados do Ambiente Virtual do Experimento realizado. Foram utilizados os operadores *ReadDatabase, SelectAtribute, SetRole, X-Validation, W-J48, ApplyModel e Performance (Classification)* (Figura 8).

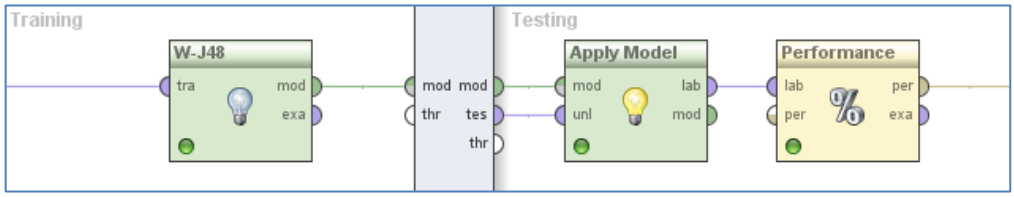
Figura 8: Operadores do RapidMiner®.



Fonte: *print screen* da aplicação.

O operador *ReadDatabase* faz a conexão com o Banco de Dados onde encontra-se o conjunto de dados do repositório; o *NumericalToPolynominal* modifica o tipo de atributo numérico selecionado para um valor categórico, utilizado no experimento para alterar o tipo das variáveis ID do aluno e ID da Turma; *SelectAtribute* é utilizado para selecionar apenas os atributos úteis ao estudo; *SetRole* é utilizado para definir o rótulo (variável preditiva) necessário ao operador W-J48; *X-Validation* é o operador que aplica a validação cruzada, método escolhido para avaliar o desempenho preditivo do classificador induzido. Neste estudo foi utilizada a validação cruzada *Leave-Out-One*, visto a disponibilidade de poucas amostras, ou exemplos (ver 5.8).

Figura 9: Operadores do RapidMiner® internos ao Operador X-Validation.



Fonte: *print screen* da aplicação.

Dentro do operador *X-Validation* (Figura 9) foram utilizados os operadores: W-J48, para construir a árvore de decisão a partir de um conjunto de dados; *ApplyModel*, para gerar alguns atributos especiais adicionais a serem utilizados pelos operadores de desempenho, como por exemplo a precisão e a confiança; Performance (*classification*), para medir o desempenho do modelo baseado em diversas medidas como precisão (*accuracy*), *kappa*, erro do classificador, etc.

4 Trabalhos relacionados

Diversas soluções vêm sendo desenvolvidas com o estudo de métodos da Mineração de Dados Educacionais na literatura, com aplicações para: apoiar os tutores e responsáveis educacionais, ao fornecer estatísticas de uso do sistema e *feedback* sobre a iteração dos alunos (SENECHAL, 2013); identificar as situações em que um tipo de abordagem instrucional proporciona melhores benefícios educacionais ao aluno e.g. aprendizagem individual ou colaborativa (FABIELI; CHARAO, 2011); agrupar alunos por meio dos padrões de interação, visando a recomendação de conteúdo para outros estudantes que possuam padrões similares de aprendizagem (PAIVA *et al*., 2013); rever o desempenho dos alunos ao classifica-los recorrendo a suas características (GODWIN *et al*., 2013); modelar emoções do estudante, e. g. verificar se um aluno está desmotivado ou confuso, objetivando a melhoria do Design Instrucional (BAKER *et. al.* 2011b, tradução nossa); minerar padrões sequenciais de acesso para detectar comportamentos indesejáveis no sistema (GOTTARDO, 2012); analisar comportamento do aluno em redes sociais para modelar seu perfil de aprendizado e iteração; minerar textos agrupando documentos por assunto (MACHADO, 2010); aplicar técnicas de visualização para análise e visualização de dados (ROMERO; VENTURA, 2010).

No tocante a aplicação de Mineração de Dados Educacionais para o estudo da evasão, Manhães *et al*. (2011) realizaram um estudo para identificar a eficiência de dez algoritmos de classificação na previsão de estudantes com risco de evasão utilizando dados acadêmicos das primeiras notas semestrais referentes a alunos de graduação da universidade brasileira UFRJ. Os resultados mostraram que é possível identificar com precisão de 80% a situação final do aluno no curso. Porém a pouca descrição do trabalho realizado dificultou a reprodução do estudo.

Kampff (2008) propõe uma arquitetura para um sistema de acompanhamento das iterações em um Ambiente Virtual de Aprendizagem, identificando por intermédio da Mineração de Dados o comportamento e as características de alunos propensos à reprovação e à evasão. O sistema alerta o professor, permitindo estabelecer comunicação personalizada e contextualizada com os alunos. Ao final, foi possível comprovar que as intervenções realizadas pelo professor, a partir dos alertas, contribuíram para a melhoria dos índices de aprovação e para redução dos

índices de evasão dos alunos na disciplina. Entretanto, o sistema de acompanhamento desenvolvido utiliza como parâmetro de entrada um arquivo .cvs, exportado do banco de dados, o que pode gerar ruídos e a não integração entre o banco de dados do Ambiente Virtual de Aprendizagem e o sistema de acompanhamento.

O processo de Mineração de Dados Educacionais apresentado nos trabalhos citados para diferentes contextos educacionais carece de uma metodologia de aplicação. Nesse estudo encontra-se a descrição de nove passos para esse processo, baseados no trabalho de Fayyad, Piatetsky-Shapiro e Smyth (1996), o que diferencia a pesquisa realizada dos estudos similares, para auxiliar no processo de mineração de dados em ambientes educacionais.

A proposta apresentada neste estudo de caso objetiva disponibilizar estimativas de desempenho escolar dos estudantes do Departamento Regional do SENAI Paraíba recorrendo às iterações realizadas no AVA. O conjunto de atributos considerados para esse estudo não contemplam todos os aspectos da aprendizagem normalmente avaliados pela metodologia formação profissional do SENAI baseada em competências, entretanto, foi considerado suficiente para alcançar a inferência prevista. Trabalhou-se com um conjunto de atributos que pudesse representar as principais atividades desenvolvidas pelos estudantes em um Ambiente Virtual. Esse estudo ofereceu explicações e entendimentos acerca do problema da evasão, possibilitando o desenvolvimento de um sistema de acompanhamento dos alunos propensos a evasão.

5 Metodologia para Minerar Dados Educacionais

Esse estudo de caso para o processo de minerar dados educacionais é baseado na metodologia proposta por Fayyad, Piatetsky-Shapiro e Smyth (1996). Os 9 passos descritos a seguir, e aplicados no Capítulo 6, sugerem uma metodologia bastante expansível:

5.1 Primeiro: Captura e compreensão dos dados

Para realizar trabalhos de análise e mineração de dados, uma boa aquisição desses dados, considerando sua relevância para o estudo, é de suma importância. É bastante comum que o analista implante no sistema, ou adapte, seu próprio esquema de captura e armazenagem de informações de uso; alguns trabalhos consistem inclusive no desenvolvimento de formas eficientes para registro das interações dos usuários com os ambientes de ensino, como o de Cardieri (2004). Um estudo financiado pela fundação Bill & Melinda Gates[9] consistiu no desenvolvimento de pulseiras com resposta galvânica da pele - sensores sem fio que monitoram reações fisiológicas, a serem utilizadas nas escolas para medir o envolvimento dos alunos em sala de aula.

Em um sistema sem preparo prévio, muitos pesquisadores na literatura direcionam seus trabalhos a lidar com o estudo de arquivos de *log* gerados pelo acesso dos estudantes ao servidor. Um desafio nesse tipo de trabalho é capturar as emoções dos estudantes, visto que tal sensibilidade é mais evidente no contato face-a-face. Godwin *et al.* (2013) e Baker *et al.* (2011b) utilizaram dados de *log* para detectar padrões no comportamento dos alunos (sem utilizar nenhuma câmera ou sensor). Nessas pesquisas os estudantes, após um período ausente do sistema tratado como comportamento *off-task*, retornam distraídos, ao mudar o foco do estudo com outras atividades, ou avançam na compreensão da atividade, pois o período ausente foi assim destinado à discutir com professores ou outros estudantes sobre o assunto estudado no Ambiente Virtual. Pardos *et al*. (2013) descrevem ainda detectores de tédio, frustração, confusão e comprometimento para prever se um estudante do ensino médio está propenso a futuramente ir para uma faculdade, considerando apenas seus dados de log. Contudo, dados de *log* de acesso

[9] http://www.gatesfoundation.org/

constituem uma fonte limitada de informações, e exigem um amplo esforço de formatação e limpeza dos dados para prepara-los para a mineração.

Após definir os dados disponíveis para a mineração, é necessário um estudo do sistema educacional e das ferramentas e ambientes disponíveis. É preciso entender como e onde esses programas captam e armazenam seus dados, assim como quais são os dados mantidos. Um estudo do formato dos registros é necessário para identificar as limitações dos registros do sistema e as possíveis formas de contorná-las, além de apresenta-las em um formato de fácil exploração. Entender sobre o domínio dos dados é naturalmente um pré-requisito para extrair algo útil.

Os dados provenientes de um AVA (definido no Capítulo 2.3) oferecem embasamento sobre diversos aspectos do comportamento e do aprendizado de cada estudante, assim como a possibilidade de comparar esses resultados com outros estudantes no contexto, identificando os que possuem um desvio significativo no padrão de acesso do material se comparado com o restante da turma. O estudo desses registros pode ajudar a explicar um resultado ruim obtido por parte destes alunos ou mesmo uma forma mais eficiente de estudo.

5.2 Segundo: Identificação do Problema

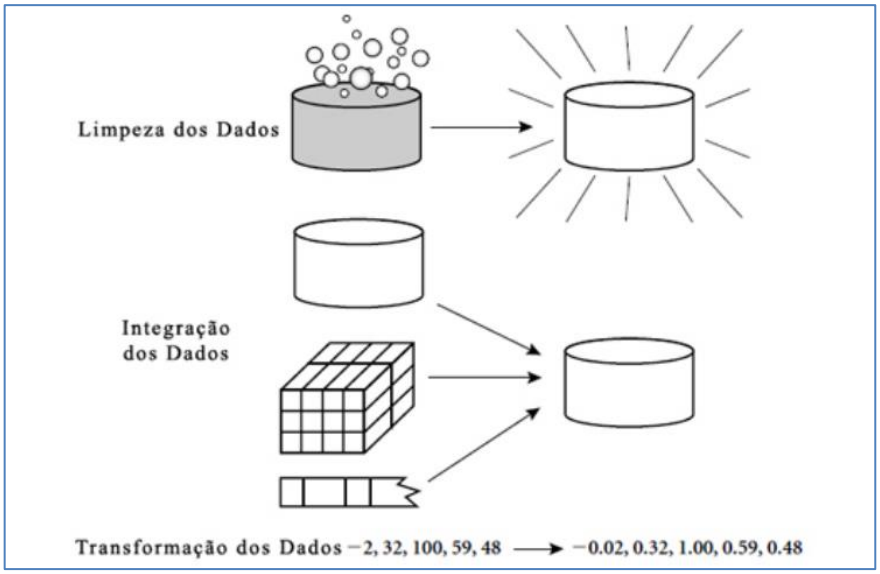
Ao ser definido o formato dos registros disponíveis ao estudo, assim como os próprios registros, é preciso traçar uma lista de metas esperadas pelo processo, bem como os critérios de desempenho que serão importantes, e as restrições impostas pelo contexto. Esses aspectos podem ser sintetizados em: qual informação será buscada, meta que servirá para validar os resultados encontrados.

Enquanto a compreensão do sistema é uma função geralmente desempenhada pelos especialistas responsáveis por manter o *software*, é comum que o educador estabeleça as metas a serem cumpridas. Estas metas podem envolver tentativas de prever comportamentos específicos dos alunos diante do curso ou dos testes, assim como identificar padrões de acesso às páginas ou recursos disponíveis. De forma análoga, é possível analisar aspecto do próprio ambiente Virtual de Aprendizagem, como encontrar pontos do curso que precisem ser revisados e melhorados, entre outras possibilidades.

5.3 Terceiro: Transformações necessárias

Algumas medidas para remoção de dados imprecisos ou falhos geralmente são necessárias (Figura 9), dependendo de fatores como o AVA utilizado, o tipo do registro escolhido como fonte de informação assim como o formato e o local em que esses dados estão armazenados.

Figura 10: Atividades de pré-processamento.



Fonte: HAN; KAMBER; PEI, 2011.

Sobre a Integração, os dados encontrados podem estar disponíveis em diferentes formatos e sistemas, se fazendo necessária a integração das fontes de dados para posterior entrada nos algoritmos de extração de padrões.

Referente à limpeza de dados, os dados podem apresentar problemas advindos do processo de coleta, como erros de leitura por sensores, erros de digitação, valores inválidos, ruídos, etc. A qualidade dos dados é um fator extremamente importante. Esse tipo de transformação inclui decisões sobre as estratégias para lidar com campos de dados faltantes e representação de informações sequenciais no tempo, bem como decidir questões de SGBD referente a tipos de dados, esquema e mapeamento de valores desconhecidos. Elimina-se dados espúrios e registros incompletos ou errôneos que possam dificultar a análise ou até mesmo alterar os resultados significativamente.

Como aspecto falho, e dificilmente controlável, Falci Jr. e Ricarte (2009) citam a susceptibilidade dos ambientes virtuais de aprendizagem a problemas inerentes da

conexão do usuário na rede de computadores, tais como queda de conexão e velocidade de transmissão dos dados, que podem gerar uma grande quantidade de informações incompletas nos registros, atrapalhar o fluxo de navegação do usuário, e causar ruído nos registros de acesso ao incitar o usuário a múltiplos pedidos de atualização de uma página tentando terminar de carregá-la. Desta forma, mesmo registros bem organizados podem necessitar de limpeza ou de uma reorganização em novas tabelas e bases de dados, facilitando a associação de informações, e concentrando apenas as informações desejáveis para a análise a ser realizada.

A transformação visa adequar os dados para a entrada nos algoritmos de mineração de dados escolhidos. É importante salientar que a execução das transformações deve ser guiada pelas metas traçadas para a extração, possibilitando que os dados resultantes dessa etapa apresentem características necessárias na obtenção dos resultados desejados. Este é um passo desenvolvido em conjunto na presença do especialista responsável pelo sistema e o especialista em mineração de dados. O primeiro, por conhecer bem os dados do sistema e o segundo, por conhecer bem o formato necessário nos algoritmos de extração do conhecimento.

5.4 Quarto: Escolha do escopo e seleção de variáveis

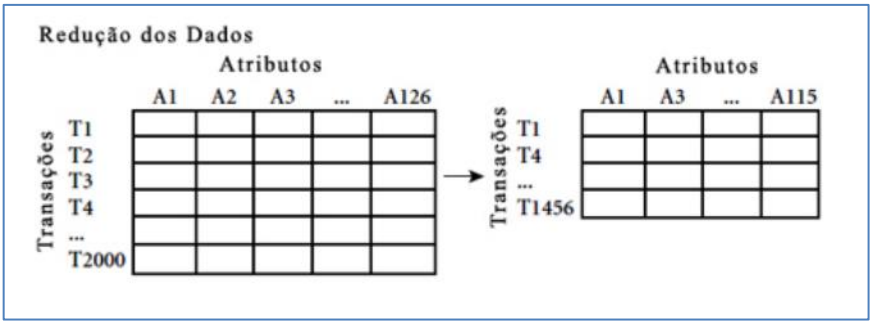
Tendo compreendido o contexto e realizado as transformações necessárias deve-se selecionar apenas os dados que são significativos para a análise pretendida (Figura 10). Diante da grande quantidade de informações armazenadas pelo AVA, é primordial a pré-seleção dos dados mais relevantes a serem explorados nos passos seguintes, reduzindo o número de variáveis consideradas. Essa escolha deve levar em conta os tipos de dados presentes e o potencial que eles têm para descrever padrões relevantes ao objetivo proposto. Etapas de seleção, filtragem e formatação destes dados são realizadas neste ponto do processo.

Para a aplicação das técnicas de mineração aos dados, é oportuna a criação de uma base de dados que servirá como objeto de busca da mineração, recorrendo à seleção de um conjunto de variáveis ou um subconjunto de dados no qual a

descoberta deve ser realizada. Nessa etapa é admitida a concepção de um conjunto de dados, uma visão ou um *Datawarehouse*[10].

Também é aqui que se define o escopo do estudo, ou em qual granularidade as informações serão detectadas (observações sobre o aluno, sobre um período de tempo específico, sobre uma atividade disponível, etc.). A definição do escopo remete a qual nível de importância cada informação tem.

Figura 11: Atividades de pré-processamento.



Fonte: HAM; KLAMBER, 2006.

Por exemplo: Para obter informações em um Ambiente Virtual com granularidade no nível do curso, realiza-se uma média de todas as observações sobre o aluno, obtendo o percentual das suas ações no curso. Já no nível de uma atividade, realiza-se uma média de todas as observações sobre o aluno em determinada atividade. Este mapeamento prevê os dados no mesmo nível da granularidade: se a granularidade for ao nível de aula, as observações serão feitas tomando como referência cada aula, sendo possível obter resultados sobre a interação do aluno aula a aula, e também possibilitando intervir sobre suas ações em uma frequência por aula. Outras granularidades significativas: nível do estudante, nível anual, nível da turma, etc. (PARDOS, 2013).

5.5 Quinto: Escolha do método para minerar dados

Com os dados adequados, chega o momento de processá-los em busca dos resultados estabelecidos pelas metas definidas nos passos anteriores. Este é o

[10] Estrutura utilizada para armazenamento e análise de grandes volumes de dados de forma consolidada, favorecendo a obtenção de informações estratégicas para a tomada de decisão e a previsão de eventos futuros.

ponto crucial da contribuição do especialista em mineração de dados para o desenvolvimento da metodologia, uma vez que é desempenhado somente por ele.

A escolha do método é feita de acordo com os objetivos desejáveis para a solução a ser encontrada. As técnicas possíveis de um algoritmo de extração de padrões podem ser agrupadas em métodos, e isso inclui decidir o propósito do modelo derivado do algoritmo de mineração de dados.

Na Tabela 4 é possível identificar os principais métodos para minerar dados Educacionais e suas respectivas aplicações. Os métodos estão apresentados conforme categorização na taxonomia proposta por Baker *et al.* (2010), melhor definidos no Capítulo 3.

Nesta etapa do processo é também importante utilizar de forma incremental os métodos de seleção e visualização de dados, na busca de padrões para decidir quais técnicas e parâmetros podem ser apropriados. Por exemplo: os modelos de dados categóricos são diferentes dos modelos de dados numéricos, influenciando na escolha do algoritmo e consequentemente do método.

5.6 Sexto: Escolha da técnica e do algoritmo

Uma vez eleito o método a ser empregado, existe uma variedade de técnicas e algoritmos par executá-lo. Nessa etapa o especialista em Mineração de dados, a partir dos objetivos propostos e dos tipos de dados disponíveis, escolhe o melhor algoritmo para determinada finalidade, validando os resultados encontrados.

Algumas das técnicas conhecidas na literatura são: Arvore de decisão (para valores categóricos), Arvore de regressão (para valores numéricos), Redes Neurais, Algoritmos Genéticos, lógica Fuzzy, regras de decisão, classificadores baseados em instâncias, entre outros. Dentre os algoritmos é possível citar: J48, C4.5, K-means, Apriori, Arvore M5 (com equações lineares em cada folha da árvore), entre outros.

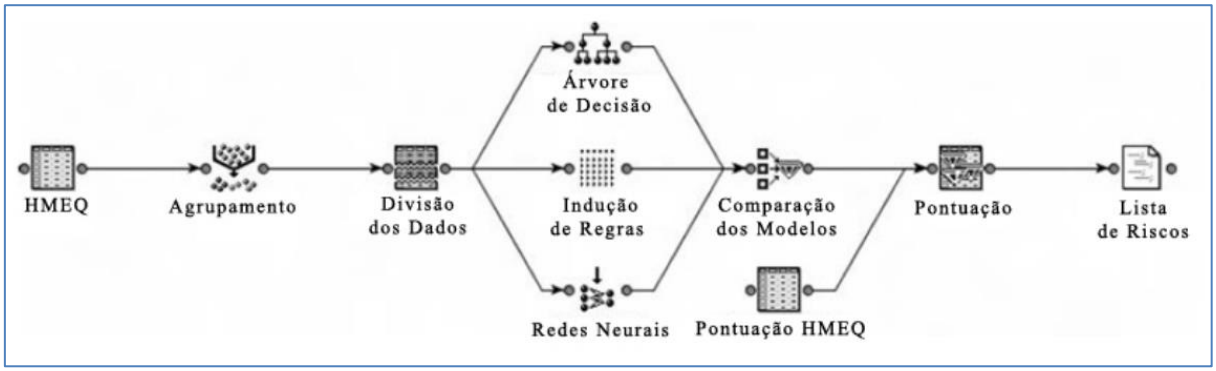
Tabela 3: Taxonomia das principais subáreas da EDM.

Categoria do Método	Objetivo do Método	Aplicação em Educação
Predição	Inferir um único aspecto dos dados (variável preditiva) a partir de uma combinação de outros aspectos dos dados (variáveis preditoras).	Detectar comportamentos dos alunos (trapaças no sistema, comportamento *off-task*[11]); Desenvolver modelos de domínio; Prever e entender o desempenho na aprendizagem dos alunos.
Agrupamento (Clustering)	Encontrar pontos de dados que naturalmente se agrupam, racionando os dados em diferentes categorias e/ou grupos.	Investigar semelhanças no comportamento considerando diversos contextos (escola, turma, material didático); Relacionar grupos de alunos considerando o comportamento apresentado na interação, ou grupos de escolas.
Mineração de Relações	Descobrir relações entre as variáveis.	Descobrir estratégias para melhorar a aprendizagem; Melhorar a disposição curricular do curso associando competências e requisitos nas tarefas; Encontrar relação entre a dificuldade do aluno, e a abordagem da disciplina.
Descoberta com modelos	Desenvolver um modelo a partir de métodos como predição, agrupamento, ou engenharia do conhecimento, e utilizá-lo como ponto de partida em outra análise.	Descobrir relações entre os comportamentos dos alunos e suas características; Analisar a pesquisa em toda a variedade de contextos.

[11] Quando um aluno sai do Ambiente Virtual de Aprendizagem, ou não faz nenhum comportamento relacionado com o ambiente educacional.

É possível aplicar vários algoritmos para realizar a tarefa desejada durante o processo de mineração. Diversas técnicas devem ser testadas e combinadas levando a obtenção de diversos modelos a fim de que comparações possam ser feitas e então a melhor técnica, ou um conjunto delas, seja utilizada, conforme proposto por Kohavi, Sommerfield e Dougherty (1996) e Mccue (2007). Na Figura 11 visualiza-se um exemplo de combinação dessas técnicas.

Figura 12: Realização da tarefa com várias técnicas.



Fonte: MCCUE 2007, *apud* Camilo, 2009, tradução do autor.

Por fim, os resultados obtidos pelos algoritmos de mineração de dados devem oferecer respostas aos objetivos nos passos anteriores. As etapas propostas são interativas. Sendo assim, não encontrando bons resultados em uma primeira tentativa requer um retorno à análise de todo o processo para identificação de pontos problemáticos que precisem ser alterados ou repensados para análise adequada.

5.7 Sétimo: Interpretação dos resultados

Após executado o modelo, o próximo passo é interpretar os padrões descobertos e, eventualmente, voltar a qualquer um dos passos anteriores. Essa etapa inclui visualização dos padrões extraídos, remoção de padrões redundantes ou irrelevantes, e tradução em termos compreensíveis aos usuários. Pode ser necessária uma melhor apresentação dos resultados, com técnicas de visualização de dados, facilitando a compreensão dos especialistas de domínio (educadores). A compreensibilidade de um dado conjunto de regras é um fator importante para a qualidade do modelo descoberto.

O trabalho dessa etapa pode também ter um foco em organizar e filtrar informações úteis presentes em meio aos dados, não analisando as mesmas, mas destacando-as para que o educador possa ter fácil acesso a elas e possa traçar suas próprias conclusões (FALCI JR; RICARTE, 2009).

5.8 Oitavo: Escolha dos parâmetros para validação

Para a confirmação do trabalho realizado, faz-se necessário saber se o modelo inferido é confiável. A utilização de parâmetros de validação dos resultados mede a confiabilidade do modelo proposto.

Um desses parâmetros é a Análise de Concordância (*Kappa*), que é baseada no número de respostas concordantes. Segundo Baltar e Okano (2005) *Kappa* é uma medida de concordância entre observadores e mede o grau de concordância além do que seria esperado tão somente pelo acaso. Esta medida de concordância tem como valor máximo o 1, onde este valor 1 representa total concordância e os valores próximos e até abaixo de 0, indicam nenhuma concordância, ou a concordância foi exatamente a esperada pelo acaso. Um eventual valor de *Kappa* menor que zero, negativo, sugere que a concordância encontrada foi menor do aquela esperada pelo acaso, e, portanto, a discordância. Landis e Koch (1977) consideram a interpretação desses valores conforme Tabela 5.

Tabela 4: Interpretação dos valores *Kappa*.

Valores de *Kappa*	Interpretação
<0	Nenhuma concordância
0-0.19	Pouca concordância
0.20-0.39	Concordância razoável
0.40-0.59	Concordância moderada
0.60-0.79	Concordância considerável
0.80-1.00	Concordância quase perfeita

Fonte: Próprio autor.

Outra importante medida na Classificação é a taxa de erro de um classificador h, também conhecida como taxa de classificação incorreta e denotada por err(h). Usualmente, a taxa de erro é obtida utilizando a fórmula 1, a qual compara a classe verdadeira de cada exemplo com o rótulo atribuído pelo classificador induzido. O operador ||E|| retorna 1 se a expressão E for verdadeira e zero caso contrário, e n é o número de exemplos. O complemento da taxa de erro, a precisão do classificador,

denotada por acc(h) é dada pela fórmula 2. Em geral, o erro calculado sobre o conjunto de exemplos de treinamento - erro aparente - é menor que o erro calculado sobre o conjunto de exemplos de teste - erro verdadeiro (MONARD *et al.* 2003).

$$\mathrm{err(h)} = \frac{1}{n}\sum_{i=1}^{n} || \, y_i \tag{1}$$

$$\mathrm{acc(h)} = 1 - \mathrm{err(h)} \tag{2}$$

É importante estimar o desempenho futuro do modelo induzido utilizando o conjunto de exemplos dado. Esse procedimento visa obter a confiabilidade nas previsões (HAN; KAMBER; PEI, 2011). A validação cruzada é uma das técnica amplamente empregada em modelagem de predição para esse fim. É baseada em amostragem para avaliar a capacidade de generalização de um modelo, a partir de um conjunto de dados. O conceito central da técnica é o particionamento do conjunto de dados em subconjuntos, e posteriormente, a utilização de alguns destes subconjuntos para a estimação dos parâmetros do modelo (dados de treinamento). Os demais subconjuntos (dados de validação ou de teste) são empregados na validação do modelo.

Monard e Baranauskas (2003) sugerem diversas formas de realizar o particionamento dos dados, sendo as três mais utilizadas na literatura: o método *holdout*, o *k-fold* e o *leave-one-out*.

Holdout: O estimador *holdout* divide os exemplos em uma porcentagem fixa de exemplos p para treinamento e (1 − p) para teste, considerando normalmente p > 1/2. Valores típicos são p = 2/3 e (1 − p) = 1/3, embora não existam fundamentos teóricos sobre estes valores.

K-fold: Este estimador é um meio termo entre os estimadores *holdout* e *leave-one-out*. Os exemplos são aleatoriamente divididos em k partições (*folds*) de tamanho aproximadamente igual a n/k exemplos. Os exemplos nos (k − 1) *folds* são usados para treinamento e a hipótese induzida é testada no *fold* remanescente. Este processo é repetido k vezes, cada vez considerando um *fold* diferente para teste. O erro na é a média dos erros calculados em cada um dos k *folds*.

Leave-one-out: O estimador *leave-one-out* é um caso especial. Por ser computacionalmente dispendioso é frequentemente usado em amostras pequenas. Para uma amostra de tamanho n uma hipótese é induzida utilizando (n − 1) exemplos; a hipótese é então testada no único exemplo remanescente. Este processo é repetido n vezes, cada vez induzindo uma hipótese e deixando de considerar um único exemplo. O erro é a soma dos erros em cada teste dividido por n.

5.9 Nono: Utilização do conhecimento

Por fim, incorpora-se o modelo adquirido ao sistema de aprendizagem, tomando ações decisórias baseadas nos conhecimentos obtidos por meio dos dados, ou simplesmente documentam-se as descobertas à disposição dos interessados.

Problemas podem ocorrer ao longo de todo o processo, se fazendo necessário escolher novos dados, novas metas ou novos algoritmos para processamento dos dados. Cada incremento permite obter respostas melhores e mais significativas ao final de todo o processo.

6 Caracterização do Problema

6.1 Contextualização

O Governo Federal criou vários programas de incentivo para levar o ensino a distância nos mais diversos pontos do país, assim como para ampliar a oferta de cursos profissionalizantes. O Serviço Nacional de Aprendizagem Industrial (SENAI), instituição de ensino analisada no experimento desse Estudo de Caso, iniciou em 2011 uma ação intitulada “PN-EAD SENAI” visando ampliar ainda mais sua abrangência. Essa ação foi voltada para o desenvolvimento de cursos e materiais de apoio instrucional na modalidade de Educação a Distância, assim como para os treinamentos e capacitações dos responsáveis por essa modalidade nas Unidades Operacionais do SENAI. Os treinamentos foram executados por intermédio do Ambiente AV@S (SIQUARA; BRAGA; ALMEIDA, 2012).

No Estado da Paraíba o Ambiente Virtual do SENAI é baseado no Sistema Moodle. Nosso estudo de caso buscou estratégias para minimizar as causas da evasão escolar desse contexto na missão de contribui com as ações do Departamento Regional da Paraíba para promover a educação profissional e tecnológica com padrão de excelência e tratar desse problema.

Os experimentos foram realizados no Banco de dados do Ambiente Virtual do Serviço Nacional de Aprendizagem Industrial (SENAI), instituição privada, de interesse público, sem fins lucrativos , estruturado em base federativa com ampla gama de programas de formação profissional. Os dados referem-se ao Departamento Regional do SENAI Paraíba.

O SENAI, concentrando seus esforços na ampliação do número de matrículas no ensino profissionalizante para suprir a alta demanda de profissionais no país, em 2011 iniciou uma ação nacional intitulada “PN-EAD SENAI” – Programa Nacional de Oferta de Educação Profissional na Modalidade a Distância – voltada para o investimento em cursos de Educação a Distância, por compreender que essa modalidade é capaz de atrair um alto volume de interessados visto os inúmeros benefícios que o Ensino a Distância oferece para os estudantes, dentre eles: redução da mobilidade, baixo custo, flexibilidade de horários, aceitação de mercado – visto o perfil favorável dos profissionais formados nessa modalidade: organização,

atenção, dedicação, pontualidade. O público alvo desses cursos são jovens estudantes e trabalhadores em geral que, por inúmeros motivos, encontram dificuldades em se inserir no ensino integralmente presencial.

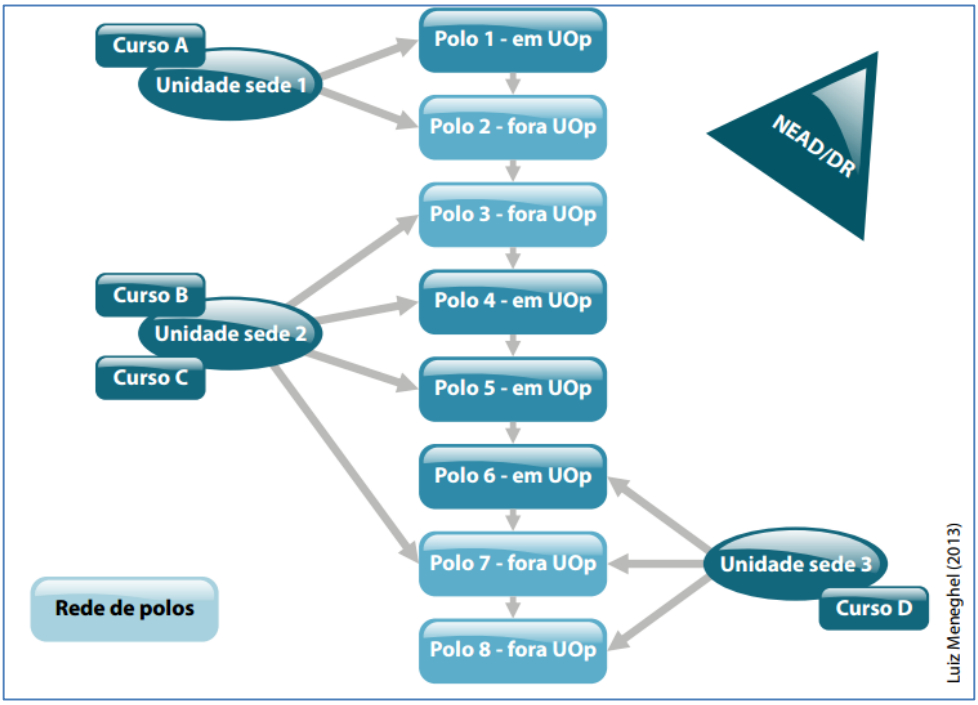
Os cursos do PN-EAD são projetados para serem realizados em um Ambiente Virtual de Aprendizagem (AVA), com materiais on-line que orientam os alunos a realizarem atividades virtuais e presenciais, além de contar com apoio de livros didáticos e acompanhamento educacional sistemático. Os cursos do PN-EAD seguem a metodologia de formação baseada em competências, por meio das Situações de Aprendizagem, que são desafios propostos aos alunos para solucionar problemas, tomar decisões, testar hipóteses e aplicar o que aprenderam a outro contexto.

Figura 13: Materiais on-line do Ambiente Virtual do SENAI.

As equipes dos Departamentos Regionais (DR) desenvolvedores são formadas por muitos profissionais, destacando-se o Conteudista - Profissional com competência reconhecida na área de conhecimento do curso; Revisor técnico - Profissional que, com competência reconhecida na área de conhecimento do curso, coordena tecnicamente a elaboração dos livros didáticos e dos materiais on-line, Designer educacional - Profissional cujas atividades principais estão relacionadas à elaboração do projeto pedagógico do curso a distância e, juntamente com conteudistas e o grupo de produção multimídia, ao desenvolvimento do curso a distância; Grupo de produção multimídia - Conjunto multidisciplinar de profissionais responsáveis por produzir os livros didáticos e materiais on-line conforme encomendado pelos conteudistas, revisores técnicos e designers educacionais. Em alguns DR's incluem-se ainda programadores de animações, jogos digitais e banco de dados, designers gráficos, web designers, web desenvolvedores, ilustradores, roteiristas, diretores de vídeo, cinegrafistas, técnicos de áudio e vídeo, entre outros.

Analogamente, para oferecer cursos nessa modalidade, cada Departamento Regional se estrutura a partir do seu Núcleo de Educação a Distância (NEAD), que, entre outras funções, faz a interlocução com o Departamento Nacional (DN), conforme exposto na Figura 14.

Figura 14: Estrutura organizacional nos Departamentos Regionais do SENAI para o PN-EAD.

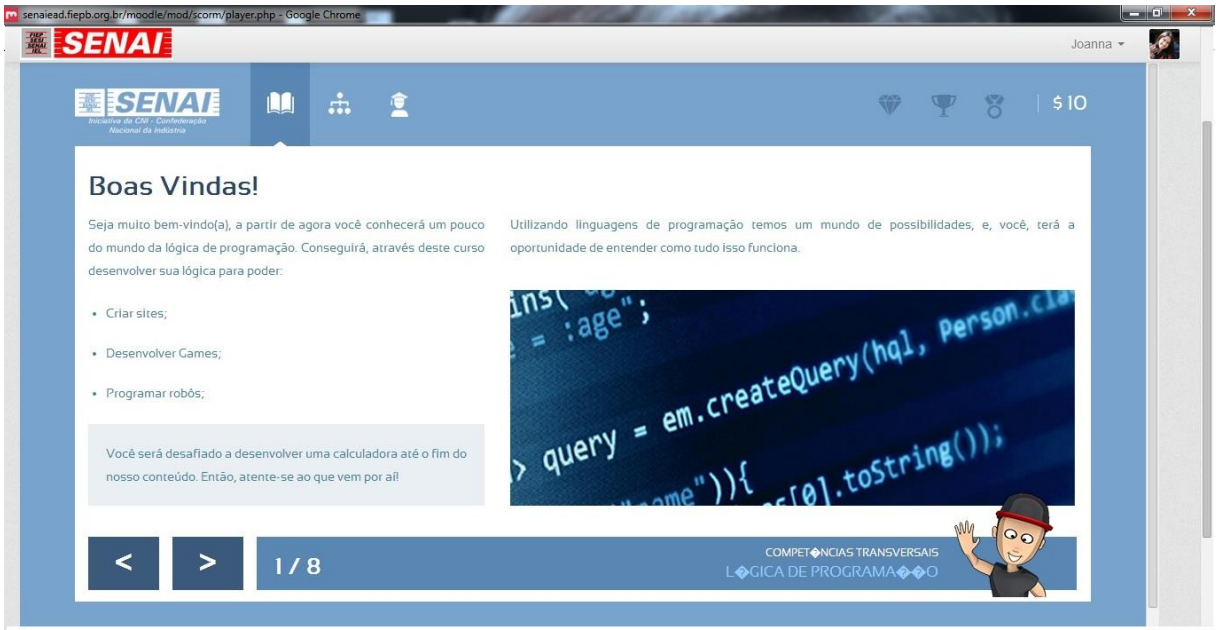


Fonte: Sabino *et al.,* 2011.

A equipe executora em cada DR encontra-se dividida em seis papéis profissionais previstos para o PN-EAD: Gestor – Gerencia todos processos relacionados ao Núcleo de Educação a Distância, assegurando a implantação do PN-EAD no DR; Tutor – Domina o conteúdo da área tecnológica do curso e a metodologia de ensino, visto que interage com os alunos por meio do AVA e, conforme a configuração da equipe no DR, atua também nas práticas presenciais; Monitor – Orienta os alunos em questões técnicas e administrativas, tanto no AVA quanto presencialmente; Coordenador pedagógico – Orienta a atuação da tutoria e da monitoria e cuida dos aspectos didático-pedagógicos intra e intercursos, conforme definido nos Planos de Ensino e nos Planos de Situação de Aprendizagem; Coordenador técnico do curso – Orienta o tutor tecnicamente e assegura a qualidade da execução do curso conforme definido no Plano de Curso; Responsável pelo polo – Organiza e monitora a execução das atividades e encontros presenciais.

Ambiente Virtual do SENAI

Figura 15: Ambiente Virtual de Aprendizagem do SENAI.



Fonte: *Print screen* do Ambiente Virtual de Aprendizagem do SENAI Paraíba.

De acordo com o papel de cada integrante na equipe executora, e sua permissão de acesso, é possível verificar relatórios do andamento das atividades dos alunos no curso, demonstrando a quantidade de tópicos visitados, acessos realizados, mensagens enviadas no chat e no fórum, notas nas atividades, etc. A Figura 15 mostra a tela inicial de um dos cursos do ambiente Virtual do SENAI.

6.2 Experimento

Esse capítulo apresenta a aplicação de uma metodologia para minerar dados educacionais em sistemas de ensino a distância visando traçar o perfil de acesso dos estudantes em um Ambiente Virtual de Aprendizagem (AVA). Para a realização deste trabalho, um cenário de uso foi construído utilizando registros oriundos do Ambiente Virtual do SENAI Paraíba, um sistema educacional desenvolvido pelo Departamento Regional do Serviço Nacional de Aprendizagem Industrial, baseado no Moodle, um sistema de Gestão da Aprendizagem amplamente utilizado e com expressivas citações na literatura.

Foram selecionadas turmas na base de dados dos cursos realizados na modalidade à distância, que atendesse os seguintes requisitos experimentais: maior quantidade de estudantes por turma, turmas concluídas com disponibilidade do status final do aluno, maior número de dados referentes aos recursos do AVA utilizados. Com base nos critérios apresentados para essa experimentação, escolheram-se dados de *Log* dos alunos de Qualificação nos cursos de Supervisor Inovador, Desenhista de Produtos Gráficos Web e Operador de Computador, contendo um total de 80 estudantes e suas mais de 70.000 interações com o sistema. Não foram considerados estudantes que desistiram da turma logo nos primeiros acessos ao Ambiente Virtual. A partir da identificação da turma objeto desse estudo, um conjunto de atributos possíveis de serem extraídos de um AVA, significativos para a definição e estruturação do modelo de previsão proposto, foi admitido.

A hipótese proposta nesse experimento buscou relacionar os comportamentos dos alunos registrados no Ambiente Virtual de Aprendizagem (*log* de acesso, materiais acessados, atividades realizadas), considerando também as realidades sociais dos mesmos (faixa etária, sexo, curso, bairro da residência), em busca de padrões de acesso, de forma a traçar o perfil dos alunos que evadem ou reprovam cursos na modalidade a Distância, assim como auxiliar os responsáveis do SENAI pelo serviço de relações com o Mercado na identificação de estudantes com características adequadas para os cursos EAD.

O estudo se baseia na Predição, recorrendo a métodos da sua subcategoria Classificação. A predição é utilizada para estimar o valor de um atributo, que nesse

estudo de caso foi o status final do aluno na turma. A técnica escolhida foi Arvore de Decisão, por constituir uma técnica muito poderosa e amplamente empregada. O algoritmo foi o J48, um dos mais disseminados na literatura, devido a bons resultados nas avaliações científicas.

Limitação dos dados utilizados no experimento

Os registros próprios de um sistema costumam ser gravados seguindo algum padrão de formatação. No entanto, os dados registrados do cenário desenvolvido para esse trabalho, no servidor Moodle, continham informações limitadas quando ao status do aluno no sistema, causadas por problemas técnicos durante a execução das turmas. Com essa limitação em vista, se fez necessário completar os dados do sistema com dados coletados por meio de registros manuais e em base de dados externa (sistema de gerenciamento Escolar). Esse tipo de registro requer esforço muito específico de análise e transformação, o que dificultou o estudo de caso.

6.2.1 Metodologia

Apesar das limitações impostas pela quantidade de informações, o experimento pôde ser realizado considerando a metodologia proposta no Capítulo 5, em nove passos:

Primeiro passo – Compreensão dos dados

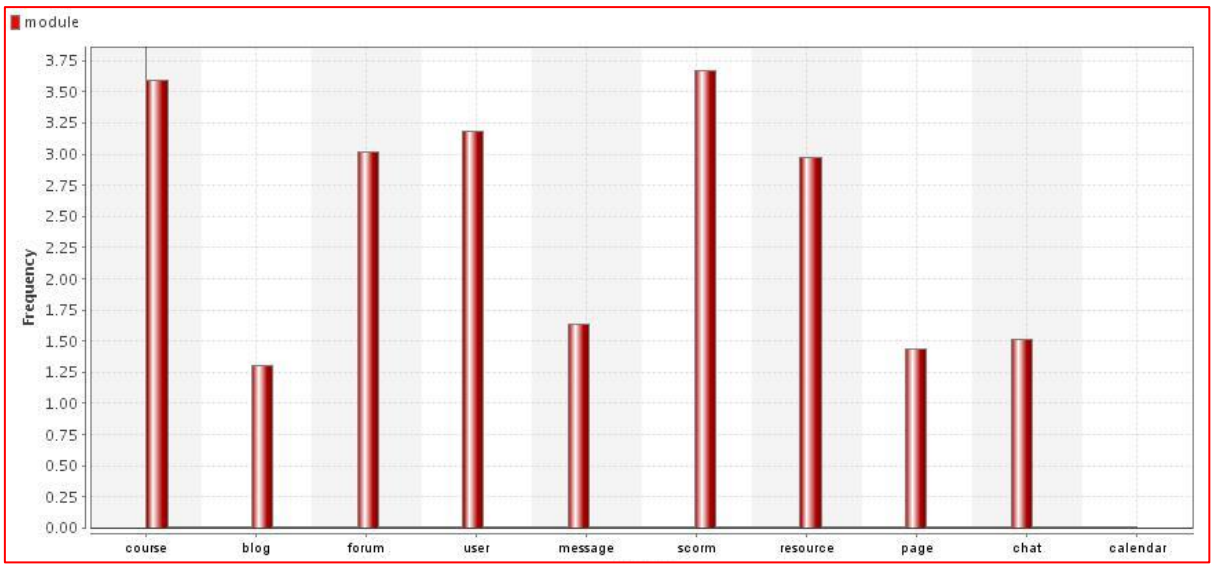
Inicialmente foi necessário um estudo do Ambiente Virtual de Aprendizagem do SENAI, das ferramentas disponibilizadas e como o gerenciador Moodle capta e armazena os dados mantidos pelos mesmos. Nessa etapa foram envolvidos o responsável pela aplicação (especialista no Ambiente Virtual do Senai) e o responsável pelo processo da Mineração de Dados (especialista em Mineração de Dados).

O escopo do estudo compreendeu informações sobre alunos de uma turma de qualificação do curso de Supervisor Inovador, que haviam concluído o curso, impossibilitando adaptações no sistema de captura dos dados a serem utilizados. Fez-se necessária a utilização apenas dos dados que o ambiente dispunha, assim como uma pré-seleção das fontes de dados mais promissoras a serem exploradas

nos passos seguintes, visando o potencial das mesmas para descrever os comportamentos e hábitos dos alunos.

Dentre os diversos aspectos comportamentais dos estudantes, considerou-se relevante a detecção de padrões de acesso oriundos de alunos que evadiram ou reprovaram o curso, em equivalência aos que concluíram com aprovação, visto que altos índices de evasão e reprovação são um problema recorrente para a modalidade de cursos a distância.

Figura 16: Frequência de acesso dos usuários referente a cada módulo do curso.



Fonte: *Print screen* da aplicação RapidMiner®.

O curso em questão ofereceu diversos mecanismos de interação com os estudantes como Fórum, *Chat*, Blog, animações em *flash* disponíveis em pacotes *scorm*, mensagens, entre outros. Na Figura 15 é possível analisar a frequência de acesso dos alunos aos módulos disponíveis – os valores encontram-se representados em escala logarítmica.

Segundo passo – Identificação do problema

Segundo o censo EAD Brasil (2013) duas das principais causas apontadas pelos alunos para a evasão em cursos na modalidade a distância foram: falta de tempo para o estudo e para participar do curso, e a falta de adaptação à metodologia. Atentos a essas questões, considerou-se essas afirmativas como base para a pesquisa sobre as causas dessa evasão, durante o experimento proposto, buscando padrões ao relacionar os comportamentos dos alunos registrados no

Ambiente Virtual de Aprendizagem às realidades demográficas destes, de forma a traçar o perfil dos alunos que evadem ou reprovam cursos na modalidade a Distância.

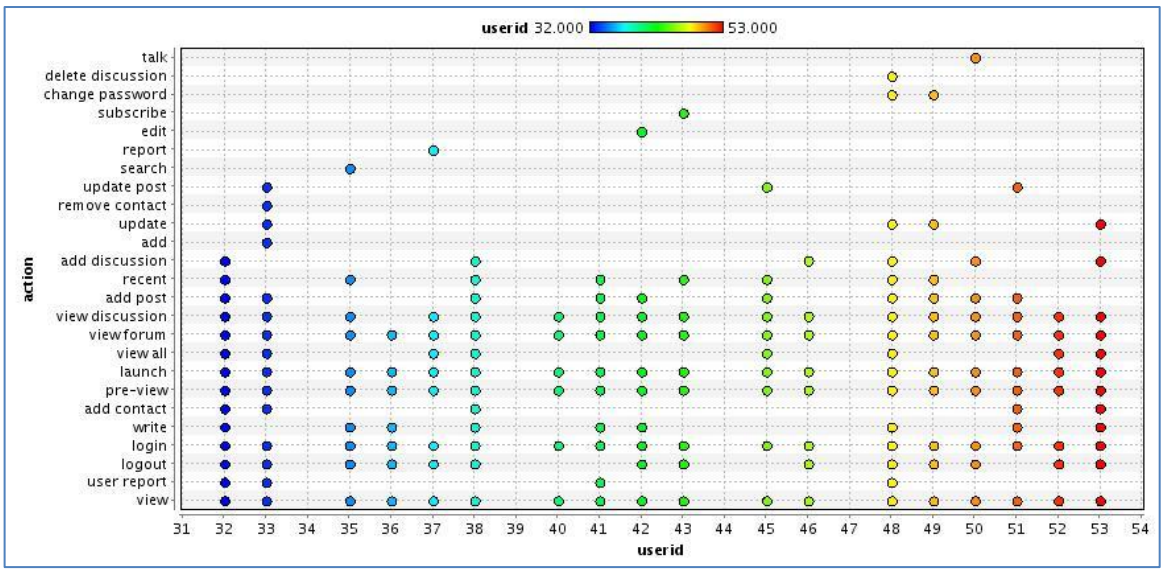
Uma etapa fundamental foi estipular quais informações sobre o aluno, e as respectivas atividades realizadas, são parâmetros importantes para delinear seu perfil de acesso. Nessa etapa foram envolvidos os responsáveis pelo acompanhamento dos alunos (monitor e tutor) e o responsável pelo processo da Mineração de Dados (especialista em Mineração de Dados). Os atributos propostos nesse estudo são organizados em quatro grupos, descritos a seguir. O objetivo foi contemplar diversos aspectos de uso e interação do estudante no Moodle:

- **Dados do Aluno:** nesse grupo de atributos o objetivo é identificar dados que representem aspectos sociais do aluno, e seu perfil de uso na realização do curso.
- **Interação**: com esse grupo de atributos pretende-se destacar a interação dos estudantes uns com os outros, usando as ferramentas disponíveis (fórum, *chat*, envio ou recebimento de mensagens), assim como entre o estudante e o professor (interação dos tutores, monitores e responsáveis pedagógicos com os estudantes no contexto do AVA). Espera-se, com estes atributos, identificar a existência de colaboração e cooperação na aprendizagem individual.
- **Aproveitamento**: esse grupo de atributos representa o cumprimento do aluno às atividades previstas, dependendo dos requisitos definidos pelo professor para cada uma delas, e o registro do seu aproveitamento (pontuação em atividades avaliativas, cumprimento de forma satisfatória das atividades ou situações problema, etc).
- **Dedicação**: nesse grupo de atributos definiram-se indicadores gerais de quantidade de acessos aos recursos do AVA.

Os gráficos gerados pela ferramenta RapidMiner® foram utilizados para melhor compreensão da disposição dos dados. Foi realizado um estudo sobre as tabelas para compreensão de quais dados estavam disponíveis para o propósito deste estudo.

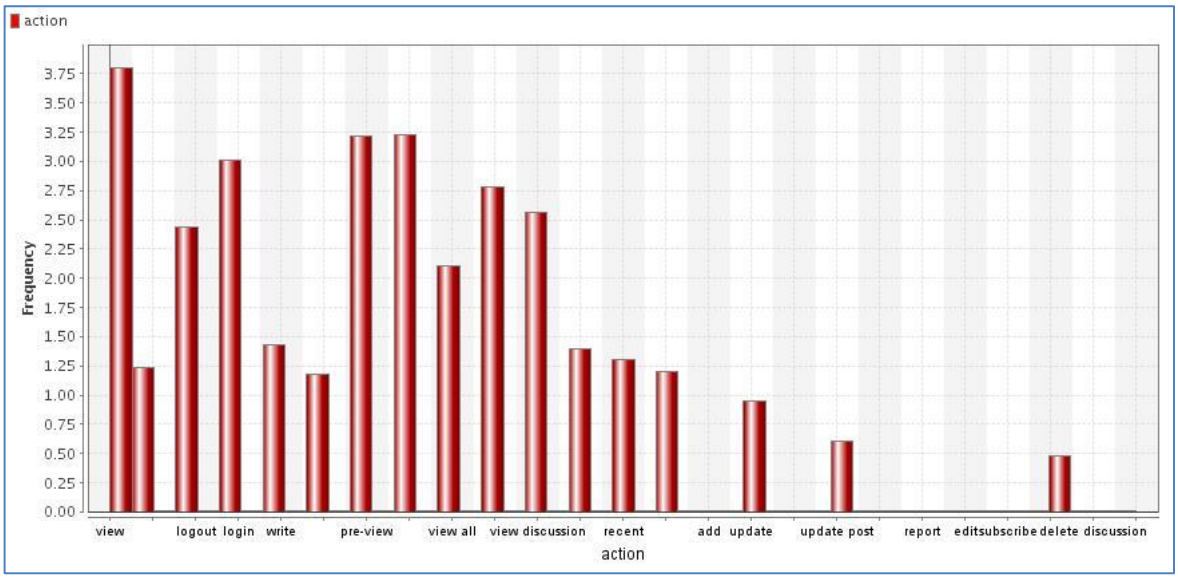
A primeira fonte de dados considerada foram os registros de acesso. Na Figura 16 encontra-se a ocorrência de registros no arquivo de log do sistema de uma das turmas do estudo de caso, com as ações possíveis de serem realizadas. Analogamente, na Figura 17 encontram-se dados considerando a frequência de ações realizadas pelos alunos da turma no curso em questão.

Figura 17: Ocorrência de registros dos usuários nas Ações disponibilizadas pelo sistema.



Fonte: *Print screen* da aplicação RapidMiner®.

Figura 18: Frequência de acesso agrupado por Acões.



Fonte: *Print screen* da aplicação RapidMiner®.

A Tabela 6 apresenta o nome das tabelas da base de dados do Moodle selecionadas para o estudo, permitindo avaliar como cada usuário interage com o ambiente. A Tabela 7 apresenta os atributos selecionados, de acordo com os grupos de informações propostos.

Tabela 5: Tabelas da base de dados do Moodle.

Tabela	Especificação
mdl_scorm_scoes_track	Dados de visualização do material didático online.
mdl_message	Registros de mensagens enviadas e recebidas.
mdl_forum	Registros de interações na ferramenta fórum.
mdl_log	Arquivo de *log* com dados referentes ao uso do sistema.

Fonte: Próprio autor.

Terceiro passo – Transformações necessárias nos dados

Nessa etapa foi realizada uma cópia das tabelas necessárias para o servidor local, durante consultas codificadas com a linguagem SQL. Houve restrições quanto à utilização de Visões no Banco de Dados e a conexão com a ferramenta RapidMiner®. O sistema inicialmente possuía dados de *log* de todos os usuários do ambiente virtual, o que inclui dados de professores e administradores além dos dados dos alunos. Uma seleção de tuplas foi realizada, considerando informações apenas referentes a turma em questão, assim como os usuários com o perfil de acesso intitulado “estudante” (registrado na tabela *mdl_role*).

Para viabilizar o processo de Mineração de dados, algumas transformações nos dados foram necessárias. Os dados registrados no servidor Moodle continham informações limitadas quando ao *status* do aluno no sistema, e alguns dados pessoais dos alunos (data de nascimento e situação civil). Esses dados foram informados com o auxílio de uma planilha complementar, retirada do sistema de gerenciamento Escolar da organização, o que tornou o processo de integração uma transformação imprescindível.

Tabela 6: Atributos escolhidos para descrever as interações dos estudantes.

Grupo	Atributo	Descrição
Dados do aluno	Total_login	Quantidade total de acessos realizado pelo aluno no AVA.
	Ultimo_acesso	Data do ultimo acesso ao sistema.
	Situacao_civil	Situação civil do aluno.
	Idade	Idade do aluno.
	Bairro	Bairro em que o aluno reside.
Interação	Total_post_forum	Soma das interações do usuário quanto ao adicionar postagens no fórum.
	Total_discus_forum	Soma das interações do usuário quanto ao adicionar um novo tema de discussão no fórum.
	nr_msg	Número de mensagens enviadas ao professor/tutor e aos alunos.
	nr_msg_rec	Número de mensagens recebidas do professor/tutor e dos alunos.
Dedicação	Total_interacao_scorm	Quantidade total de acesso ao modulo scorm.
	Total_acessos_forum	Quantidade total de acesso ao fórum.
	Total_view	Total de visualizações do material didático.
	Total_msg_view	Total de mensagens visualizadas.
	Aproveitamento	Valor binário, representando se o aluno finalizou o curso com aproveitamento. **Atributo objetivo da previsão.**

Fonte: Próprio autor.

Após a integração das fontes de dados, foi realizada a indução de novos atributos, a partir dos dados originais. Para obter os atributos *Nr_msg*, *Nr_msg_rec*, *Total_acessos, Total_interacao_scorm, Total_acesos_forum, Total_view_sistema,* e *Total_msg_view*, que representam quantas ações de cada tipo o usuário executou no ambiente, contabilizou-se o total de ocorrências para cada aluno na tabela *mdl_log*. Esse processo também ocorreu para obter os atributos *Total_post_forum, Total_discus_forum*, da tabela *mdl_forum* e para obter *Total_msg* e *Total msg_rec* da tabela *mdl_msg*. Na Tabela 7 é possível analisar as consultas foram realizadas no Banco de Dados utilizando a linguagem SQL que para esse fim. Outros atributos

foram derivados, como é o caso do atributo *idade*, criado a partir do atributo *Data_nascimento*. Na Tabela 8 encontram-se as consultas realizadas para induzir novos atributos.

Os registros com alunos que não obtiveram nenhuma interação foram desconsiderados. Ruídos causados por rotinas administrativas no sistema também foram removidos.

Tabela 7: Consultas realizadas para obter a integração dos dados.

Consulta utilizando a linguagem SQL	**Ação realizada**
CREATE VIEW TabelaDinamica AS SELECT @atributos FROM mdl_log WHERE mdl_userid IN (SELECT mdl_role_assigments.id FROM mdl_role_assigments WHERE mdl_role_assignments`.`roleid` = @Alunos)	-- criando tabela dinâmica com os atributos para o servidor local.
CREATE TABLE dados (@atributosEmTabela);	-- criando tabela para importar dados manuais do SGE.
LOAD DATA INFILE 'TabelaDoExcel.csv' INTO TABLE dados FIELDS TERMINATED BY ';' ENCLOSED BY '' LINES TERMINATED BY ',';	-- importando TabelaDoExcel.csv com separação de atributos por ponto vírgula, e separação de tuplas por vírgula.
CREATE VIEW TabelaDinamica AS SELECT * FROM (TabelaDinamica INNER JOIN dados ON TabelaDinamica.id = dados.id_user)	-- JOIN entre a tabela dinâmica criada, e os dados importados do SGE, para unificação dos dados.

Fonte: Próprio autor.

Tabela 8: Consultas realizadas para induzir novos atributos.

Consulta utilizando a linguagem SQL	Ação realizada
(SELECT @tabelaGerada . @totalDeOcorrencias FROM (SELECT COUNT(*) AS @TotalDeOcorrencias, userid FROM mdl_log WHERE action = '@atributo' GROUP BY userid) AS @tabelaGerada WHERE (mdl_user.id = @tabelaGerada.userid)) AS @atributo	-- contando ocorrências de cada um dos atributos da tabela gerada.

Fonte: Próprio autor.

Quarto passo: Escolha do escopo e seleção das variáveis

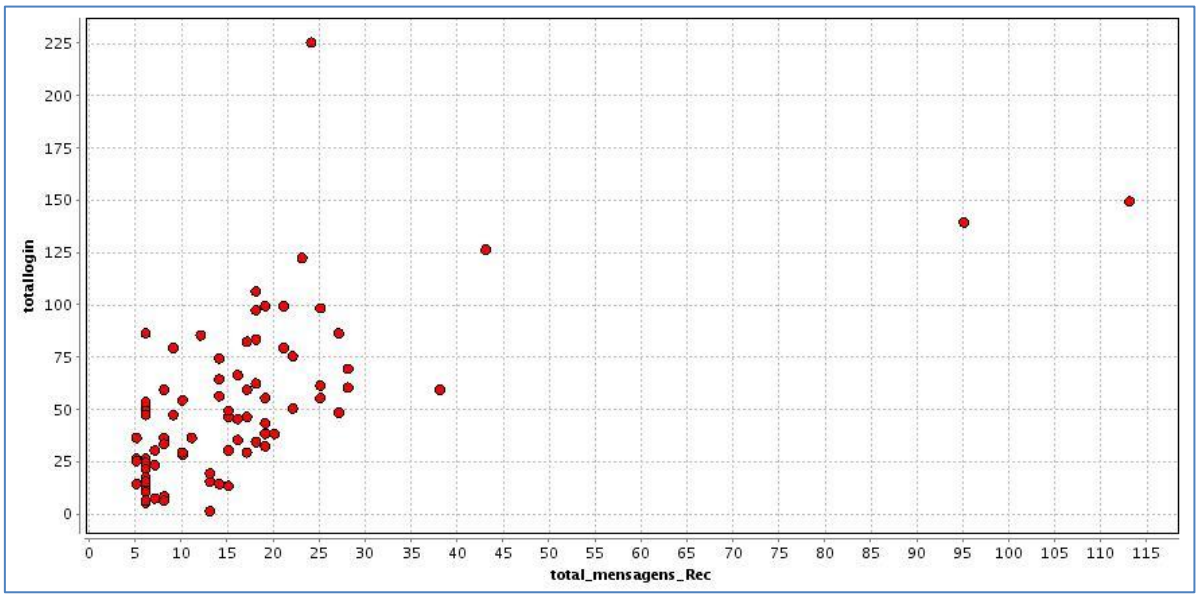
A granularidade escolhida foi no nível de cada estudante, considerando o perfil de acesso individual, ao considerar que o atributo a ser inferido será para cada novo estudante que entrar no sistema. Os atributos selecionados para caracterizar o estudante foram descritos na tabela 6. Nessa etapa foram envolvidos o responsável pela aplicação (especialista no Ambiente Virtual do SENAI) e o responsável pelo processo da Mineração de Dados (especialista em Mineração de Dados).

Para minerar os dados foram selecionados um total de 14 atributos, visto que alguns deles foram agrupados em novos atributos ou foram desconsiderados após interações posteriores – como exemplo o total de interações com a ferramenta chat, desconsiderado por não haver quantidade de interações relevantes no contexto da turma, justificado por serem trabalhadores em geral e não possuíam horários disponíveis para interações síncronas. Os dados foram organizados em uma tabela única que permitisse uma visão integrada dos registros, em que cada linha representasse a totalidade das informações dos alunos e as colunas apresentassem cada um de seus atributos.

Outro ponto observável foi a relação estreita entre a quantidade de mensagens enviadas para o aluno com a quantidade acessos ao Ambiente Virtual

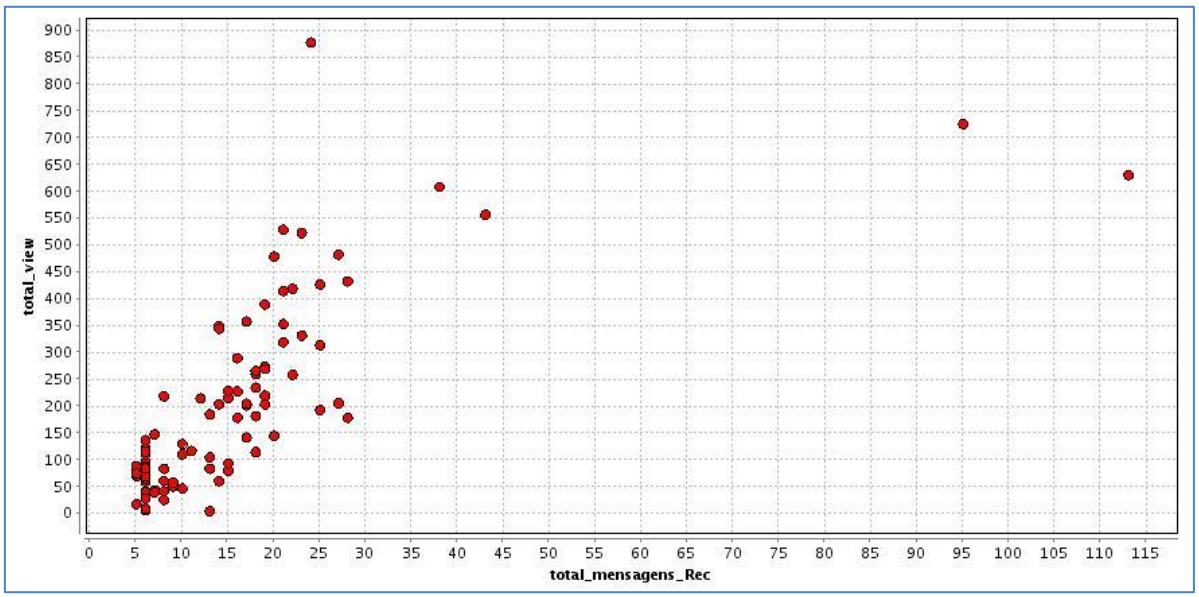
(Figura 19), reforçando ainda mais a metodologia de acompanhamento que o SENAI vêm empregando em seus cursos. Analogamente, interagir constantemente com o aluno pode motivá-lo a visualizar o material didático (Figura 20).

Figura 19: Relação entre Quantidade de Mensagens e Número de acessos.



Fonte: *Print screen* da aplicação RapidMiner®.

Figura 20: Relação entre Quantidade de Mensagens e Número de visualização do material didático.



Fonte: *Print screen* da aplicação RapidMiner®.

Quinto passo: escolha do método para minerar dados

O objetivo proposto para esse experimento focou principalmente na predição de uma variável categórica e binária, o atributo "*status* do aluno", podendo gerar registros como "aprovado" ou "reprovado" (por evasão ou por não cumprimento das métricas avaliativas estabelecidas). Para esse propósito, a literatura considera viável o método de predição, de forma a inferir o atributo categórico.

Sexto passo: escolha da técnica e do algoritmo

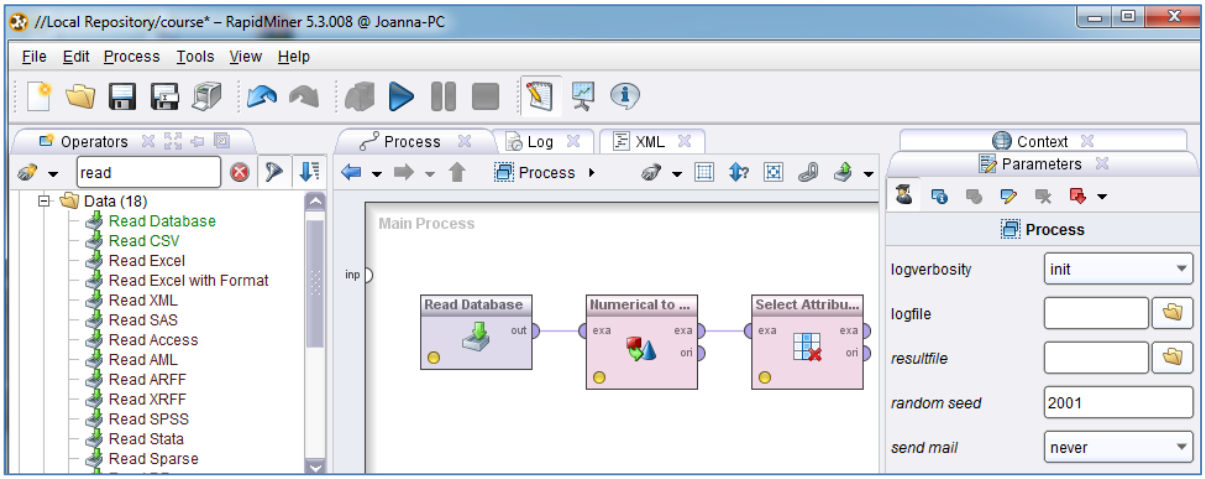
No estudo de caso desenvolvido, o algoritmo de mineração de dados empregado foi o C45/J48, preferidos por estarem bem documentados no estudo dos relatos encontrados na literatura, disporem de ferramentas e implementações acessíveis (CONTI, 2011) (KAMPFF, 2008). O algoritmo J48 faz parte da ferramenta WEKA, e é derivado dos algoritmos ID3 e C4.5 (QUINLAN, 1993), sendo estes fundamentados na descoberta do conhecimento pela estrutura de uma árvore de decisão. No Capítulo 3 o algoritmo é abordado com mais detalhes.

Preferiu-se utilizar a ferramenta RapidMiner® versão 5.3, justificada pela sua interface gráfica intuitiva e fácil de utilizar, além de possibilitar um desenvolvimento incremental durante o processo. Outro fator relevante foi a familiaridade da equipe de pesquisa com a ferramenta. Para utilizar o algoritmo J48 no RapidMiner® se fez necessário instalar o pacote de expansão WEKA.

No RapidMiner® todos os operadores são criados por meio de uma interface amigável (Figura 21). Basta selecionar o operador no menu lateral esquerdo, e arrastá-lo para a área de trabalho principal. Para a leitura, essa ferramenta oferece suporte para carregar dados de diferentes formas (conexão com banco de dados, arquivo .csv, planilha Excel).

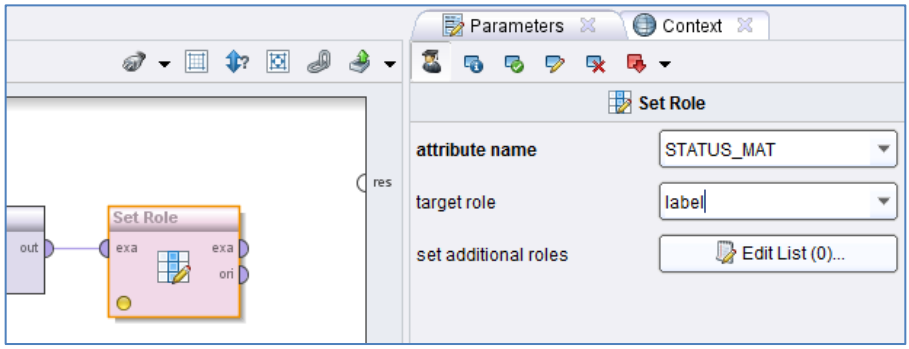
Após carregar os dados no sistema, realizou-se as transformações necessárias nos dados e a seleção dos atributos utilizados na predição inserindo e configurando operadores disponíveis no RapidMiner®. Determinou-se o atributo Status_Matrícula como variável alvo da predição, mediante o operador *SetRole* (Figura 22).

Figura 21: Operador de Leitura do RapidMiner®.



Fonte: *Print screen* da aplicação.

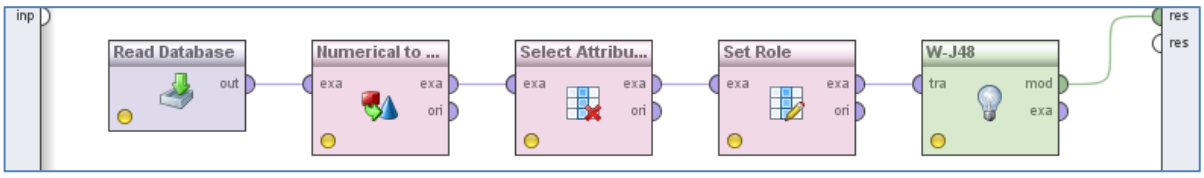
Figura 22: Operador SetRole do RapidMiner®.



Fonte: *Print screen* da aplicação.

O próximo operador inserido foi o W-J48, responsável por implementar a Arvore de decisão. Nesse momento a disposição dos operadores encontraram-se conforme Figura 23.

Figura 23: Disposição de Operadores no RapidMiner®.



Fonte: *Print screen* da aplicação.

As Figuras 24 e 25 demonstram os resultados desse experimento, apresentando uma representação textual dos resultados e a Arvore de Decisão gerada pelo algoritmo, respectivamente.

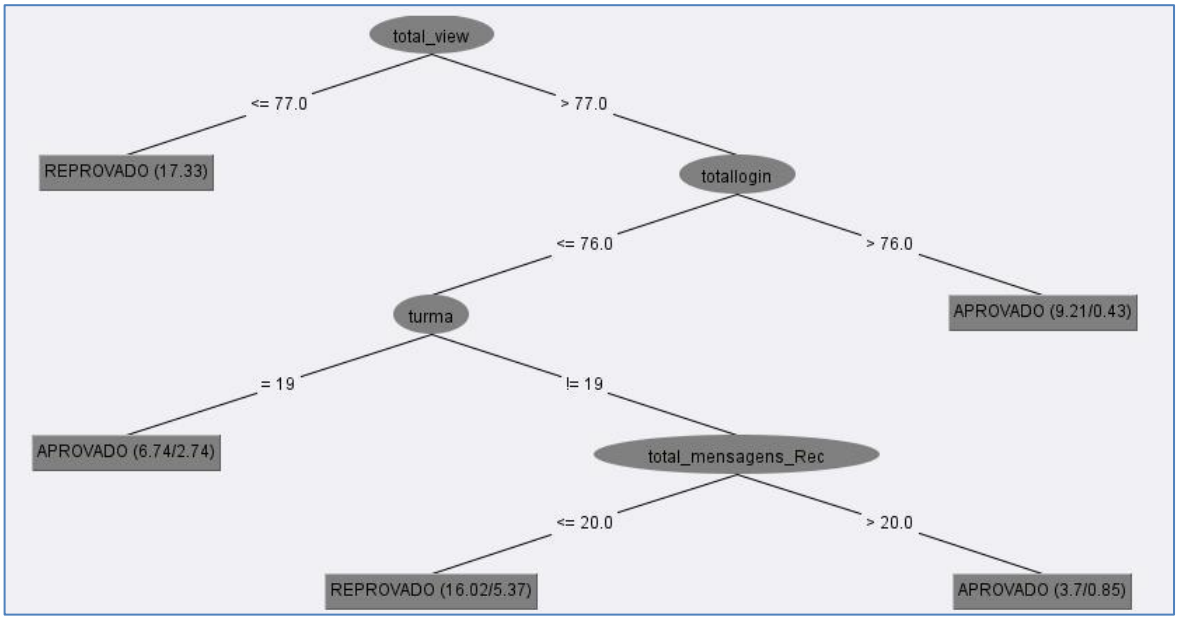
Figura 24: Regra de decisão resultada da experimentação.

W-J48

```
J48 pruned tree
------------------

total_view <= 77.0: REPROVADO (17.33)
total_view > 77.0
|   totallogin <= 76.0
|   |   turma = 19: APROVADO (6.74/2.74)
|   |   turma != 19
|   |   |   total_mensagens_Rec <= 20.0: REPROVADO (16.02/5.37)
|   |   |   total_mensagens_Rec > 20.0: APROVADO (3.7/0.85)
|   totallogin > 76.0: APROVADO (9.21/0.43)

Number of Leaves  :     5

Size of the tree :      9
```

Fonte: *Print screen* da aplicação RapidMiner®.

Figura 25: Arvore de Decisão resultada da experimentação.



Fonte: *Print screen* da aplicação RapidMiner®.

Sétimo passo – Interpretação dos resultados

A árvore de decisão obtida apontou para uma forte contribuição dos atributos: total_view (número de visualizações do Material Didático) e total_login (quantidade de login realizado no AVA durante o período do curso) como aspectos importantes para prever o status final do aluno na turma. Dois grandes indicadores nas regras classificaram os alunos como possíveis reprovados: baixa frequência de acesso ao sistema, e pouca visualização do material didático disponível. Sugere-se então criar abordagens cada vez mais eficazes para envolver os alunos em seus estudos, de forma que participem das discussões da turma, enviando mensagens motivacionais e que sugiram a importância das leituras complementares e a participação nos fóruns de discussão. Os demais atributos não obtiveram relevância na construção do modelo.

Oitavo passo – Escolher parâmetros de validação

Alguns parâmetros foram utilizados para validar os dados obtidos, segundo critérios definidos por Monard e Baranauskas (2003):

Análise de Concordância (Kappa): a concordância obtida foi de 0.459, valor interpretado por Landis e Koch (1977) como uma concordância moderada. É importante ressaltar que o coeficiente Kappa é influenciado pelo conjunto de dados utilizado. No contexto da Mineração de Dados Educacionais a concordância dificilmente será analisada por estar entre um “número mágico”, pois os alunos podem agir de forma própria, sem nenhum padrão de comportamento, dentro do mesmo Ambiente Virtual de Aprendizagem, influenciando no valor final desse coeficiente.

Erro do Classificador: o Erro do Classificador para o dado conjunto de teste foi de 27.85%, indicando a média do afastamento de todos os valores fornecidos pelos classificadores e o seu real valor, denotada por err(h).

Precisão (accuracy): A precisão obtida foi de 72.15%. Calculada pelo complemento da taxa de erro, a precisão do classificador acc(h) é dada por (1 – err(h)).

Validação Cruzada – Para estimar o desempenho futuro do modelo utilizou-se a técnica da Validação Cruzada (definida no Capítulo 5), amplamente empregada na literatura. O método de particionamento utilizado foi o *leave-one-out*, devido a baixa quantidade de amostras disponíveis. A quantidade de amostras também influenciou no custo computacional aceitável para a aplicação desse método.

Nono passo – Usar o conhecimento adquirido

O modelo de Predição fruto desse experimento, visando novos alunos que estejam propensos a evasão, apresentou uma confiabilidade menor que a esperada devido a baixa quantidade de amostras utilizadas. Porém essa técnica auxiliará no desenvolvimento de um sistema inteligente para uso em atividades instrucionais do SENAI, aplicado para modelar os alunos tendenciosos à evasão logo nos primeiros acessos, organizando e filtrando informações úteis para o educador traçar suas próprias conclusões. Uma maior quantidade de amostras, referente aos alunos, melhor induzir o modelo, gerando conclusões mais precisas sobre o aspecto comportamental dos exemplos. Um futuro sistema de alerta pode notificar o professor sobre alunos com tendência a reprovação, e sugerir que seja feito contato com os estudantes.

7 Conclusões

Essa pesquisa buscou uma solução capaz de predizer quando um aluno possui características tendenciosas para a evasão, nos cursos de Modalidade a Distância do SENAI, dado o registro das suas características sociais e sua iteração no Ambiente Virtual de Aprendizagem. Além disso, evidenciou os fatores intermediários que influenciam na evasão, como, por exemplo, a quantidade de mensagens direcionadas ao aluno.

Durante o estudo foi possível observar a relação existente entre o total de mensagens enviadas para o aluno, com a quantidade de visualização do material didático e login realizado no AVA. Esses resultados demonstram a importância da iteração constante entre aluno e professor em um ambiente virtual de ensino.

Devido o baixo número de alunos contidos no experimento, a hipótese gerou resultados muito específicos para o conjunto de dados utilizado, sendo pouco confiável deduzir o comportamento de novos alunos para outros contextos com o modelo gerado. Melhores resultados podem ser gerados aumentando o total de exemplos no estudo.

Os resultados limitam-se também a considerar só o quantitativo das iterações e as avaliações numéricas de aprendizagem dos alunos, tornando incompleto o diagnóstico do ponto de vista de uma análise pedagógica. Para aprimorar esse aspecto, é possível considerar a qualidade das contribuições dadas pelos alunos por sua relevância com o assunto abordado no curso, não sendo esse o foco do nosso estudo de caso.

Destacam-se como contribuições desse trabalho o estudo das características inerentes aos alunos do Ambiente Virtual do SENAI, e o domínio metodológico concebido pelos desenvolvedores do SENAI, na utilização do RapidMiner® como ferramenta para Mineração de dados para uma posterior integração do modelo gerado ao AVA da instituição, possibilitando predizer o valor de alguma variável.

Apresentam-se alguns desafios futuros:

Criar um sistema que analise os dados em uma frequência diária, a fim de identificar os alunos que estejam propensos a desistir do curso, ou que não possuem condições para cursar uma turma a distância, enquanto o número de faltas for inferior a 25% do total do curso, possibilitando a intervenção do educador em tempo hábil para evitar a reprovação, com antecedência suficiente para que sejam tomadas as medidas necessárias.

Proporcionar uma ferramenta adequada ao uso desse contexto, com necessidade de maior simplicidade, com interface intuitiva e amigável, permitindo que educadores leigos possam utilizá-la com facilidade.

Oferecer uma análise qualitativa das interações dos alunos, de forma a não deixar de lado a qualidade no aprendizado.

Referências Bibliográficas

BAKER, R.S.J.d. (2010) Data Mining for Education. In McGaw, B., Peterson, P., Baker, E. (Eds.) International Encyclopedia of Education (3rd edition), vol. 7, pp. 112-118. Oxford, UK: Elsevier.

BAKER, R.S.J.d., ISOTANI, S., de CARVALHO, A. (2011) Mineração de Dados Educacionais: Oportunidades para o Brasil. Revista Brasileira de Informática na Educação, 19 (2), 3-13.

BAKER, R.S.J.d., MOORE, G., WAGNER, A., KALKA, J., KARABINOS, M., ASHE, C., YARON, D. (2011b) The Dynamics Between Student Affect and Behavior Occuring Outside of Educational Software. Proceedings of the 4th bi-annual International Conference on Affective Computing and Intelligent Interaction.

Baltar, V. T., OKANO, V., Laboratório de Epidemiologia e Estatística - Análise de Concordância - Kappa. 2005; Disponível em: < http://www.lee.dante.br>. Acesso em: 01 de abr. 2014.

BOGDAN, R., BIKLEN, S. - Características da investigação qualitativa. In: Investigação qualitativa em educação: uma introdução à teoria e aos métodos. Porto, Porto Editora, p.47-51, 1994.

CALDAS, Maria Aparecida Esteves. Estudos de revisão da literatura: fundamentação e estratégia metodológica. São Paulo: Hucitec; Brasília: Instituto Nacional do Livro, 1986. 69p.

CAMILO, Cássio Oliveira; SILVA, João Carlos da; Mineração de Dados: conceitos, tarefas, métodos e ferramentas. Goiânia: Instituto de Informática/UFG, 2009.

CARDIERI, M. A. C. d. A. Mecanismo de monitoramento do uso de recursos web para apoio à avaliação de ambientes. Dissertação de Mestrado, Faculdade de Engenharia Elétrica e de Computação — UNICAMP, 2004.

CensoEAD (2013) CensoEAD.BR:2012. Relatório analítico da aprendizagem a distância no Brasil. ABED – São Paulo: Pearson Education do Brasil, 2013.

COLE, J.; Foster H.; Using Moodle: Teaching with the Popular Open Source Course Management System. Second Edition. O'Reilly Community Press: Printed in the United States of America, November 2007.

COSTA, E, Baker, R. S.J.D., Amorim, L., Magalhães, J., Marinho, T. (2012) Mineração de Dados Educacionais: conceitos, técnicas, ferramentas e aplicações. Jornada de Atualização em informática na Educação – JAIE.

CUSTÓDIO, C. A. Avaliação da Usabilidade do Ambiente de Ensino à distância Moodle sob a Perspectiva de Professores. Dissertação de Mestrado, UNIMEP, Piracicaba, 2008.

CONTI, F., Charao, A. S.; Análise de Prazos de Entrega de Atividades no Moodle: um Estudo de Caso Utilizando Mineração de Dados. RENOTE. Revista Novas Tecnologias na Educação. v. 9, p. 1-10, 2011.

FAYYAD, Usama; Piatetski-Shapiro, Gregory; Smyth, Padhraic (1996) The KDD Process for Extracting Useful Knowledge from Volumes of Data. In: Communications of the ACM, pp.27-34, Nov. 1996.

FALCI Jr, G. R., Ricarte, I. L. M. (2009) Metodologias de Mineração de Dados aplicadas a Ambientes Educacionais Online, 4. In: Encontro dos Alunos e Docentes do Departamento de Engenharia de Computação e Automação Industrial, 2009, Campinas. Anais. 2009. p. 89-92.

GARRISON, Randy, ANDERSON, Terry (2003). eLearning in the 21st Century: A Framework for Research and Practice. London & New York: RoutledgeFalmer.

GODWIN, K.E., ALMEDA, M. V., PETROCCIA, M., BAKER, R.S., & FISHER, A.V. (2013). Classroom activities and off-task behavior in elementary school children. In M. Knauff, M. Pauen, N. Sebanz, and I. Wachsmuth (Eds.), Proceedings of the 35th Annual Meeting of the Cognitive Science Society, 2428-2433

GOTTARDO, E. ; KAESTNER, C ; NORONHA, V. R. . Previsão de Desempenho de Estudantes em Cursos EAD Utilizando Mineração de Dados: uma Estratégia Baseada em Séries Temporais.. In: 23º Simpósio Brasileiro de Informática na Educação (SBIE), 2012, Rio de Janeiro, RJ. Congresso Brasileiro de Informática da Educação, 2012.

HAN, J., KAMBER, M., PEI, J. Data Mining: Concepts and Techniques. Morgan Kaufmann Publishers Inc., San Francisco, CA, USA, 2011.

H. MANNILA. Data mining: machine learning, statistics, and databases. In Proceedings of the 8th International Conference on Scientific and Statistical Database Management, Stockholm, pages 1–6, 1996.

KAMPFF, A. J. C.; Reategui, E.; Lima, J. V.; Mineração de dados educacionais para a construção de alertas em ambientes virtuais de aprendizagem, como apoio a prática docente. RENOTE. Revista Novas Tecnologias na Educação, v. 6, p. 1-9, 2008.

KOCK Jr., N. F.; MCQUEEN, R. J.; BAKER, M. Learning and process improvement in knowledge organizations: A critical analysis of four contemporary myths. The Learning Organization, 1996. p. 31–40.

KOHAVI, R., SOMMERFIELD, D., DOUGHERTY, J., Data Mining using MLC++, a Machine Learning Library in C++. International Journal of Artificial Intelligence Tools, Vol. 6, No. 4, 1997, p. 537-566.

LANDIS, J.R., 81 KOCH, G.G. The measurement of observer agreement for categorical data. Biometrics, 33, 1977, p. 159-174.

MACHADO, A. P., FERREIRA, R., BITTENCOURT, I. I., ELIAS, E., BRITO, P., COSTA, E. B.; Mineração de texto em Redes Sociais aplicada à Educação a Distância. Colabor@ (Curitiba), v. 6, p. 132, 2010.

MANHÃES, L. M. B., Cruz, S. M. S., Macário Costa, R. J., Zavaleta, J., Zimbrao, G.; Previsão de Estudantes com Risco de Evasão Utilizando Técnicas de Mineração de Dados. In: 22º Simpósio Brasileiro de Informática na Educação - SBIE2011, 2011, Aracajú - Sergipe.

MCCUE, C., Data Mining and Predictive Analysis - Intelligence Gathering and Crime Analysis. Elsevier, 2007.

MONARD, M. C., BARANAUSKAS, J. A., Conceitos Sobre Aprendizado de Máquina. In: Solange O. Rezende. (Org.). Sistemas Inteligentes -- Fundamentos e Aplicações. 1 ed. Barueri-SP: Editora Manole Ltda, 2003, v. 1, p. 89-114.

MOODLE Brasil. Ambiente de Aprendizagem Moodle Brasil. Disponível em: http://www.moodlebrasil.net/moodle/. Acesso em: 08 mai. 2007.

MOORE, M. G. (1989). Three types of interaction. American Journal of Distance Education, 3(2), 1-7.

NAVEGA, S. Princípios essenciais do data mining. In: INFOIMAGEM. 2002. Anais. Cenadem, nov. 2002. Disponível em: <www.intelliwise.com/snavega>. Acesso em: 03 mar. 2014.

PARDOS, Z., BAKER, R.S.J.d., SAN PEDRO, M.O.Z., Gowda, S.M., and Gowda, S. 2013. Affective states and state tests: Investigating how affect throughout the school year predicts end of year learning outcomes. In Proceedings of the 3rd International Conference on Learning Analytics and Knowledge, 117-124.

PAIVA, R. O. A. ; BITTENCOURT, I. I. ; SILVA, A. P. . Uma Ferramenta de Autoria para Recomendação Pedagógica Baseada em Mineração de Dados Educacionais. In: Congresso Brasileiro de Informática na Educação, 2013, Campinas. Anais dos Workshops do Congresso Brasileiro de Informática na Educação, 2013.

POZZER, C. T. Aprendizado por Árvores de Decisão: Disciplina de Programação de Jogos 3D. Notas de Aula. UFSM, 2006.

QUINLAN, J. R. C4.5: Programs for Machine Learning. Morgan Kaufmann Publishers, 1993.

REZENDE, S. O. ; PUGLIESI, J. B. ; MELANDA, E. A. ; PAULA, M. F. . Mineração de Dados. In: Solange Oliveira Rezende. (Org.). Sistemas Inteligentes -- Fundamentos e Aplicações. 1 ed. Barueri, SP: Editora Manole Ltda, 2003, v. 1, p. 307-336.

ROMERO, C., VENTURA S., GARCÍA E., Data mining in course management systems: Moodle case study and tutorial. Computers & Education, 51(1): 368–384, 2008.

ROMERO, C., VENTURA, S., Educational Data Mining: A Review of the State-of-the-Art. IEEE Transaction on Systems, Man and Cybernetics, Part C: Applications and Reviews. 40(6), 601-618, 2010.

SABINO, R. F. ; ROCHA, F. G. ; GUALTER, Atanásio Júnior . A educação a distância no SENAI: do Telecurso ao PRONATEC. In: V Colóquio Internacional Educação e Contemporaneidade, 2011, São Cristóvão. São Cristóvão: Universidade Federal de Sergipe, 2011. v. 1.

SAO PEDRO, M.A., Baker, R.S.J.d., Gobert, J., Montalvo, O. Nakama, A. (2013) Leveraging Machine-Learned Detectors of Systematic Inquiry Behavior to Estimate and Predict Transfer of Inquiry Skill. To appear in User Modeling and User-Adapted Interaction. 23: 1-39 Mar. 2013.

SENECHAL, A. C. L., Análise e Pré-processamento de Dados Utilizando Técnicas de Mineração de Dados para o Moodle. 2013. Trabalho de Conclusão de Curso. (Graduação em Ciência da Computação) - Universidade Federal de Lavras. Orientador: Eric Fernandes de Mello Araújo.

Serviço Nacional de Aprendizagem Industrial. Departamento Nacional. Modelo de execução do programa nacional de educação a distância (PN-EAD) : módulo 1 / Serviço Nacional de Aprendizagem Industrial. Departamento Nacional, Serviço Nacional de Aprendizagem Industrial. Departamento Regional de Santa Catarina. Brasília : SENAI/DN, 2013.

SIQUARA, E.C. ; BRAGA, P. F. ; ALMEIDA, F. B. C. . AVAS: Uma solução para gestão de EAD baseada na integração de instalações moodle. In: 18º CIAED - Congresso Internacional ABED de Educação à Distância, 2012, São Luiz.

VERGARA, S. C, Métodos de pesquisa em administração. São Paulo: Atlas, 2005.