UNIVERSIDADE DO VALE DO ITAJAÍ
PROGRAMA DE MESTRADO ACADÊMICO EM COMPUTAÇÃO APLICADA

ARQUELAU PASTA

APLICAÇÃO DA TÉCNICA DE DATA MINING NA BASE DE DADOS DO AMBIENTE DE GESTÃO EDUCACIONAL: UM ESTUDO DE CASO DE UMA INSTITUIÇÃO DE ENSINO SUPERIOR DE BLUMENAU-SC.

DISSERTAÇÃO DE MESTRADO

São José (SC), Março de 2011

UNIVERSIDADE DO VALE DO ITAJAÍ
CURSO DE MESTRADO ACADÊMICO EM COMPUTAÇÃO APLICADA

ARQUELAU PASTA

por

Arquelau Pasta

Dissertação apresentada como requisito parcial à obtenção do grau de Mestre em Computação Aplicada.
Orientador: Prof. Raimundo Celeste Ghizoni Teive, Dr.

São José (SC), Março de 2011.

FOLHA DE APROVAÇÃO

Esta página é reservada para inclusão da folha de assinaturas, a ser disponibilizada pela Secretaria do Curso para coleta da assinatura no ato da defesa.

Dedico este trabalho a duas pessoas que durante este período de ausência, entenderam que esta ausência era necessária.
Para vocês: Silvana, minha esposa e Nadine, minha filha.
Dedico também ao meu pai, que não teve tempo de esperar.

A UM AUSENTE

Tenho razão de sentir saudade,
tenho razão de te acusar.
Houve um pacto implícito que rompeste
e sem te despedires foste embora.
Detonaste o pacto.
Detonaste a vida geral, a comum aquiescência
de viver e explorar os rumos de obscuridade
sem prazo sem consulta sem provocação
até o limite das folhas caídas na hora de cair.

Antecipaste a hora.
Teu ponteiro enlouqueceu, enlouquecendo nossas horas.
Que poderias ter feito de mais grave
do que o ato sem continuação, o ato em si,
o ato que não ousamos nem sabemos ousar
porque depois dele não há nada?

Tenho razão para sentir saudade de ti,
de nossa convivência em falas camaradas,
simples apertar de mãos, nem isso, voz
modulando sílabas conhecidas e banais
que eram sempre certeza e segurança.

Sim, tenho saudades.
Sim, acuso-te porque fizeste
o não previsto nas leis da amizade e da natureza
nem nos deixaste sequer o direito de indagar
porque o fizeste, porque te foste.

CARLOS DRUMMOND DE ANDRADE

AGRADECIMENTOS

Muitos fizeram, muitos fazem e muitos irão fazer parte das várias etapas de minha vida. Alguns contribuíram de forma peculiar, outros às vezes com um simples: “Legal, vamos lá”. Quero agradecer a todos, sem cometer a injustiça de esquecer alguém.

Iniciando por meus pais, Dona Nair e Seu Orlando (In Memoriam) afinal sem eles não estaria escrevendo isto.

Minha esposa Silvana e filha Nadine, que tiveram vários momentos de nosso convívio destinados a elaboração deste trabalho.

À minha família, pelo incentivo nos momentos difíceis, fazendo com que eu seguisse em frente. À todos sem distinção, irmão, cunhados, cunhadas, sobrinhos e sobrinhas.

Aos meus amigos de mestrado, Vital e Pedro, pelas idas e vindas, pelas conversas e trabalhos trocados.

A todas as pessoas ligadas ao Programa de Pós-Graduação em Computação Aplicada da Universidade do Vale do Itajaí (UNIVALI), professores, coordenação. Em especial a nossa sempre atenciosa e prestativa Maria de Lurdes, pelos muitos documentos emitidos e enviados.

Ao meu orientador Professor Dr. Raimundo Celeste Ghizoni Teive pela sua amizade e apoio, por sua vocação inequívoca, pelo seu espírito inovador, intelectual e empreendedor na tarefa de multiplicar seus conhecimentos, por ser um verdadeiro mestre.

A todos os professores e aos profissionais do Instituto Blumenauense de Ensino Superior (IBES). Pelo apoio, incentivo e presteza no auxílio às atividades e discussões sobre o trabalho apresentado.

E por fim à Deus, o grande criador, pela oportunidades e privilégios a mim concedidos.

APLICAÇÃO DA TÉCNICA DE DATA MINING NA BASE DE DADOS DO AMBIENTE DE GESTÃO EDUCACIONAL: UM ESTUDO DE CASO DE UMA INSTITUIÇÃO DE ENSINO SUPERIOR DE BLUMENAU-SC.

Arquelau Pasta

Março / 2011

Orientador: Raimundo Celeste Ghizoni Teive, Dr.

Área de Concentração: Computação Aplicada

Linha de Pesquisa: Inteligência Aplicada

Palavras-chave: Mineração de Dados. Gestão da Informação. Gestão de Instituições de Ensino Superior.

Número de páginas: 153

RESUMO

Para que o conhecimento seja gerado não basta apenas ter a informação. As Instituições de Ensino Superior (IES) podem hoje serem consideradas como organizações. Uma das funções das Instituições de Ensino é a geração e disseminação de conhecimento, obtido através do processo de ensino e aprendizagem e para que este processo aconteça numa forma dinâmica e eficaz, as IES estão cada vez mais buscando subsídios, ferramentas e técnicas, para que seus alunos adquiram o conhecimento. Conseqüentemente, todo este conhecimento acumulado pode e deve ser utilizado para que cada vez mais as instituições busquem estreitar o contato com seus alunos e disponibilizar aos seus gestores, informações precisas e eficazes para tomada de decisões. A contribuição desta dissertação de mestrado refere-se a aplicação de técnicas de Data Mining em ambientes de gestão educacional. Para tanto foram aplicadas as técnicas de Associação, Classificação e Clusterização nesta base de dados. A pesquisa aborda por meio de um levantamento bibliográfico os conceitos sobre Gestão da Informação, Sistemas de Informação, Data Warehouse, Data Mining com suas técnicas e tarefas, finalizando com a ferramenta de mineração WEKA. A aplicação das técnicas de Data Mining, segue a metodologia CRISP-DM, na qual são observados desde o conhecimento sobre o negócio até a implementação dos resultados. Dessa forma, como um dos resultados obtidos na pesquisa, viu-se que a aplicação de uma ferramenta de Data Mining pode ser um poderoso instrumento para a gestão das informações nas IES.

APPLICATION OF DATA MINING TECHNIQUE IN THE DATABASE OF THE EDUCATIONAL MANAGEMENT ENVIRONMENT: CASE STUDY OF A HIGHER EDUCATION INSTITUTE IN BLUMENAU-SC.

Arquelau Pasta

March / 2011

Supervisor: Raimundo Celeste Ghizoni Teive, Dr

Area of Concentration: Applied Computer Science

Line of Research: Applied Intelligence

Key words: Data Mining. Information Management. Management of Institutes of Higher Education.

Number of pages: 153

ABSTRACT

For knowledge be generated, just having information is not enough. Institutions of Higher Education (IES) nowadays may be considered as organizations. One of the functions of Teaching Institutions is to generate and disseminate the knowledge obtained through the teaching and learning process, and to enable this process to occur in a dynamic and effective form, IESs are increasingly looking for support, tools and techniques that will enable their students to acquire knowledge. Consequently, all this accumulated knowledge can and should be used to enable institutions to form closer contact with their students and provide their managers with accurate and effective information for decision-making. The contribution of this master's degree dissertation is the application of data mining techniques in educational management environments. The techniques of Association, Classification and Clusterization were used in this database. The research uses bibliographical research to search for concepts on Information Management, Information Systems, Data Warehouse, and Data Mining, with their techniques and tasks, concluding with the mining tool WEKA. The application of Data Mining techniques follows the CRISP-DM methodology, taking into account from business knowledge through to the implementation of the results. Thus, one of the results obtained in the research was that the application of a Data Mining tool can be a powerful tool for managing information in the IES.

LISTA DE ILUSTRAÇÕES

LISTA DE TABELAS E QUADROS

LISTA DE GRÁFICOS E EQUAÇÕES

LISTA DE ABREVIATURAS E SIGLAS

ACP	Análise de Componentes Principais
AD	Árvore de Decisões
AG	Algoritmos Genéticos
AGE	Ambiente Gestão Educacional
ANS	Aprendizagem Não Supervisionada
AS	Aprendizagem Supervisionada
ARFF	Extensão do arquivo utilizado pelo WEKA
AVA	Ambiente Virtual de Aprendizagem
DRA	Descoberta de Regras de Associação
EAD	Educação à Distância
EEP	Empregado de Empresa Privada
EI	Extração da Informação
FNQ	Fundação Nacional da Qualidade
FPU	Funcionário Público
IES	Instituição de Ensino Superior
KDD	Knowledge Discovery in Database
MCA	Mestrado em Computação Aplicada
MD	Mineração de Dados
NFA	Negócio Familiar
NPR	Negócio Próprio
NTR	Não Trabalha
OUT	Outro
RBC	Raciocínio Baseado em Casos
RNA	Redes Neurais Artificiais
SEI	Sistema de Extração da Informação
SI	Sistemas de Informação
SIE	Sistemas de Informação para Executivos
SIG	Sistemas de Informação Gerencial
SM	Salário Mínimo
SPT	Sistemas de Processamento de Transações
SSTD	Sistemas de Suporte a Tomada de Decisão
TI	Tecnologia da Informação
TIC	Tecnologia da Informação e Comunicação
UNIVALI	Universidade do Vale do Itajaí
WEKA	Waikato Environment for Knowledge Analysis

SUMÁRIO

1 INTRODUÇÃO

Num ambiente em que a velocidade das mudanças e a necessidade de adequação as estas é cada vez maior, a análise de informações em grandes bases de dados, torna-se um processo que exige o uso de técnicas e ferramentas que tornem a atividade de coleta, análise e utilização das informações, mais ágil e confiável.

Lucas (2002, p. 13) acredita que:

> A transformação da informação em conhecimento pode fazer com que as organizações sobrevivam neste mercado globalizado, pois esta transformação fornecerá informações que após serem analisadas de forma correta possam ser utilizadas para tomada de decisões mais seguras, aliadas a adequação da postura estratégica da organização, na qual o conhecimento passa a fazer parte, antevendo as mudanças pelas quais a organização a de passar em função da competitividade do mercado.

A partir do crescimento do volume de informações que as corporações manipulam, gera-se a necessidade urgente de técnicas e ferramentas que transformem dados em conhecimento útil de forma inteligente e automática. A solução para esta necessidade das organizações de obterem conhecimento de grandes volumes de dados está na utilização de técnicas de mineração de dados para extrair as informações implícitas existentes nos Bancos de Dados destas organizações.

Dalfovo (2007, p. 57) define que a utilização da informação de forma eficaz e eficiente, torna-se um elemento primordial para o sucesso das organizações, sendo incorporado inclusive em seu patrimônio. O saber que a informação é um dos principais recursos estratégicos que a organização dispõe, requer que estas informações estejam estruturadas, disponíveis e sejam íntegras, condições estas que se fazem possível somente com o uso de tecnologias computacionais, comumente designadas de Tecnologia da Informação e Comunicação (TIC), ou Sistemas de Informação (SI).

O grande desafio das organizações é estruturar e disponibilizar para seus gestores, as informações geradas por elas mesmas, e que estes utilizem estas informações como recurso estratégico, objetivando a obtenção de vantagem competitiva sustentável. A Figura 1 demonstra como a informação encontra-se disponibilizada na maior parte dos bancos de dados, desta forma a tarefa de filtrar a informação é dificultosa.

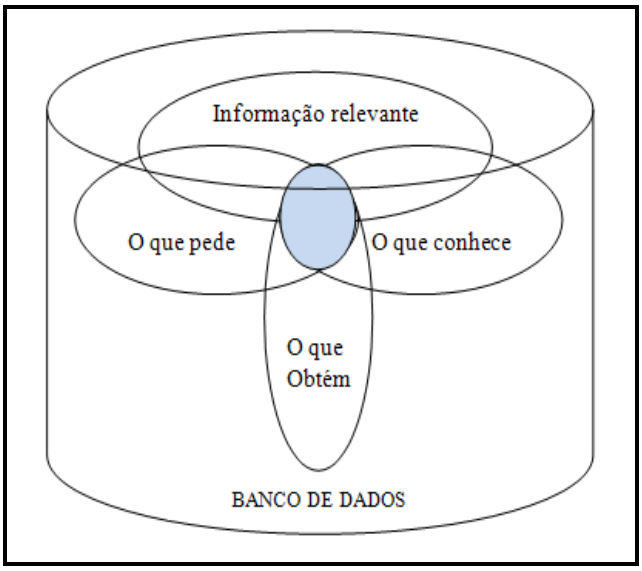


Figura 1 - Distribuição da informação
Fonte: Adaptado de Aguilar (apud STAREC 2005, p. 50)

Também Foguel e Souza (1993 apud Maccari, 2002, p.20) ao analisarem a os diversos setores econômicos, relatam que:

> [...]a Universidade, como instituição, está inserida na era organizacional. Como as demais organizações, atingiu, ao longo do tempo, um grau de complexidade significativo, obrigando os seus administradores a rever suas funções e apresentar propostas para acelerar o seu desenvolvimento.

Uma enorme mudança tem sido observada a partir da última versão da Lei de Diretrizes e Bases da Educação (LDB, 1996). Pode-se observar que o setor educacional passou a ser visto como uma grande oportunidade de negócios para os empreendedores. Isto pode ser confirmado através da análise do Censo da Educação Superior realizado pelo Instituto Nacional de Estudos e Pesquisas Educacionais Anísio Teixeira (Inep/MEC) no ano de 2008, no qual observa-se um aumento no número de IES no país.

O grande desafio das IES é estruturar e disponibilizar para seus gestores, as informações geradas por seus diversos sistemas, e que estes utilizem estas informações como recurso estratégico, objetivando a obtenção de vantagem competitiva sustentável, uma vez que os SI ainda são pouco utilizados pelas IES.

Os dados fornecidos pelos Ambientes de Gestão Educacional (AGE's) são analisados sob a ótica de informações meramente estatísticas, sobre o acesso aos cursos, conteúdos, quantidade de acessos, restringindo e limitando assim a capacidade de compreensão implícita

nas informações sobre as mais variadas tendências de utilização e a percepção das possibilidades de vantagens competitivas que possam ser obtidas com base em seu conteúdo.

A extração de informações que sejam relevantes aos interesses dos gestores, está se tornando complexa diante da quantidade de dados armazenados. Denomina-se *Knowledge Discovery in Databases* – KDD (Descoberta de Conhecimento em Bases de Dados), a atividade de "garimpar" a informação contida nestes dados. Apesar de ser comum usar os termos KDD (*Knowledge Discovery in Database*) e *Mineração de Dados* com o mesmo significado, Fayyad et al.(1996) definem o KDD como sendo o processo da extração de conhecimento dos dados como um todo, e *Mineração de Dados,* como apenas uma etapa em particular do KDD, sendo que nesta etapa a extração de padrões dos dados é realizada através do uso de algoritmos específicos.

Descobrir o conhecimento oculto nas grandes bases de dados das mais diversas organizações, seja de forma automática ou semi-automática é o objetivo do *Mineração de Dados,* além de permitir uma maior agilidade no processo de tomada de decisão por parte dos gestores.

O ato de coletar e armazenar os dados, em si, não traz nenhuma contribuição para a melhoria na estratégia de qualquer organização. Deve-se fazer uma análise, criando indicadores com intuito de descobrir padrões de comportamento implícito na base de dados, bem como suas relações de causa e efeito. Assim, as informações contidas nestas bases de dados, processadas e analisadas de forma correta, tornam-se requisitos primordiais na tomada de decisões.

Chiara (2003, p. 1) reforça que "para a aplicação de técnicas de Mineração de Dados, é necessário que se tenha uma coleção de dados disponível. Entretanto, o problema é conseguir dados relevantes para se extrair deles conhecimento potencialmente útil"

Dentre estas grande bases de dados, cita-se os Ambientes de Gestão Educacional (AGE) utilizados pelas IES para interagirem com seus os alunos. Considera-se uma necessidade fundamental para as IES a missão de gerir as informações, haja vista a existência de uma crescente demanda e atualização das tecnologias da informação, sendo este recurso de vital importância para a permanência das IES no mercado.

O gestor que “possui, domina e usa” a informação de forma estratégica possui papel fundamental no desenvolvimento de qualquer organização, da qual faça parte. O gestor deve trabalhar a informação de forma que sirva como elemento base para a tomada de decisões, desde que esta informação seja precisa, segura, confiável e esteja a disposição, informações. estas que se encontram no AGE das IES.

Os recursos do AGE da IES em questão não passaram por muitas análises quanto ao seu uso estratégico. Com isto, pretendeu-se analisar a utilização das informações armazendas neste ambiente, a fim de promover uma melhor eficácia na gestão das informações e oferecer novos dados para que a IES possa melhor explorar as tecnologias e auxiliar nas tomadas de decisões por parte de seus gestores.

1.1 PROBLEMA DE PESQUISA

Mesmo observando o crescimento da utilização dos dados disponíveis nos AGE’s pela IES, não se percebe muita preocupação em analisar a utilização destas plataformas para possibilitar a extração de informações, que podem ser utilizadas pelos gestores objetivando a obtenção de vantagem competitiva sustentável.

Observa-se um problema que é a não utilização das informações disponíveis nos AGE’s. Exemplo disto pode ser citado como o relatório de acesso ao ambiente, que no momento não tem informação nenhuma sendo extraída dele. Os sistemas informatizados nesta IES geram relatórios com formatação complicada, falta de informações ou informações às vezes inconsistentes ou redundantes. Estes problemas de dispersão e inconsistência das informações contribuem para erros na tomada das decisões por parte dos seus gestores, ou as tornam menos eficazes.

A utilização de ferramentas que auxiliem na busca, seleção e extração de informações relevantes em grandes bases de dados, tem recebido cada vez mais importância nas organizações, uma vez que estas ferramentas têm como principal objetivo minimizar o trabalho manual e a disponibilização de informações corretas aos gestores destas organizações.

Dentre estas técnicas cita-se a Descoberta de Conhecimento em Base de Dados (KDD, abreviatura do termo em inglês, *Knowledge Discovery in Database*), a qual utiliza a técnica de extração de informações conhecida como Mineração de Dados. Trata-se de um processo da

extração de padrões, considerados interessantes e não corriqueiros, a partir de uma base de dados. O foco do problema desta dissertação encontra-se na Gestão das Informações Gerenciais pelos gestores da IES, como apoio para o planejamento estratégico, devido a sua importância nas tomadas de decisões.

A aplicação da técnica de Mineração de Dados se dará na IES, haja vista a mesma estar inclusa num setor que está atraindo cada vez mais investidores, seja por meio da aquisição ou da fusão entre as IES e a manutenção da competitividade deste setor faz com que as mesmas revejam seus planejamentos estratégicos, a fim de se manterem competitivas. O problema consiste em averiguar se a utilização da técnica de Mineração de Dados, em face da enorme disponibilidade de dados armazenados no AGE da IES, deve-se à ausência de uma metodologia adequada para a qual resulte em informações úteis para os gestores.

Diante desta problemática, cabe levantar a seguinte questão que foi norteadora da pesquisa de campo: De que forma as técnicas de extração de informações podem auxiliar os gestores da IES? Mais pragmaticamente, como a gestão da informação obtida pelo uso de técnicas de extração de informação pode ajudar os profissionais da IES, a auxiliarem na tomada de decisões estratégicas para o gerenciamento de sua instituição?

1.1.1 Solução Proposta

Uma forma de auxiliar o gestor a resolver o problema anteriormente mencionado é disponibilizar ferramentas que o auxiliem nessa mineração dos dados contidos em seus repositórios. Diante deste contexto, é de significativo interesse que se possua uma ferramenta que lhe forneça uma melhor visualização das informações mais importantes para a tomada de suas decisões.

Dentre os objetivos do *Mineração de Dados*, está a descoberta de forma automática ou semi-automática do conhecimento que encontra-se "oculto" nas grandes quantidades de dados que as organizações possuem, permitindo de forma ágil e rápida a tomada de decisões.

Isto vem ao encontro de Cardoso e Machado (2008, pg. 497) que definem o *Mineração de Dados* como:

> [...] uma técnica que faz parte de uma das etapas da descoberta de conhecimento em banco de dados. Ela é capaz de revelar, automaticamente, o conhecimento que está implícito em grandes quantidades de informações armazenadas nos bancos de dados

> de uma organização. Essa técnica pode fazer, entre outras, uma análise antecipada dos eventos, possibilitando prever tendências e comportamentos futuros, permitindo aos gestores a tomada de decisões baseada em fatos e não em suposições.

Para tanto, serão aplicadas as técnicas de Mineração de Dados, a fim de se obter informações necessárias, confiáveis e de qualidade, para que os gestores tomem suas decisões. As informações que aqui foram utilizadas passaram pelos processos de seleção, análise e disseminação, visando um direcionamento estratégico da IES.

As ferramentas de Mineração de Dados podem ser empregadas como ferramenta complementar no processo de tomada de decisões, visando facilitar ao gestor a busca pela informação correta dentro da grande massa de dados que os SI das IES oferecem. Aliada a rapidez na busca por esta informação e podendo gerar como conseqüência uma vantagem competitiva.

A técnica de Mineração de Dados, que faz parte das ferramentas de KDD, tem por objetivo agilizar o processo de mineração das informações, facilitando a busca e minimizando as dificuldades de se procurar informações em grandes bases de dados.

Furtado (2004) sustenta que:

> Os problemas relacionados ao entendimento, resumo e tratamento de informações foram inicialmente resolvidos na área do "Knowledge Discovery from Databases" – KDD-, que busca descobrir co-relacionamentos e dados implícitos nos registros de um Banco de Dados, extraindo-os para obter conhecimento novo, útil e interessante, ou seja, enfoca o processo global de descoberta do conhecimento de dados, incluindo como os dados são armazenados e acessados.

Mesmo que autores que têm por objeto este assunto determinem mais tarefas, abordar-se-ão as que serão utilizadas neste estudo, sendo elas: associação, classificação e clusterização (*clustering)*. Estas técnicas foram escolhidas por serem próximas entre si nos seus objetivos e por serem de maior compreensão por parte do gestor da IES.

Associação: Tem por objetivo a combinação de itens considerados importantes, sendo que a presença de tal item indica implicitamente na presença de outro item na mesma transação. Este processo teve como precursor Agrawal, em 1993. (AGRAWAL, IMIELINSKI e SWAMI, 1993)

Classificação: Classes de objetos são criadas para agrupar objetos com características semelhantes. São utilizados dados sobre o passado de determinada base para encontrar padrões com valores significativos, aos quais irão levar a regras sobre o futuro destes objetos.

Clusterização: Os dados heterogêneos são reagrupados em grupos com características semelhantes, método conhecido como *clustering.* A clusterização é definida por Berry (1997) como sendo "a tarefa de segmentar uma população heterogênea em um número de subgrupos (ou clusters) mais homogêneos possíveis, de acordo com alguma medida". O que diferencia a clusterização da classificação é a não existência de grupos pré definidos.

Pode-se aplicar as tarefas e técnicas da MD aos dados gerados pelos AGE, nos quais podem ser encontradas relações entre os dados disponíveis, segundo Kampff (2009, p. 79):

> Processos de MD podem ser empregados, também, para descobrir características e comportamentos em alunos que indiquem risco de evasão ou reprovação e, então, essa descoberta pode contribuir para a atuação docente, de forma a evitar esses resultados indesejados. Utilizar técnicas de MD, portanto, possibilita identificar padrões de acesso, de realização de atividades e de interação dos alunos que os levam a obter êxito (ou não) e dessa forma, oferecer embasamento para a construção de ferramentas que auxiliem na prática docente, buscando a redução dos índices de evasão e reprovação.

Quando aplicada em sistemas de ensino a MD, freqüentemente está apoiada nas mesmas técnicas utilizadas em aplicação comerciais, fazendo uma analogia entre a navegação do aluno pelas páginas do curso com a navegação de um potencial cliente nas páginas do produto ou de empresa. A pesquisa de padrões de comportamento em ambientes educacionais se dá principalmente pelas técnicas de Descoberta de Regras de Associação (DRA) e ou pela aplicação de tarefas de associação, classificação ou clusterização.

Os resultados da MD podem ser utilizados para obter uma melhor compreensão dos processos subjacentes de ensino, para a geração de recomendações e conselhos aos alunos, para melhorar a gestão de objetos de aprendizagem.

A técnica de Descoberta de Regras de Associação tem por finalidade descobrir padrões de acesso às páginas dos cursos pelos acadêmicos ou encontrar associações entre as diversas páginas por eles visitadas. Enquanto as tarefas objetivam agrupar os acadêmicos pelo comportamento de acesso, procurando por similaridades entre eles. A avaliação do desempenho, a adaptação e recomendação de conteúdos tendo como base o comportamento dos alunos, também são outras áreas de aplicação da MD.

A MD pode ser aplicada nos AGE's, não fazendo a identificação dos acadêmicos, apenas identificando suas características, uma vez que se pode fazer a mineração sobre uma base de dados de acadêmicos matriculados em determinada disciplina ou na base dos acadêmicos matriculados nos cursos á distância oferecidos pela IES.

O emprego das técnicas de Mineração de Dados, permite as IES's criarem parâmetros capazes de entender o comportamento dos dados armazenados, permite também a identificação das afinidades existentes entre estes dados, além de proporcionar a previsão de comportamentos e hábitos dos dados.

Será aplicada a técnica de mineração de dados na IES, sendo esta instituição de ensino privado, haja vista, que este setor vem despertando o interesse de novos investidores. Isto faz com que as IES repensem em seus métodos, buscando novas tecnologias e ferramentas que possam auxiliá-las a manterem-se neste mercado altamente competitivo.

1.1.2 Delimitação de Escopo

Com base no pressuposto de que um dos fatores principais para a garantia de sobrevivência das organizações está fortemente vinculado a eficácia na gestão de seus custos operacionais, optou-se por utilizar nesta dissertação de uma ferramenta de Mineração de Dados com distribuição gratuita.

Com intuito de atender a solução proposta nesta dissertação, a análise dos dados foi feita num Ambiente de Gestão Educacional, no qual as informações consideradas pessoais, como: nomes, endereços de e-mail, telefones foram omitidos para preservação da integridade de seus proprietários.

Uma vez que este trabalho objetiva-se na utilização da técnica de Mineração de Dados, com o intuito de municiar os gestores com informações confiáveis, relevantes e de qualidade para a tomada de decisões estratégicas, foram consideradas unicamente as informações constantes na base de dados do Ambiente de Gestão Educacional.

Embora seja reconhecido que as IES e alguns outros setores do mercado sofram impacto direto do uso da informação, suas dimensões e comportamentos não são a essência deste trabalho.

1.1.3 Justificativa

Assim como as demais organizações, as IES não se excluíram dos avanços gerados pela TI, que vão além do simples conjunto de recursos computacionais. Elas estão buscando extrair destes recursos o máximo de informações e com o uso destas informações gerirem suas atividades.

As atividades desenvolvidas pelas IES, seja desde o ensino básico ou superior, devem ser entendidas como uma atividade empresarial semelhante a qualquer outra. Diante disto as instituições de ensino estão sujeitas às mesmas pressões que aflige aos demais mercados.

Furtado (2004, p. 4) destaca que:

> [...]o setor educacional vem atraindo um número crescente de novos atores e o mercado educacional de novos integrantes, que passam a disputar o domínio deste mercado com as instituições tradicionais. Estas, por sua vez, vêem-se forçadas a rever suas práticas e métodos até então utilizados, como condição para que possam continuar tendo relevância em seus serviços prestados e que sobrevivam em um cenário altamente competitivo.

Devido ao aumento do número de IES no Brasil, estudos, pesquisas e discussões estão sendo elaboradas sobre os mecanismos desta evolução e como esta vem influenciando o desenvolvimento educacional do país.

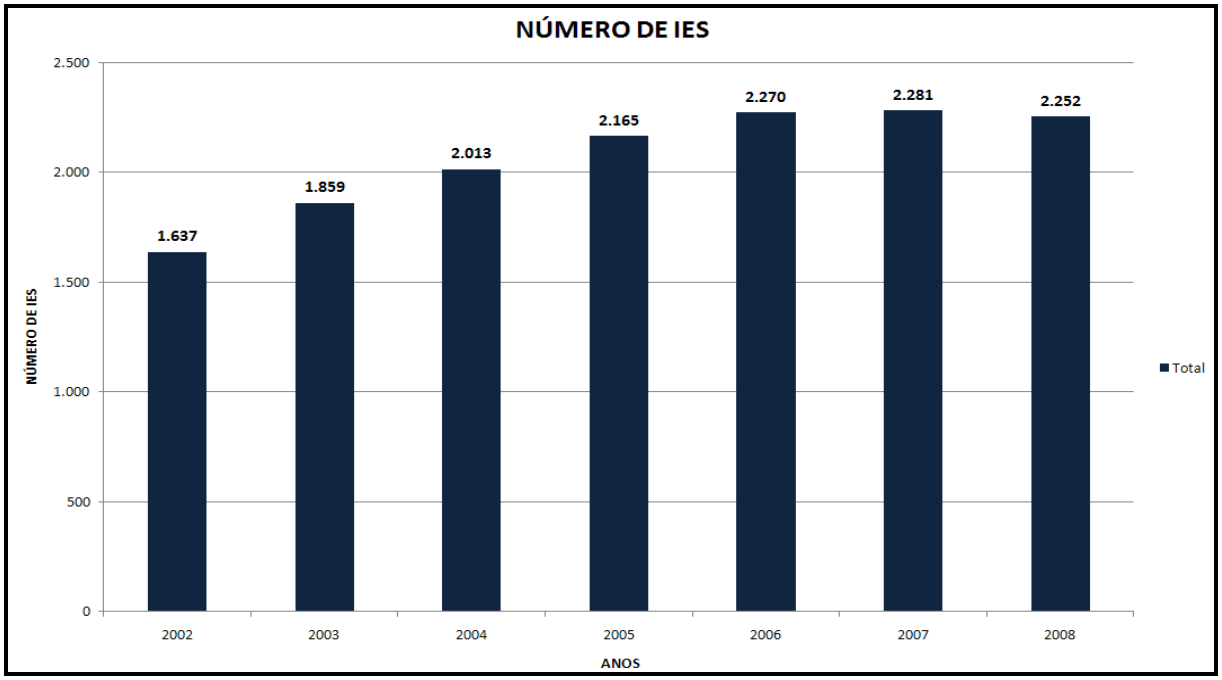


Gráfico 1 - Número de IES no Brasil
Fonte: Adaptado de INEP (2010).

Segundo dados do Censo da Educação Superior do ano de 2008, divulgados pelo Instituto Nacional de Estudos e Pesquisas Educacionais Anísio Teixeira, (INEP/MEC), houve uma redução de vinte e nove IES no Brasil, apresentado no Gráfico 1, neste período. Esta redução deu-se em virtude da integração de instituições, por fusão ou compra, que vinha sendo observada nos últimos anos.

O Censo da Educação Superior de 2008, mostra que houve um crescimento de alunos ingressantes, onde 1.936.078 novos alunos ingressaram no ensino superior, o que corresponde a um aumento de 8,5% a mais em relação ao ano de 2007. O número total de matrículas foi 10,6% maior em relação ao ano de 2007, totalizando 5.808.017 alunos matriculados nos curso de graduação, observado no Gráfico 2. (INEP, 2010).

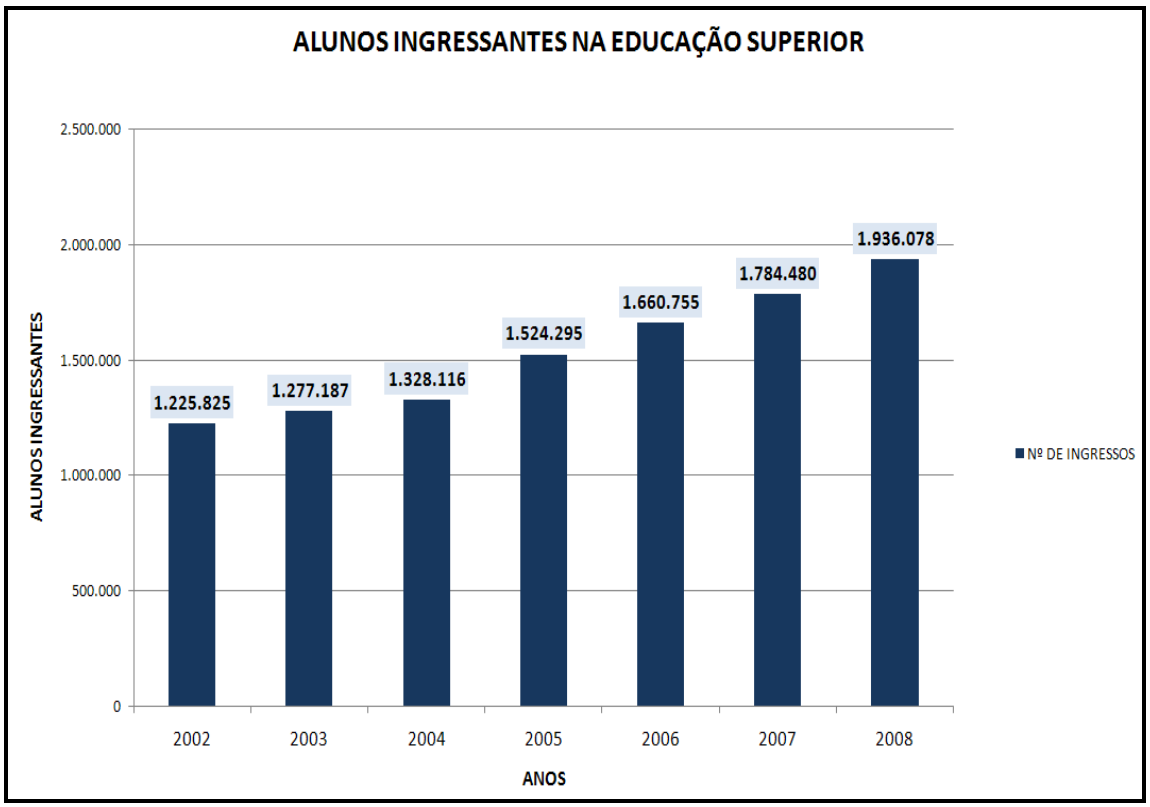


Gráfico 2 - Relação de alunos ingressantes no Ensino Superior
Fonte: Adaptado de INEP (2010).

Esta realidade não é diferente para as IES, ainda mais quando leva-se em consideração as fusões que estão ocorrendo no mercado de ensino. Isto faz com que a concorrência se torne mais agressiva, transformando a informação disponibilizada aos discentes, docentes e colaboradores das IES um bem precioso.

Outro fato importante a ser observado é o crescente número de alunos inscritos na modalidade de ensino à distância, conforme dados do Censo da Educação Superior de 2008 e melhor representado no Gráfico 3, houve um aumento de 42% no número de alunos inscritos nesta modalidade de ensino, em relação ao ano de 2007.

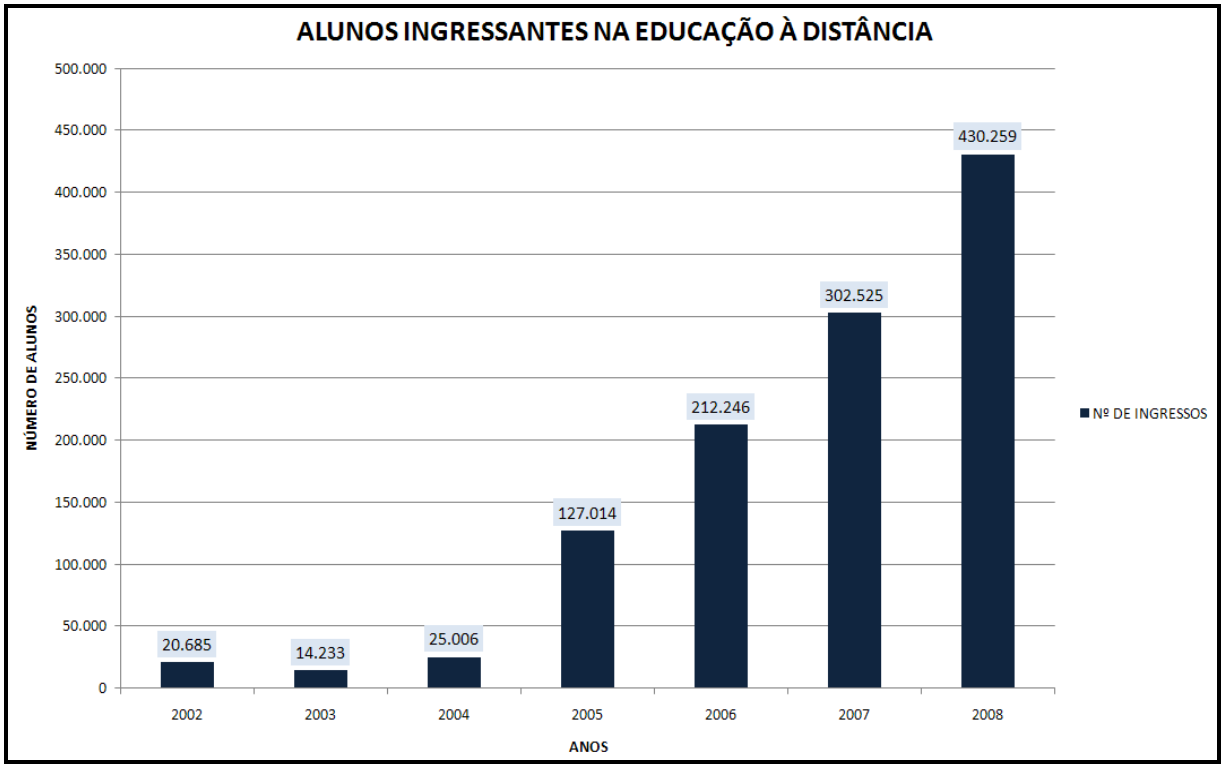


Gráfico 3 - Alunos ingressantes na Educação superior à Distância
Fonte: Adaptado de INEP (2010).

Com relação as IES privadas e estaduais o INEP/MEC relata que:

> Quanto ao número de cursos, houve um incremento de 1.231 (5,2%) novos cursos de graduação presencial nas IES brasileiras e apenas as IES estaduais não registraram crescimento em relação a 2007, com um decréscimo de 1,6% nos cursos ofertados. Do mesmo modo, houve o aumento de 7,3% (cerca de 319 mil) no número de vagas ofertadas em graduação presencial e a distância. As instituições privadas foram responsáveis pela oferta de cerca de 4 milhões de vagas em 2008, apresentando aumento de 4% em relação a 2007. Em 2008 foram ofertadas 463.969 vagas nos cursos de Educação Tecnológica, com um aumento de 17,8% em relação a 2007. As IES privadas são responsáveis por cerca de 94% dessa oferta. (INEP, 2010).

Com base nestas informações retiradas dos relatórios do MEC/IMEP, se observa que existe uma defasagem entre a oferta e a procura, o que indica que a concorrência neste setor vem se tornando mais intensa. Vivencia-se uma nova era na gestão das IES, com base nas novas regras, portarias e leis que regularizam o setor educacional no país.

Frente a argumentação anterior, são duas as justificativas da relevância do tema: a crescente importância da Gestão da Informação em todas as organizações e a pouca disponibilidade de estudos e pesquisas voltadas para esta área tendo como foco a aplicação destes conceitos nas IES, uma vez que estas são responsáveis por gerar o conhecimento necessário a sua utilização.

A quantidade de dados, dos mais variados tipos e sua falta de estruturação, aliada a quantidade de informações que estão sendo disponibilizadas aos gestores, acabam se tornando elementos que dificultam o processo de tomada de decisões.

Os dados foram disponibilizados em duas planilhas do Excel, onde uma das planilhas contém as respostas dos itens propostos no questionário sócio-educacional aplicado aos candidatos no ato da inscrição ao processo seletivo (Vide Anexo A). Noutra planilha encontram-se os itens propostos no questionário sócio-economico aplicado aos egressos da IES (Vide Anexo B).

Decidir qual a melhor oportunidade, o melhor momento, a melhor prática para se trabalhar a informação, visando a obtenção da vantagem competitiva sustentável nas IES, vem tornando-se cada vez mais o objetivo a ser alcançado pelos gestores que administram esta organização.

Não obstante as IES defrontam-se com certas dificuldades na transformação dos milhares de dados que são por ela produzidos diariamente, em informações realmente estratégicas que auxiliem nas tomadas de decisões.

Acredita-se que com a utilização das técnicas de *Mineração de Dados* no AGE das IES estimule a criação e utilização de informações de caráter realmente útil para os gestores, visando a identificação de novas oportunidades, formas de uso e auxiliando na tomada das decisões estratégicas. A aplicação de técnicas de *Mineração de Dados* nas IES's vem reforçar seu "arsenal" de estratégias para enfrentarem o mercado.

Furtado (2004) salienta que "Ferramentas que auxiliem na busca, seleção e extração de informações específicas e relevantes na Web - e não somente oriundas dela - têm cada vez mais recebido maior importância, de forma a minimizar o trabalho manual do usuário".

Kampff (2009, p. 79) destaca que a utilização de técnicas de MD:

> [...]possibilita identificar padrões de acesso, de realização de atividades e de interação dos alunos que os levam a obter êxito (ou não) e, dessa forma, oferecer embasamento para a construção de ferramentas que auxiliem na prática docente, buscando a redução dos índices de evasão e reprovação.

Um fator considerado como crítico para a aceitação de qualquer ferramenta de TI é a facilidade de uso da mesma. A mineração de dados por meio de técnicas de Data Mining suporta funções muito sofisticadas, funções que se encontram embutidas nos softwares, desta forma fazendo com que os usuários não necessitem serem conhecedores das técnicas de mineração para obterem seus resultados sejam em telas ou por meio de relatórios impressos.

Já em virtude da participação no projeto de pesquisa em Sistemas de Informações aprovado junto ao CNPq e coordenado pelo Professor Dr. Oscar Dalfovo, tem-se a pretensão de ampliar o escopo do trabalho, averiguar a utilização das técnicas de Mineração de Dados e disponibilizar um instrumento como ferramenta auxiliar para os gestores que possibilite a tomada de decisões realmente estratégicas.

Diversas técnicas de mineração de dados têm sido aplicadas com sucesso em diferentes tipos de dados educacionais e têm ajudado a enfrentar muitos problemas usando a classificação tradicional, técnicas de agrupamento e análise de associação.

Baseado na preocupação que existe entre o baixo índice de matriculados e as altas taxas de evasões, esta pesquisa visa buscar conhecimentos sobre o processo de inscrição dos acadêmicos e dos egressos da IES em questão. Os resultados obtidos através desses estudos poderão auxiliar os gestores da IES na tomada de decisões em relação ao projeto acadêmico a ser desenvolvido junto ao setor responsável pelo marketing da IES.

O conhecimento a ser gerado é de suma importância, não só para o setor responsável pela divulgação do processo seletivo da IES, mas também para os coordenadores dos cursos, que podem melhor definir as políticas administrativas para os ingressantes em seus respectivos cursos.

1.2 OBJETIVOS

Dentro deste cenário, os objetivos da proposta em questão são a seguir descritos.

1.2.1 Objetivo Geral

O objetivo geral deste projeto visa um estudo e aplicabilidade da técnica de Mineração de Dados na base de dados do Ambiente de Gestão Educacional de um IES de Blumenau-SC, para auxiliar os gestores na descoberta da informação e conhecimento.

1.2.2 Objetivos Específicos

Para a concretização do objetivo geral, elencam-se a seguir os objetivos específicos desta proposta:

a) identificar, na IES, o que os gestores entendem por gestão da informação e descoberta da informação e conhecimento;

b) levantar junto a IES quais são as principais informações e conhecimento do AGE para disponibilizar aos gestores da IES;

c) identificar ferramentas que utilizam técnicas de Mineração de Dados, que possam ser aplicadas nas bases de dados do AGE da IES, para auxiliar os gestores na descoberta da informação e conhecimento;

d) aplicar nas bases de dados do AGE da IES, uma ferramenta de Mineração de Dados na extração da informação e conhecimento, para auxiliar os gestores em futuras decisões estratégicas a respeito dos futuros ingressantes e egressos.

1.3 METODOLOGIA

A pesquisa tem como uma de suas definições mais simples, a obtenção de respostas por meio do uso de métodos científicos, para as questões ou problemas propostos. Tendo como ponto de partida uma dúvida levantada ou um problema a ser resolvido e fazendo uso de um ou mais métodos científicos, a pesquisa tem o objetivo de buscar uma solução ou resposta para a dúvida levantada ou o problema em questão.

1.3.1 Metodologia da Pesquisa

A metodologia de pesquisa tem por objetivo apresentar uma sequência de etapas que deverão ser seguidas para alcançar os objetivos propostos, com resultados coerentes (WAZLAWICK, 2008).

Em vista disso, a fim de atingir o objetivo geral e os específicos deste projeto, primeiramente foi feito um levantamento bibliográfico de trabalhos contemplando a área de estudo, sendo que os trabalhos de Lucas (2002), Zambenedetti (2002), Furtado (2004), Alves (2005), Martinhago (2005), Alvarenga (2006), Cella (2006), Scoss (2006), Galucci (2007), Kampff (2009) serviram de base para a aquisição de elementos que definiram este projeto.

Tendo como base a literatura traçou-se um quadro teórico a fim de sustentar o desenvolvimento da pesquisa, alinhando-o com os objetivos deste projeto. O conhecimento e a comparação puderam ser feitos com base nos trabalhos relacionados ao tema da pesquisa.

O método empregado nesta pesquisa foi o método indutivo, o qual dá privilégio a observação para se alcançar os objetivos. No método indutivo se todas as premissas são verdadeiras, a conclusão é provavelmente será verdadeira, mas não necessariamente.

Sob o ponto de vista da natureza, a pesquisa pode ser classificada como aplicada. Na pesquisa aplicada, procura-se a partir da geração de conhecimento para a aplicabilidade prática, direcionando para a solução de determinados problemas.

Quanto ao ponto de vista dos procedimentos técnicos foram utilizadas fontes bibliográficas, para o desenvolvimento da fundamentação teórica, ou seja, a elaboração do conteúdo teórico da pesquisa deu-se a partir de material já publicado, constituído de livros, artigos de periódicos e também de material disponibilizado na Internet (GIL, 2002).

Sob o ponto de vista da abordagem do problema, a pesquisa é classificada como qualitativa. Para Chizzotti (1995, p.89) "[...] a finalidade de uma pesquisa qualitativa é intervir em uma situação insatisfatória, mudar condições percebidas como transformáveis, onde pesquisador e pesquisados assumem, voluntariamente, uma posição reativa."

Para Neves (1996) a pesquisa qualitativa:

> Compreende um conjunto de diferentes técnicas interpretativas que visam a descrever e a decodificar os componentes de um sistema complexo de significados. Tem por objetivo traduzir e expressar o sentido dos fenômenos do mundo social; trata-se de reduzir a distância entre indicador e indicado, entre teoria e dados, entre contexto e ação.

Nesta modalidade de pesquisa os dados coletados em suas várias etapas estão em constante processo de análise e avaliação, sendo que na análise as novas descobertas serão

novamente analisadas para orientar uma nova ação que possa modificar as condições consideradas indesejadas.

Os dados foram disponibilizados em duas planilhas do Excel, onde uma das planilhas contém as respostas dos itens propostos no questionário sócio-educacional aplicado aos candidatos no ato da inscrição ao processo seletivo (Vide Anexo A). Noutra planilha encontram-se os itens propostos no questionário sócio-economico aplicado aos egressos da IES (Vide Anexo B).

Perante o ponto de vista dos objetivos, a pesquisa pode ser classificada como exploratória. Na exploração procurou-se maior familiaridade com o problema objetivando-o torná-lo explícito.

1.3.2 Procedimentos Metodológicos

Para o estudo aqui em questão a metodologia adotada pode ser enquadrada, conforme sugerem Roesch (1999) e Vergara (2003), como um estudo de caso, já que trata da aplicação de uma técnica de Mineração de Dados no AE de uma IES específica. Desta forma, as conclusões do estudo não podem ser generalizadas para outras IES, devido as características peculiares de cada.

Para Merriam (1988, apud GODOI, BANDEIRA-DE-MELLO e DA SILVA, 2006, p. 119), um estudo de caso qualitativo é “uma descrição (holística e intensiva) de um fenômeno bem delimitado (um programa, uma instituição, uma pessoa, um grupo de pessoas, um processo ou uma unidade social”).

O estudo de caso tem com característica fundamental uma maior complexidade na coleta dos dados, levando-se em consideração outras modalidades de pesquisa, uma vez que possui mais de uma técnica de coleta de dados. A qualidade dos resultado deve ser assegurada por procedimentos utilizados na obtenção e coleta dos dados. Os dados para o estudo de caso podem ser obtidos por diferentes fontes, como: a análise de documentos, entrevistas, depoimentos pessoais e questionários, sendo o estudo de caso considerado o mais completo tipo de delineamento de pesquisa. (GIL, 2002).

A intenção do estudo de caso é de revelar a interação entre o interno e externo que são característicos de um mesmo fato. Tratando-se de uma pesquisa qualitativa, o método de "estudo de caso" foi escolhido por possibilitar a observação do contexto a ser pesquisado.

O projeto como um todo foi envolvido na pesquisa bibliográfica. A elaboração do conteúdo teórico da pesquisa deu-se a partir de material já publicado, constituído de livros, artigos de periódicos nacionais e estrangeiros e também de material disponibilizado na Internet. Houve também reuniões com o responsável pelo ambiente de aprendizagem da IES, com intuito de melhor compreender o ambiente a ser pesquisado.

Quando da revisão bibliográfica, a busca pela solução do problema passou pela escolha da técnica de mineração de dados que melhor atendesse o objetivo determinado. Para o problema de pesquisa deste trabalho a técnica mais adequada foi Mineração de Dados.

1.4 ESTRUTURA DA DISSERTAÇÃO

O trabalho está organizado em 05 (cinco) capítulos correlacionados. O Capítulo 1, Introdução, apresentou por meio de sua contextualização o tema proposto neste trabalho. Da mesma forma foram estabelecidos os resultados esperados por meio da definição de seus objetivos e apresentadas as limitações do trabalho permitindo uma visão clara do escopo proposto. Apresentou-se ainda a Metodologia da pesquisa utilizada.

O segundo capítulo apresenta a Fundamentação Teórica que orienta a investigação, complementada por trabalhos científicos (monografias, dissertações, teses e artigos científicos específicos), como outras fontes literárias (livro, periódicos, internet, banco de dados virtuais, entre outros) com aderência ao mesmo. Neste capítulo são abordados os assuntos pertinentes ao trabalho como: a definição de Sistemas de Informação, a Informação e suas características, Gestão da Informação, Extração da Informação, Mineração de Dados, Algoritmos de Classificação e a Ferramenta WEKA assim como as Ferramentas de Gestão de IES.

O Capítulo 3 apresenta os Trabalhos Relacionados, onde os três primeiros apresentam temas relacionados ao domínio de Gestão de IES. Na seqüência outros trabalhos envolvendo Mineração de Dados, na sequencia apresentam-se trabalhos que abordam o uso de Mineração de Dados em Ambientes de Gestão Educacional.

O quarto capítulo traz os resultados dessa implementação e aplicação das técnicas de Associação, Classificação e de Clusterização dos dados analisados.

No Capítulo 5, são tecidas as conclusões do trabalho, relacionando os objetivos identificados inicialmente com os resultados alcançados. São ainda propostas possibilidades de continuação da pesquisa desenvolvida a partir das experiências adquiridas com a execução do trabalho.

2 REFERENCIAIS TEÓRICOS

A fundamentação teórica serve de base para fundamentação da pesquisa, em termos teóricos e empíricos, servindo também de auxiliar nos instrumentos de coleta de dados utilizados para a pesquisa realizada.

Neste item apresenta-se como estado da arte o embasamento sobre Sistemas e Informação, Mineração de Dados Com suas técnicas e tarefas, Ferramentas para a Mineração de dados, Gestão de IES e finalizando com Ferramentas de Gestão.

2.1 SISTEMAS E INFORMAÇÃO

As possibilidades de uso das tecnologias da informação ficaram mais claras com a utilização crescente do computador. A velocidade de processamento disponível permite que as organizações alcancem seus objetivos com mais facilidade, fazendo uso de informações precisas, no tempo certo e local adequado. Diante destas possibilidades muitos conceitos novos surgiram.

Dalfovo (2007, p. 19) identifica que:

> Para se definir Sistemas de Informação, é preciso ter em mente algumas definições ou conhecimentos sobre o computador, hardware, software e telecomunicação. Existem diversas definições sobre sistemas de informação. Algumas definições baseiam-se no modelo comportamental, outras no modelo técnico.

Stair e Reynolds (2008, p. 14) reforçam que

> Um sistema de informação baseado em computadores (CBIS – *computer based information system)* é composto por hardware, software, bases de dados, telecomunicações, pessoas e procedimentos configurados para coletar, manipular, armazenar e processar dados em informações.

Rezende (1999, apud REZENDE e ABREU, 2009, p.38) enfatiza que “Todo sistema, usando ou não recursos de Tecnologia da Informação, que manipula e gera informação pode ser genericamente considerado como Sistema de Informação”.

Os SI encarregam-se do papel principal no suporte a tomada de decisão, com base nas ferramentas de análise, seja na apresentação, no armazenamento, gerenciamento e na recuperação da informação, através de Data Marts, Data Warehouse, que utilizam alguma

técnica de mineração de dados ou processamento de transações, podendo disponibilizar a informação através dos ambientes de redes.

Sistemas de Informação tendem a ser a solução para muitas organizações desde que estas organizações tenham certeza do que necessitam e saibam aonde querem chegar com o uso de suas informações.

2.1.1 Informação

Ter o conceito de informação definido é essencial para o entendimento correto dos SI e como o processamento por eles realizado gera a informação necessária para a tomada de decisões.

Buckland (1991) considera a "informação como coisa", ou seja, transforma a informação em algo que pode ser alcançado, possível de ser expressado.

Capurro (2003, apud CELLA, 2006, p. 135) define que:

> O conceito de informação refere-se a processos cognitivos humanos ou a seus produtos objetivados em documentos, evidencia uma vez mais os limites de todo o paradigma ou modelo, nesse caso do paradigma social, no momento em que a relação entre informação e significado torna-se problemática, quando se deseja transportá-la para sistemas não sociais.

Outro autor Robredo (2003, p. 1) cita em sua obra, a qual traz uma definição de um compêndio inglês, na qual a informação é "um conjunto de dados organizados de forma compreensível, registrado em papel ou em outro meio e suscetível de ser comunicado".

Para Le Coadic (2004, p. 4) informação "é um conhecimento inscrito (registrado) em forma escrita (impressa ou digital), oral ou audiovisual, em um suporte".

A informação está diretamente ligada ao conceito de transformação, no qual os dados armazenados nas bases de dados das organizações passam por diversos processos (limpeza, filtragem, codificação, agrupamento, entre outros) para gerarem a informação.

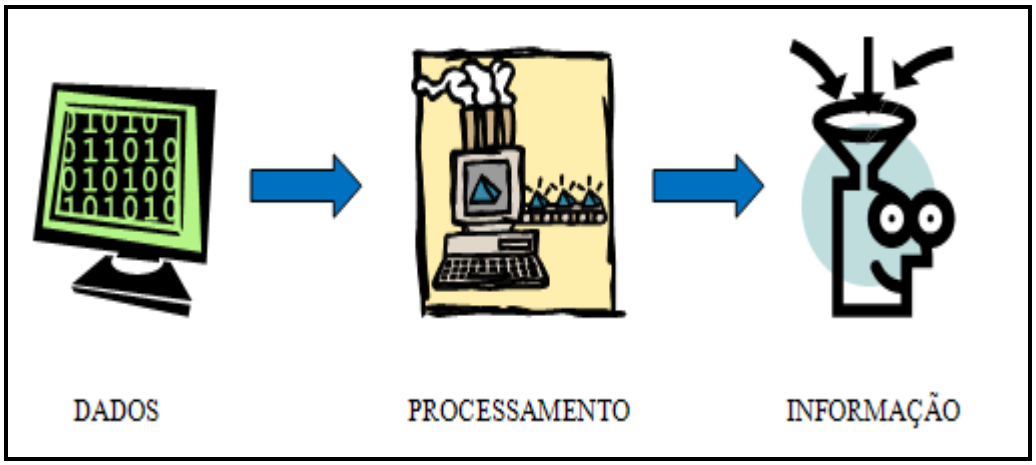


Figura 2 - Transformação de dado em informação
Fonte: Adaptado de Stair e Reynolds (2008)

Stair e Reynolds (2008, p. 4) consideram a informação como "um conjunto de fatos organizados de modo a terem valor adicional, além do valor dos fatos propriamente ditos".

Rezende e Abreu (2009, p. 38) enfatizam que:

> [...] informação é todo o dado *trabalhado, útil, tratado,* com valor significativo atribuído ou agregado a ele e com sentido natural e lógico para quem usa a informação. O dado é entendido com um elemento da informação, um conjunto de letras, números ou dígitos, que, tomado isoladamente não transmite nenhum conhecimento, ou seja, não contém significado claro.

O desejo em deter, controlar e manipular a informação está presente na história desde os tempos passados, contudo após o término da Segunda Guerra Mundial, e mais recentemente com o avanço das tecnologias de comunicação, e com a redução nos custos da TI, observa-se um aumento significativo na produção de informações.

O uso dos meios de comunicações e da grande rede mundial, a Internet, tornou-se um elemento primordial nas organizações, por conseqüência, reduziu o custo com o processamento da informação, assim possibilitando que mais e mais pessoas pudessem ter acesso a mesma.

Barreto (2002) descreve que no período compreendido entre 1945 e 1980, a gerência da informação era considerada um problema, pois a ordenação, organização e disseminação de informações não atingiriam seus objetivos uma vez que as teorias e os instrumentos da época não tinham a capacidade requerida para tal situação.

Desde aquela época até os dias atuais, a gerência da informação tem assumido um papel importante, ademais perante a economia globalizada, persuadindo os gestores a demonstrar interesse maior em poder gerir a informação de forma mais sistemática.

Período	Conceito de Informação	Importância atrelada
1950-1960	Elemento Burocrático indispensável	Tentativa de eliminação do processamento realizado através de papéis.
1960-1970	Apoio aos objetivos gerais	Auxílio na gestão das atividades da organização.
1970-1980	Domínio da gestão da organização	Agilização no processo de tomada de decisão
1980-2000	Utilizada como recurso estratégico	Obtenção da vantagem competitiva e manutenção da organização no mercado.

Figura 3 - Evolução do conceito de informação
Fonte: Adaptado de Laudon e Laudon (1996)

Hommerding (2001, p. 28) destaca que:

> As tecnologias da informação devem ser consideradas ferramentas básicas de trabalho, instrumento para qualquer tipo de unidade de trabalho/informação, uma vez que o processamento, o gerenciamento, a recuperação e a disseminação da informação, por meio dessas tecnologias, são mais eficientes e eficazes.

Resumindo os conceitos apresentados, diz-se que a informação é o conhecimento registrado. De encontro ao exposto, tem-se que informação hoje é também considerada como fonte de transformações, assim como relata Silveira (2008) ”É a partir da informação que as pessoas podem modificar suas vidas, controlar suas inseguranças e frustrações, se situar no tempo e no espaço, evoluir mental e espiritualmente e ajudar a melhorar a vida de seus semelhantes”.

Alvarenga (2006, p. 31) relata que seja qual for o profissional que fará uso da informação, ele deve saber distinguir as informações que lhe são apresentadas e quais realmente são necessárias para suas necessidades.

Embora a informação seja um ativo que precisa ser administrado, tal qual os demais bens da organização, ela tem uma característica diferente do ponto de vista de sua utilização: ela é infinitamente reutilizável, não se deteriora nem se deprecia, e seu valor é determinado apenas pelo usuário.

Dalfovo (2007, p. 22) destaca que:

> O uso eficaz da informação nas organizações passa a ser um patrimônio, em que é considerado um fator chave para o sucesso das organizações. Este fator torna-se mais expressivo quando as organizações defrontam-se com as mudanças de mercado e avanços das tecnologias.

Dadas as atuais abordagens e dimensões da informação, há a exigência de um novo profissional, apto e habilitado a utilizar os recursos tecnológicos, com a incumbência de disseminar a informação, promovendo o compartilhamento desta para todos os usuários.

Tornar a organização mais competitiva, proporcionar aos seus gestores informações com maior valor agregado, transformar informações em conhecimento, para apoiar o planejamento estratégico das organizações, estas são as funcionalidades da gestão da Informação.

A medida que se conceitua a informação, tem-se uma maior dificuldade baseada nas mudanças sociais e tecnológicas que recriam a cada dia uma nova realidade, seja ela pessoal ou empresarial. Diante desta nova perspectiva, surge um norte a ser seguido, no qual o conhecimento adquirido, resultado da capacidade de recordação de fatos, torna-se um diferencial a ser utilizado. Este conhecimento por vezes desprezado é de muita valia diante as adversidades enfrentadas pelas organizações.

2.1.1.1 A importância da informação

A informação, quando usada de forma eficaz, passa a ser considerada um patrimônio das organizações, vindo a ser considerada também com um fator chave para o sucesso destas. Isto é mais visível quando as organizações enfrentam mudanças no mercado em que atuam.

Carvalho (2001, p. 27) afirma que:

> A tecnologia da informação é a ferramenta utilizada pelo executivo, tomador de decisão para fazer da informação o recurso estratégico. Então devem-se estudar as três partes – a tecnologia necessária, o perfil do executivo, a qualidade da informação – para que os objetivos das organizações sejam alcançados de forma eficaz e eficiente.

Para Castro (2000, p. 28) “A informação é um requisito básico para a sobrevivência do ser humano. Permite o necessário intercâmbio entre o homem e o ambiente em que ele vive”.

Dalfovo (1998, p. 23) destaca que “O mercado não se limita somente ao conhecimento da informação. De alguma forma a informação é o prolongamento do produto na prestação de serviço. A informação é tão importante que passa a ser o centro das atividades nas empresas”.

Observa-se a importância da informação em qualquer nível de atividade realizada pelo homem. Na sociedade pós-industrial, chamada de sociedade da informação, esta tem lugar de destaque, considerada como elemento indispensável para a tomada de decisões.

Stair e Reynolds (2008, p.6) destacam as qualidades da informação para que a mesma seja considerada útil nas tomadas de decisões. Para os autores, a informação é considerada valiosa quando ela é:

a) **Precisa**: quando a informação está isenta de erros;
b) **Completa**: quando a informação contém todos os fatos relevantes;
c) **Econômica**: a geração da informação tem um custo considerado baixo;
d) **Flexível**: pode ser utilizada em diversos fins;
e) **Confiável**: a fonte produtora da informação é confiável;
f) **Relevante**: tem importância para o tomador de decisões;
g) **Simples**: a informação demasiadamente complexa pode confundir o tomador de decisões;
h) **Apresentada no momento exato**: informações apresentadas após a ocorrência dos fatos não traz novidade;
i) **Verificável**: deve se possível verificar se realmente a informação está correta;
j) **Acessível**: os usuários com permissão de acesso podem tê-la no momento em que precisam e;
k) **Segura**: seu acesso só deve ser permitido a quem tem permissão.

Os mesmos autores reforçam que “o valor da informação está diretamente ligado a como ela auxilia os tomadores de decisões atingirem seus objetivos organizacionais”.

Rezende e Abreu (2009, p. 36-37) ressaltam que informação e planejamento são palavras-chaves para a organização das organizações, citando ainda que as “informações

personalizadas e oportunas são fundamentais para a inteligência empresarial ou organizacional".

Rezende (2001, p.3) destaca que "a formulação estratégica de qualquer negócio sempre é feita a partir das informações disponíveis, portanto, nenhuma estratégia pode ser melhor que a informação da qual é derivada". Nesse contexto, verifica-se que, a chance da organização tornar competitiva está fortemente influência pela gestão da informação.

Para Cella (2006, p. 136)

> As instituições podem obter vantagens competitivas por intermédio do uso da informação através da realização de investimentos em informação e tecnologia da informação, do uso estratégico da informação agregando-a a seus produtos e serviços e da aprendizagem organizacional.

Mesmo a informação sendo de fundamental importância para as organizações, deve ser transformada em informações úteis para os gestores, pois é com estas informações, com valores agregadas a ela que se pode obter uma vantagem competitiva sustentável, mantendo-se a frente de seus concorrentes.

Por intermédio do uso estratégico da informação, associando-a aos produtos e serviços, juntamente com o investimento em informação as IES podem obter vantagens competitivas sustentáveis. A agregação de valores a informação ultrapassa os métodos de consulta, pesquisa e disseminação tradicionais, aos usuários das organizações.

Cella (2006, p. 141) ressalta o valor da informação dentro das IES.

> O valor da informação e a tomada de decisão por parte dos gestores são afetados pela qualidade da mesma, ou seja, quando a informação não tem qualidade ou é deficiente, os gestores não conseguem tomar as melhores decisões, afetando todo o processo de gestão da organização. Uma informação tem qualidade quando é relevante, precisa, acessível, concisa, clara, quantificável e consistente.

Uma grande quantidade de informações é tratada pelas organizações diariamente, estas informações são extraídas, processadas, armazenadas e disseminadas à todos os usuários das organizações, tanto internos quanto externos. Uma parte destas informações destina-se ao apoio das operações diárias das organizações e a outra parte será utilizada para auxiliar os gestores em suas tomadas de decisões em todos os níveis da organização.

Para Castro (2002, p. 29),

> A criação, captação, organização, distribuição, interpretação e comercialização da informação são processos fundamentais, enquanto que, a tecnologia utilizada para apoiar estes processos pode ser considerada menos importante do que a informação contida nos sistemas. A informação é dinâmica e capaz de criar grande valor para as organizações.

Beuren (2000, p.67-68 apud CELLA, 2006, p. 146) destaca que: "... para assegurar o valor estratégico da informação, na fase de execução dos planos organizacionais, precisa haver um processo coordenado de todas as etapas do gerenciamento da informação.".

Em vista disto, os responsáveis pelos projetos de tecnologia da informação devem atender às novas regras de negócio promovendo as alterações organizacionais necessárias para alcançarem seus objetivos. A criatividade e inovação nas formas de identificar as fontes de informação são de extrema importância uma vez que a informação que se deseja nem sempre está disponível nas tradicionais fontes de informação.

2.2 SISTEMAS DE INFORMAÇÃO

Devido a rápida evolução tecnológica, as mudanças e pressão infringidas pelo mercado, torna-se essencial que os gestores das IES tenham agilidade e versatilidade em suas decisões, porém, ressalta-se que para isto, os gestores necessitam de informações cada vez mais precisas e atualizadas.

Para Beuren (2000, p.39) "... o sistema de informação é o encarregado de prover informações, em todas as etapas do processo de gestão (planejamento, execução e controle), para os diferentes níveis hierárquicos e áreas funcionais da empresa.".

Uma forma de manter-se preparado, tendo uma visão integrada da organização, é fazer uso do SI. A crescente evolução das tecnologias tem possibilitado a criação de SI, preocupados como processo de geração das informações.

Para Cella (2006, p. 149)

> O sistema de informações é dependente do sistema de gestão de uma instituição, todos os esforços para o desenvolvimento da arquitetura e do sistema de informações devem concentrar esforços na identificação das informações necessárias para o processo de gestão empresarial e na determinação dos respectivos subsistemas que darão suporte a gestão.

A criação de um ambiente organizacional em que as informações sejam confiáveis e tenham fluência na estrutura da organização é o maior objetivo quando se utilizam de SI.

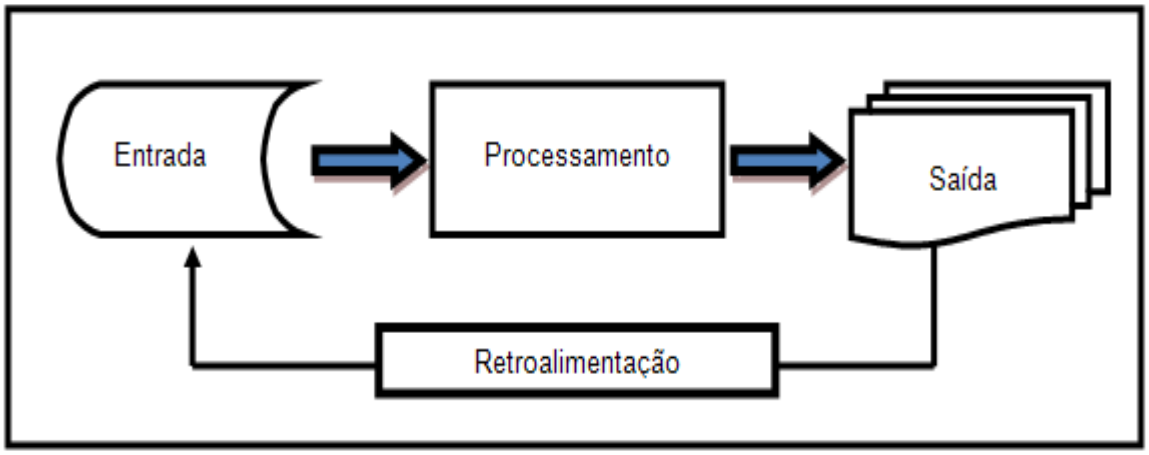


Figura 4 - Integração dos Sistemas de Informação
Fonte: Adaptado de Stair e Reynolds (2008, p. 6).

Para Stair e Reynolds (2008, p. 12) um SI "é um conjunto de elementos ou componente inter-relacionados que coletam (entrada), manipulam (processo) e disseminam (saída) dados e informações e oferecem um mecanismo de realimentação para atingir um objetivo", como observado na Figura 4.

Cella (2006, p. 146) corrobora enfatizando que:

> [...] um sistema de informação deve estar devidamente compatibilizado com a estrutura de autoridade, de decisões e de responsabilidade pela execução de atividades estabelecidas pela organização, de tal forma que, as informações destinadas a formular os planos, executar as funções e avaliar o desempenho sejam estruturadas de acordo com os objetivos das unidades organizacionais e comunicadas em tempo hábil às pessoas certas.

Para Laudon e Laudon (2001, p. 21), os SI são divididos nos seguintes níveis:

Nível Operacional: Neste nível se encontram os SI que gerenciam as atividades primárias e transacionais das organizações. Estes SI têm como objetivo responder a questões de rotinas e fluxo de transações;

Rezende e Abreu (2009, p.111) destacam que o SI deste nível:

> Cria condições para a adequada realização de trabalhos diários da empresa, onde o nível operacional de influência considera uma parte bem específica da estrutura organizacional da empresa. Neste caso, o nível de informação é *detalhada* (analítica), contemplando pormenores específicos de um dado, de uma tarefa ou atividade.

Nível Especialista: aqui se encontram os SI que auxiliam os funcionários especializados de uma organização. Seu objetivo é auxiliar a organização na aquisição e

integração de novos conhecimentos aos seus negócios e a organizar o fluxo dos papéis dentro da organização;

Gouveia e Ranito (2004, p. 58-59) consideram que os SI deste nível são:

> São sistemas de informação que suportam o trabalho de quem lida com dados e com conhecimento. Têm que permitir a integração de novo conhecimento no negócio, logo devem ser muito flexíveis, bem como permitir o controlo de fluxo do trabalho, sendo assim, fáceis de utilizar e não obrigarem não obrigarem a grande «desvios» do trabalho normal para que se faça a recolha de informação. Caso contrário, as pessoas tendem a não os usar, o que deita por terra todo o interesse dum sistema deste tipo.

Nível Administrativo: Os SI deste nível trabalham com as atividades administrativas de nível médio dentro da organização e têm como objetivo gerir e monitorar a informação para os gerentes deste nível;

Fialho (2001, p.68) menciona que neste nível "os sistemas de informações gerenciais servem às funções de planejamento e tomada de decisão. Apresentam relatórios sumarizados com informações condensadas".

Nível Estratégico: Neste nível se encontram os SI que auxiliam as atividades de planejamento de longo prazo e o objetivo destes SI é adequar as mudanças ocorridas no ambiente externo com a capacidade organizacional existente.

Neste nível os SI, vão além das informações gerenciais tradicionais, nas quais são produzidos apenas relatórios. Os SI deste nível fornecem auxílio imediato na resolução de problemas complexos e que não podem ser assistidos pelos SI do nível administrativo, sugerindo alternativas e possibilitando condições ideais às tomadas de decisões finais.

Laudon e Laudon (2001, p. 27) classificam os sistemas de informação de acordo com o tipo de problema organizacional que eles resolvem:

> Sistemas de nível estratégico são sistemas de informação utilizados para o nível de decisão, contribuindo para o planejamento estratégico da organização. Seu propósito é contabilizar as mudanças no ambiente externo com as capacidades organizacionais existentes; Sistemas táticos são sistemas de suporte gerencial, usados para resolver questões que envolvem controles e avaliação do processo de atingimento de objetivos; Sistemas de conhecimento são usados para resolver questões que envolvem conhecimento de especialidades técnicas, dando suporte aos funcionários especializados com o propósito de ajudar a empresa a integrar novos conhecimentos ao negócio; Sistemas operacionais são os sistemas usados para resolver problemas relacionados à operação, serviço e produção, respondendo as questões de rotina e fluxo de transações.

Os SI tendem a serem flexíveis, uma vez que as funcionalidades neles implementadas devem ser parametrizáveis, de forma a garantir uma atualização contínua as necessidades das organizações, sem que sejam necessárias substituições ou a reescrita destes SI. Devem ainda suportar a tomada de decisões individuais ou coletivas, abrangendo as várias competências e conhecimentos dos gestores envolvidos no processo.

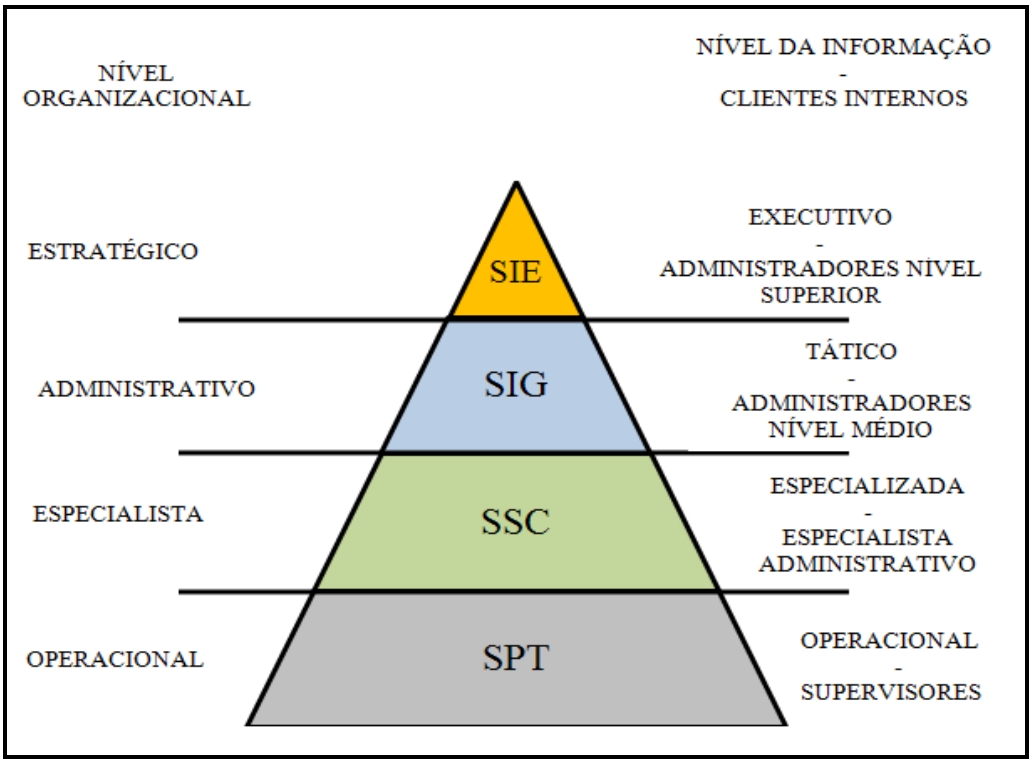


Figura 5 - Relação entre SI e seus níveis de abrangência dentro das organizações
Fonte: Adaptado de Laudon e Laudon (2001)

Na medida em que o nível organizacional é elevado, as informações tornam-se mais refinadas e com propósitos mais estratégicos dentro da organização. Quanto mais se aproximam do nível estratégico da organização, mais alto é o impacto das informações nos objetivos e maior é a especificidade dos problemas.

2.2.1 Sistemas de Informação e seus tipos

Existe uma concordância sobre a divisão dos tipos de SI em relação as suas funções administrativas. Autores como Laudon e Laudon (2001) , Rodrigues (1996), Dalfovo (2007), Stair e Reynolds (2008) classificam o SI em:

a) Sistema de Processamento de Transações (SPT);

b) Sistema de Automação de Escritórios (SAE);

c) Sistema de Informação Gerencial (SIG);

d) Sistema de Informação de Suporte à Tomada de Decisão (SSTD) e

e) Sistema de Informação para Executivos (SIE).

A atenção cairá sobre as informações gerenciais, em virtude da dificuldade de definição o que as torna um grande aliado na gestão de qualquer organização.

2.2.1.1 Sistema de Processamento de Transações (SPT)

Stair e Reynolds (2008, p.20) conceituam o SPT como: "um conjunto de pessoas, procedimentos, softwares, base de dados e dispositivos usados para registrar as transações completas de negócios.

Para Rezende e Abreu (2009, p. 114)

> Nos Sistemas de Informações Operacionais, cada transação empresarial envolve a entrada e a alimentação de dados, o processamento e o armazenamento, e a geração de documentos e relatórios. Com suas inúmeras características, como o grande volume de dados, muitas saídas de informações, envolvendo alto grau de repetição e computação simples.

O SPT é utilizado para registrar transações diárias de negócios, automatizar rotinas de gestão administrativa que anteriormente representavam tarefas consideradas intensas, abrangendo o processamento de grandes massas de dados.

Para Maccari e Sauaia (2004) "Este tipo de sistema implementa procedimentos e padrões para assegurar uma consistente manutenção dos dados e tomada de decisão". Ele garante que os dados trocados sejam consistentes e estejam a disposição de que deles necessitar. É utilizado para efetuar as transações entre clientes e a organização. Exemplo: Folha de pagamento, o qual foi considerado o primeiro SPT a ser utilizado nas organizações.

2.2.1.2 Sistema de Automação de Escritórios (SAE)

São sistemas direcionados aos funcionários que trabalham no "escritório" da organização. São sistemas informatizados tais como: processadores de texto, planilhas eletrônicas, sistemas de correio e agendamento eletrônico. Suas principais características são o aumento da produtividade dos funcionários e a troca de dados.

Para Gouveia e Ranito (2004, p. 59) “são sistemas de computador destinados ao aumento da produtividade do trabalhador de dados – pessoal administrativo – que tende a processar informação em vez de a criar (incluindo o seu uso, manipulação e disseminação)”.

2.2.1.3 Sistema de Informação Gerencial (SIG)

O SIG é direcionado à tomada de decisões estruturadas. A coleta de seus dados é feita internamente na organização e tem por base os dados primários existentes na organização. Sua principal característica é a utilização somente de dados estruturados.

De acordo com Stair e Reynolds (2008, p. 21) o foco de um SIG “é basicamente a eficiência operacional”. Ele pode auxiliar a organização a atingir suas metas, dando suporte ao nível gerencial por meio de relatórios e gráficos, de modo que seus gerentes possam ter controle e planejar as operações de forma mais eficiente.

Oliveira (1992, apud DALFOVO 2007, p. 29), relata que os SIG “são voltados aos administradores de empresas que acompanham os resultados das organizações semanalmente, mensalmente e anualmente, eles não estão preocupados com os resultados diários”.

Os SIG tornam o plano de atuação organizacional mais fortalecido, haja vista que por meio do recebimento dos dados e da geração destes em informações úteis, o processo de tomada de decisão possibilita a gestão da organização de forma mais estratégica e, por conseguinte resulta em vantagem competitiva sustentável em relação as organizações concorrentes.

Oliveira (2000, p. 183) define o SIG como:

> [...] um método formal de tornar disponíveis para a administração, oportunamente, as informações precisas necessárias para facilitar o processo de tomada de decisão e para dar condições para que as funções de planejamento, controle e operacionais da organização sejam executadas eficazmente. O sistema fornece informações sobre o passado, o presente e o futuro projetado e sobre eventos relevantes dentro e fora da organização.

Rezende e Abreu (2009, p. 114) corroboram relatando que os SIG ”trabalham com os dados agrupados (ou sintetizados) das operações funções empresariais da empresa, auxiliando a tomada de decisão do corpo gestor ou gerencial das unidades departamentais, em sinergia com as demais unidades”.

Direcionado para os níveis gerenciais e para as unidades de negócios, os SIG têm como foco a eficiência operacional, incorporando informações comuns coletadas nas bases de dados alimentadas pelo SPT. As entradas do SIG são dados internos ou externos à própria organização.

Os dados internos, são dados pertinentes a cada área da organização, sendo que são importantes para a integração das atividades no geral. Já os dados externos, podem-se citar como dados de oriundos de fornecedores, clientes, instituições financeiras, concorrentes, dentre outros.

Heizmann (2002, p. 42) enfatiza que:

> As informações geradas por este sistema, são voltadas aos administradores e gerentes e se apresentam na forma de relatórios resumidos de rotinas sobre o desempenho da empresa, sendo utilizados para acompanhar os resultados das operações da organização, trazendo benefícios como: a melhoria da produtividade e serviços e redução de custos, além de possibilitar previsões futuras.

O objetivo dos SIG é o fornecimento de informações aos gerentes de nível médio, a fim de que estes possam tomar decisões sobre suas áreas de atuação. Estes sistemas normalmente fornecem relatórios pré-programados com informações oriundas dos STP.

Resume-se então o SIG ao processo de transformação de dados em informações. E, uma vez que esse processo esteja direcionado para a geração de informações que são necessárias e utilizadas no processo decisório da organização, diz-se que esse é um sistema de informações gerenciais.

2.2.1.4 Sistema de Informação de Suporte à Tomada de Decisão (SSTD)

Dalfovo (2007, p.32) classifica estes sistemas como:

> São sistemas voltados para Administradores, tecnocrátas especialistas, analistas e tomadores de decisão. São sistemas de acesso rápido, interativos, orientados para ação imediata. As características são flexíveis, com respostas rápidas; permitem um controle para municiar a entrada e saída dos dados; e um instrumento de modelagem e análise sofisticado.

Os SSTD funcionam como base na tomada de decisões, possuem uma grande quantidade de dados e diversas ferramentas para manipulação destes, o que permite uma

flexibilização e adaptação ao meio em que se encontra, proporcionando uma capacidade maior nas respostas oferecidas.

O uso da Tecnologia de Informação afeta diretamente o desempenho organizacional, para garantir o sucesso no meio, as organizações estão dependentes destas ferramentas, tendo como conseqüência a utilização dos SI alinhados com o planejamento estratégico das organizações, os benefícios competitivos gerados por sua utilização, fato este que leva seus concorrentes a se automatizarem, caso queiram permanecer no mercado.

2.2.1.5 Sistema de Informação para Executivos (SIE)

Por fim, os SIE, também chamados de Sistemas de Suporte à Decisão Estratégica (SSDE), *Decision Suport System (DSS)* ou ainda Sistemas de Apoio à Decisão (SAD), que são sistemas que dão suporte as atividades do nível estratégico.

Segundo Gouveia (2009, p. 22)

> [...] o processo de tomada de decisão com auxílio de computadores iniciou na década de 70, onde os processos começaram a ser informatizados e as informações passaram a ser pré-definidas e selecionadas por meio dos Executive Information Systems (EIS).

Hadda (2007, p. 62) relata que os SIE "podem alterar radicalmente o processo de tomada de decisão e aumentar a produtividade e a acuracidade das decisões tomadas pelos gestores". Esta alteração dá-se em função das informações apresentadas, as quais são oriundas das diversas áreas da organização.

Stair e Reynolds (2008, p. 393) definem o SIE como:

> [...] uma coleção organizada de pessoas, procedimentos, softwares, bases de dados e dispositivos utilizados no apoio a decisões e à resolução de problemas específicos. O foco de um DSS é na eficiência da tomada de decisões diante de uma situação em que são apresentados problemas não estruturados ou semi-estruturados.

Para Rezende e Abreu (2009, p. 115) os SIE "contemplam o processamento de grupos de dados operacionais e transações gerenciais, transformando-os em informações estratégicas". O uso deste tipo de sistema é importante devido ao fato de que as ferramentas de apoio que visam alavancar o crescimento dos negócios das organizações são cada vez mais necessárias os gestores.

Os SIE são direcionados aos gestores que tenham pouco ou quase nenhum contato com SI automatizados. Suas características consistem na combinação de dados internos e externos; a apresentação de relatórios muitas vezes em forma de gráficos; acesso a banco de dados internos e externos.

De acordo com Dalfovo (2007, p. 27)

> As informações necessárias que os executivos precisão são visualizadas no E.I.S. através de formas numéricas, textual, gráficas ou por imagens. Com a utilização do E.I.S. pode-se visualizar estas informações desde o nível operacional até nível analítico, de uma forma segura e rápida, possibilitando um melhor conhecimento e controle da situação, possibilitando uma maior agilidade e segurança no processo decisório.

Os SIE permitem aos gestores fazerem o acompanhamento diário dos resultados, elaborando por meio de tabulações de dados de todas as áreas funcionais da organização, finalizando com a exibição destes resultados em forma de gráficos. O que antes dos SI levava-se dias para ser feito, agora com o uso dos SIE pode ser obtido em poucos segundos.

Vedovelli (2005, p. 59) baseado em Stair (1998) e Pozzebon (1997), descreve algumas características desejáveis aos SIE:

a) **Facilidade de uso:** Os SIE devem ser fáceis no aprendizado e em sua utilização;

b) **Manipulação de dados externos e internos, qualitativos e quantitativos:** as informações fornecidas são extraídas tanto do ambiente interno como do externo e contem dados estruturados ou não;

c) **Execução de análises de dados:** as análises e simulações são efetuadas sobre metas a serem alcançadas;

d) **Alto grau de especialização:** as informações devem estar em formatos específicos, de acordo com a necessidade dos gestores;

e) **Fornecimento de flexibilidade:** Os SIE devem permitir alterações em razão das alterações ocorridas nos ambientes interno e externo;

f) **Recursos de comunicação:** a disseminação das informações entre gerentes e gestores deve ser instantânea e precisa, sendo que deve estar disponível a qualquer instante e lugar.

As organizações têm enfrentado um grande desafio que é a previsão dos problemas e a concepção de soluções práticas, com o intuito de alcançarem seus objetivos. Sobrevivem as organizações que estão bem informadas a respeito dos ambientes nos quais estão inseridas.

O uso de SIE é um dos fatores de melhoria na tomada das decisões estratégicas, o que permite que se obtenha uma vantagem competitiva sustentável em relação aos seus concorrentes.

Dalfovo (2007, p. 26) corrobora com o contexto dizendo que,

> Não é uma questão de modernidade para comandar a empresa por meio de computadores em vez de papéis, mas principalmente de flexibilidade e rapidez. Em função da complexidade do mercado, as empresas estão sendo obrigadas a agilizar seu processo de decisão.

Observa-se que existem diferentes tipos de SIS, para diferentes necessidades dentro das organizações. Estes diferentes tipos de SI auxiliam a organização na descrição e diagnóstico de suas operações, transações e servem de base para a tomada de decisões, assegurando de forma conjunta com a infra-estrutura de suporte, a função de captura, processamento e disseminação das informações.

Os gestores devem ter confiança e segurança quando da tomada de decisões e por meio do uso de SIE, o impacto a ser causado pode ser minimizado, pois será feito tendo como base informações mais precisas e coerentes com a decisão tomada.

Stair (2002, p. 19), reforça que:

> O foco de um Sistema de Suporte à Decisão incide sobre a eficácia da tomada de decisão. Enquanto um Sistema de Informação Gerencial ajuda a organização a "fazer as coisas certas", um SSD ajuda o gerente a "fazer a coisa certa", naquele momento.

Os SIE são sistemas que possibilitam a realização de simulações das situações reais a serem enfrentadas pelas organizações, o que tende a tornar a tarefa dos gestores mais precisa e confiável, haja vista a realização de "experiências virtuais", evitando possíveis erros nas tomadas de decisões.

Ressalta-se que os SI devem estar alinhados com as metas de negócio definidas, satisfazendo as necessidades das decisões. Estes SI devem também permitir o planejamento

de longo prazo, no qual haja a integração entre dados dos diversos níveis da organização, com objetivo maior de obter a vantagem competitiva sustentável.

2.3 GESTÃO DA INFORMAÇÃO

A fim de conceituar o termo, de acordo com Valetim (2006, p.18), a gestão da informação é um "conjunto de atividades para prospectar/monitorar, selecionar, filtrar, tratar, agregar valor e disseminar informação, bem como para aplicar métodos, técnicas, instrumentos e ferramentas que apóiem esse conjunto de atividades".

Dalfovo (2007, p. 57), retrata o estado em que se encontram os administradores e a necessidade do uso de sistemas de informação para o desenvolvimento estratégico das organizações:

> O desafio que os administradores enfrentam nos dias atuais, é o de prever os problemas e conceber soluções práticas para eles, a fim de realizar os anseios objetivados pela empresa. Os administradores precisam estar bem informados, pois a informação é a base para toda e qualquer tomada de decisão. Os sistemas de informação têm um papel fundamental e cada vez melhor em todas as organizações de negócios. Os sistemas de informação eficazes podem ter um impacto na estratégia corporativa e no sucesso organizacional. As empresas em todo o mundo estão desfrutando maior segurança, melhores serviços, maior eficiência e eficácia, despesas reduzidas, aperfeiçoamento no controle e na tomada de decisões devido aos sistemas de informação.

A gestão da informação é um instrumento que promove a compreensão da realidade dos mercados, das técnicas, dos concorrentes e da sua cultura, intenções e de sua capacidade, além de possuir uma relação estreita com a produtividade da organização. Assim, tornando a organização mais competitiva.

Conforme Stair (2006, apud DALFOVO, 2007, p. 58):

> [...] os Sistemas de Informação, hoje, são a última moda no mercado, ou seja, o recente aprimoramento da moda é utilizado nas estruturas de decisões da empresa e, quando corretamente aplicado, trará, certamente, resultados positivos às empresas. Caso contrário, torna-se difícil sua implementação até mesmo por seu alto custo. É necessário, porém, saber, antes de tudo, ao certo, aonde quer chegar, a necessidade os Sistemas de Informação, para que possam ser bem elaborados e desenvolvidos, tornando-se sistemas fundamentais e capacitados para a tomada de decisões da empresa.

A utilização dos recursos computacionais, como os sistemas de informações, vem a agregar mais valor ao "produto final" de qualquer organização.

O filósofo Aristóteles já afirmava que a necessidade de informação é derivada do desejo de saber, da curiosidade humana, portanto, determinada em função do conhecimento já adquirido.

Com o crescente avanço tecnológico, a informação está tendo um tratamento mais cauteloso, do que visto há alguns anos atrás. A sua disseminação e utilização não dependem única e exclusivamente daqueles que fazem uso da mesma. O uso de tecnologias deve ser tratado como mais profissionalismo e com mais seriedade pelos meios empresariais, educacionais, dentre tantos.

Em relação a estes cuidados, tem-se uma nova geração de técnicas, ferramentas computacionais aliadas, extraídas das ciências já existentes dentre as quais cita-se a Gestão do Conhecimento, conforme relata Dalfovo (2007, p. 63):

> Ela é utilizada atualmente para manusear, transformar, concatenar, aprimorar difundir informações entre as pessoas que fazem uso da mesma. Uma vez aprimorada, a informação transpôs vária barreiras e alcançou a rede mundial de computadores, de onde foi difundida para pessoas espalhadas por todo o mundo e adquiriu status de negócio eletrônico (e-business) e comércio eletrônico (e-commerce), permitindo assim ás empresas fazerem uso das mesmas sem a necessidade de deslocamentos desnecessários.

São exatamente estas técnicas que têm por objetivo o aprimoramento dos conceitos referentes à aquisição, assimilação e disseminação das informações nos meios educacionais, comerciais e empresariais.

Em um mercado onde a lei da competição predomina, é muito importante que existam fontes de conhecimento e informação eficientes, já que eles são o caminho para a melhoria contínua de todas as organizações.

Para Stata (1997, p. 392), a interconexão de alguns elementos, faz com que haja a geração do conhecimento:

> Sistemas de informações gerenciais transformam dados em informações e depois ajudam os gerentes a transformar informações em conhecimento, e conhecimento em ação. O desafio está em decidir que informação e conhecimento – e em que forma – são necessários. Se tivermos a aprendizagem organizacional em mente como um dos objetivos no desenho dos sistemas de informação, teremos maior probabilidade de gerar informações e o conhecimento que os gerentes necessitam para tomar ações efetivas".

Informação e Gestão da Informação, hoje são o foco de diversas abordagens e discussões de como as organizações podem obter vantagens sobre seus concorrentes, a chamada vantagem competitiva.

Para as organizações da esfera privada, o uso destas abordagens significa melhorias na qualidade de seus produtos e serviços, aumento dos índices de satisfação de seus clientes, inovação e elevação da produtividade, gerando aumento nos índices de rentabilidade e desempenho das organizações.

Devido a natureza competitiva atual do mercado globalizado, fica evidente para os gestores que não há mais espaço para erros. As conseqüências relativas a execução de uma estratégia errônea ou a implantação e gestão incorretas de novos negócios, sem o apoio da inteligência competitiva podem ser extremamente graves.

A gestão da informação representa hoje uma área onde o conhecimento está imergindo, focando sua concentração cada vez mais nas dimensões econômicas, auxiliando a novos e velhos negócios a manter e ou conquistar suas vantagens competitivas.

Alvarenga (2006, p. 45) faz uma síntese das principais características da Gestão da Informação, conforme se observa no Quadro 1, onde resume em sete tópicos as características necessárias para a gestão organizacional com base nas informações.

Características	Parâmetros importantes
1) Necessidade de informação diante das escolhas.	Quanto maior o grau de informações acerca das decisões que deverão ser tomadas, maior a probabilidade de acerto na decisão.
2) A informação como medida de organização de um sistema.	Sistemas ditos organizados devem preocupar-se com as informações necessárias e não com a quantidade de informações neles inseridas.
3) Informações confiáveis	A identificação das fontes informacionais e sua qualidade precisam ser analisadas antes da utilização destas informações na tomada de decisão.
4) Comunicação como apoio para a disseminação da informação.	O processo de comunicação é fundamental no processo de desenvolvimento e disseminação da informação.
5) Sistemas que proporcionem agilidade e segurança.	O cenário tecnológico atual obriga a disseminação das informações por meio de mecanismos ágeis e seguros. Apoio nas tecnologias de comunicação (informática).
6) Necessidades de conhecimento dos profissionais acerca dos conceitos de informação.	A falta de conhecimento dos conceitos ocasiona dificuldade na definição das informações corretas para o desenvolvimento das funções, excesso de trabalho e falta de compreensão para com outros usuários.

Continua...

Conclusão.

7) Capacidade de definição das informações necessárias.	Nem toda informação que a empresa gera ou adquire é importante para determinada situação ou decisão. Conhecer as informações que a empresa possui proporciona agilidade na tomada de decisão e evita perda de foco.
8) Necessidade de conhecimento das informações gerenciais.	Por serem utilizadas especificamente para a tomada de decisões, precisam ser conhecidas, principalmente pelo nível estratégico da organização (Tomadores de decisão).

Quadro 1 - Características da informação
Fonte: Alvarenga (2006, p. 45)

A realização destes processos e adaptação aos conceitos da organização traz benefícios não apenas ao fator competitivo, mas a organização acaba recebendo uma boa imagem no mercado e perante os consumidores.

Os gestores das organizações têm a sua disposição diversas ferramentas para análise da informação, dentre as quais destacam-se: Workflow, Data Mining, Data Mart, Data Warehouse, CRM, OLAP, além de softwares desenvolvidos especificamente para suas organizações.

Observa-se que muitos dos processos básicos utilizados numa ação organizacional são dependentes das informações que são tratadas por estes processos.

Em virtude disto muitas organizações têm destinado grandes esforços, sejam financeiros ou por meio de capital humano para gerir estas informações, fazendo para isto uso de tecnologias da informação.

2.4 A IMPORTÂNCIA DOS SIG NA GESTÃO ESTRATÉGICA

O constante processo de mudanças gerado pela sociedade contemporânea exige que as decisões administrativas sejam cada vez mais rápidas e precisas, com isto as IES vêem seus paradigmas passarem por modificações constantes.

As IES são consideradas como organizações com um elevado grau de complexidade, devido ao fato de realizarem e ensinarem atividades de múltiplas finalidades. Estas atividades têm relação com o tripé: ensino, pesquisa e extensão, o que torna as IES uma das organizações mais complexas.

Para Alves (2005, p. 58) o impacto destas mudanças afeta a infra-estrutura das IES, e para a autora:

> A infra-estrutura de tecnologia da informação representa todos os recursos de hardware, software, telecomunicações e pessoal que podem ser partilhados em uma organização. É importante ressaltar que o projeto e a sua implementação devem conter os recursos tecnológicos necessários para dar suporte aos trabalhos a serem realizados.

Os SIE manipulam dados de diversas áreas da organização e têm como resultados informação tanto quantitativas como qualitativas, sendo estas utilizadas para avaliação de resultados e alcance dos objetivos estratégicos definidos.

A Fundação Nacional da Qualidade (FNQ) (2007, p. 8) salienta que:

> Além dos sistemas de indicadores de desempenho (informações que indicam, quantitativamente, a evolução e o nível de desempenho), são utilizados, freqüentemente, sistemas de informação que produzem informações qualitativas para avaliação de desempenho e tomada de decisão, como relatórios de auditoria, pareceres e avaliações especializadas, laudos técnicos e pesquisas de opinião e de monitoramento. O que caracteriza tais sistemas, portanto, é o seu emprego para a tomada de decisão.

Atualmente no mercado encontram-se várias ferramentas que são utilizadas para a análise de informações gerenciais e auxílio nas tomadas de decisões, dentre as quais se destacam: *Business Intelligence (BI), Customer Relationship Management (CRM), Enterprise Resource Planning (ERP), Workflow, Data Warehouse, Data Mining.*

Para Rezende e Abreu (2009, p. 164) “os gestores das organizações necessitam de Sistemas de Informações efetivos, ou seja, que processem grande volume de dados e produzam informações válidas, úteis e oportunas”.

Os SIG vêm se tornando ferramenta de fundamental importância nas organizações para a tomada de decisões, tendo como objetivos obter, estruturar e disseminar a informação gerada pela própria organização, o que antes era feito sem nenhuma estrutura e ao alcance de poucos, hoje se encontra estruturado e disponível a qualquer pessoa dentro da organização.

Beruen (2000, p.59), reforça este conceito dizendo que:

> Uma vez que a empresa reconhece o papel positivo que a informação pode representar, cabe a ela refletir sobre questões primordiais relativas à criação de processos eficazes de gestão da informação. Tal esforço poderia resultar no

> desenvolvimento e implementação de uma arquitetura da informação, que promova uma postura eficaz no atendimento das necessidades de informações dos gestores.

A qualidade afeta diretamente no valor da informação e a tomada de decisões, haja vista que com uma informação sem qualidade os gestores não conseguem elaborar suas decisões de forma eficaz, isto afeta diretamente todo o processo de gestão.

Tanto as IES como qualquer outra organização, devem desfrutar dos benefícios oferecidos pela tecnologia e mais especificamente pelo uso dos SIG. Para tanto devem abandonar velhos hábitos de trabalho e gestão e adotar as novidades trazidas pelas ferramentas e sistemas a disposição.

Bernardes e Abreu (2004) reforçam a idéia sustentando que "Os sistemas de informação devem proporcionar às universidades um embasamento quantitativo e qualitativo nos seus planejamentos, nos processos de tomada de decisão e no estabelecimento das atividades no plano operativo".

A informação, sendo utilizada como um recurso estratégico, deve estar em constante interação com todos os níveis da IES. A qualidade, o valor e a segurança da informação passam a ser fundamentais para a IES nos processos de tomada de decisão.

De Mori (2008, p. 63) destaca que:

> O processo administrativo apresenta como elemento básico a tomada de decisão e, para que este processo seja adequado, é necessário dispor de um sistema de informações eficiente. Portanto, fica claro que as empresas que possuem um SIG adequado podem ter uma vantagem competitiva em relação às suas concorrentes, diminuindo o nível de risco, que é parte integrante e inseparável das decisões estratégicas, táticas e operacionais nas empresas.

A rápida evolução dos recursos tecnológicos e dos meios de comunicação, juntamente com a redução dos custos envolvidos no processo de armazenagem, processamento e disseminação das informações, torna a implantação de um SIG cada vez mais acessível, a fim de proporcionar soluções cada vez melhores para o gerenciamento da informação nas IES.

Uma maneira de contrapor o insucesso de uma IES é fazendo uso das tecnologias da informação e de suas ferramentas. Assim a gestão das informações e a tomada de decisões estarão baseadas em instrumentos tecnológicos que conseguem extrair de forma mais rápida e precisa as informações necessárias para a gestão estratégica da IES.

Oliveira Junior e Castro (2006) enfatizam que "A gestão estratégica da organização se tornará concreta pela empresa a partir do uso de ambientes, recursos e tecnologias específicas que possibilitem a sua execução."

Para os autores os recursos da TI, dentre os quais citam-se os SIG merecem maior atenção por parte dos gestores, pois é por meio destes recursos que se tem a geração e gestão da informação que servirá para a tomada das decisões.

Uma vasta quantidade de estudos e projetos abordando a Gestão da Informação, o uso de técnicas de *Datamining*, dentre outros assuntos relacionados a este projeto, são gerados nas IES, por meio dos trabalhos das disciplinas, trabalhos de conclusões de curso, mas observa-se que poucos têm sido utilizados para a gestão da IES que gerou estes estudos.

Ressalta-se que o sucesso ou fracasso de um SIG, está diretamente relacionado com a contextualização da IES na qual será utilizado, desde haja um ambiente onde sua implantação e uso sejam aceitos pelos seus usuários.

2.5 EXTRAÇÃO DA INFORMAÇÃO

Com a crescente quantidade de informações disponibilizadas diariamente, vê-se a disseminação do uso de técnicas e ferramentas de extração e manipulação de informações, os chamados Sistemas de Extração de Informações (SEI), para lidarem com este crescente volume de informações, nos seus diversos formatos.

A Extração de Informações (EI) tem por objetivo a localização e extração de informações consideradas relevantes em um documento ou coleção de documentos, a fim de estruturar estas informações dentro de um padrão de saída, geralmente em um banco de dados, para facilitar sua manipulação e posterior análise.

Para Zambedenetti (2002, p. 25-26)

> A extração de informações tem muitas aplicações potenciais. Por exemplo, a informação disponível em textos não-estruturados pode ser armazenada em bancos de dados tradicionais e usuários podem examiná-las através de consultas padrão.

Para se extrair a informação de um texto qualquer, percorre-se o texto em busca de determinados eventos que identifiquem elementos da busca. Procura-se nestes elementos outras informações que caracterizem o evento. Toma-se como exemplo, a evasão acadêmica:

busca-se no texto em referência eventos que indiquem o índice de evasão, os principais motivos que levam os alunos a evadirem das IES.

Para Barion e Lago (2008, p. 133) "O processo de extração de informação identifica palavras dentro de conceitos específicos e ainda contém um processo de transformação que modifica a informação extraída em um formato compatível com um banco de dados."

Scarinci (1997, p.22) explica o funcionamento de um SEI, no qual o SEI analisa um texto por várias vezes com objetivo de extrair do texto informações bem específicas, sendo que nesta busca da informação o SEI é condicionado a buscar informações consideradas relevantes, deixando de lado as que não têm importância para o usuário.

Para Cordeiro (2003, p. 17) a EI:

> [...] pretende identificar elementos relevantes no interior de determinados documentos, os quais já sabemos que contém a informação que nos interessa. Os elementos relevantes extraídos serão depois armazenados em alguma estrutura previamente definida, por exemplos numa tabela de uma Base de Dados.

A EI presta um grande serviço à mineração de dados, uma vez que por meio da EI as informações extraídas de uma base de dados são as consideradas mais relevantes para o usuário. O que resulta em tomadas de decisões com maior grau de certeza, haja vista que as informações desnecessárias e/ou redundantes são descartadas.

De acordo com Souza (2006, p. 163):

> "[...] há que se distinguirem os sistemas de recuperação de informações (SRI) dos sistemas de gestão de bancos de dados (SGBD). [...] No sentido estrito do conceito, nenhum programa de computador lida, sob o ponto de vista da máquina, com informações, a não ser que possua alguma capacidade de arrazoamento, e, assim mesmo, a utilização do termo dá margem a discussões. No uso corrente, porém, ambos os termos são utilizados para sistemas, apesar das diferenças entre os sistemas de recuperação de informações e sistemas de recuperação de dados, como os SGBDs".

Deve-se separar os dois conceitos, uma vez que o SGBD trata com tabelas e a forma de interação, consulta, é feita por meio de uma linguagem específica para tal, *Data Manipulation Language* (DML), Linguagem de Manipulação de Dados, trazendo como resultado apenas duas possíveis respostas, existe ou não existe um conjunto de dados que atendam a consulta. Enquanto no SRI, é passível que não exista apenas uma única resposta à consulta realizada, em virtude da incerteza associada ao documento analisado.

Portanto a informação tem hoje uma importância que cresce a cada dia. Ela tornou-se o elemento base para a organização, desde a aquisição, transformação até a sua utilização nas tomadas de decisões.

2.6 MINERAÇÃO DE DADOS

Tem-se observado um crescimento demasiadamente rápido do volume de dados armazenados nas mais diversas corporações, dados estes armazenados em banco de dados, destinados para os mais diversos fins. Toma-se como exemplo o banco de dados do projeto Genoma Humano que estima-se já conter uma quantidade de registros em torno de 10^9 objetos armazenados.

Devido à melhora da tecnologia da informação e o crescimento da Internet, as organizações são capazes de coletar e armazenar enorme quantidade de dados. Pessoas gradualmente estão percebendo que os dados não são iguais a informação, que os dados devem ser analisados e extraídos.

Profissionais são treinados para analisar e interpretar os dados, mas os aumentos na quantidade de dados, tipo de dados, e dimensões de análise, têm dificultado estas ações. A TI tem ido além do armazenamento, transmissão e processamento. Os dados precisam ser convertidos em informação e conhecimento para apoiar a tomada de decisão.

O principal desafio é como fazer com que os dados armazenados nos bancos de dados sejam convertidos de dados aparentemente sem sentido em informações úteis. Este desafio é crítico, porque as organizações estão cada vez mais contando com uma análise eficaz das informações simplesmente para se manterem competitivas.

Pela descoberta de conhecimento em bases de dados, o conhecimento interessante, regularidades, e informações de alto nível podem ser extraídos dos conjuntos de dados relevantes em bases de dados e estes elementos serem investigados a partir de diferentes perspectivas.

Diante deste contexto as aplicações em mineração de dados (MD) podem dar condições significativas às organizações, uma vez que lhes são oferecidos conhecimento e informações que permitem uma melhor tomada de decisão, ou seja, a MD é uma ferramenta de grande valor quando utilizada em questões de análise de informações gerenciais.

Analogamente a classificação dos SI, a mineração de dados está situada mais especificamente no nível organizacional de decisões gerenciais, conforme Figura 6.

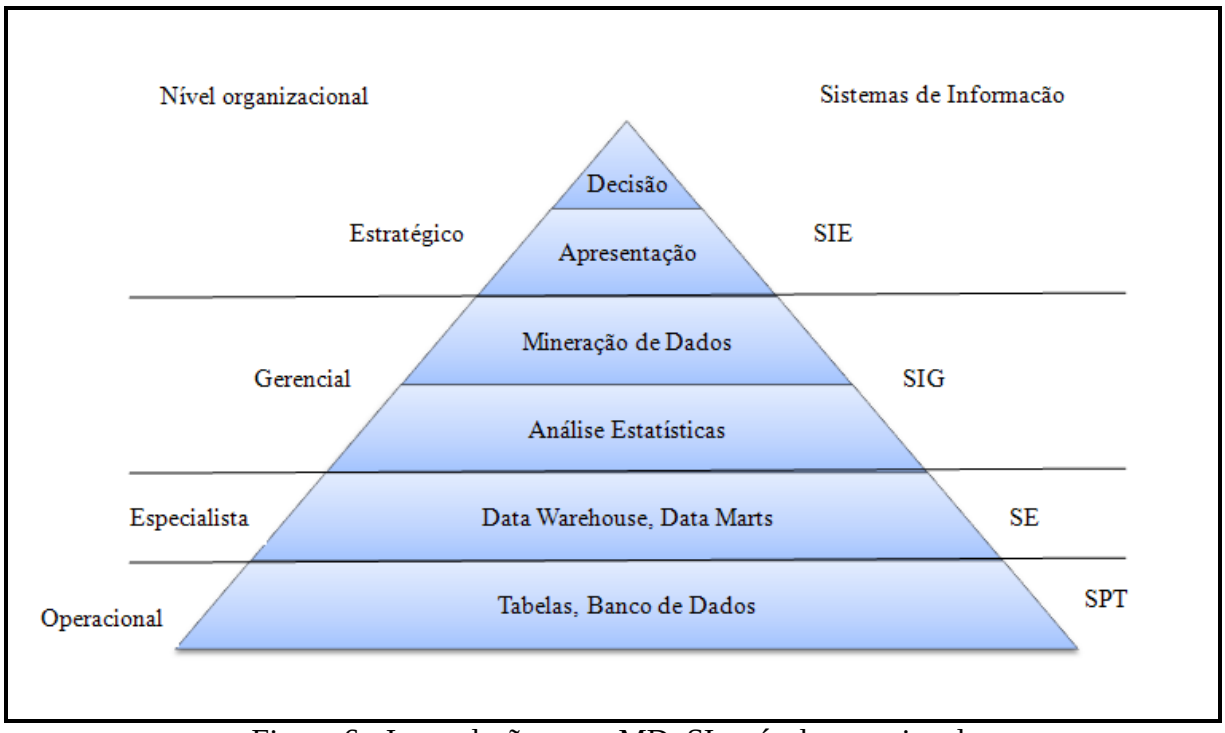


Figura 6 - Interrelação entre MD, SI e nível operacional
Fonte: Adaptado de Santos (2008).

A tecnologia tem proporcionado o que chamam de "A Era da Informação", cada vez mias milhares de informações estão sendo armazenadas nos bancos de dados das organizações em todo o mundo. Essas informações, que servem de base para a tomada das decisões, encontram-se implícitas no meio dos milhões de dados armazenados.

Shiba (2008) relata que esta capacidade de armazenagem é dada em função da oferta de recursos tecnológicos, onde a capacidade de armazenamento está cada vez maior, aliado ao desenvolvimento de softwares que dão suporte a esta função.

A autora cita ainda que:

> A alta disponibilidade de recursos para armazenamento de dados também permitiu às organizações um aumento significativo nos investimentos para a capacitação de seus ambientes no que se refere à captura, transformação e retenção de informações, dotando-os de softwares funcionalmente capazes de suportar todo o fluxo das transações de negócio. (SHIBA, 2008, pg. 21)

Os dados fornecidos pelos ambientes de aprendizagem virtual são analisados sob a ótica de informações meramente estatística, sobre o acesso aos cursos, conteúdos, quantidade de acessos, etc., restringindo e limitando assim a capacidade de compreensão implícita nas

informações sobre as mais variadas tendências de utilização e a percepção das possibilidades de vantagens competitivas que possam ser obtidas com base em seu conteúdo.

As inovações tecnológicas na área de armazenamento de dados, bem como sua utilização, vêm crescendo proporcionalmente em relação aos avanços das novas tecnologias de informação e comunicação, as chamadas TIC's. A extração de informações que sejam relevantes aos interesses dos gestores, está se tornando complexa diante da quantidade de dados armazenados. Denomina-se *Knowledge Discovery in Databases* – KDD (Descoberta de Conhecimento em Bases de Dados), a atividade de "garimpar" a informação contida nestes dados.

Apesar de ser comum usar os termos KDD (Knowledge Discovery in Databases) e Mineração de Dados com o mesmo significado, Fayyad et al.(1996) definem o KDD como sendo o processo da extração de conhecimento dos dados como um todo, e *Mineração de Dados*, como apenas uma etapa em particular do KDD, sendo que nesta etapa para a extração de padrões dos dados é realizada através do uso de algoritmos específicos.

Descobrir o conhecimento oculto nas grandes bases de dados das mais diversas organizações, seja de forma automática ou semi-automática é o objetivo do *Mineração de Dados*, além de permitir uma maior agilidade no processo de tomada de decisão por parte dos gestores.

"KDD é um processo não trivial para identificar padrões válidos, novos, potencialmente úteis e compreensíveis em dados existentes". (FAYYAD, PIATETSKY-SHAPIRO E SMYTH, 1996 apud BATISTA, 2004, p.32).

O KDT (descoberta do conhecimento em textos (*KDT - Knowledge Discovery from Text)*) é igual ao KDD, ou seja, é iterativo e interativo transformando dados de baixo nível em conhecimento de alto nível. (FURTADO, 2004, p. 29).

A diferença entre estes dois conceitos é feita da seguinte forma: o *KDD* utiliza-se de uma base de dados, tabulados e estruturados, para extrair o conhecimento, enquanto o *KDT* extrai o conhecimento de dados não tabulados e estruturados.

Várias atividades estão relacionadas ao KDD, que por sua vez contribuiu em várias áreas, dentre as quais se destacam: a estatística, o aprendizado de máquina, a área de banco de dados e a inteligência computacional. (GOLDSCHMIDT e PASSOS, 2005)

"A descoberta de conhecimento em bases de dados é multidisciplinar e, historicamente, se origina de diversas áreas, dentre as quais podem ser destacadas a estatística, inteligência computacional, reconhecimento de padrões e banco de dados." (BOENTE, 2006 apud BOENTE, OLIVEIRA E ROSA, 2007, p. 3).

Goldschmidt e Passos (2005, p. 6) ordenam as atividades anteriormente citadas do KDD em três grupos:

a) desenvolvimento tecnológico: reúne as fases e concepção e desenvolvimento de algoritmos, ferramentas e tecnologias, com o objetivo de serem empregadas em bases de dados para a aquisição de novos conhecimentos;

b) execução de KDD: resume-se na busca do conhecimento;

c) aplicação dos resultados: refere-se ao uso das informações obtidas pelo processo de KDD.

Por ser um processo contínuo e cíclico o *KDD*, permite que os seus resultados sejam refinados e melhorados a medida que são analisados. Para esta melhoria, alguns autores estabelecem os passos as serem seguidos, dentre os quais destacam-se Fayyad, Piatetsky-Shapiro e Smyth, ressalta-se que apesar dos passos serem seqüenciais, pelo fato do processo ser interativo e iterativo, pode-se rever cada etapa a qualquer momento, dando ao processo uma maior flexibilidade e consequentemente uma melhoria nos resultados, conforme visto na Figura 7.

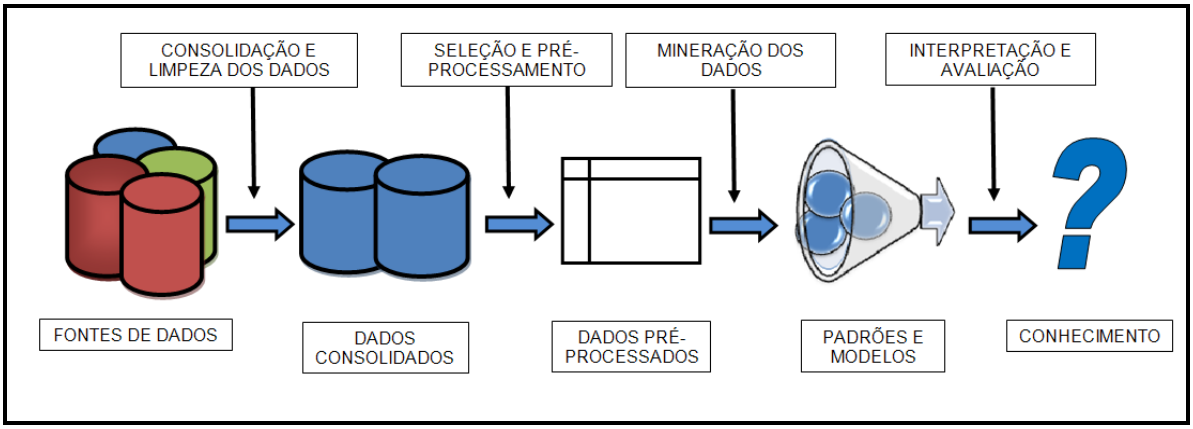


Figura 7 - Etapas do KDD
Fonte: Adaptado de Figueira (1998, p. 8.)

"Os passos da Mineração de Dados podem ser vistos como um subprocesso dentro do KDD. Eles consistem em uma preparação mais refinada dos dados provenientes das etapas

anteriores, na aplicação de algoritmos apropriados e na pré-avaliação dos resultados" (HORST e MONARD 2000, apud CHIARA 2003, p. 8).

Ferreira (2008, p. 29) destaca as etapas do desenvolvimento do KDD:

> KDD é um processo desenvolvido em três etapas básicas: pré-processamento, que objetiva a análise, integração, transformação e limpeza dos dados; Data Mining, que se refere à aplicação de algoritmos de extração de padrões; pós-processamento, que consiste na seleção e ordenação das descobertas, representação inteligível do conhecimento e geração de relatórios.

Na etapa denominada de **consolidação dos dados**: tem-se a coleta e consolidação dos dados que dará início ao processo de extração do conhecimento.

Nesta etapa se deve definir quais são as perguntas e decisões que se encaminham para a fase de garimpagem das informações. É uma fase muito importante, na qual deve-se manter o foco nas informações estratégicas, de interesse, assim como a adaptação à realidade dos usuários.

Para Shiba (2008, p. 36):

> Esta etapa também é conhecida como Preparação de Dados, e por envolver uma série de atividades até a sua finalização, que envolve inclusive o estudo de processos, acaba se tornando a etapa que exige maior esforço dentro de um projeto de extração de conhecimento.

Na etapa de **seleção e pré-processamento**, o objetivo é a melhora na qualidade e transformação dos dados, afim de evitar possíveis distorções na extração do conhecimento, se necessário os dados devem ser transformados, a fim de facilitar e eliminar possíveis barreiras para a etapa seguinte, a de mineração de dados.

Scoss (2006, p. 24) define que: "...neste processo realiza-se uma avaliação da base de dados que será trabalhada, verificando as inconsistências das informações ali armazenadas, como por exemplo: dados duplicados, faltantes, impossíveis de serem analisados, entre outros".

Shiba (2008, p. 36) defende que "o pré-processamento deve eliminar a diferença de tipos nas variáveis que representam um mesmo conceito, ou seja, uniformizar os atributos, que muitas vezes foram extraídos de bancos de dados distintos".

Na etapa de **mineração de dados,** são definidos quais algoritmos serão utilizados na extração do conhecimento, sendo que estes podem várias tarefas, tais como: classificação, agrupamento, regressão, associação e sumarização.

Chiara (2003, p. 8) destaca:

> ...por ser considerado um dos passos cruciais e mais complexos do KDD, a Mineração de Dados também pode ser considerada como um processo que, por sua vez, pode ser dividido em várias sub-etapas. Basicamente, os algoritmos a serem utilizados devem ser escolhidos de acordo com o problema que está sendo atacado (dados categóricos × dados reais; modelos descritivos × modelos preditivos). Normalmente existem vários métodos para um mesmo objetivo de KDD e a fase de Mineração de Dados inclui a aplicação de diversas técnicas assim como a avaliação e a comparação dos resultados obtidos.

Kanashiro (2007, p.21) relata que: "Alguns parâmetros, como o tipo de tarefa de mineração de dados e a forma como os padrões serão representados, são determinados pelos interesses do usuário final e conseqüentemente influenciará na escolha do algoritmo".

Na última etapa, **interpretação e avaliação,** busca-se analisar os resultados obtidos para o julgamento do modelo obtido da fase anterior. Nesta etapa também busca-se criar uma forma de interpretar os resultados visando a leitura direta dos mesmos.

Nesta etapa são avaliados os resultados obtidos da mineração quanto a sua qualidade e utilidade e relevância. Uma nova filtragem é feita removendo as informações consideradas irrelevantes e redundantes, para serem utilizadas pelo usuário final, conseqüentemente o conhecimento extraído é disponibilizado para os gestores usarem em suas decisões.

Isto vem ao encontro de Hiraghi (2006, p. 26), que conclui que:

> A avaliação é o momento de mensurar a qualidade da mineração de dados realizada, a partir da análise de performance dos modelos obtidos. Também são verificados se os objetivos do negócio foram alcançados. Normalmente, com base nos resultados obtidos na avaliação o processo de mineração é revisado podendo ser retomadas fases anteriores.

Finalizando esta etapa, os resultados obtidos são disponibilizados por meio de visualizações nas mais diversas formas.

Kanashiro (2007, p. 19) destaca que:

> A descoberta de conhecimento em bases de dados também é considerada um processo interativo, no qual existe a necessidade do conhecimento de domínio da

> aplicação do usuário que é utilizado desde a preparação dos dados na etapa de pré-processamento, na execução do processo de mineração de dados e na validação do conhecimento extraído.

O emprego das técnicas de *Mineração de Dados*, permite as organizações criarem parâmetros capazes de entender o comportamento dos dados armazenados, permite também a identificação das afinidades existentes entre estes dados, além de proporcionar a previsão de comportamentos e hábitos dos dados.

De acordo com Quoniam et al (2001):

> As ferramentas *Data Mining* identificam todas as possibilidades de correlações existentes nas fontes de dados. Através das técnicas para exploração de dados, pode-se desenvolver aplicações que venham a extrair, dos bancos de dados, informações críticas, com o objetivo de subsidiar plenamente o processo decisório de uma organização.

Com base no anteriormente exposto, vislumbra-se a necessidade de um acompanhamento mais profundo por parte das instituições e aqueles que participam do processo ensino-aprendizagem, onde o acompanhamento das atividades, a análise das informações contidas nas bases de dados, pode direcionar para quais medidas devam ser tomadas com relação ao andamento dos cursos, da participação dos alunos e professores. Em afirmação ao que dizem Quoniam et al (2001):

> O objetivo é estimar possíveis mudanças e melhorias necessárias no conteúdo e estrutura do curso, e de suas atividades, com o intuito de minimizar desorientações que poderão ocorrer durante o acesso às páginas e recursos do curso online, além de descobrir modelos de aprendizagem similares.

Alinhado ao pensamento de Fayyad (1996) a descoberta de conhecimento pode ser obtida por meio de complexas interações realizadas entre homem e uma base de dados, geralmente por meio do uso de uma série heterogênea de ferramentas.

A combinação de mecanismos como *Data Warehouse*, fluxo de trabalho, controle de versão e banco de dados propicia aos usuários um ambiente de trabalho único, a partir do qual se tem acesso não apenas aos objetos do Processo de Desenvolvimento de Software (conhecimento explícito), assim como a todos os documentos gerados, além de acesso a informações sobre os indivíduos adequados para realizarem determinadas tarefas. (DINGSOYR, 2002).

Fayyad, Piatetsky-Shapiro e Smyth (1996) afirmam que os algoritmos de mineração consiste basicamente de algum mix específico de três componentes:

a) O modelo: Há dois fatores relevantes. Eles são a função do modelo e a forma de representação do modelo. O modelo contém parâmetros que são determinados a partir dos dados;

b) O critério de preferência: A base para a preferência de um modelo ou conjunto de parâmetros sobre o outro, dependendo dos dados fornecidos. O critério é geralmente algum tipo de "função de bondade" de ajuste do modelo aos dados, talvez temperado por uma suavização, para evitar o excesso de montagem, ou gerando um modelo com muitos graus de liberdade a ser condicionada pelos dados fornecidos;

c) O algoritmo de pesquisa: A especificação de um algoritmo para encontrar modelos particulares e parâmetros, dados apresentados, um modelo, e um critério de preferência.

A escolha de quais técnicas de mineração de dados aplicar, depende da tarefa de mineração a ser realizada. As exigências das tarefas de mineração e as suas características influenciam a viabilidade entre os métodos de mineração e os problemas de negócio.

2.7 METODOLOGIA DE MINERAÇÃO DE DADOS

A metodologia a ser utilizada nesta dissertação baseia-se na metodologia CRISP/DM. Para um melhor entendimento a seguir esta metodologia será descrita.

A metodologia CRISP-DM (Cross Industry Standard Process For Data Mining, *Processo Padrão Inter-Indústrias para Mineração de Dados*) foi desenvolvida por um consórcio formado por NCR Systems Engineering Copenhagen, DaimlerChrysler AG, SPSS Inc. e OHRA Verzekeringen en Bank Groep B.V em 1996 (CRISP-DM, 2010).

É proposta a utilização dessa metodologia mediante a uma adaptação ao contexto da IES, visando a criação de um fluxo de trabalho que permitirá a IES a extração de conhecimentos úteis para a tomada de decisões, integrando o conhecimento aos seus gestores.

A metodologia CRISP-DM é formada por um conjunto de fases e processos padrões utilizado para o desenvolvimento de projetos de Mineração de Dados, independente de

ferramentas e da área de negócios. Seus principais objetivos são: converter as necessidades de negócios em tarefas de Mineração de Dados, promover transformações nos dados e nas técnicas, fazer uso de métricas para avaliação da qualidade dos resultados e elaborar a documentação do projeto.

A metodologia CRISP-DM de mineração de dados é descrita em termos de um modelo de processo hierárquico, que consiste em conjuntos de tarefas descritas em quatro níveis de abstração (do geral para o específico): a fase, a tarefa genérica, tarefas especializadas e instância de processo (CRISP-DM, 2010).

O modelo atual processo de mineração de dados fornece uma visão geral do ciclo de vida de um projeto de mineração de dados. Ele contém as fases de um projeto, suas tarefas respectivas e as relações entre essas tarefas. Neste nível de descrição, não é possível identificar todas as relações.

O processo de Mineração de Dados é considerado como um projeto com um ciclo de vida cuja interatividade em suas fases, faz com que a sequencia não seja rigorosa, porém dependente do resultado obtido em cada fase anteriormente trabalhada. Este ciclo abrange seis fases conforme vistos na Figura 8.

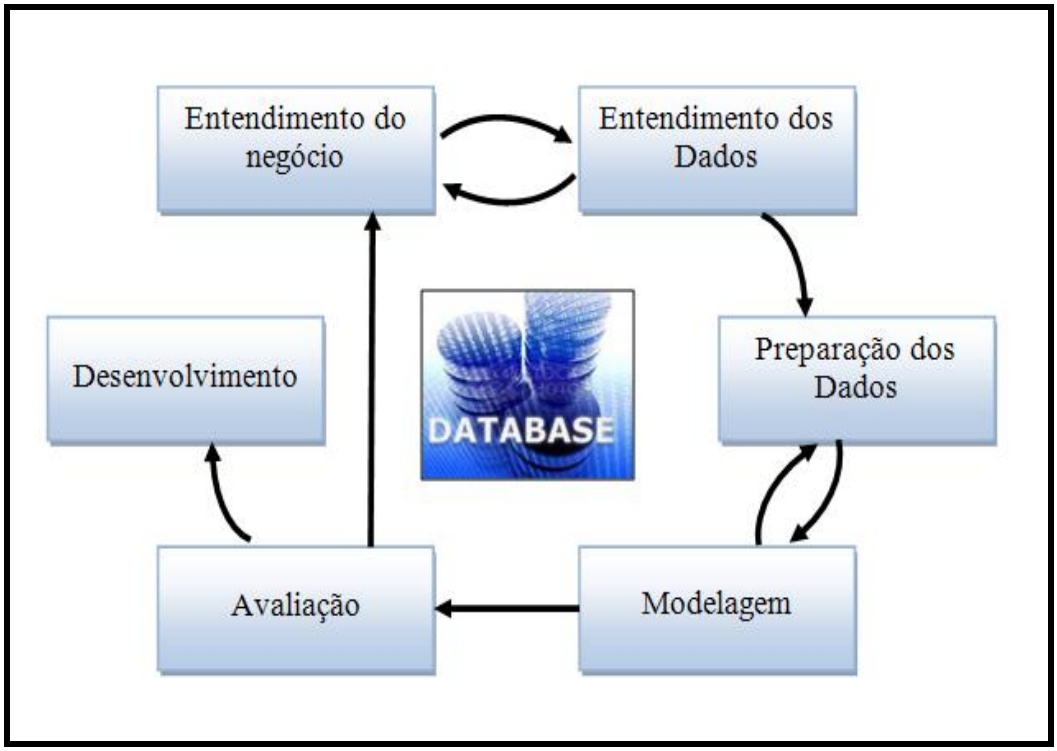


Figura 8 - Fases do modelo de referência CRISP-DM
Fonte: Adaptado de CRISP-DM (2010).

Define-se a seguir cada fase do modelo apresentado na Figura 8.

Na fase de Entendimento do Negócio (Business Understanding), que é considerada a fase inicial, tem-se por meta o entendimento dos objetivos do projeto e os requisitos a partir de uma perspectiva de negócios, a seguir, tendo com base o conhecimento adquirido, define-se o problema e um plano preliminar deve ser projetado para atingir os objetivos. (CRISP-DM, 2010, tradução nossa).

O objetivo desta fase é deixar bem definidos os objetivos e os requisitos do projeto, tendo sempre a visão do domínio a ser tratado.

Na fase do Entendimento dos Dados (Data Understanding), o início se dá com a coleta inicial de dados e segue com atividades, com intuito de promover uma maior familiarização com os dados, objetivando a identificação de problemas, a qualidade e utilidades dos dados e a detecção de subconjuntos de dados interessantes a formulação de hipóteses e descoberta de informações ocultas. (CRISP-DM, 2010, tradução nossa).

As tarefas efetuadas nesta fase são: a coleta inicial de dados, exploração e verificação das qualidades dos dados.

Na fase de Preparação de Dados (Data Preparation) é realizada a construção do banco de dados que será submetido a ferramenta de mineração. Os dados oriundos deste banco passam pelos processos de seleção, limpeza, transformação, integração e formatação dos dados.

Hiragi (2008, p.26) destaca que:

> O resultado desta fase será o conjunto de dados que servirá de subsídio para mineração dos dados. Aqui ocorre a seleção de atributos, o tratamento de valores faltantes, erros nos dados, integração de fontes de dados, formatações, divisão dos dados em, pelo menos, um conjunto de treinamento e um conjunto de avaliação, entre outras.

Na fase Modelagem (Modelling) são definidas as técnicas de modelagem dos dados que serão utilizadas e seus parâmetros são ajustados. Como existem diversas técnicas de mineração para o mesmo problema, faz-se necessário as vezes retornar a fase de preparação dos dados.

Para Hiragi (2008, p. 26)

> A modelagem é a parte que envolve processos de inteligência artificial e estatística de forma mais significativa. Inicialmente devemos escolher a tarefa de mineração de dados a ser usada, sempre com base no domínio de conhecimento e tipos de dados. Então de posse da tarefa (por exemplo, classificar, estimar, descrever ou visualizar) iremos selecionar a ferramenta de inteligência artificial ou estatística que implemente a técnica escolhida.

As tarefas desta fase são: a seleção da técnica de modelagem, a geração de testes do projeto, a construção e validação do modelo.

A fase Avaliação (Evaluation) tem por objetivo verificar a existência de algum problema em relação aos objetivos do negócio. Nesta fase é realizada a mensuração da qualidade dos dados minerados.

Para Dias (2001, p.24) "O principal objetivo é determinar se existe alguma questão de negócio importante que não foi suficientemente considerada. Nesta fase, uma decisão sobre o uso dos resultados de mineração de dados deverá ser alcançada".

Ao término desta fase, uma decisão sobre a utilização dos resultados da mineração de dados deve ser atendido. Possui como tarefas: a análise dos resultados, a revisão dos processos e a definição dos próximos passos.

Na sexta e última fase**, a do Desenvolvimento** (Deployment), o modelo com a melhor performance criado é distribuído ao cliente para que este o coloque em prática. Para Hiragi (2008, p. 27): "A colocação em uso pode ser vista como utilizar resultados obtidos pela aplicação (a um novo conjunto de dados) do modelo selecionado para apoiar uma tomada de decisão por parte do decisor que o utiliza".

Esta fase possui como tarefas: a elaboração de plano de distribuição, a criação de um plano de monitoramento e manutenção, a elaboração do relatório final e por último a revisão do projeto.Para um melhor entendimento das fases e suas respectivas tarefas e saídas, o Quadro 2 mostra todos os conceitos anteriormente apresentados.

FASE	TAREFAS	SAÍDAS
Entendimento do Negócio	• Determinar os objetivos do negócio;	• . Background; • Os objetivos do negócio; • Critérios de sucesso do negócio.
	• Avaliar a situação;	• Inventário dos recursos; • Requisitos, premissas e restrições; • Riscos e contingências; • Terminologia; • Custos e benefícios.

Continua...

Conclusão.

Entendimento do Negócio	• Determinar as metas da Mineração de Dados;	• Metas da Mineração de Dados; • Critérios de sucesso da Mineração de Dados.
	• Produzir o plano do projeto	• Plano do projeto; • A avaliação inicial de ferramentas e técnicas.
Entendimento dos Dados	• Coletar os dados iniciais;	• Relatório da coleta inicial dos dados.
	• Descrever os dados;	• Relatório da descrição dos dados.
	• Explorar os dados;	• Relatório da exploração dos dados.
	• Verificar a qualidade dos dados.	• Relatório da qualidade dos dados.
Preparação dos Dados	• Selecionar os dados;	• Justificativa para inclusão/exclusão.
	• Limpar os dados;	• Relatório de limpeza dos dados.
	• Construção dos dados;	• Atributos derivados;
		• Registros gerados.
	• Integrar os dados;	• Dados mesclados.
	• Formatar os dados	• Dados reformatados.
Modelagem	• Selecionar a técnica de modelagem;	• Técnica de modelagem;
		• Modelagem de pressupostos.
	• Gerar o design do teste;	• Design do teste.
	• Construir o modelo;	• As definições de parâmetros;
		• Modelos;
		• Descrição do modelo resultante.
	• Avaliar o modelo.	• Modelo de avaliação;
		• Parâmetros revisados.
Avaliação	• Avaliar os resultados;	• Avaliação dos resultados de mineração de dados no que diz respeito aos critérios de sucesso empresarial;
		• Modelos aprovados.
	• Processo de revisão;	• Revisão do processo.
	• Determinar os próximos passos.	• Lista de ações possíveis;
		• Decisão.
Desenvolvimento	• Implantação do plano;	• Plano de implantação.
	• Plano de manutenção e monitoramento;	• Plano de manutenção e monitoramento.
	• Produzir o relatório final;	• Relatório final;
		• Apresentação final.
	• Projeto de revisão.	• Documentação da experiência.

Quadro 2 - Constructo das fases do modelo CRISP-DM
Fonte: Adaptado de CRISP-DM (2010).

CRISP-DM foi projetado para fornecer orientação para os iniciantes em mineração de dados e para fornecer um modelo de processo genérico que pode ser especializada de acordo com as necessidades de qualquer ramo de atividade ou da empresa.

A MD pode ser desenvolvida de modo não-sistemático, se qua haja nehnhum cuidado em seu desenvolvimento, o que não é recomedado, pois acarreta em resultados não esperados ou imprecisos. Com intuito de evitar este tipo de situação o uso de uma metodologia vem a

garantir que o processo da MD seja desenvolvido de modo sistemático e padronizado, o que acarretará em resultados precisos e confiáveis.

A metodologia CRISP-DM tem seu sucesso devido ao fato de ter sido desenvolvida à prática, não estar atrelada a nenhuma ferramenta específica de mineração de dados, mas sim a junção das melhores práticas que são utilizadas em um projeto de mineração de dados, aliada ao fato de atuar sobre todo o processo de MD.

2.8 TAREFAS DE MINERAÇÃO DE DADOS

Conforme o objetivo pretendido, várias tarefas de MD podem ser realizadas. Conceitua-se tarefa de Mineração de Dados o modo como as informações serão mineradas, trata-se de uma funcionalidade.

O objetivo a ser alcançado pode ser obtido pelo uso de mais de uma tarefa e esta pode se utilizar de diversas abordagens. Conhecidas como técnicas, essas abordagens podem se utilizar de diversos tipos de algoritmos para a implementação de determinada tarefa.

A Figura 9 demonstra a interação entre esses elementos.

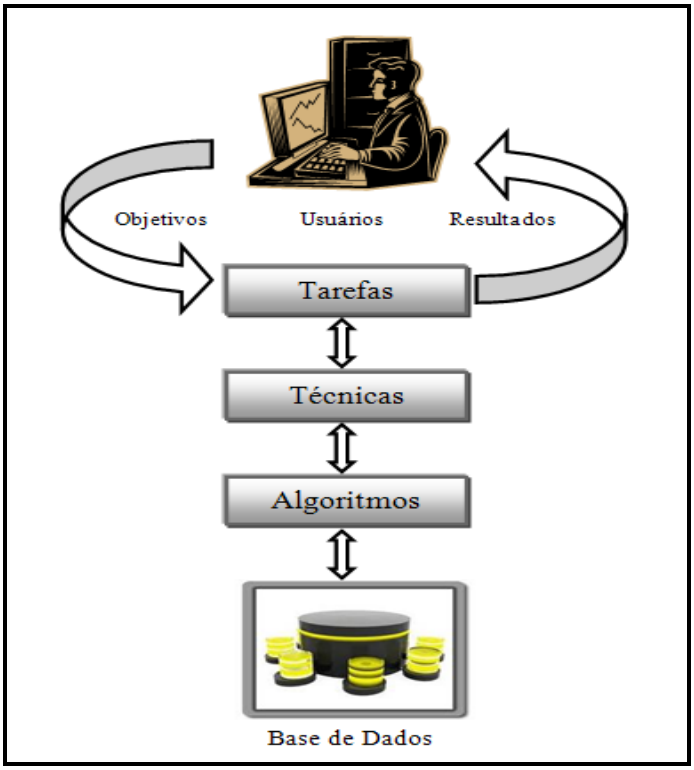


Figura 9 - Interação entre os elementos da MD

A MD tem dois principais tipos de tarefas: a atividade preditiva e a descritiva. Classificação e Regressão são consideradas tarefas de atividade preditiva, enquanto as atividades de Associação, Clusterização e Sumarização são as principais atividades descritivas.

2.8.1 Classificação

Classificar é um conceito já muito utilizado pelo ser humano. Esta tarefa consiste na criação de classes previamente definidas de acordo com as semelhanças de algumas características.

A tarefa de classificação é considerada como uma tarefa preditiva, haja vista que suas classes não são definidas, essa tarefa determina um conjunto de classes (padrões) que podem ser usadas para classificar novos objetos. Rabelo (2007, p. 27) reforça que "Ela busca uma função que permite associar corretamente cada registro (x) de um banco de dados a um único rótulo categórico de (y) chamado de classe".

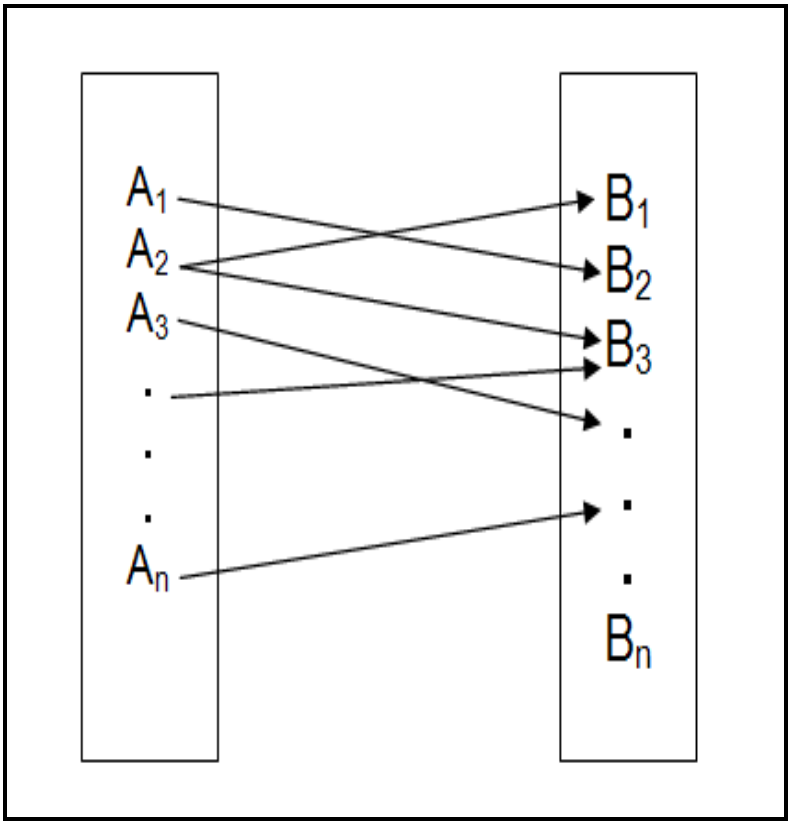


Figura 10 - Ligação entre dados e classes
Fonte: Rabelo (2007, p. 27)

Para Cardoso e Machado (2008, p. 506) a classificação:

> [...] é o processo de criar modelos (funções) que descrevem e distinguem classes ou conceitos, baseados em dados conhecidos, com o propósito de utilizar esse modelo para predizer a classe de objetos que ainda não foram classificados. O modelo construído baseia-se na análise prévia de um conjunto de dados de amostragem ou de treinamento, contendo objetos corretamente classificados. Exemplo: grupos de pesquisas já definidos contendo alguns professores e, a partir da análise de dados das pesquisas de outros professores que não pertencem a esses grupos, sugerir a sua entrada.

A tarefa de classificação tem por princípio a descoberta de algum tipo de relacionamento entre os atributos preditivos e o atributo meta, com intuito de se descobrir um novo conhecimento, o qual possa ser utilizado na previsão de uma nova classe, ainda desconhecida.

A classificação consiste na previsão de uma variável categórica, ou seja, para descobrir uma atividade que irá mapear um conjunto de registros em um conjunto de variáveis predefinidas chamadas classes. Esta atividade pode ser aplicada a novos registros, de modo a prever a classe em que esses registros se enquadra. Diversos algoritmos são aplicados nas tarefas de classificação, mas os que mais aparecem são as Redes Neurais, Back-Propagation, Classificadores Bayesianos e Algoritmos Genéticos.

Para prever se um acadêmico irá ou não evadir-se da instituição em função de sua situação financeira, a instituição necessita de alguns dados sobre o acadêmico em sua base de dados. A partir desses dados, um algoritmo de classificação pode descobrir regras que prevêem se um novo acadêmico irá ou não evadir-se. Essa informação é então armazenada em um novo atributo, nesse caso o atributo objetivo. Seu valor pode assumir dois possíveis valores: SIM, significando a evasão, ou NÃO, caso contrário. De posse do atributo determinado, o passo seguinte é selecionar um subconjunto de atributos preditivos entre todos os atributos dos acadêmicos no banco de dados.

Um algoritmo de classificação pode analisar os dados da Tabela 1 a fim de determinar quais os valores dos atributos preditivos devem ser relacionados, com cada um dos atributos objetivos. Com base neste conhecimento gerado pode-se aplicar então para a previsão das futuras evasões por parte dos acadêmicos.

Tabela 1 - Entrada de dados para a tarefa de classificação

Sexo	Idade	Auxílio	Evasão
Masculino	26	Sim	Não
Feminino	19	Não	Sim
Masculino	19	Não	Sim
Masculino	30	Não	Não
Feminino	20	Sim	Não
Feminino	29	Não	Não
Masculino	18	Não	Sim

Fonte: Da pesquisa (2010)

A representação do conhecimento descoberto é representada na forma de regras do tipo SE-ENTÃO. A interpretação destas regras é fetia da seguinte maneira: "SE os atributos preditivos satisfazem a uma condição no antecedente da regra, ENTÃO a a classe é indicada no consequente da regra. A Figura 11 mostra as regras extraídas de um algoritmo de classificação, tendo como atributos os dados da Tabela 1.

SE (Sexo=Masculino e Idade >20) ENTÃO Evasão=Não
SE (Sexo=Feminino e Idade > 25) ENTÃO Evasão=Não
SE (Sexo=Masculino e Idade<20 e Auxílio=Não) ENTÃO Evasão=Sim
SE (Sexo=Feminino e Idade<20 e Auxílio=Sim) ENTÃO Evasão=Não

Figura 11 - Regras de classificação
Fonte: Da pesquisa (2010)

Torna-se necessário fazer experimentos com os algoritmos disponíveis a fim de verificar qual melhor se adequa a aplicação em questão. (SCOSS, 2006)

2.8.2 Regressão

A tarefa de regressão é semelhante à tarefa de classificação, ela busca funções que fazem o mapeamento dos registros contidos em uma base de dados. Por lidar com resultados contínuos, esta tarefa pode ser utilizada como uma tarefa de classificação, estabelecendo-se que diferentes faixas de valores correspondem a diferentes classes.

Para Scoss (2006, p. 29) "A estimação ou regressão é similar a tarefa de classificação, porém restringe-se a atributos numéricos. Ela busca por funções, sejam lineares ou não, que possam mapear registros de um banco de dados". Esta regra tem por objetivo a definição de

um valor numérico de alguma variável desconhecida a partir dos valores de variáveis já são conhecidas.

Na regressão, há uma busca por uma função linear ou não, bem como a variável que está sendo prevista consiste de um atributo numérico (contínua), presente em bases de dados com valores reais. A fim de implementar a tarefa de regressão, os métodos de Estatística e Redes Neurais são utilizadas.

2.8.3 Associação

A regra de associação foi desenvolvida para analisar os dados de uma base de dados num ambiente de marketing, no qual os dados de entrada são os compostos de cada transação efetuada por um cliente e os dados de saída são composições obtidas por meio de regras. Exemplificando, quando um cliente compra um produto "X", em N% das vezes, ele compra o produto "Y" também. Esta regra teria a representação vista no Quadro 3:

SE sexo = Masculino "E" Estado civil = Casado "E" Renda > 1800 Implica que Consumo = Roupa de grife, perfume nacional.

Quadro 3 - Representação da Regra de Associação

Esta regra tem por objetivo a localização de tendências que facilitem a compreensão de padrões em grandes bases de dados. Os seus algoritmos procuram por relações entre os itens das transações, analisando os que ocorrem simultaneamente, dando possibilidade de entendimento de novos modelos.

Barioni (2001, p. 17) define que "A tarefa dessa técnica envolve a descoberta de regras de associação que indiquem correlações interessantes entre objetos de um dado banco de dados". Ela estuda um padrão de relacionamento existente entre itens de um dado.

Para Motta (2010, p. 8) “Uma regra de associação é uma implicação da forma: $A \rightarrow B$, onde $A \subseteq I$, $B \subseteq I$ e $A \cap B = ø$. Neste caso, lê-se A implica em B, onde A é chamado antecedente e B é o conseqüente da regra”.

A quantidade de regras de associação que podem ser encontradas numa aplicação de associação é extensa e muitas destas regras não são consideradas relevantes para os analistas. Uma forma de resolver esta questão é a introdução de medidas de interesse, que fazem a distinção entre as regras relevantes e as não relevantes. Estas medidas são chamadas de suporte e confiança. (BARIONI, 2002, p. 17)

O primeiro algoritmo eficiente de regras de associação foi o algoritmo Apriori desenvolvido por Agrawal e Srikant em 1993. O primeiro passo deste algoritmo é a pesquisa de conjuntos de itens freqüentes. O usuário dá um limite mínimo para o apoio e o algoritmo de pesquisa todos os conjuntos de itens que aparecem com um apoio superior a esse limite. O segundo passo é a construção de regras a partir de conjuntos de itens encontrados na primeira etapa. O algoritmo calcula a confiança de cada regra e mantém apenas aqueles em que a confiança é maior que um limiar definido pelo usuário.

A tarefa de associação consiste em identificar e descrever as associações entre as variáveis no mesmo item ou associações entre os itens diferentes que ocorrem simultaneamente, de uma forma freqüente em bases de dados.

A busca de associações entre os itens durante o intervalo temporal é também comum. Assim, os algoritmos Apriori e GSP (Generalized Sequential Patterns), entre outros, são os mais utilizados para implementar a descoberta da tarefa de associação.

2.8.4 Clusterização ou Segmentação

A tarefa de clusterização faz a identificação da classe de cada objeto de modo que, os objetos contidos numa mesma classe apresentem um alto grau de similaridade entre si e um baixo grau de similaridade em relação a objetos de outras classes. Esta tarefa também é conhecida como “agrupamento”, uma vez que agrupa os objetos em classes com o grau de similaridade mais próximo.

Para Martinhago (2005, p. 22)

> Um cluster pode ser definido como um conjunto de objetos agrupados pela similaridade ou proximidade e, a segmentação pode ser definida como a tarefa de segmentar uma população heterogênea em um número de subgrupos (ou clusters) mais homogêneos possíveis, de acordo com alguma medida.

Para Macedo e Matos (2010, p. 26) "A análise de cluster tem como objetivo verificar a existência de diferentes grupos dentro de um determinado conjunto de dados, e em caso de sua existência, determinar quais são eles"

A clusterização pode ser considerada como uma tarefa que identifica um conjunto finito de categorias com intuito de descrever os dados. Seu objetivo principal é fazer a partição da base de dados em um número determinado de clusters, nos quais as instâncias destes clusteres sejam similares, conforme visto na Figura 12.

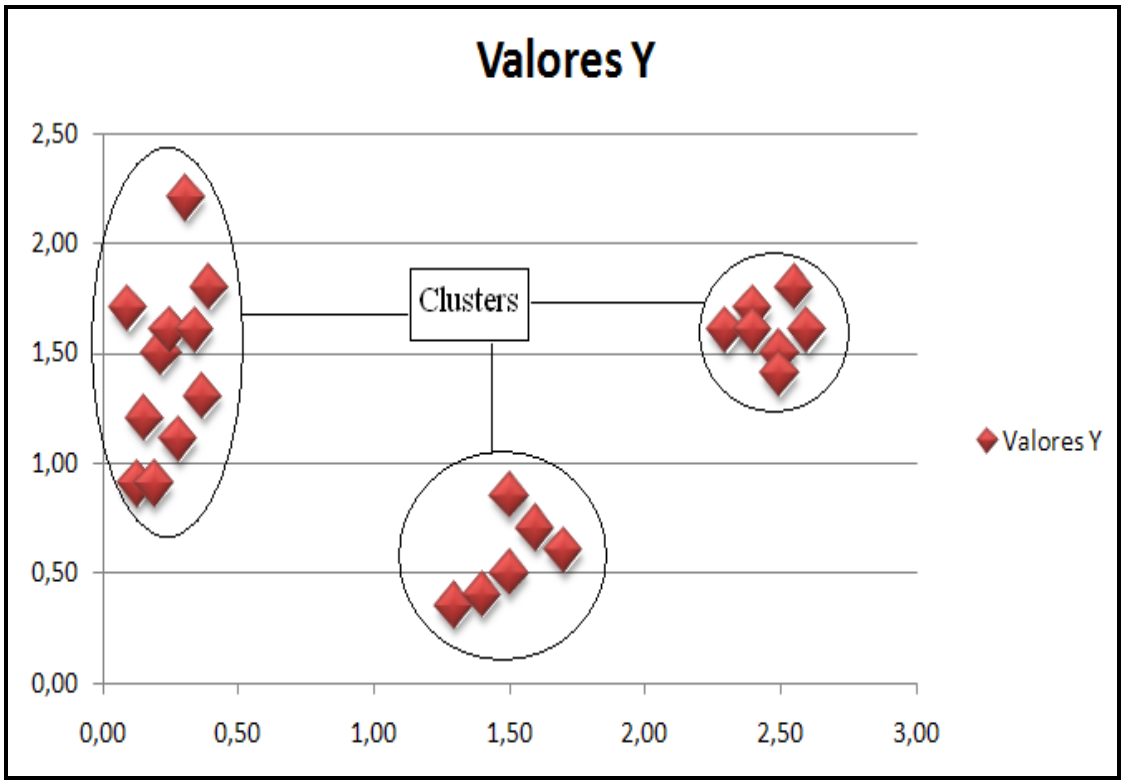


Figura 12 - Exemplo da visualização de clusters

.

Os dados podem ser agrupados em classes ou clusters de elementos similares. Não é passada nenhuma informação ao sistema sobre a existencia de determinadas classes. A descoberta das classes é feita pelo prórpio algoritmo, que agrupa os dados em classes com as características semelhantes. Diferente da classificação, na clusterização não há classes pré-definidas.

2.8.5 Sumarização

A tarefa de sumarização tem por objetivo a identificação e apresentação das principais características dos dados, de forma concisa e compreensível. É considerada uma tarefa descritiva.

Conforme Fayyad (1996 apud DIAS, 2001, p. 10), a tarefa de sumarização envolve métodos para encontrar uma descrição compacta para um subconjunto de dados. Um simples exemplo esta tarefa poderia ser tabular o significado e desvios padrão para todos os itens de dados. Métodos mais sofisticados envolvem a derivação de regras de sumarização.

A sumarização visa identificar e indicar as características comuns entre um conjunto de dados. Esta tarefa é aplicada nos clusters obtidos na tarefa de clusterização, com a Lógica Indutiva e Algoritmos Genéticos são exemplos de tecnologias que podem ser implementadas na sumarização. As técnicas de sumarização são sempre aplicadas à análise exploratória de dados e à geração automática de relatórios.

Dias (2001, p. 11) sintetiza as principais tarefas de Mineração de Dados, suas descrições e exemplifica-as, conforme pode ser visto na Tabela 2:

Tabela 2 - Sintese das tarefas de Mineração de Dados

TAREFA	DESCRIÇÃO	EXEMPLOS
Classificação	Constrói um modelo de algum tipo que possa ser aplicado a dados não classificados a fim de categorizá-los em classes.	• Classificar pedidos de crédito; • Esclarecer pedidos de seguros fraudulentos; • Identificar a melhor forma de tratamento de um paciente.
Estimativa ou Regressão	Usada para definir um valor para alguma variável contínua desconhecida.	• Estimar o número de filhos ou a renda total de uma família; • Estimar o valor em tempo de vida de um cliente; • Prever a demanda de um consumidor para um novo produto.
Associação	Usada para determinar quais itens tendem a co-ocorrerem (serem adquiridos juntos) em uma mesma transação.	• Determinar quais os produtos costumam ser colocados juntos em um carrinho de supermercado.
Segmentação ou Clusterização	Processo de partição de uma população heterogênea em vários subgrupos ou grupos mais homogêneos.	• Agrupar clientes por região do país; • Agrupar clientes com comportamento de compra similar; • Agrupar seções de usuários Web para prever comportamento futuro de usuário.
Sumarização	Envolve métodos para encontrar uma descrição compacta para um subconjunto de dados.	• Tabular o significado e desvios padrão para todos os itens de dados; • Derivar regras de síntese.

Fonte: Dias (2001, p. 11)

2.9 TÉCNICAS DE MINERAÇÃO DE DADOS

Devido ao extenso número de problemas de Mineração de Dados, não há uma técnica que possa ser utilizada para a resolução de todos eles. Cada problema possui suas peculiaridades, assim sendo, diferentes técnicas são utilizadas para a resolução de problemas com propósitos diferentes.

Para Dias (2001, p. 12) "A familiaridade com as técnicas é necessária para facilitar a escolha de uma delas de acordo com os problemas apresentados. A seguir são descritas as técnicas de Mineração de Dados normalmente usadas".

Há um número relativo de técnicas para a extração do conhecimento em bases de dados que podem ser aplicados as tarefas de *Mineração de Dados.* Estas técnicas consistem na aplicação de um ou mais algoritmos, implementados em ferramentas acadêmicas ou comerciais, com propsito de descobrirem conhecimento a partir da base de dados a ser explorada.

Chiara (2003, p. 9) destaca que,

> Um ponto a ser observado é que algumas técnicas são melhores para determinados problemas e domínios de conhecimento que outras. Portanto, não há um método universal de Mineração de Dados. A escolha de um algoritmo particular para um determinado problema deve ser analisado empiricamente.

Martinhago (2005, p. 24) ressalta:

> Um ponto importante é que cada técnica tipicamente resolve melhor alguns problemas do que outros, não há um método universal e a escolha é uma arte. Para as aplicações, grande parte do esforço vai para a formulação do problema, ou seja, a especificação de que tipo de informações o algoritmo de mineração deve procurar no conjunto de dados disponíveis.

As técnicas de Mineração de Dados utilizadas atualmente são extensões naturais ou generalizações de métodos analíticos já conhecidos. A novidade consiste na possibilidade de applicação destas técnicas buscando auxiliar os gestores no processo decisório e com o objetivo de encontrarem novas estratégias para os negócios, isto se deve ao aumento da capacidade de armazenamento de informações e à redução nos custos de processamento.

Pode-se citar como técnicas mais conhecidas: algoritmos genéticos, árvores de decisão, descoberta de regras de associação, raciocínio baseado em caso e redes neurais

artificiais, entre outros. Nesta pesquisa têm-se as tarefas de clusterização e classificação. A seguir são descritas as principais técnicas de *Mineração de Dados.*

2.9.1 Técnicas Estatísticas

Várias técnicas estatísticas têm sido aplicadas à tarefa de Mineração de Dados, com vistas a determinação de possíveis correlações entre variáveis do problema, associações e levantamento das variáveis mais significativas que descrevem o fenômeno. Pode-se citar:

a) Coeficiente de Correlação Linear de Pearson;

b) Coeficiente de Determinação Múltipla (R^2);

c) Multicolinearidade;

d) Análise de Componentes Principais.

Como no problema desta dissertação, os dados, tanto de ingressantes como de egressos, são na sua maior parte constituídos de informações qualitativas e categóricas, das três primeiras técnicas não puderem ser aplicadas. Assim, a fundamentação teórica, no que tange as técnicas estatísticas, vai ser concentrada na técnica de Análise de Componentes Principais, a qual será descrita a seguir.

2.9.1.1 Análise de componentes principais (ACP)

Com a Análise de Componentes Principais (ACP) se obtém um novo conjunto de coordenadas que pode ser utilizado para descrever os dados de forma simplificada. Utilizando a ACP é possível reduzir o número de dimensões nos dados de forma a não perder informações importantes sobre os dados (SCHMITT, 2005).

Segundo Johnson e Wichern (2002), a técnica de componentes principais busca a redução da dimensionalidade e interpretação do conjunto de dados. Os autores destacam que a partir de um conjunto de m variáveis, serão obtidas outras m variáveis, não correlacionadas, que são combinações lineares do conjunto original de variáveis.

Schmitt (2005) destaca que:

> [...] geometricamente, as componentes principais representam um novo sistema de coordenadas, obtidas por uma rotação do sistema original, que fornece as direções de máxima variabilidade, e proporciona uma descrição mais simples e eficiente da estrutura de covariância dos dados.

Com a aplicação da ACP é possível determinar um número mínimo de variáveis que expliquem a maior parte da variação dos dados e então reduzir a dimensionalidade do conjunto, retirando algumas componentes principais sem causar uma grande perda de informação.

A ACP consiste em, a partir da matriz de correlação das variáveis, obter os autovalores e autovetores que representarão a variabilidade explicada dos dados por cada componente principal, e os coeficientes das componentes principais, respectivamente. Existem exatamente m autovalores, não negativos, e também m autovetores correspondentes a cada autovalor (LIRA, 2004; SCHMITT, 2005).

Para determinar o número ideal de componentes principais, existem vários critérios práticos. Nesse trabalho serão apresentados o critério do scree plot e critério de Kaiser ao decorrer do exemplo a seguir.

Para facilitar o entendimento da técnica, será demonstrado um exemplo presente em Schmitt (2005, apud REIS, 2010). Considerando um conjunto de dados em que 3 variáveis (peso, altura e idade) foram pesquisadas em relação a 8 pessoas. Os dados estão presentes na Tabela 3.

Tabela 3. Conjunto de dados com 8 observações e 3 variáveis

Observação	**Peso (X_1)**	**Altura (X_2)**	**Idade (X_3)**
1	55	164	25
2	90	185	18
3	79	179	47
4	60	172	45
5	83	177	49
6	83	176	50
7	95	189	65
8	54	160	23

Fonte: Schmitt (2005)

Para obter as componentes principais é recomendado padronizar os dados para calcular a matriz de correlações, para que variáveis com unidades que representem uma

grande variação nos dados não implique na análise de forma incorreta (LIRA, 2004; SCHMITT, 2005).

A matriz de correlações amostrais, R, após a padronização dos dados é dada por:

$$R = \begin{bmatrix} 1{,}000 & 0{,}949 & 0{,}499 \\ 0{,}949 & 1{,}000 & 0{,}526 \\ 0{,}499 & 0{,}526 & 1{,}000 \end{bmatrix}$$

A partir da matriz R, podem ser obtidos os autovalores λ_1, λ_2 e λ_3, e os autovetores v_1, v_2 e v_3.

$$\lambda_1 = 2{,}3412$$
$$\lambda_2 = 0{,}6088$$
$$\lambda_3 = 0{,}0500$$
$$v_1 = \begin{bmatrix} 0{,}6193 \\ 0{,}6248 \\ 0{,}4755 \end{bmatrix}$$
$$v_2 = \begin{bmatrix} -0{,}3572 \\ -0{,}3150 \\ 0{,}8793 \end{bmatrix}$$
$$v_3 = \begin{bmatrix} 0{,}6992 \\ -0{,}7144 \\ 0{,}0281 \end{bmatrix}$$

Segundo Schmitt (2005 apud REIS, 2010), o **critério do *scree plot*** pode ser utilizado para determinar a quantidade mínima de componentes principais necessárias para explicar a variação dos dados. O critério consiste em representar a porcentagem de variância dos dados explicada num gráfico e, quando a curva gerada pelos pontos passa a ser quase paralela ao eixo das abscissas, as componentes devem ser desconsideradas. No caso de apenas 3 variáveis, como o do exemplo, este critério não é muito indicado. O gráfico gerado a partir do exemplo pode ser observado na Figura 13.

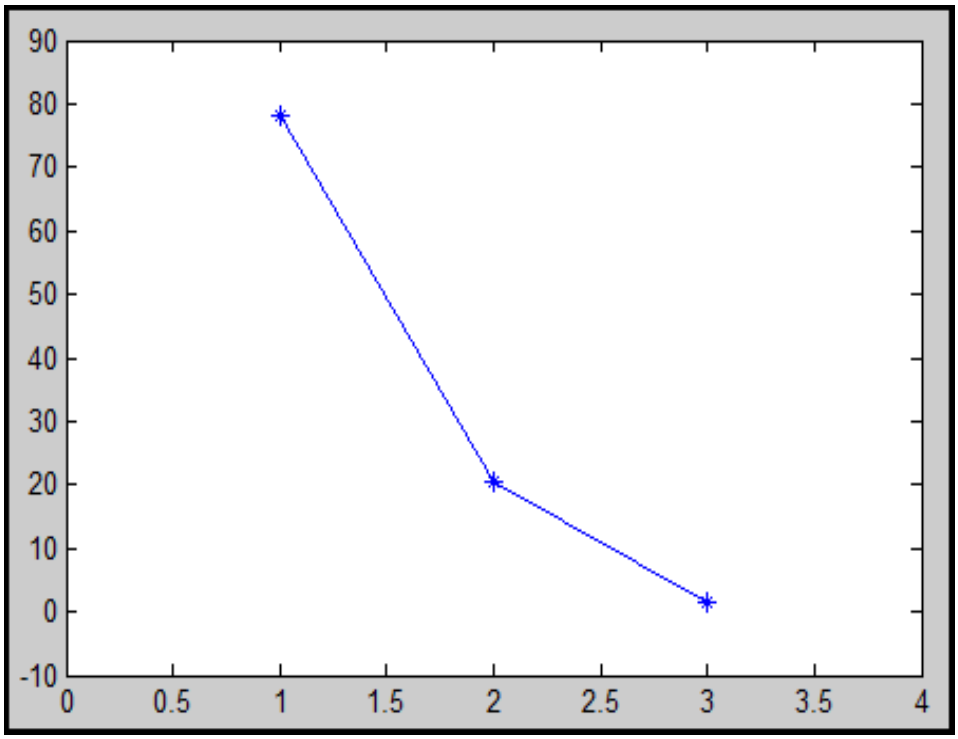


Figura 13 - *Scree plot*
Fonte: Reis (2010)

2.9.2 Exemplo de utilização de ACP na Mineração de Dados

A ACP pode ser utilizada para facilitar as tarefas de classificação e clusterização, pois diminuindo a quantidade de variáveis envolvidas sem perder informações relevantes, o cálculo tende a ficar mais rápido e preciso. (SCHMITT, 2005 apud REIS, 2010).

Mais detalhes sobre as tarefas de Mineração de Dados podem ser encontradas na Subseção 2.2.2 deste trabalho.

Considerando os dados do exemplo, foi gerado o gráfico presente na Figura 14. O gráfico utiliza apenas as duas primeiras componentes principais, pois elas explicam 98,33% da variação dos dados. Percebe-se que mesmo tendo diminuído o número de dimensões, não foram descartadas informações relevantes. Na Figura 14 é possível observar a distribuição dos elementos em 3 *clusters* distintos.

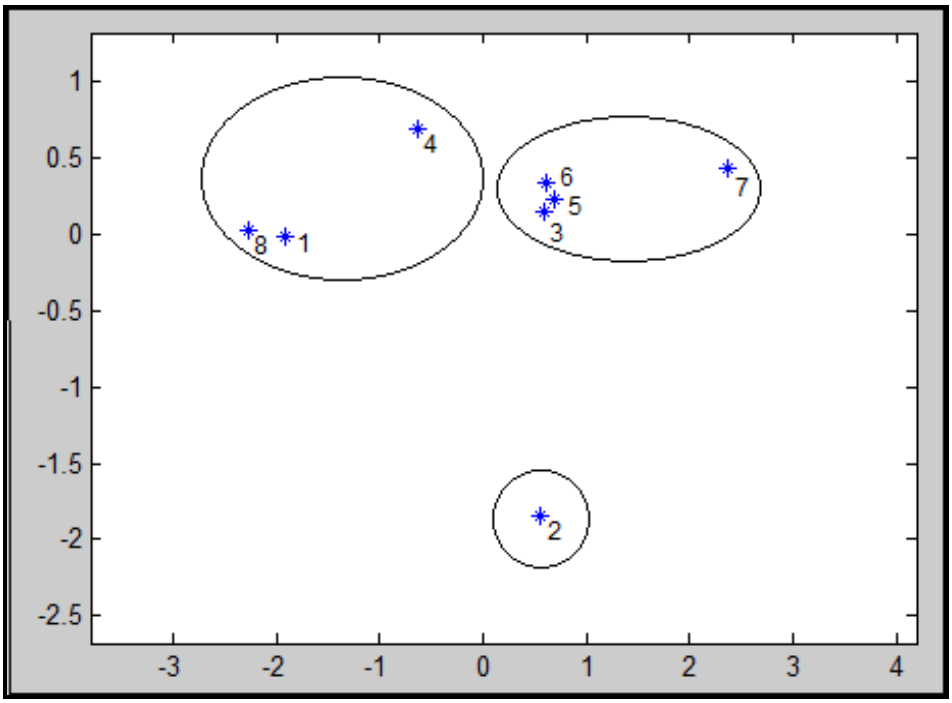


Figura 14 - Resultado da clusterização após utilizar ACP
Fonte: Reis (2010).

2.9.3 Algoritmos Genéticos (AG)

Nos anos 60 John Holland inventou os Algoritmos Genéticos (AG) e seus alunos na Universidade de Michigam os desenvolveram em meados de 1970. Holland tinha como objetivo o estudo formal dos fenômenos da evolução, tal qual ocorrem na natureza e o desenvolvimento de formas de importal tais fenomenos aos sistemas de computação.

Segundo Harrison (1998 apud DIAS, 2001, p. 13) “Os algoritmos genéticos usam os operadores de seleção, cruzamento e mutação para desenvolver sucessivas gerações de soluções”. Para Scoss (2006, p. 46) os AG são “baseados no conceito de evolução, os algoritmos genéticos usam processos de combinações genéticas, mutações e seleção natural”. São ações de melhoria que utilizam processos como a combinação genética, mutação e seleção natural, com base nos conceitos da evolução das espécies.

Para Almeida (2006, p. 4) “Os AG imitam o processo natural, na forma de um sistema artificial, por meio de operações que se equivalem aos mecanismos genéticos da natureza”. Um AG é um procedimento repetitivo para transformações sucessivas de uma população de organismos e é utilizado na MD na formulação de hipóteses sobre a dependência entre variáveis. A técnica de AG é indicada para as tarefas de classificação e segmentação.

2.9.4 Árvore de Decisões (AD)

Como o próprio nome diz, a técnica de Árvore de Decisões (AD) tem sua estrutura semelhante a de uma árvore, na qual suas ramificações representam as decisões possíveis.A partir destas decisões as regras que classificam um conjunto de dados são geradas.

Para Bispo (1998, p. 90):

> A sua estrutura é muito fácil de entender e de assimilar. Dividem os dados em subgrupos, com base nos valores das variáveis. O resultado é uma hierarquia de declarações tipo "Se ... então ..." que são utilizadas, principalmente, para classificar dados.

Rabelo (2007, p. 29) descreve a técnica de AD como sendo uma "Técnica que utiliza a recursividade para particionamento da base de dados na construção de uma árvore de decisão. Cada nó não terminal desta árvore representa um teste ou decisão sobre o item de dado". Tem como objetivo a separação das classes e tuplas de classes diferentes a fim de serem alocadas em subconjuntos diferentes, cada qual com suas regras. A técnica de AD é indicada para as seguintes tarefas: classificação e regressão e tem como exemplos de algoritmos: CART, CHAID, C4.5, C5.0, Quest, ID-3, SLIQ e SPRINT. (DIAS, 2001)

2.9.5 Descoberta de Regras de Associação (DRA)

Os algoritmos para a Descoberta de Regras de Associação (DRA) têm com objetivo procurar relações entre os dados de em conjunto de dados, que ocorrem com determinda frequencia. Esta técnica é muito utilizada na área do comércio, na busca de padrões de compra com intuito de orientar as ações dos gestores de vendas.

Martinhago (2005, p. 26) define assim a DRA:

> A regra de associação é uma expressão representada na forma X => Y (X implica em Y) , em que X e Y são conjuntos de itens da base de dados e $X \cap Y = \emptyset$; X é o antecedente da regra (lado esquerdo) e Y é o conseqüente da regra (lado direito) e pode envolver qualquer número de itens em cada lado da regra.

Confiança e suporte são dois parâmetros básicos na DRA. Estes limitam a quantidade de regras a serem extraídas e faz uma descrição da qualidade destas regras. Dias (2001, p. 12) define estes parâmetros como: "Uma regra de associação tem a forma geral X1 ^ ... ^ Xn => Y [C,S], onde X1,..., Xn são itens que prevêem a ocorrência de Y com um grau de confiança C e com um suporte mínimo de S e "^" denota um operador de conjunção (AND)".

Kampff (2009, p. 65) destaca que :

> A ordem de apresentação das regras estabelece uma lista de decisão, a ser aplicada em seqüência. A regra que aparece primeiro na lista tem maior prioridade para predizer a classe. Quando um registro é classificado, nenhuma outra regra posterior de classificação será aplicada sobre ele.

Como já mencionado, a aplicação desta regra é utilizada com frequência na área de comércio, sendo conhecida como análise de cesta de mercado, como exemplo, a regra pode descobrir que, quando qualquer cliente compra um produto "A", em N% das vezes, ele compra também o produto "B". A técnica de DRA é indicada para a tarefa de associação. Alguns algoritmos que implementam regras de associação tem-se: Apriori, AprioriTid, AprioriHybrid, AIS, SETM , DHP, DIC, Eclat, Maxclique e Cumalte. (DIAS, 2001)

2.9.6 Raciocínio Baseado em Casos (RBC)

Quando se tenta resolver algum problema, uma das primeiras soluções está apoiada em experiências passadas. O Raciocínio Baseado em Casos (RBC) faz uso de soluções já utilizadas para a solução de determinado problema, procurando um caso mais similar ao proposto.

Para Dias (2001, p.12) o RBC "Tenta solucionar um dado problema fazendo uso direto de experiências e soluções passadas. A distância dos vizinhos dá uma medida da exatidão dos resultados".

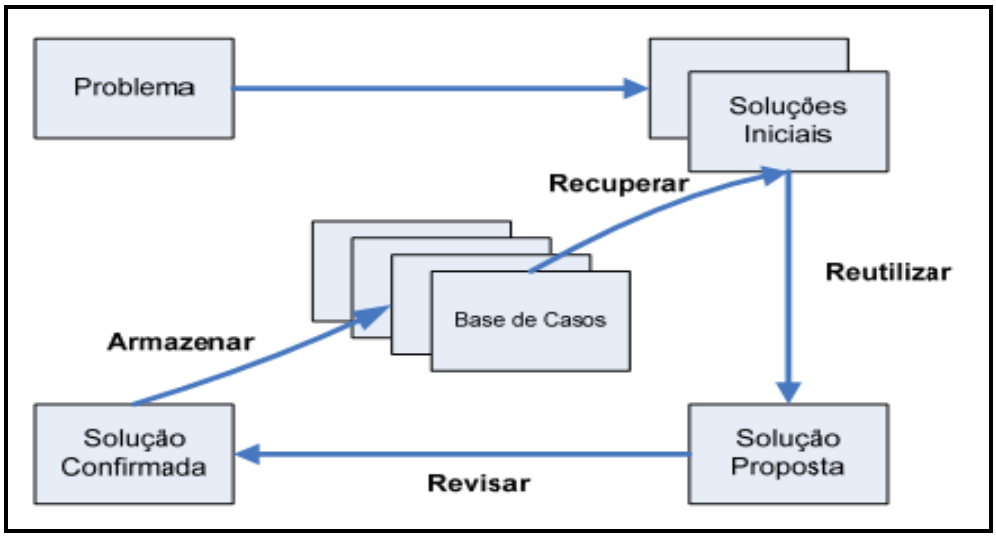


Figura 15 - Ciclo clássico do RBC
Fonte: Piva Junior (2006, p. 27).

Fonseca (2008, p. 16) descreve o contexto de aplicação do RBC, o que é observado na Figura 15.

> Na resolução de problemas, aplicando o RBC, uma solução para um novo caso é obtida recuperando casos similares anteriormente analisados e derivando suas respectivas soluções de modo a se adequar ao novo problema. O processo se realiza quando um novo caso é apresentado ao sistema. Em face do novo problema, utiliza-se um conjunto de métricas de similaridade para determinar quais casos anteriores mais se assemelham ao caso proposto, bem como se determinam as características-chave utilizadas nessa comparação.

O RBC possibilita ao gestor o uso do conhecimento no apoio as tomadas de decisões, haja vista a compatibilidade entre este tipo de sistema e os sistemas administrativos utilizados pelas IES. Eles proporcionam a extração, organização e o reuso do conhecimento utilizado na resolução de situações anteriores desta forma prmitindo o aprimoramento das soluções.

Para Von Wangenheim e Von Wangenheim (2003):

> Raciocínio Baseado em Casos é um enfoque para a solução de problemas e o aprendizado baseado em experiência passada. RBC resolve problemas ao recuperar e adaptar experiências passadas - chamadas casos - armazenadas em uma base de casos. Um novo problema é resolvido com base na adaptação de soluções de problemas similares já conhecidas.

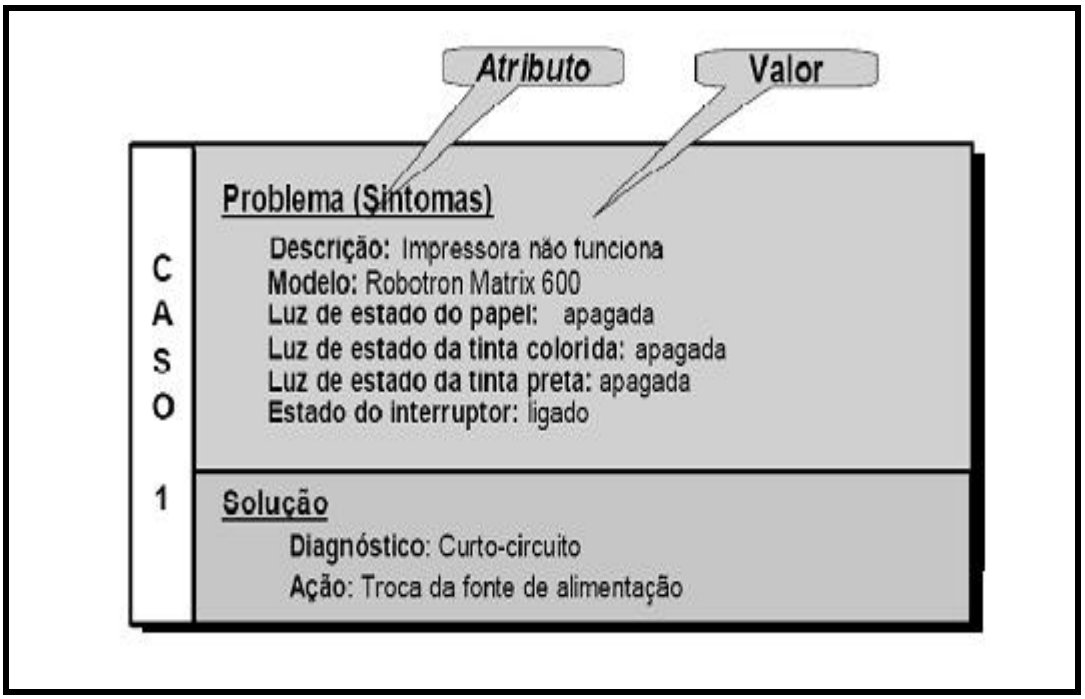


Figura 16 - Exemplo de um registro de RBC armazenado
Fonte: Von Wangenheim e Von Wangenheim (2003)

O uso do RBC tem como limitador o acesso às bases de dados completas, corretas e confiáveis que possuam em seus registros, a descrição completa dos problemas e das soluções

anteriormente utilizadas e armazenadas. A técnica de RBC é indicada para as tarefas de classificação e segmentação e os seguintes algoritmos mais conhecidos que implementam esta técnica são: BIRCH, CLARANS e CLIQUE. (DIAS, 2001)

2.9.7 Redes Neurais Artificiais (RNA)

As Redes Neurais Artificiais (RNA) são uma técnica computacional que constrói o modelo matemático inspirado no cérebro humano para o reconhecimento de imagens e sons, com capacidade de conhecimento, generalização, associação e abstração, constituída por sistemas paralelos distribuídos compostos de simples unidades de processamento.

As unidades de processamento são uma ou mais camadas interligadas por um grande número de ligações, na maioria dos modelos, essas conexões estão associadas a pesos, que, após o processo de aprendizagem, armazenam o conhecimento adquirido pela rede. (Kovacks, 2002).

Segundo Almeida (2009, p. 31)

> A tecnologia de Redes Neurais procura imitar o processo de resolver problemas do cérebro. Assim como o ser humano aplica conhecimento adquirido de experiências passadas para resolver novos problemas ou situações, de igual modo uma rede neural trabalha com exemplos previamente resolvidos para construir um sistema de "neurônios" que tomem novas decisões ou façam classificações e previsões.

De acordo Ferreira (2008, p. 50):

> Rede Neural Artificial, um termo raro há cerca de duas décadas na literatura científica, representa hoje uma vigorosa área de aplicação multidisciplinar, constituindo genuinamente uma ferramenta para o estudo de fenômenos complexos. A modelagem de dados para melhor entender fenômenos complexos, multidimensionais, bem como a tentativa de estimar uma variável dependente em função de outras de mais fácil obtenção, tem levado ao desenvolvimento de várias técnicas de análise. Uma das ferramentas mais exploradas e que tem apresentado bons resultados nas mais diferentes áreas do conhecimento é a técnica das Redes Neurais Artificiais (RNAs).

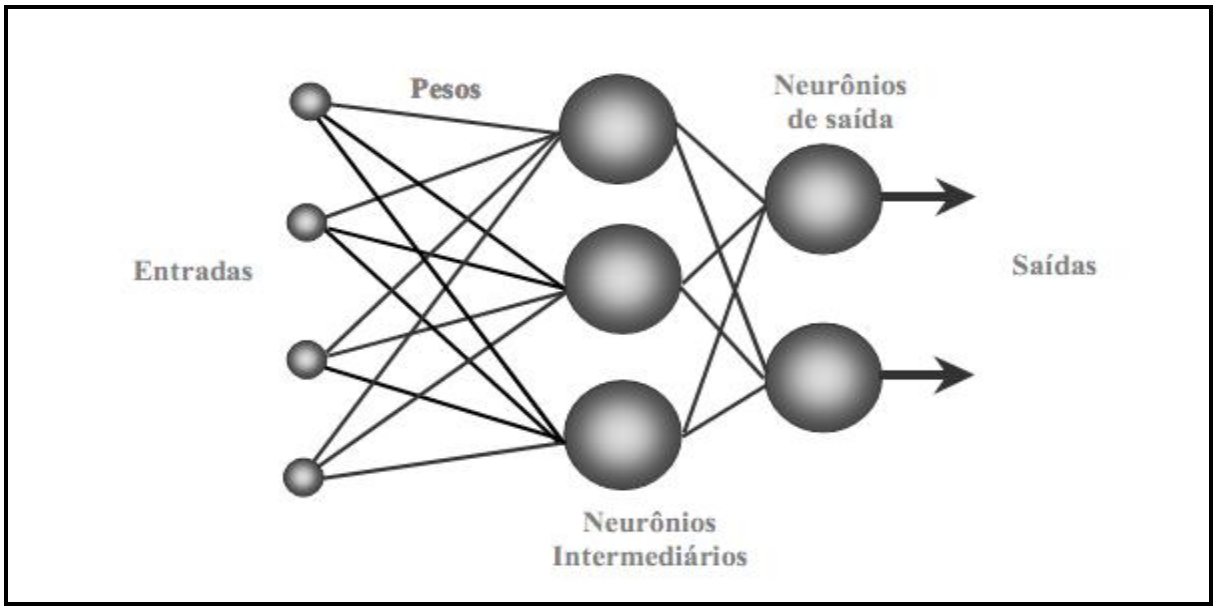


Figura 17 - Exemplo de uma Rede Neural Artificial de multiplas camadas
Fonte: Furtado (1999, apud FERREIRA 2008, p. 54)

As RNA's têm sido usadas com sucesso para as relações do modelo envolvendo séries temporais complexas em várias áreas do conhecimento. A maior vantagem das RNA's em relação aos métodos convencionais é que eles não exigem informações detalhadas sobre os processos físicos do sistema a ser modelado, com ele sendo descrito explicitamente na forma matemática e ainda por ser fortes e têm uma alta taxa de acurácia preditiva.

Dias (2001, p. 14) ressalta que "Uma das principais vantagens das redes neurais é sua variedade de aplicação, mas os seus dados de entrada são difíceis de serem formados e os modelos produzidos por elas são difíceis de entender.

Uma melhor definição de RNA é dada por Costa (2010, p. 24), na qual descreve que:

> A Rede Neural Artificial (RNA) foi desenvolvida e utilizada como uma ferramenta de resolução de problemas em vários campos. RNAs são generalizações de modelos matemáticos de sistema biológico nervoso em nosso cérebro e uma das principais vantagens da RNA é a capacidade de construir um modelo do problema utilizando os dados a partir de medições experimentais do domínio do problema. Ao invés de ser programado por um usuário em uma percepção tradicional, RNAs adquirem os seus conhecimentos aprendendo as relações das variáveis de dados e construção de um modelo, implicitamente, para relacionar as variáveis de entrada e saída para o problema.

A técnica de RNA é indicada para a resolução de problemas que envolvam: classificação, estimativa e clusterização e os algoritmos desenvolvidos para esta técnica são: Perceptron, Rede MLP, Rede Hopfield, Rede BAM, Redes ART, Rede IAC, Rede LVQ, Rede Counterpropagation, Rede RBF, Rede PNN, Rede Time Delay, Neocognitron, Rede BSB

(DIAS, 2001). A autora apresenta um resumo das principais técnicas de MD juntamente com suas tarefas e os algoritmos mais utilizados:

Tabela 4 - Técnicas de MD, Tarefa e Algoritmos

Técnica	Descrição	Tarefas	Algoritmos
Algoritmos Genéticos	Métodos gerais de busca e otimização, inspirados na Teoria da Evolução, onde a cada nova geração, soluções melhores têm mais chance de ter "descendentes".	Classificação; Segmentação.	Algoritmo Genético Simples; Genitor, CHC; Algoritmo de Hillis; GA-Nuggets; GA-PVMINER.
Árvores de Decisão	Hierarquização dos dados, baseada em estágios de decisão (nós) e na separação de classes e subconjuntos.	Classificação; Regressão.	CART, CHAID, C4.5, C5.0, Quest, ID-3, SLIQ e SPRINT.
Descoberta de Regras de Associação	Estabelece uma correlação estatística entre atributos de dados e conjuntos de dados	Associação	Apriori, AprioriTid, AprioriHybrid, AIS, SETM e DHP.
Raciocínio Baseado em Casos	Baseado no método do vizinho mais próximo, combina e compara atributos para estabelecer hierarquia de semelhança.	Classificação; Segmentação.	BIRCH, CLARANS e CLIQUE
Redes Neurais Artificiais	Modelos inspirados na fisiologia do cérebro, onde o conhecimento é fruto do mapa das conexões neuronais e dos pesos dessas conexões.	Classificação; Segmentação.	Perceptron, Rede MLP, Redes de Kohonen, Rede Hopfield, Rede BAM, Redes ART, Rede IAC, Rede LVQ, Rede Counterpropagation, Rede RBF, Rede PNN, Rede Time Delay, Neocognitron, Rede BSB.

Fonte: Dias (2001, p. 14)

As aplicações de softwares, sejam elas para quais finalidades forem, devem levar em consideração fatores de decisão como o domínio da aplicação, a linguagem na qual a aplicação será desenvolvida, a plataforma do sistema operacional, dentre tantos outros.

Desta forma, as aplicações desenvolvidas para MD, podem ser classificadas sob diversas perspectivas, dependendo de suas técnicas e quais tarefas irão realizar. A seguir é realizado um breve estudo das principais ferramentas de Mineração de Dados disponíveis no mercado.

2.10 FERRAMENTAS DE MINERAÇÃO DE DADOS

Tendo em vista que uma análise exaustiva de todas as ferramentas existentes seria em princípio demasiada trabalhosa e fora de foco desta dissertação, optou-se por analisar as principais ferramentas mencionadas em trabalhos correlatos, em especial o trabalho de dissertação de Cruz (2007), no qual o autor faz uma descrição das ferramentas mais utilizadas.

Em seu trabalho Cruz (2007, p. 42) contabilizou 159 ferramentas de Mineração de Dados, eliminado ainda as que não trabalhassem com RNA ou Máquina de Vetores de Suporte (outra tarefa de Mineração), chegando ao final com 36 ferramentas. Ele caracterizou as ferramentas mediante alguns critérios, o que pode ser observado na Tabela 5:

a) **versão** - final (F) ou beta (B);

b) **licença** - comercial (C), freeware e shareware (F) ou pública (P);

c) **disponibilidade** – se é ou não disponibilizada uma versão de demonstração (Demo) ou a ferramenta é totalmente operacional para download (Download);

d) **aplicação de uso** - academica (A) ou comercial (C) e

e) **a arquitetura** - Stand alone (S), Cliente/Servidor (C/S) ou Processamento Paralelo (PP).

Tabela 5 - Ferramentas segundo as características

Ferramenta	**Versão**	**Licença**	**Disponibilidade**	**Uso**	**Arquitetura**
Alyuda Neuro Inteligence	F	C	S	C	S
BrainMaker	F	C	N	A/C	S
BSVM	F	F	S	A	S
Clementine	F	C	N	C	S/C S
DTREG	F	C	S	A/C	S
EQUBITS Foresight (tm)	F	C	S	A/C	S
EWA Systems	F	C	N	A/C	S/C S
GhostMiner	F	C	N	A/C	S
Gist	F	F	S	A	S
Gornik	F	C	N	C	S/C S
Insightful Miner	F	C	S	A/C	S/C S
Kernel Machines	F	F	S	A	S
Knowledge Miner	F	C	S	A/C	S
KXEN	F	C	N	C	S/C S
LIBSVM	F	F	S	A	S
MATLAB NN Toolbox	F	C	S	A	S
MCubiX from Diagnos	F	C	N	C	S
MemBrain	F	F	S	A	S

Continua...

Conclusão.

NeuralWorks Predict	F	C	S	C	S
NeuroSolutions	F	C	S	A/C	S/C S
NeuroXL	F	C	N	C	S
IPNNL Software	B	F	S	A	S
Oracle Data Mining	F	C	S	C	S,CS,PP
Orange	F	F	S	A	S
PcSVM	B	P	S	A	S
R	F	P	S	A	S
SAS Enterprise Miner	F	C	S	A/C	CS
StarProbe	F	C	S	A/C	S/C S
STATISTICA NN	F	C	S	A	S/C S
SvmFu 3	B	P	S	A	S
SVM-light	F	F	S	A	S
TANAGRA	F	F	S	A	S
HhinkAnalitics	F	C	N	C	CS
Tiberius	F	C	S	A/C	S/C S
Weka	F	P	S	A	S
XLMiner	F	C	S	A/C	S

Fonte: Cruz (2007, p. 45)

Diante desta grande quantidade de ferramentas de Mineração de Dados disponíveis, torna-se necessário fazer uma nova seleção a fim de se tornar prático o trabalho aqui pretendido e para que os objetivos sejam atendidos. Assim sendo os critérios que levaram a escolha da ferramenta foram: a aplicabilidade da tarefa de Descoberta de Regras de Associação, a utilização da técnica de Associação, aliado ao fato da ferramenta ser de licença Livre.

2.11 WEKA

Nessa pesquisa se fez o uso de um software que atende-se os critérios anteriormente citados e em especial que fosse de licença livre. O software escolhido foi Weka, do acrônimo (Waikato Environment for Knowledge Analysis). O software WEKA tem sido bastante utilizado no meio acadêmico em pesquisas que envolvam a área de MD. Sua escolha se justifica por causa de sua ampla aplicabilidade – já que lida com atributos numéricos (reais e inteiros), nominais e caracteres (*string*).

O WEKA foi desenvolvido por universitários da Universidade de Waikato, na Nova Zelândia, no ano de 1999 e sua licença é General Public Licence (GPL), o que significa que é um programa de distribuição e difusão livre [1].

Este software é formado por um conjunto de algoritmos que implementam várias técnicas que são utilizadas para a resolução de problemas reais de MD. O WEKA foi desenvolvido na linguagem Java, cuja principal característica é sua portabilidade, assim sendo podendo ser executado em diversas plataformas, dentre as quais, Windows, MAC Os X e Linux. O único requisito é que o computador possua a máquina virtual Java instalada (MORATE, 2010).

O software WEKA é composto por dois pacotes: um pacote autônomo, para manipulação direta dos algoritmos, usando o formato de dados próprio, e um pacote de classes em Java que implementam estes algoritmos. Nessa segunda forma, é possível desenvolver uma aplicação em linguagem Java que faça uso destes algoritmos e aplicá-los em quaisquer bancos de dados através de uma conexão JDBC (Java DataBase Connectivity).

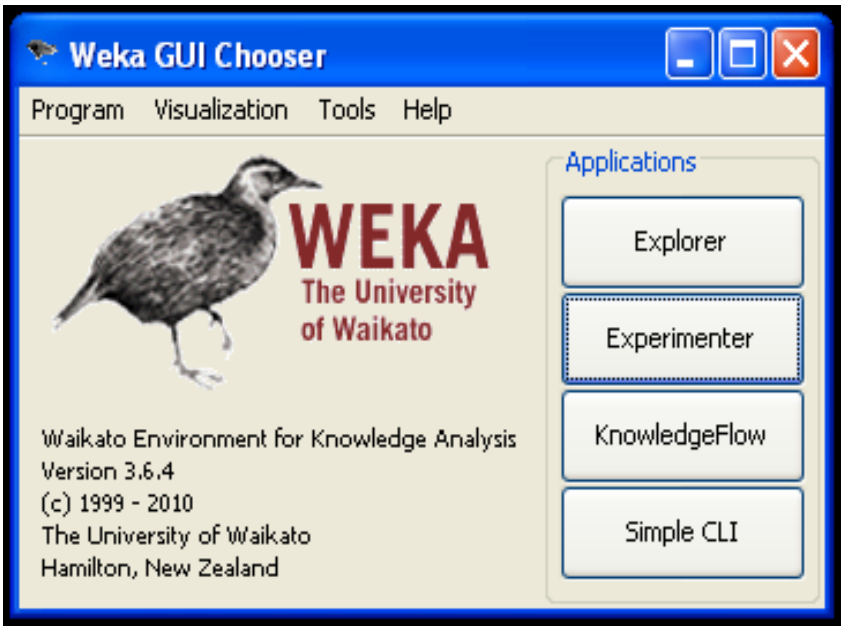


Figura 18 - Tela inicial do software WEKA

Fonte: WEKA (2010).

[1] Disponível em: http://www.cs.waikato.ac.nz/~ml/weka

Os módulos de tarefas disponíveis no WEKA e que serão utilizados para a aplicação na solução proposta são os de Preprocess e Associate, este último utilizando o Algoritmo Apriori para a tarefa de Descoberta de Regras de Associação com a técnica de Associação. Pode-se aplicar os algoritmos diretamente a um conjunto de dados ou fazer uma chamada de seu próprio código Java.

O software possui as seguintes tarefas e técnicas de MD implementadas (MORATE 2010):

a) Tarefas: Pré-processamento de dados e aplicação de filtros, associação, classificação, clusterização, seleção de atributos e visualização de dados;

b) Técnicas: bayes, functions, lazy, meta, trees, rules, cobweb, farthestfirst, makedensity based clusterer, simple k-means, apriori, predictive apriori, tertius, entre outros.

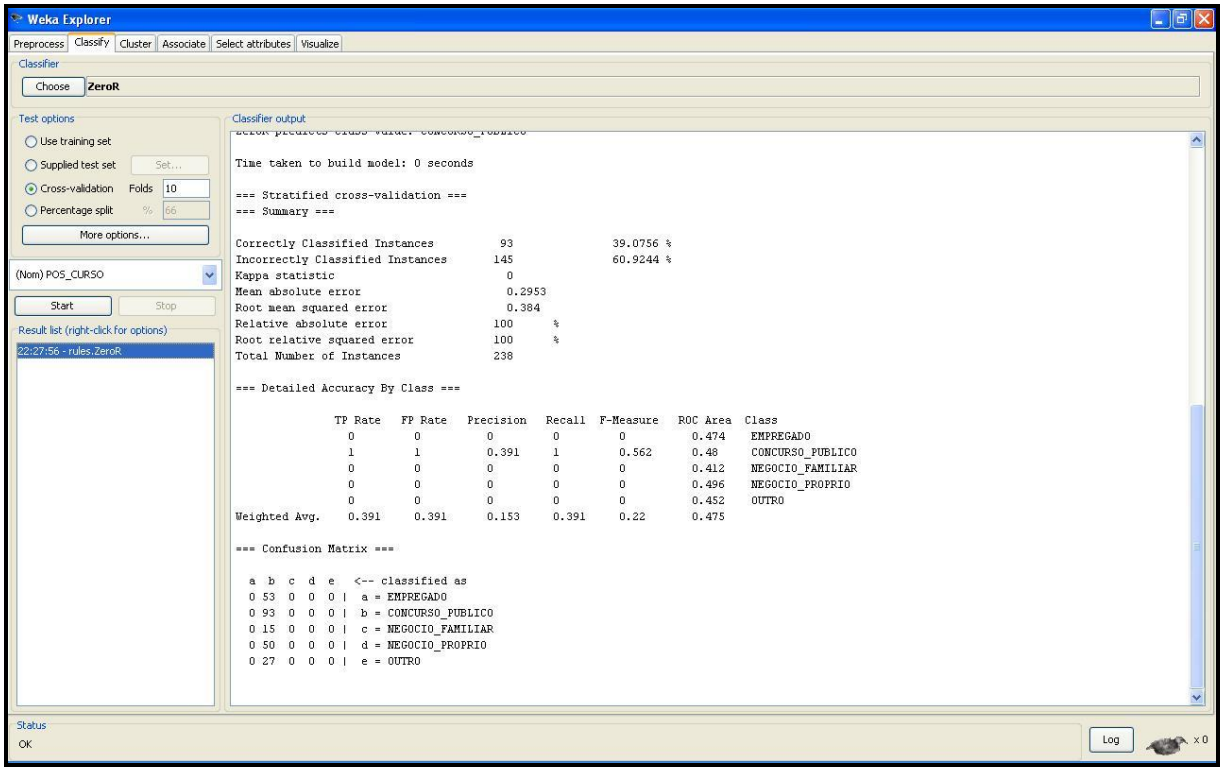


Figura 19 - Exemplo da aplicação da tarefa de classificação

Fonte: WEKA (2010)

Conforme Scoss (2006) e Morate (2010) o WEKA pode ser utilizado de diversas formas, em função do mesmo possuir quatro diferentes interfaces implementadas, que são elas:

a) Explorer: Nesta interface são aplicadas as tarefas e técnicas de MD sobre a base de dados;

b) Experimenter: Esta interface é útil para a aplicação de um ou mais técnicas de classificação sobre uma grande base de dados e em seguida fazer comparações estatísticas sobre elas;

c) Knowledge-flow: Esta é considerada a interface que mais apresenta o funcionamento da ferramenta, uma vez que tem sua representação de forma gráfica;

d) Simple client: Esta interface oferece um local para inserção de comandos. Mesmo possuindo uma aparência considerada simples, é nela que realiza qualquer operação suportada pelo WEKA.

O software WEKA trabalha com um formato de arquivo próprio, denominado ARFF (Attribute Relation File Format). Neste arquivo, que pode ser no formato de texto, estão contidas: a definição do domínio dos atributos e as instâncias, que representam os dados que serão trabalhados.

De acordo com Morate (2010, p. 3), um arquivo ARRF é composto por uma estrutura definida em três partes:

a) Cabeçalho;

b) Declaração dos atributos e

c) Seção de dados.

No cabeçalho tem-se a definição do nome da relação, a declaração dos atributos contém uma lista de todos os atributos (um atributo por linha), com os nomes dos atributos e seus tipos. O WEKA trabalha com os seguintes tipos de dados:

a) Numérico: trabalhando com números reais e decimais;

b) Inteiros: números sem decimais;

c) Datas;

d) String: com ressalva para a substituição de espaços em branco por *underline e;*

e) Enumerados: nos quais os tipos são previamente definidos pelo usário, Ex.: Sexo {Masc, Fem}.

O software WEKA consegue importar dados em arquivos nos formatos ARFF, CSV, C4.5 e binário. O WEKA consegue também acessar os dados de uma URL ou de um banco de dados, utilizando a linguagem SQL (Structured Query Language) por meio da conexão JDBC. A Figura 20 mostra um exemplo de um arquivo final no formato ARRF,

```
% Arquivo Exemplo. Comentário

@relation exemplo

@attribute matricula NUMERIC
@attribute semestre REAL
@attribute ano REAL
@attribute sexo {M, F}
@attribute data_acesso DATE "dd-MM-yyyy HH:mm"

@data

20091234,1,2009,M,"12-04-2009 12:23"
20091334,1,2009,F,"22-03-2009 09:35"
20092625,2,2009,M,"05-09-2009 16:32"
20092289,1,2009,M,"01-03-2009 18:10"
20092513,2,2009,M,"29-08-2009 17:45"
.
.
.

20092689,2,2009,f,"23-10-2009 15:56"
```

Figura 20 - Exemplo de arquivo no formato ARRF

Fonte: Adaptado de Morate (2010, p. 5)

Nesta pesquisa utilizou-se a ferramenta WEKA para as tarefas de Associação, Classificação e Clusterização. O fator que mais influenciou na escolha da ferramenta foi o fato da mesma possuir as tarefas e técnicas definidas para a pesquisa e possuir uma interface gráfica para a visualização dos resultados.

2.12 GESTÃO DE IES

As IES foram criadas e mantidas muitas vezes por empreendedores que advinham grande parte do segmento da educação, e outros que viam neste nicho de mercado uma boa oportunidade de negócios.

Isto traz uma reflexão sobre a gestão adotada por ambos, primeiro sobre os que são oriundos da educação, detêm o saber da sala de aula, o conhecimento e ensino das metodologias, inclusive as de administração, mas isto não lhes garante o sucesso administrativo. Segundo sobre os oriundos de outros setores, que apesar de serem experientes nas questões de organização, enfrentam a complexidade da razão de ser das IES.

O sistema universitário brasileiro tem se expandido nos últimos anos, o que trouxe consigo uma série de novas exigências para as IES. Citam-se como novas exigências a criação e desenvolvimento de novas competências atribuídas aos gestores, os quais têm que atuar nos diversos setores da instituição. Estas alterações vêm de encontro as normas estabelecidas pela Lei de Diretrizes Básicas da Educação Nacional (LDB) – lei nº 9.394/96.

A valorização do planejamento estratégico, a definição de metas mais claras, a administração de projetos e as novas medidas que visam os objetivos finais de qualquer organização, tiveram forte influência do aumento da competitividade gerado pela globalização.

Para Alves (2005, p. 37) "A gestão baseada nos objetivos estratégicos tornou-se fundamental para o pleno desenvolvimento de qualquer tipo de organização, estando aí incluídas as IES particulares".

Colenci Jr et. al (2008, p. 3) destaca que:

> Com o crescimento do número de instituições particulares, o ensino superior passou de um direito a um negócio de prestação de serviços, com fins lucrativos. Ocorre que muitas instituições colocam seu foco na rentabilidade da empresa, em detrimento da sua responsabilidade social de preparação de um cidadão consciente de seus atos e comprometido com o desenvolvimento da sociedade sobre as quais deverão estabelecer suas bases administrativas.

Nobrega (2004 apud BRAGA; MONTEIRO, 2005, p. 150) trata a gestão como sendo um processo prático, levando apenas em consideração o resultado obtido. O mesmo autor retrata a gestão como uma pesquisa de critérios a serem utilizados na tomada de decisões.

Corroborando com isto Colceni Jr et. al. (2008, p. 4), dizem que:

> A importância do planejamento estratégico é de preparar a empresa para o crescimento em direção à visão da empresa ou da instituição em longo prazo. Comumente, muitos gestores agem sem estruturar suas decisões, sem saber quais as verdadeiras vocações e não conseguem estabelecer as políticas e as diretrizes

Outra postura a ser observada é conforme afirmam Tachizawa e Andrade (2006): gestão tem seu conceito ampliado, mediante a junção de atividades de controle, quer sejam internas ou externas, incluindo indicadores de gestão, de desempenho e qualidade.

Neste contexto Alves (2006, p. 30) considera que:

> Esse modelo sistêmico permite que a IES proceda a uma análise do meio ambiente para definir a sua estratégia em longo prazo, a partir de um provável cenário e dos objetivos institucionais. A identificação dos processos sistêmicos-chave fornece o suporte e estabelece as condições indispensáveis ao delineamento estratégico. Na abordagem sistêmica, valoriza-se o todo da organização, ou um conjunto de partes em constante interação.

Tachizawa (2006) tenta estabelecer, através de seu modelo de gestão, um entendimento dos processos sistêmicos, baseado no uso do instrumento analítico fluxo básico da instituição. Para o autor, processo sistêmico é um conjunto de atividades que produzem resultados, podendo ser controlado por uma ou mais ferramentas e/ou técnicas, efetuadas por várias pessoas.

Alves (2005, p. 33) entende que:

> Além de criar condições para que os processos-chave se estabeleçam, o modelo sistêmico atua na revisão da configuração organizacional da instituição, ou seja, nos aspectos que não estão relacionados à atividade fim. Entre eles se destacam as atividades administrativas e os demais recursos que necessitam serem revisados na busca de convertê-los em produtos que, no caso em tela, são os serviços educacionais.

O enfoque sistêmico direciona o entendimento do ambiente externo como sendo um fator importante a ser considerado na gestão das IES, mediante aos desafios apresentados por este ambiente.

O gestor de qualquer IES tem várias responsabilidades, sejam elas pedagógicas, administrativas ou financeiras, necessitando controlar e coordenar todos os ambientes integrantes da IES, a fim de transformar o ambiente num ambiente de trabalho contínuo e próspero.

2.12.1 Ferramentas de Gestão

As mudanças impostas pela nova ordem da economia ao setor de ensino superior levaram muitas IES a traçarem novas estratégias para se manterem no mercado. O funcionamento das IES está sob influência direta ou indiretamente das alterações sofridas pelo contexto externo em que se encontram e o controle destas alterações afeta a sua gestão, o que pode ser minimizado com o uso de boas ferramentas de gestão e a utilização correta das informações.

Chiavenato (2000, p. 599) sustenta que "as organizações precisam adaptar-se e incorporar tecnologia que provém do ambiente geral para não perderem a sua competitividade."

Para Almeida e Almeida (2006, p. 104):

> [...] o uso das tecnologias na gestão escolar revela novos papéis dos seus profissionais - como organizadores de informações, criadores de significados e líderes - na tomada compartilhada de decisões. Esses profissionais encontram nas tecnologias, especialmente naquelas de Informação e Comunicação, o suporte adequado para o desenvolvimento de suas atividades, apoiadas em informações provenientes de fontes distintas, internas ou externas ao sistema, e na colaboração com seus pares e com a comunidade escolar.

As IES têm uma questão a resolver que é a de estruturar e disponibilizar para seus gestores as informações geradas pelos seus sistemas de gestão, possibilitando a transformação destas informações em tomadas de decisões estratégicas.

Moran (apud VIEIRA, ALEMIDA e ALONSO, 2003, p. 153) retrata que:

> Os principais colégios e universidades do Brasil utilizam esses programas integrados de gestão. Diminuem a circulação de papéis, formulários, ofícios, tão comuns nas escolas públicas e convertem todas as informações em arquivos digitais que vão sendo catalogados, organizados em pastas eletrônicas por assunto, assim como o fazemos na secretaria, só que ficam armazenados num computador principal, chamado servidor.

Observa-se que a maioria das IES não está apta para enfrentar o cenário atual do ensino superior. Sua sobrevivência em meio a este ambiente incerto está presa aos resultados gerados pela gestão estratégica.

O mesmo Moran salienta que:

> Existem no mercado programas de gestão tecnológica que têm como princípio integrar todas as informações que dizem respeito à escola. Eles possuem um banco de dados com todas as informações dos alunos, famílias, professores, funcionários, fornecedores e, do ponto de vista pedagógico, bancos de informações para as aulas, para as atividades de professores, dos alunos, bibliotecas virtuais, etc. Todo esse conjunto de informações costuma circular primeiro numa rede interna, chamada Intranet, à qual alunos, professores e pais podem ter acesso, em diversos níveis, por meio de senhas. Num segundo momento, a Intranet se conecta com a Internet, abre-se para o mundo através de uma página WEB, uma página na Internet, que tem como finalidade imediata a divulgação da escola - marketing -, e como finalidade principal, facilitar a comunicação entre todos os participantes da comunidade escolar. Moran (apud VIEIRA, ALEMIDA E ALONSO, 2003, p. 152).

Entende-se que a gestão da informação é um processo que consiste no ato da geração, coleta, assimilação e aproveitamento da informação, tornando a IES mais inteligente e competitiva, visando obter os melhores resultados em produtividade e capacidade de inovação das IES.

Tanto o ambiente empresarial quanto o das IES, possuem ao seu dispor diversas ferramentas para gestão, entre as quais podem ser citadas, o planejamento estratégico, os programas de qualidade total, a Gestão Participativa, a reengenharia, downsizing, a terceirização, o empowerment, Customer Relationship Management (CRM), Business Inteligence, Balanced Scorecard, os Sistemas de Informações, todas elas visando segundo seus defensores a solução dos problemas gerenciais.

3 TRABALHOS RELACIONADOS

Esta dissertação apresenta o uso de técnicas de Mineração de Dados em Ambientes de Gestão Educacional, com intuito de auxiliar os gestores destes ambientes nas tomadas de decisões, relativas ao uso dos mesmos. Para tanto nas seções seguintes apresentam-se trabalhos voltados para a Gestão da Informação em IES e suas ferramentas, o uso da Mineração de Dados em diversos ambientes e especificamente em Ambientes de Gestão Educacional, a aplicação das Tarefas de Associação, Classificação e Clusterização, conjuntamente com suas respectivas técnicas.

3.1 GESTÃO DA TECNOLOGIA DA INFORMAÇÃO EM IES

Tendo como objetivo a aplicação de técnicas de Inteligencia Competitiva, aliada a mineração de textos, o trabalho de Furtado (2004) faz uso da ferramenta comercial Coppernic[2] para busca de informações a respeito do mercado de Intituições de Ensino Superior na cidade do Rio de Janeiro e sua região metropolitana, e a software TEMIS[3] para a realização das tarefas de clusterização e categorização dos documentos encontrados pela ferramenta anterior.

O modelo de Inteligência Competitiva proposto pela autora foi realizado em seis etapas assim definidas:

a) Entendimento do mercado das Instituições de Ensino Superior Privado: onde procurou obter informações sobre o mercado das universidades, com base em parâmetros previamente determinados, para a criação de documentos aos quais posteriormente passaram por um processo de extração de valores nos textos;

b) Busca de dados: por meio da busca em ambientes governamentais, associações da área de Ensino Superior, instituições de pesquisa, revistas e jornais de grande circulação on-line;

[2] Disponível em: http://www.copernic.com
[3] Disponível em: http://www.temis.com

c) Solução: foi desenvolvida a construção dos atributos considerados relevantes para uso na extração das informações dispostas nos docuemtnos armazenados;

d) Aplicação do software Insight Discover^{tm} Extractor – IDE que fez a extração das informações com base nos atributos anteriormente especificados;

e) Construção da solução: nesta etapa foram utilizados os softwares Insight Discover^{TM} Clusterer – IDC e Insight Discover^{TM} Categorizer – IDK, para a categorização e clusterizaçãodos documentos;

f) Busca do conhecimento na coleção de textos: a partir dos documetos armazenados obteve-se as relações baseadas nos atributos relevantes os quais geraram algumas conclusões.

Para Furtado (2004, p. 98):

> A integração entre as áreas de Mineração de Textos com Inteligência Competitiva é possível e necessária, pois as empresas possuem uma grande quantidade de informação disponível para análise e essa análise torna-se inviável caso não seja realizada com o auxílio de técnicas e ferramentas computacionais.

Observou-se que a aplicação das técnicas de Text Mining auxiliou no processo de tomada de decisões, por meio de informações que se encontravam ocultas nos documentos armazenados, fazendo que a IES conseguisse obter vantagem competitiva.

3.2 UTILIZAÇÃO DE MINERAÇÃO DE DADOS EM GERAL

Em seu trabalho Shiba(2008) formaliza um modelo de processo de KDD, no qual foram definidas três etapas principais:

a) Pré-processamento;

b) Mineração de Dados;

c) Pós-processamento.

Na etapa de Pré-processamento, a autora trabalhou com o desenvolvimento de um programa que elaborou a geração de um arquivo com dados unificados para posterior aplicação de técnicas de MD.

Na etapa de mineração de dados foi preparardo o ambiente de análise de dados, tendo como o problema apresentado e em seguida realizou-se a escolha da técnica de Mineração de Dados. A primeira base utilizada para os testes continha 144 registros, porém devido a uma grande quantidade de registros incompletos, cerca de 10%, foi necessário a utilização de uma base com maior quantidade. Uma segunda base de dados foi utilizada para testes, esta continha 56.000 registros.

Para Shiba(2008, p. 85) "[...] Elaborar um modelo com uma amostra pequena pode ser útil quando as classes estão representadas proporcionalmente em relação a uma amostra maior[...]", isto quer dizer que se os dados a serem trabalhados numa amostra menor estiverem proporcionalmente distribuidos em relação a um amostra com maior quan tidade de registros, o resultado da aplicação da MD não sofrerá influência.

Finalizando o trabalho, a autora destaca dois objetivos: a) a avaliação do desempenho aplicada ao modelo de testes e b) a disponibilização da base de conhecimento gerada. No trabalho foi explorada a aplicabilidade de um modelo de extração do conhecimento por meio de técnicas de MD, aplicado num grupo de clientes a fim de verificar ações de retenção. Tomando por base os resultados obtidos, foi identificado um cenário evolutivo em relação a análise dos dados, o que tornará possível o planejamento de ações de vendas direcionadas ao perfil dos grupos de clientes selecionados.

Outra trabalho foi desenvolvido por Machado Filho (2006) no qual o autor propos o desenvolvimento de um ambiente de MD, utilizando dois modelos de Redes Neurais Artificiais, Multi Layer Perceptron (MLP) e Radial Basis Function (RBF), em problemas de classificação e prediçao de dados. O autor incorpoou em seu ambiente a técnica do Algoritmo Genético para a determinação da topologia da rede e na extração das regras.

Machado Filho fez uso da plataforma MS Excel, por esta apresentar algumas características que considerou importantes, como: os recursos gráficos, a utilização de funções já existentes, a importação e exportação dos dados e principalmente a integração da ferramenta com os demais aplicativos do pacote MS Office.

A utilização de técnicas de Mineração de Dados na detecção de outliers em auxílio à auditoria operacional com um estudo de caso com dados do sistema de informações hospitalares é o trabalho de Bodini Junior (2009), que propõe o uso de Algoritmos de agrupamento Nebuloso e Máquina de Vetor Suporte para a evidencicação de Outliers, que são registros encontrados em bases de dados que se destacam dos demais por sua falta de semelhança.

3.3 MINERAÇÃO DE DADOS EM AMBIENTES EDUCACIONAIS

Ramaswani e Bhaskaran (2009) em seu artigo abordam a extração da informação em ambientes educacionais com meio de avaliar o desempenho dos alunos. A presente investigação centra-se em várias técnicas de recurso de seleção, que é um dos mais importantes e frequentemente utilizados no pré-processamento de dados para Mineração de Dados. Os procedimentos gerais sobre seleção de recursos em termos de método de filtro é seguido com o efeito de técnicas de seleção de recursos em um banco de dados contendo informações de alunos do ensino secundário.

A seleção de recursos tem sido um campo ativo e fecundo da área de pesquisa em reconhecimento de padrões, aprendizado de máquina, as estatísticas e as comunidades de Mineração de Dados. O objetivo principal de seleção de recursos é escolher um subconjunto de variáveis de entrada de recursos para eliminar os que são irrelevantes ou sem informações preditivas. A seleção de recursos tem provado na teoria e na prática para ser eficaz em aumentar eficiência de aprendizagem, aumentando a precisão preditiva e reduzir a complexidade dos resultados de aprendizado. A seleção de recursos no aprendizado supervisionado tem por objetivo principal encontrar um subconjunto recurso que produz maior precisão na classificação.

Cardoso e Machado (2008) utilizaram a plataforma Lattes como base para a aplicação e análise de uma ferramenta de Mineração de Dados com o objetivo de extrair informações a respeito da produção científica de seus professores e colaboradores da Universidade Federal de Lavras (UFLA). Inicialmente foram selecionados mais de mil currículos, destes 575 foram os selecionados por dados mais específicos para a pesquisa.

As autoras utilizaram quatro exemplos para mostrar a aplicação das Regras de Associação. No primeiro exemplo fizeram a associação entre a quantidade de publicações

contidas na Plataforma Lattes, desenvolvidas por pessoas que trabalha na UFLA e as pessoas que não trabalham. Como resultado obtiveram uma amostra com 1.977 publicações das quais 55% são publicações de pessoas que não estavam trabalham na UFLA quando da publicação e o restante 45% de pessoas que estavam trabalhando na UFLA no momento da publicação.

No segundo exemplo as autoras analisam os resultados obtidos no exemplo anterior, mais especificamente os 55% de pessoas que tiveram alguma publicação, mas não estavam trabalhando. Como resultado obtiveram a quantidade de 1.062 publicações. Cardoso e Machado (2008) alertam "[...] uma pessoa, ao receber afastamento total para treinamento, fazer pós-graduação, por exemplo, não está atuando na Ufla durante o período do afastamento".

O terceiro exemplo apresentado pelas autoras faz uma relação entre as publicações cadastradas e o tempo de serviços prestados à UFLA por seus autores, tendo como resultado a caracterização de que a maioria das publicações foi realizada após o ingresso do autor na UFLA. (MACHADO e CARDOSO , 2008).

No quarto e último exemplo as autoras fazem a junção de duas situações: o local de realização de uma pós-gradução, se no exterior ou no Brasil e o número de publicações feitas. O resultado apresenta uma relação direta entre a quantidade de publicações (74) e pessoas (34) que fizeram a pós-graduação no Brasil. Machado e Cardoso (2008) ressaltam que: "[...] A média de publicações no exterior de pessoas que cursaram a pós-graduação fora do Brasil é maior numa razão de 2,71 com relação às pessoas que cursaram pós-graduação no Brasil".

Cardoso e Machado (2008) elaboraram mais quatro análises:

a) Análises de regras de associação e de padrão seqüencial: onde analisam o tempo decorrido entre a conclusão do mestrado e o início do doutorado relizado pelas pessoas que trabalham na UFLA;

b) Análises de padrões seqüenciais: duas consultas foram realizadas: a primeira, analisa a relação entre o tempo de cadastro do currículo na Plataforma Lattes e o tempo de vínculo profissional com a instituição e a segunda, analisa a relação temporal entre o tempo de serviço e o ano de início das pesquisas realizadas pelo colaborador;

c) Análises de *cluster*: através da identificação de um *cluster* considerado desconhecido, analisaram o tempo de duração das pesquisas realizadas pelos colaboradores da instituição;

d) Análise de classificação e predição: esta teve por objetivo a análise entre as atividades exercidas e as publicações realizadas, onde buscaram saber em qual nível de atividade (ensino, pesquisa e direção) ocorriam mais publicações.

Os resultados apontam que o uso de técnicas de MD traz ao gestor a possibilidade de uma gestão das informações mais eficaz, uma vez que a integração dos dados armazenados gera informações úteis para a tomada de decisões.

Dias et al (2008), no seu artigo científico "Aplicação de Técnicas de Data mining no Processo de Aprendizagem na Educação a Distância", apresentam um estudo de caso aplicado no ambiente de aprendizagem denominado LabSQL. O LabSQL de aprendizagem utilizado para o ensino da linguagem SQL.

Conforme Dias et all (2008):

> No ambiente de aprendizagem de SQL, o aprendiz visualiza o texto didático acompanhado de exemplos executáveis. Juntamente com o conteúdo são apresentadas listas de exercícios para que o aprendiz treine suas habilidades. Existem três tipos de exercícios: objetivos de múltipla escolha (ou V/F); não objetivos descritivos e exercícios de programação.

Os dados armazenados no banco de dados do ambiente serviram para a aplicação das tarefas de Mineração de Dados: Árvore de Decisões e Redes Bayesianas. Foram analisadas sete turmas, com uma média de trinta alunos por um período de dois semestres, na modalidade de ensino semi-presencial. Dentre as turmas analisada quatro delas eram de pós-graduação e as demais três de graduação.

Os autores analisaram 272 registros dos usuários, com os quais trabalharam com 18 atributos: sexo, códigos: do curso, do tipo do curso, da disciplina, da turma, do coordenador do curso, o tempo levou para inscrição na turma após o início da inscrição, se trabalhou em equipe, se uso a agenda de anotações de sistema, os totais: de pontos obtidos na resolução de problemas, de problemas resolvidos, a média de pontos dos problemas resolvidos, a quantidade de acessos as páginas do ambiente; se ficou acima da média de acessos de todas as

turmas e em sua turma (sim ou não); se ficou acima da média de pontos de todas as turmas e se ficou acima da média de pontos da sua turma, conforme **Erro! Fonte de referência não encontrada.**.

Na técnica de Redes Bayesianas os autores utilizaram software Bayesware Discoverer e observaram que existe uma forte influência na demora para inscrição no curso em relação a média da quantidade de acessos, sendo esta influenciada pelo curso no qual o aluno irá participar.

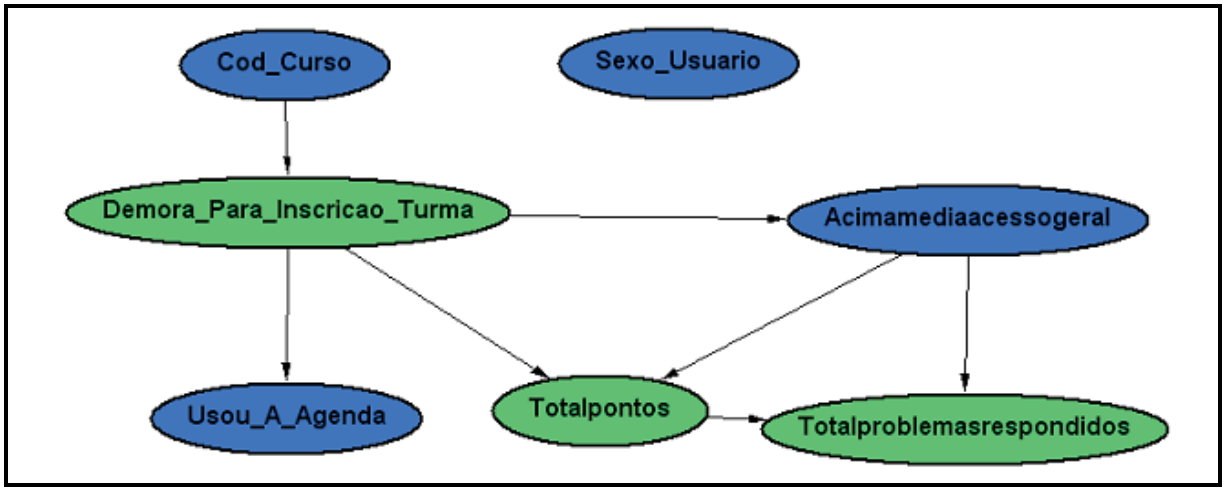


Figura 21 - Rede Bayesiana para Análise da Demora para Inscrição

Fonte: Dias et al (2008, p. 110)

Outra técnica aplicada pelos autores foi a de Árvore de Decisão, utilizando a ferramenta WEKA, a qual fez uso da tarefa de classificação, implementada com o algoritmo J48 em validação cruzada. Esta técnica foi aplicada com objetivo de verificar a precisão dos modelos de classificação utilizados, onde obtiveram uma média de 83,13% de acurácia. Para os autores a combinação de MD com ambientes de EAD, permite a análise das praticas feitas pelos usuários, trazendo benefícios para os envolvidos no processo de ensino-aprendizagem.

A descoberta do conhecimento em base de dados é o tema de Scoss (2006) em seu trabalho de especialização. A autora faz uso da tarefa de clusterização para análise do desempenho dos docentes da Universidade do Extremo Sul Catarinense, com objetivo de analisar o perfil dos docentes no contexto da Avaliação Institucional.

A autora realizou sua pesquisa fazendo uso das tarefas de clusterização e classificação, as quais foram aplicadas em uma base de dados que continha 36.672 instâncias e 21 atributos. Na tarefa de clusterização foram definidos quatro clusteres, sendo estes o número de áreas de

conhecimento disponíveis na Universidade, área de licenciatura, área de Saúde e Biológicas, área de Sociais e Aplicadas e área de Engenharia e Tecnologia.

Já na tarefa de associação a autora utilizou o algoritmo ZeroR, que caracteriza o esboço de uma única regra, tendo como base o item que mais vezes aparece na base de dados. Diante dos resultados obtidos foi possível gerar informações sobre o desempenho dos docentes da universidade e a partir destas informações a sugestão de ações a serem tomadas pelos gestores.

Amorim, Barone e Mansur (2008) demonstram em seu artigo intitulado "Técnicas de Aprendizado de Máquina Aplicadas na Previsão de Evasão Acadêmica", apresentado no XIX Simpósio Brasileiro de Informática na Educação no ano de 2008, a eficiência do uso de técnicas de MD aplicadas à evasão acadêmica, no qual os autores aplicam a técnica de aprendizado de máquina.

Os autores utilizaram como base para sua pesquisa uma IES localizada no município de Goytacazes no RJ. O universo pesquisado era composto por 8.073 matriculas que foram realizadas nos seguintes cursos oferecidos pela IES: 1.765 matrículas no curso de Administração, 1.160 no curso de Engenharia da Produção, 2.642 no curso de Fisioterapia e 2.506 matrículas no curso de Pedagogia. A **Erro! Fonte de referência não encontrada.** mostra o fluxo de interação do sistemas criado pelos autores.

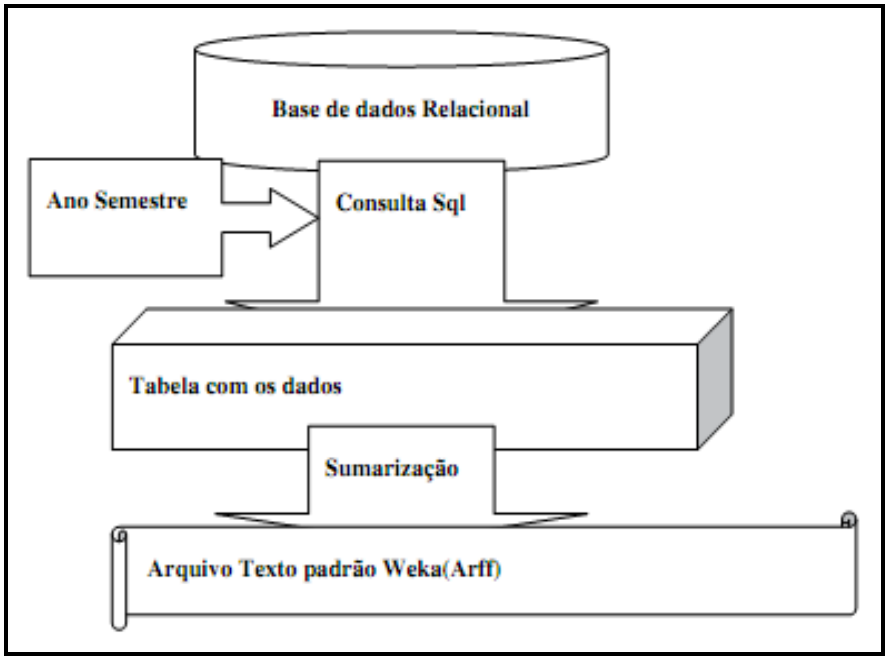


Figura 22 - Arquitetura do sistema mapeador

Fonte: Amorim, Barone e Mansur (2008).

Após extraírem os dados do banco de dados da IES, os autores elegeram os classificadores que foram utilizados. Estes classificadores encontram-se disponíveis na ferramenta de mineração WEKA e foram os seguintes:

a) J48 – baseado em árvores de decisão;

b) SMO – baseado em maquinas de vetores de suporte;

c) Bayes Net – baseados em métodos bayesianos.

Para avaliar a eficiência dos classificadores escolhidos os autores elaboraram um levantamento das matrículas realizadas no início do ano de 2002 até o segundo semestre do ano de 2006. Foram consideradas as matrículas novas, as re-matrículas e os trancamentos ocorridos neste período. A Tabela 6 resume os dados encontrados pelos autores:

Tabela 6 - Tabela de evasão por curso

Cursos	Total de Matrícula e re-matrículas	Total de trancamentos	Percentual semestral de evasão
Administração	1765	298	16,88%
Engenharia da Produção	1160	363	31,29%
Fisioterapia	2624	558	21,12%
Pedagogia	2506	563	22,47%
Geral	8073	178	22,07%

Fonte: Amorim, Barone e Mansur (2008).

A experiência foi considerada bem sucedida, fato constatado através da comparação entre a realidade encontrada na base e o resultado apresentado pelos classificadores. Cada um dos classificadores foi aplicado nos valores anteriormente mencionados e obtiveram como resultado:

Tabela 7 - Grau de acurácia dos classificadores na evasão

	Bayes Net	SMO	J48
Classificação Correta	89,7084%	91,2521%	89,6512%
Classificação Incorreta	10,2916%	8,7479%	10,3488%

Fonte: Amorim, Barone e Mansur (2008).

Observa-se que existe uma diferença bem baixa entre os três classificadores utilizados. Para os autores o artigo contribuiu para a apresentação da técnica de aprendizado de máquinas e na escolha do melhor classificador, o que propicia ao gestor da IES, novos horizontes em relação ao problema de evasão acadêmica. Com base nesta nova informação, novas estratégias de retenção podem ser tomadas.

4 APLICAÇÃO DAS TÉCNICAS DE MD EM AGE

Neste capítulo é apresentado como foram realizadas as tarefas de associação, classificação e clusterização, bem como a análise dos resultados provenientes dessas tarefas. Na Seção 4.1, são apresentadas características gerais sobre o problema a ser tratado. A seguir, na Seção 4.2, são apresentados os experimentos de associação, sendo em seguida, na Seção 4.3 apresentados os experimentos da tarefa de classificação. E por fim, na Seção 4.4, são apresentados os experimentos realizados usando a tarefa de clusterização.

Inicialmente foram colhidos 238 registros relacionados as questões de identificação dos acadêmicos ingressantes e 165 registros relacionados as questões dos egressos.

Após uma análise prévia da base de dados, foi detectado que haviam vários itens sem o devido preenchimento ou com erros de digitação ou ainda com valores redundantes. Os registros restantes foram adequados por meio de uma criteriosa avaliação manual, com exceção de alguns registros que foram excluídos da base, por não apresentarem condições de correção. Por fim, trabalhou-se com 238 registros de ingressantes e 165 registros para egressos.

4.1 CARACTERÍSTICAS DO PROBLEMA A SER TRATADO

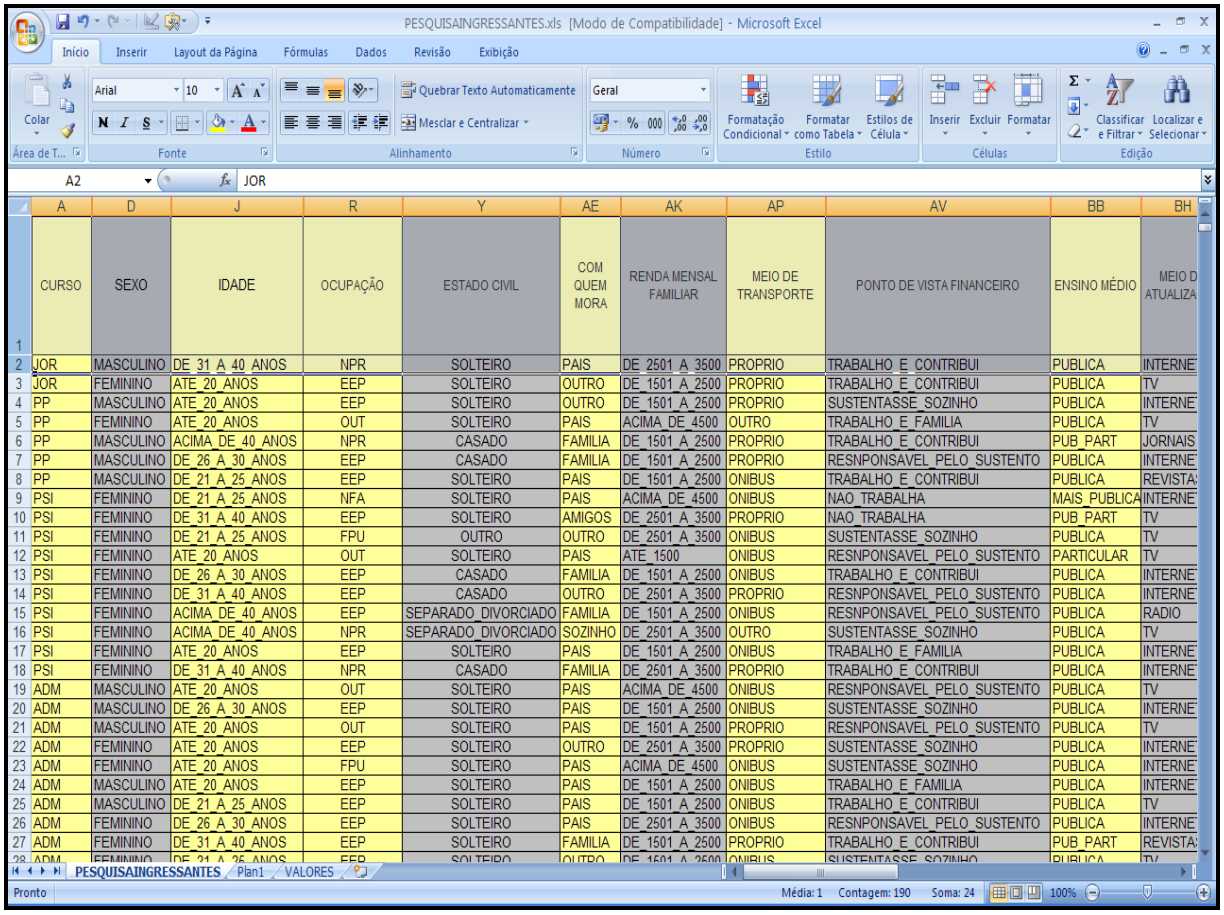
Conforme o objetivo geral e os específicos propostos no início deste trabalho e alinhados com a metodologia empregada para a aquisição das informações a serem mineradas, num primeiro momento buscou-se compreender a necessidade dos gestores da IES, a fim de poder oferecer uma solução para seus questionamentos.

Para Dias (2001) na MD existe a possibilidade de não existir um problemas real a ser solucionado, uma vez que a MD pode ser utilizada como um processo de descoberta, onde nem sempres é feito o levantamento das suposições a serem discutidas. Assim sendo o primeiro passo para se decobrir conhecimento em bases de dados é uma correta definição do problema a ser tratado.

Nesta etapa da pesquisa entrou-se em contato com a diretoria da IES a ser pesquisada, solicitando o acesso à base de dados, para que fossem realizadas as etapas da Mineração de Dados. A ela foi solicitada a permissão de acesso, mas observou-se que apenas o acesso não

permitiria à obtenção de todos os dados realmente necessários a pesquisa, o que levou a solicitação de uma cópia da base. Processo este, que vale ressaltar, foi um dos entraves na elaboração desta pesquisa, pois tratam-se de dados considerados confidenciais e estratégicos a IES.

A ferramenta WEKA, transformou os dados utilizados nesta pesquisa em regras com informações úteis e mais claras aos gestores da IES. A ferramenta tornou possível a interpretação e compreensão dos resultados por parte de todos os envolvidos na realização desta pesquisa, sendo que os resultados foram considerados satisfatórios por todos.



PESQUISAINGRESSANTES.xls [Modo de Compatibilidade] - Microsoft Excel

	A	D	J	R	Y	AE	AK	AP	AV	BB	BH
1	CURSO	SEXO	IDADE	OCUPAÇÃO	ESTADO CIVIL	COM QUEM MORA	RENDA MENSAL FAMILIAR	MEIO DE TRANSPORTE	PONTO DE VISTA FINANCEIRO	ENSINO MÉDIO	MEIO D ATUALIZA
2	JOR	MASCULINO	DE_31_A_40_ANOS	NPR	SOLTEIRO	PAIS	DE_2501_A_3500	PROPRIO	TRABALHO_E_CONTRIBUI	PUBLICA	INTERNE
3	JOR	FEMININO	ATE_20_ANOS	EEP	SOLTEIRO	OUTRO	DE_1501_A_2500	PROPRIO	TRABALHO_E_CONTRIBUI	PUBLICA	TV
4	PP	MASCULINO	ATE_20_ANOS	EEP	SOLTEIRO	OUTRO	DE_1501_A_2500	PROPRIO	SUSTENTASSE_SOZINHO	PUBLICA	INTERNE
5	PP	FEMININO	ATE_20_ANOS	OUT	SOLTEIRO	PAIS	ACIMA_DE_4500	OUTRO	TRABALHO_E_FAMILIA	PUBLICA	TV
6	PP	MASCULINO	ACIMA_DE_40_ANOS	NPR	CASADO	FAMILIA	DE_1501_A_2500	PROPRIO	TRABALHO_E_CONTRIBUI	PUB_PART	JORNAIS
7	PP	MASCULINO	DE_26_A_30_ANOS	EEP	CASADO	FAMILIA	DE_1501_A_2500	PROPRIO	RESNPONSAVEL_PELO_SUSTENTO	PUBLICA	INTERNE
8	PP	MASCULINO	DE_21_A_25_ANOS	EEP	SOLTEIRO	PAIS	DE_1501_A_2500	ONIBUS	TRABALHO_E_CONTRIBUI	PUBLICA	REVISTA
9	PSI	FEMININO	DE_21_A_25_ANOS	NFA	SOLTEIRO	PAIS	ACIMA_DE_4500	ONIBUS	NAO_TRABALHA	MAIS_PUBLICA	INTERNE
10	PSI	FEMININO	DE_31_A_40_ANOS	EEP	SOLTEIRO	AMIGOS	DE_2501_A_3500	PROPRIO	NAO_TRABALHA	PUB_PART	TV
11	PSI	FEMININO	DE_21_A_25_ANOS	FPU	OUTRO	OUTRO	DE_2501_A_3500	ONIBUS	SUSTENTASSE_SOZINHO	PUBLICA	TV
12	PSI	FEMININO	ATE_20_ANOS	OUT	SOLTEIRO	PAIS	ATE_1500	ONIBUS	RESNPONSAVEL_PELO_SUSTENTO	PARTICULAR	TV
13	PSI	FEMININO	DE_26_A_30_ANOS	EEP	CASADO	FAMILIA	DE_1501_A_2500	ONIBUS	TRABALHO_E_CONTRIBUI	PUBLICA	INTERNE
14	PSI	FEMININO	DE_31_A_40_ANOS	EEP	CASADO	OUTRO	DE_2501_A_3500	PROPRIO	RESNPONSAVEL_PELO_SUSTENTO	PUBLICA	INTERNE
15	PSI	FEMININO	ACIMA_DE_40_ANOS	EEP	SEPARADO_DIVORCIADO	FAMILIA	DE_1501_A_2500	ONIBUS	RESNPONSAVEL_PELO_SUSTENTO	PUBLICA	RADIO
16	PSI	FEMININO	ACIMA_DE_40_ANOS	NPR	SEPARADO_DIVORCIADO	SOZINHO	DE_2501_A_3500	OUTRO	SUSTENTASSE_SOZINHO	PUBLICA	TV
17	PSI	FEMININO	ATE_20_ANOS	EEP	SOLTEIRO	PAIS	DE_1501_A_2500	ONIBUS	TRABALHO_E_FAMILIA	PUBLICA	INTERNE
18	PSI	FEMININO	DE_31_A_40_ANOS	NPR	CASADO	FAMILIA	DE_2501_A_3500	PROPRIO	TRABALHO_E_CONTRIBUI	PUBLICA	INTERNE
19	ADM	MASCULINO	ATE_20_ANOS	OUT	SOLTEIRO	PAIS	ACIMA_DE_4500	ONIBUS	RESNPONSAVEL_PELO_SUSTENTO	PUBLICA	TV
20	ADM	MASCULINO	DE_26_A_30_ANOS	EEP	SOLTEIRO	PAIS	DE_1501_A_2500	ONIBUS	SUSTENTASSE_SOZINHO	PUBLICA	INTERNE
21	ADM	MASCULINO	ATE_20_ANOS	OUT	SOLTEIRO	PAIS	DE_1501_A_2500	PROPRIO	RESNPONSAVEL_PELO_SUSTENTO	PUBLICA	TV
22	ADM	FEMININO	ATE_20_ANOS	EEP	SOLTEIRO	OUTRO	DE_2501_A_3500	PROPRIO	SUSTENTASSE_SOZINHO	PUBLICA	INTERNE
23	ADM	FEMININO	ATE_20_ANOS	FPU	SOLTEIRO	PAIS	ACIMA_DE_4500	ONIBUS	SUSTENTASSE_SOZINHO	PUBLICA	INTERNE
24	ADM	MASCULINO	ATE_20_ANOS	EEP	SOLTEIRO	PAIS	DE_1501_A_2500	ONIBUS	TRABALHO_E_FAMILIA	PUBLICA	INTERNE
25	ADM	MASCULINO	DE_21_A_25_ANOS	EEP	SOLTEIRO	PAIS	DE_1501_A_2500	ONIBUS	TRABALHO_E_CONTRIBUI	PUBLICA	TV
26	ADM	FEMININO	DE_26_A_30_ANOS	EEP	SOLTEIRO	PAIS	DE_2501_A_3500	ONIBUS	RESNPONSAVEL_PELO_SUSTENTO	PUBLICA	INTERNE
27	ADM	FEMININO	DE_31_A_40_ANOS	EEP	SOLTEIRO	FAMILIA	DE_1501_A_2500	PROPRIO	TRABALHO_E_CONTRIBUI	PUB_PART	REVISTA

Figura 23 - Dados para mineração em Excel
Fonte: Da pesquisa (2010)

O próximo passo foi enviar ao gestor da IES um documento pedindo-o que elaborasse questões de seu interesse referente a gestão da instituição(Vide Anexo A). O objetivo destas questões era o de definir o tipo de informação que seria interessante de ser descoberta na base de dados e iniciar o processo de KDD, através da compreensão do domínio da aplicação e do estabelecimento dos objetivos a serem atingidos (CRISP-DM, 2010).

4.1.1 Seleção, limpeza e transformação dos dados

Nesta etapa realizou-se a seleção dos dados conforme o processo de KDD, com objetivo de analisar o conhecimento do gestor da IES em relação as informações de interesse. Entendeu-se que alguns dados seriam desnecessários, como processo de inscrição, filiação, entre outros, que foram eliminados antes do processo de limpeza dos mesmos.

Os dados após a limpeza passaram pelo processo de conversão de formato, a ferramenta WEKA utiliza o formato ARFF. Este procedimento foi realizado convertendo o arquivo gerado na ferramenta MS Excel, para o formato **Comma-separated values** (**CSV**), em português, **Valores separados por Vírgula,** em seguida os dados foram formatados no padrão de uso do arquivo ARFF, contendo o cabeçalho, a descrição dos campos e respectivos tipos de dados e por fim a sequência de registros que compuseram a amostra, conforme apresentado na Figura 24. Esta etapa foi realizada com o uso de um editor de textos, alterando apenas a extensão do arquivo salvo para ARFF.

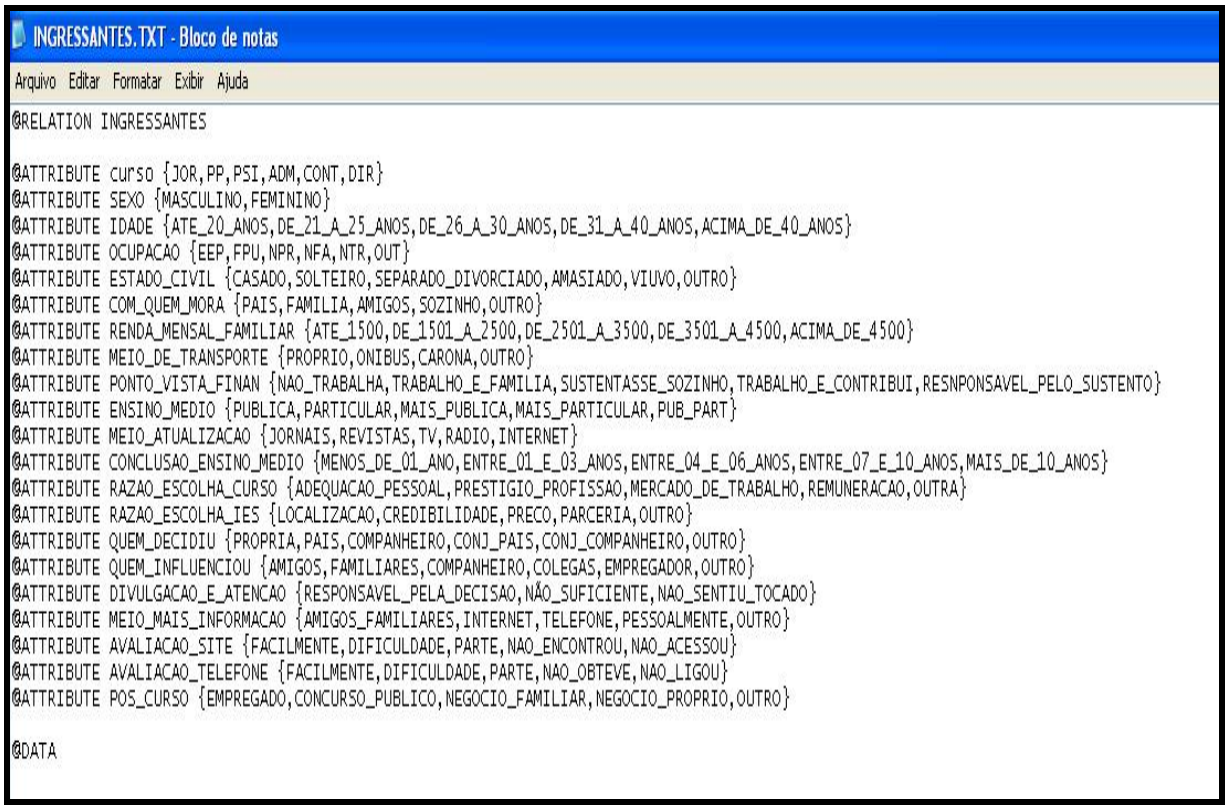


Figura 24 - Exemplo de Cabeçalho no arquivo ARFF
Fonte: Da pesquisa (2010)

Os atributos foram criados com o tipo de dados nominal, uma vez que a ferramenta WEKA não trabalha a regra de classificação com atributos do tipo numérico. Em função desta limitação os atributos numéricos foram enquadrados em faixas de valores nominais.

A escolha da base de dados onde será feita a análise e a escolha da ferramenta de mineração a ser utilizada são consideradas atividades cruciais para o êxito no trabalho. Também deve ser levado em consideração a definição dos objetivos a serem contemplados. Nesta seção foram descritas a estapas de pré-processamento, com intuito de descrever os passos necessários para aplicação das técnicas de MD junto a ferramenta WEKA.

4.1.2 Aplicação das técnicas de Mineração de Dados

Considerada como sendo o elemento principal no processo da descoberta de conhecimento em bases de dados, a etapa de Mineração de Dados, resume-se na aplicação efetiva de uma das técnicas de MD, pela aplicação do algoritmo escolhido sobre os dados a serem analisados com objetivo de descobrir padrões. Tendo a base de dados sido preparada, após a aplicação do algoritmo, ocorre a busca por padrões, associações, classificações e criação de clusters, a fim de identificar novas relações.

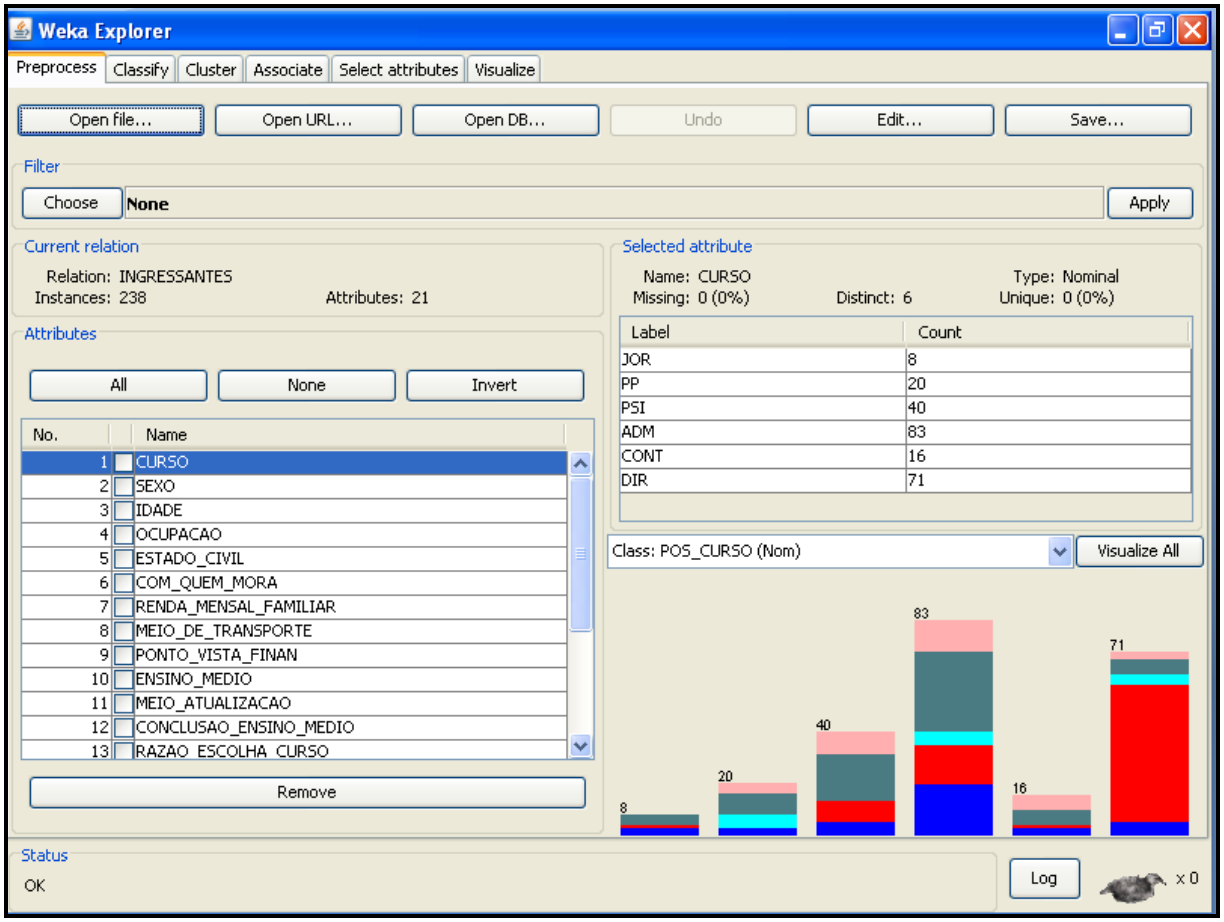


Figura 25- Instanciação dos atributos dos ingressantes para mineração
Fonte: Da pesquisa (2010)

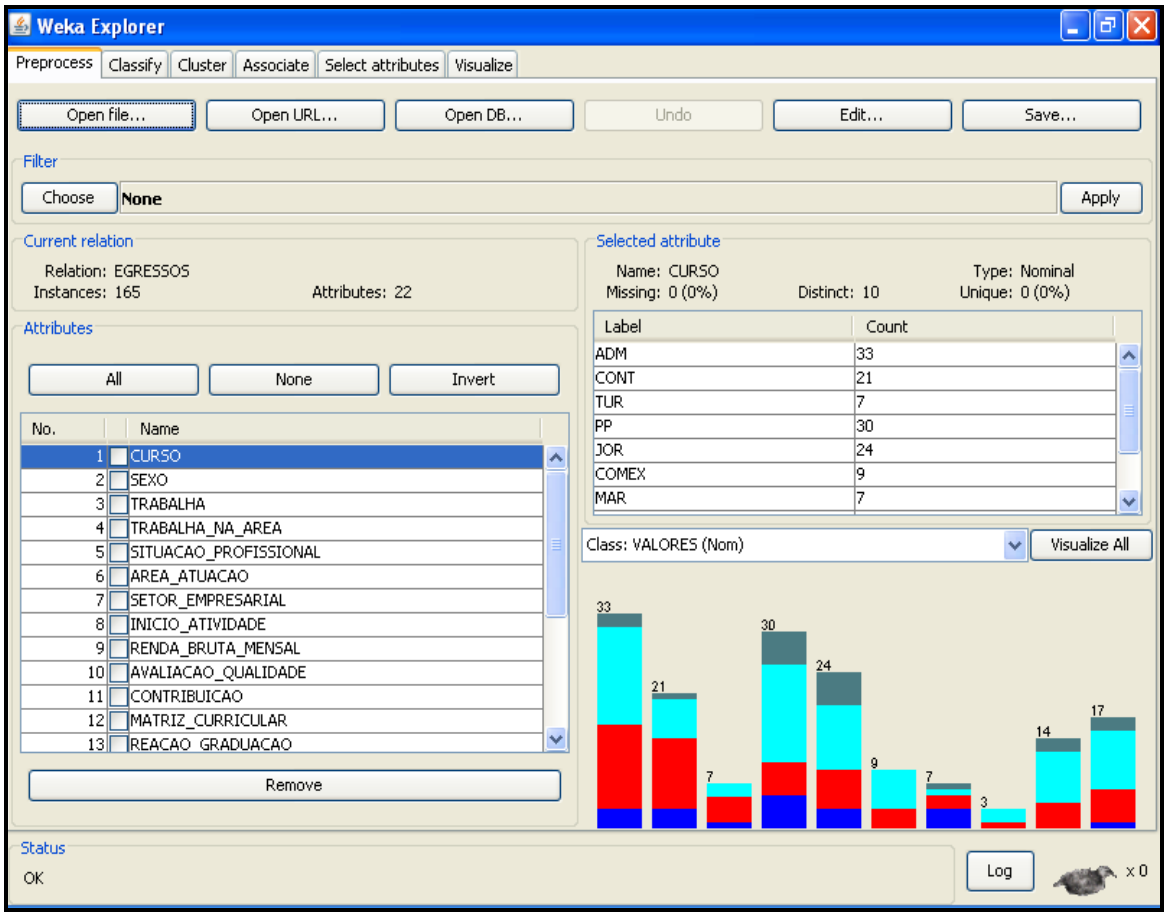


Figura 26 - Instanciação dos atributos dos egressos para mineração
Fonte: Da pesquisa (2010)

Conforme observa-se nos Anexo A e Anexo B, as questões aplicadas a cada grupo de entrevistados têm significativas diferenças.

4.1.3 Tipos de aprendizado

Todas as tarefas de MD passam por um treinamento, uma aprendizagem, sendo que nesta fase os dados processados são “apresentados” ao algoritmo de mineração que será utilizado, com o objetivo de aprender, ou seja, de identificar os padrões considerados úteis no processo de descoberta de conhecimento.

Em Mineração de Dados têm-se dois tipos de aprendizados indutivos chamados de Aprendizagem Supervisionada e Aprendizagem Não-Superivisionada. A Aprendizagem Supervisionada é direcionada a tomada de decisões e é por meio dela que se realizam inferências nos dados com objetivo de realizar predições, nas quais há o uso de atributos para previsão do valor futuro. Enquanto na Aprendizagem Não-Supervisionada as atividades são descritivas, permintindo a descoberta de padões e a geração de novos conhecimentos.

4.1.4 Aprendizagem Não Supervisionada (ANS)

Nestas tarefas o rótulo da classe a ser utilizadas para trabalhar não é conhecido bem como o número de classes que serão treinadas. O objetivo destas tarefas é identificar padrões de comportamento semelhantes nos dados armazenados. As tarefas abordadas nesta pesquisa que pertencem a esta técnica são as tarefas de Associação e Clusterização.

4.1.4.1 Associação

Na tarefa de associação, o objetivo é a descoberta de regras de associação, que são expressões $X \rightarrow Y$ (onde se lê: SE (X) ENTÃO (Y)), sendo que X e Y são conjuntos de itens, $X \cap Y = \emptyset$. Esta regra tem com significado que os conjuntos de itens X e Y ocorrem frequentemente juntos numa mesma transação (registro). (Agrawal et al 1993).

Um exemplo de uma regra do tipo $X \rightarrow Y$ poderia ser: 95% dos candidatos ingressantes que já trabalham também possuem meio próprio de transporte. O valor 95% é dito a confiança da regra, ou seja, representa o número de candidatos ingressantes que trabalham e também possuem meio próprio de transporte, dividido pelo número de candidatos ingressantes que já trabalham.

Para avaliar uma regra de associação existe outra medida que é o valor do suporte da regra, que representa a frequência de ocorrência dos itens X e Y em relação à base de dados (AGRAWAL et al.,1993)

$$FSuporte = \frac{|X \cup Y|}{N}$$

onde

X = número de ocorrências da primeira coluna

Y = número de ocorrências da segunda coluna

N = total de registros

Equação 1 - Fórmula do cálculo do suporte
Fonte:Tsunoda (2008)

A Figura 29 apresenta um exemplo de resultados gerados pelo WEKA quando da aplicação da tarefa de regras de associação é realizada, destacando-se os principais elementos:

a) As regras são ordenadas pela confiança;

b) Os valores depois de antecedentes e consequentes das regras representam o número de suas respectivas ocorrências.

```
Apriori
=======

Minimum support: 0.35 (83 instances)
Minimum metric <confidence>: 0.9
Number of cycles performed: 13

Generated sets of large itemsets:

Size of set of large itemsets L(1): 20

Size of set of large itemsets L(2): 34

Size of set of large itemsets L(3): 12

Size of set of large itemsets L(4): 2

Best rules found:

 1. COM_QUEM_MORA=PAIS ENSINO_MEDIO=PUBLICA 86 ==> ESTADO_CIVIL=SOLTEIRO 86    conf:(1)
 2. COM_QUEM_MORA=PAIS 110 ==> ESTADO_CIVIL=SOLTEIRO 106    conf:(0.96)
 3. OCUPACAO=EEP MEIO_DE_TRANSPORTE=ONIBUS 97 ==> QUEM_DECIDIU=PROPRIA 93    conf:(0.96)
 4. OCUPACAO=EEP MEIO_DE_TRANSPORTE=ONIBUS ENSINO_MEDIO=PUBLICA 92 ==> QUEM_DECIDIU=PROPRIA 88    conf:(0.96)
 5. RENDA_MENSAL_FAMILIAR=DE_1501_A_2500 ENSINO_MEDIO=PUBLICA QUEM_DECIDIU=PROPRIA 89 ==> OCUPACAO=EEP 85    conf:(0.96)
 6. COM_QUEM_MORA=PAIS QUEM_DECIDIU=PROPRIA 88 ==> ESTADO_CIVIL=SOLTEIRO 84    conf:(0.95)
 7. OCUPACAO=EEP MEIO_DE_TRANSPORTE=ONIBUS 97 ==> ENSINO_MEDIO=PUBLICA 92    conf:(0.95)
 8. OCUPACAO=EEP MEIO_DE_TRANSPORTE=ONIBUS QUEM_DECIDIU=PROPRIA 93 ==> ENSINO_MEDIO=PUBLICA 88    conf:(0.95)
 9. OCUPACAO=EEP ENSINO_MEDIO=PUBLICA 119 ==> QUEM_DECIDIU=PROPRIA 111    conf:(0.93)
10. OCUPACAO=EEP ESTADO_CIVIL=SOLTEIRO 98 ==> QUEM_DECIDIU=PROPRIA 91    conf:(0.93)
```

Figura 27 - Regras criadas para ingressantes
Fonte: Da pesquisa (2010)

A primeira regra mostra que 86 (36% do total da amostra) que moram com os pais e fizeram o ensino médio em escola pública ainda são solteiros com uma confiança de 100%. Outra regra a ser considerada é a regra 04 (Se Ocupacao=EEP e Meio_de_Transporte=Onibus e Ensino_Medio=Publica, com 92 registros, implica que Quem_Decidiu=Propria com 86 ocorrências, com um grau de confiança de 96%).

Observa-se na regra de número 7 que 97 dos ingressantes (representando 41% da amostra), quem tem como ocupação se Empregado de Empresa Privada (EEP) e tem como meio de transporte o ônibus, estudou em escola pública, com um grau de confiança de 96 %. Estas regras foram geradas tendo como base todos os atributos utilizados na pesquisa.

A fim de verificar a consistência da regra de associação, outro experimento foi realizado, desta vez reduzindo o numero de atributos pesquisados, sendo eles apenas: Conclusao_Ensino_Medio, Razao_Escolha_Curso, Razao_Escolha_IES e Pos_Curso. Estipulando um percentual mínimo de suporte de 40%, applicou-se a fórmula para os seguintes atributos: Conclusao_Ensino_Medio, Escolha_Curso e Pos_Curso.

Tabela 8 - Cálculo do Suporte Conclusao_Ensino_Medio

CONCLUSAO_ENSINO_MEDIO	NR. OCORRENCIAS	%SUPORTE
ENTRE_01_E_03_ANOS	71	30%
ENTRE_04_E_06_ANOS	44	18%
ENTRE_07_E_10_ANOS	51	21%
MAIS_DE_10_ANOS	39	16%
MENOS_DE_01_ANO	33	14%
TOTAL DA AMOSTRA	238	

Fonte: Da pesquisa (2010)

Como o resultado apresentado com maior percentual de suporte foi onde se encontra a faixa de tempo relativa ao período de ingresso no curso após a conclusão do ensino médio, que apresentou 29,83% dos casos, aplicou-se então uma segunda regra de suporte desta vez analisando o atributo Razao_Escolha_Curso.

Tabela 9 - Cálculo do Suporte Razao_Escolha_Curso

RAZAO_ESCOLHA_CURSO	NR. OCORRENCIAS	SUPORTE
ADEQUACAO_PESSOAL	39	55%
MERCADO_DE_TRABALHO	17	24%
OUTRA	4	6%
PRESTIGIO_PROFISSAO	6	8%
REMUNERACAO	5	7%
TOTAL DA AMOSTRA	71	

Fonte: Da pesquisa (2010)

Para o atributo Razao_Escolha_Curso, a faixa de opções que obteve maior expressividade foi a de Adequação Pessoal, isto indica que aqueles que logo iniciam um curso de graduação procuram adequar-se ao gosto pessoal, “trabalhar no que gosta”. O que levou a mais uma interação da fórmula, agora aplicada ao atributo Pos_Curso, cujos resultados são apresentados na Tabela 10:

Tabela 10 - Cálculo do Suporte Pos_Curso

POS_CURSO	NR. OCORRÊNCIAS	%SUPORTE
CONCURSO_PUBLICO	5	13%
EMPREGADO	13	33%
NEGOCIO_FAMILIAR	5	13%
NEGOCIO_PROPRIO	16	41%
TOTAL	39	

Fonte: Da pesquisa (2010)

A fim de validar o resultado obtido, tornou-se necessário a utilização de mais um atributo Razao_Escolha_IES, para o qual o resultado está demonstrado na Tabela 11:

Tabela 11 - Cálculo do Suporte Razao_Escolha_IES

RAZAO ESCOLHA IES	NR. OCORRÊNCIAS	SUPORTE
CREDIBILIDADE	3	8%
LOCALIZACAO	36	92%
TOTAL	39	

Fonte: Da pesquisa (2010)

De posse destes resultados foi então calculado o grau de confiança para as instâncias da união dos três atributos com maior expressividade, aplicando-se a fórmula descrita na Equação 2, a confiança é uma medida de força da regra. (AGRAWAL ET ALL., 1993):

$$FConfiança = \frac{|X \cup Y|}{X}$$

onde

X = número de ocorrências da primeira coluna

Y = número de ocorrências da segunda coluna

Equação 2 - Cálculo da Confiança
Fonte: Tsunoda (2008)

O resultado final para a confiança fica em 92% da ocorrência da união entre estes quatro atributos anteriormente descritos, uma vez que estes atributos tiveram os resultados acima do valor estipulado para o suporte que era de 40%. O que se observa com a aplicação desta regra pode ser melhor visualizado pelo resultado gerado pela ferramenta WEKA.

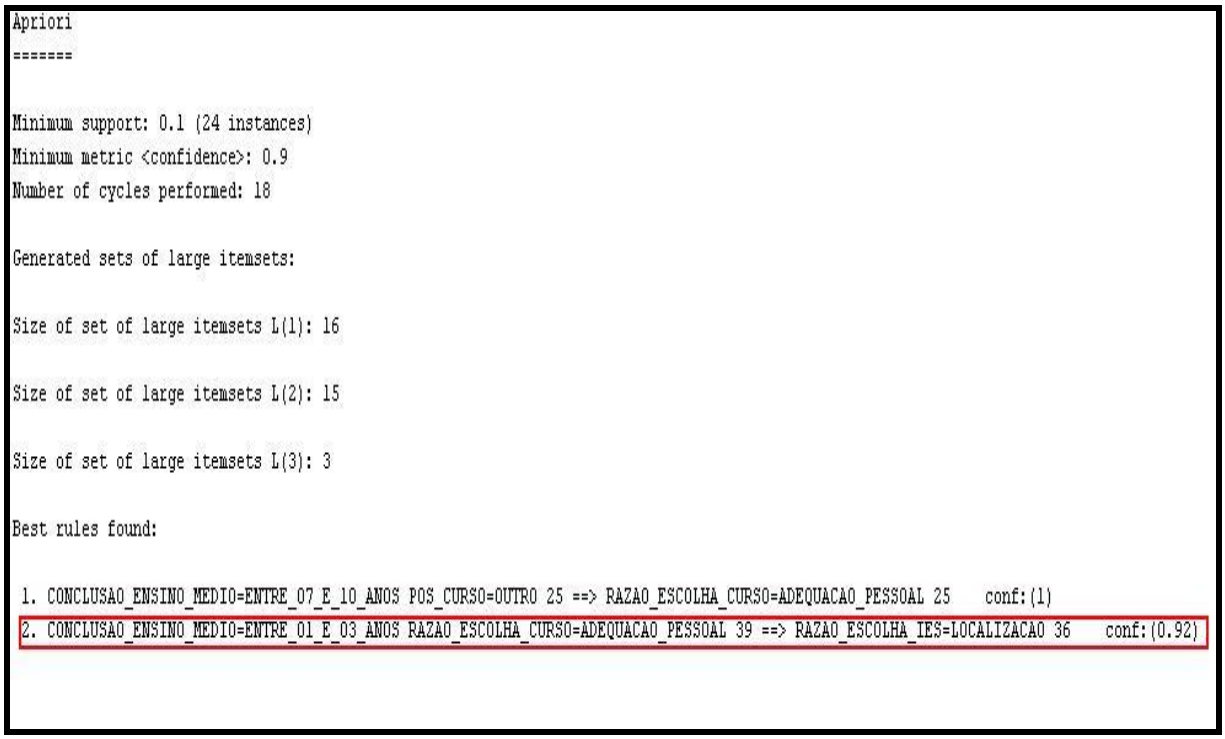


Figura 28 – Resultado da Associação feita no WEKA
Fonte: Da pesquisa (2010)

Para aplicação da regra de associação foram utilizadas as 238 instancias, nas quais observou-se que os ingressantes que concluiram o ensino médio num período entre 01 e 03 anos e escolheram o curso por motivo de adequação pessoal, com 39 instancias, são alunos que escolheram a IES por sua localização, em 36 ocorrências, haja vista a mesma localizar-se no centro da cidade. Esta regra demostra que um ponto a favor da IES está em situar no centro da cidade, tendo seu acesso facilitado em função da próximidade de pontos de ônibus.

Esta relação tem um grau de confiança de 92%. O que ficou mais evidente quando se analisou os atributos Ensino_Medio, Meio_Transporte, Razao_Escolha_IES.

As regras observadas trazem como tendência a formação dos ingressantes, que realizaram o ensino médio em escola pública, na escolha da IES, em conjunto com a renda familiar (de R$ 1.501,00 a R$ 2.500,00) sendo que estes formam a grande parte do público ingressante na IES. Os resultados também indicaram que um dos principais fatores para a escolha da IES é sua localização, que fica no centro da cidade. Outro resultado da análise demonstra que os ingressantes deste conjunto são em maioria EEP.

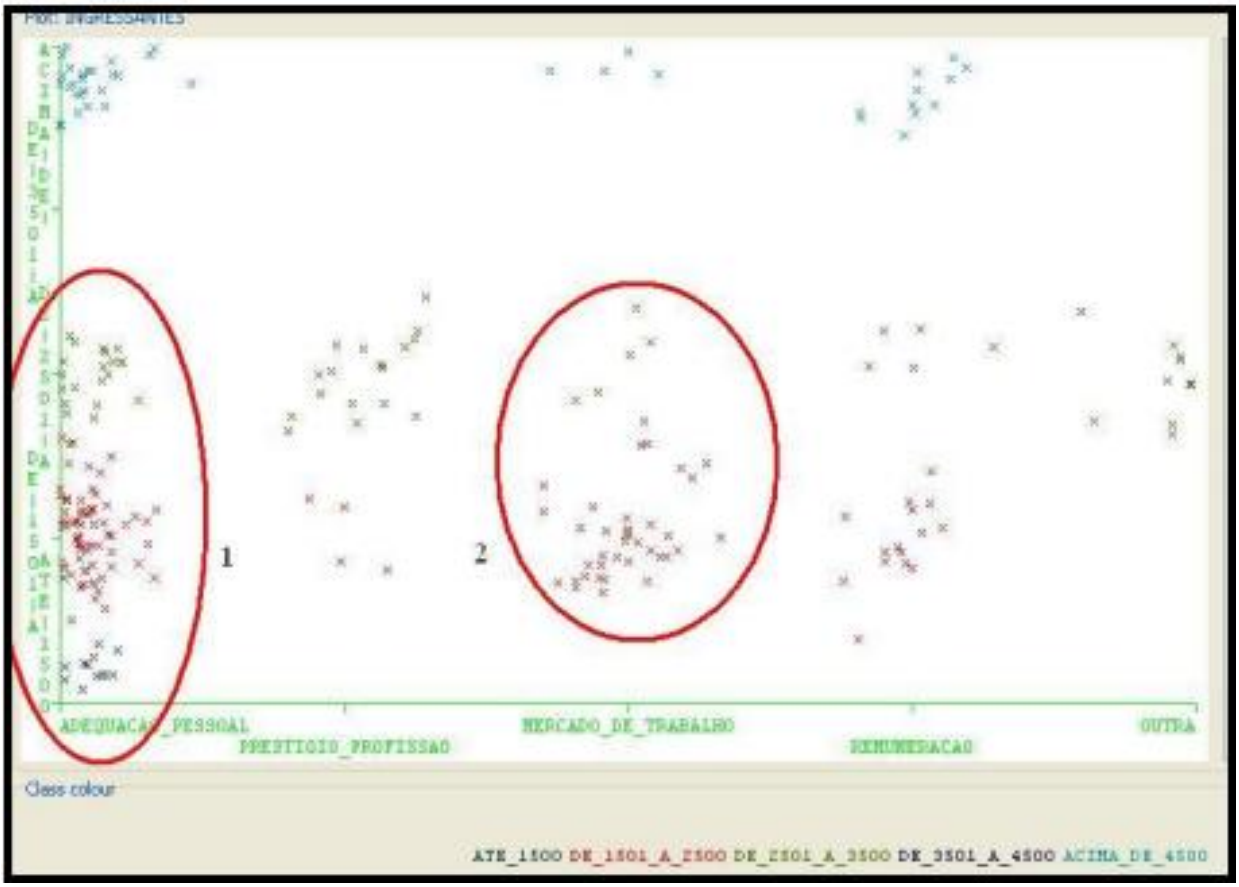


Gráfico 4 - Renda X Razao da Escolha do Curso
Fonte: Da pesquisa (2010)

Conforme análise efetuada utilizando os atributos Renda_Mensal_Familiar e Razao_Escolha_Curso, verificou que existem dois grandes grupos:

a) Ingressantes com renda bruta familiar de R$ 1.500,00 até R$ 2.500,00, cuja principal razão para escolha do curso foi a adequação pessoal;

b) Ingressantes com a mesma faixa de renda, mas com o foco voltado para o mercadode trabalho.

O que demonstra a preocupação dos ingressantes em estarem buscando através do curso superior uma forma de melhoria em sua condição de vida relativa ao seu sustento. Observa-se também que existe um vazio na faixa de renda entre R$ 3.501,00 até R$ 4.500,00, o que representa que pessoas que se encontram nesta faixa salarial, ou já possuem um curso superior, fato este que poderia ser utilizado como atrativo numa campanha, como a concessão de algum benefício para quem já possue uma graduação e quer fazer outra.

Sob a mesma ótica de análise do ponto de vista econômico dos ingressantes, foram analisadas as possíveis associações entre os atributos: Curso, Ponto_de_Vista_Financeiro e Pos_Curso, conforme expresso no Gráfico 5.

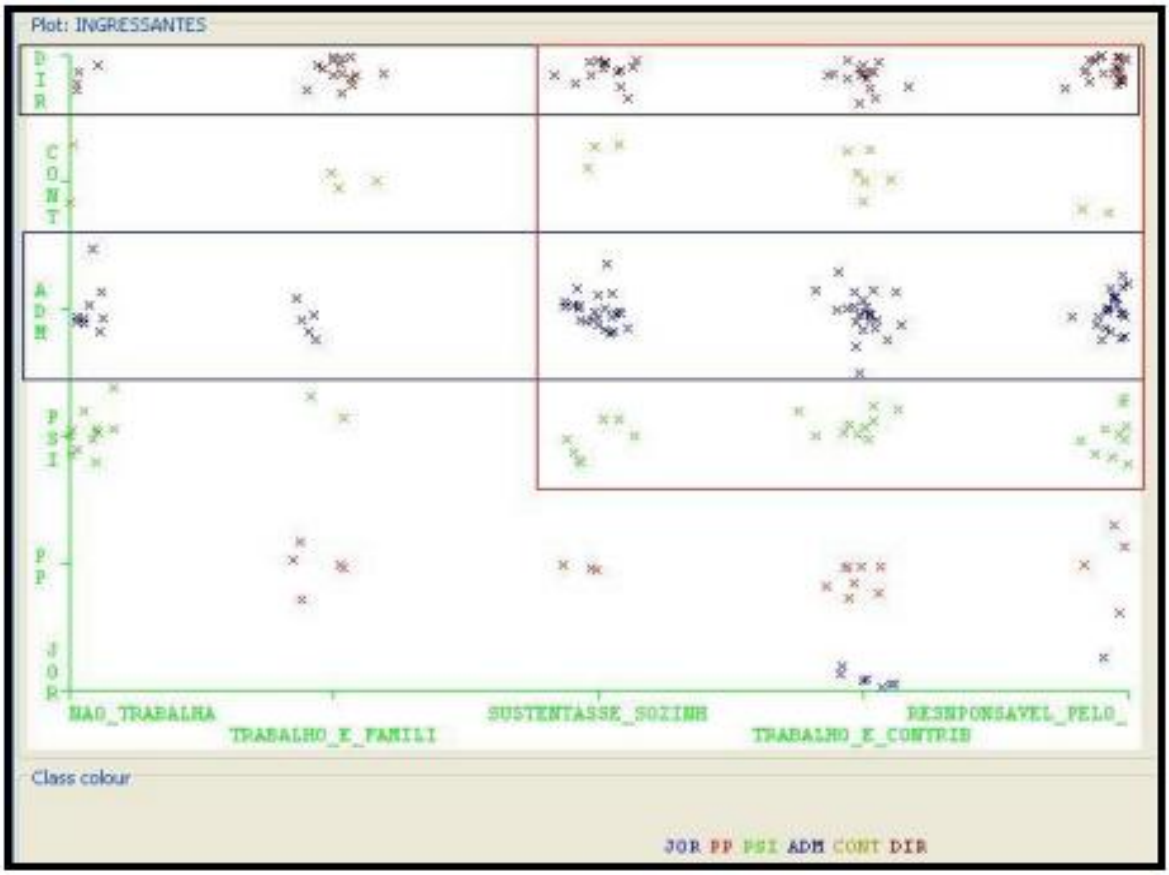


Gráfico 5 - Análise Curso X Ponto de Vista Financeiro e Pos Curso
Fonte: Da pesquisa (2010)

O que se observa no resultado desta análise é que a grande maioria dos ingressantes contribui de alguma forma monetária na renda familiar, sendo que destes, os alunos do curso de Administração, 54% são responsáveis pelo sustento da família e têm como meta Pós_Curso o ingresso numa carreira estável, por meio da realização de um concurso público, como será melhor explicitado na tarefa de Clusterização.

Para os egressos, foram analisadas num primeiro experimento 165 instancias da base de dados, com grau de confiança de 90% (automática gerada pela ferramenta), o que resultou na geração de 10 regras, conforme Figura 29.

```
Apriori
=======

Minimum support: 0.7 (115 instances)
Minimum metric <confidence>: 0.9
Number of cycles performed: 6

Generated sets of large itemsets:

Size of set of large itemsets L(1): 7

Size of set of large itemsets L(2): 11

Size of set of large itemsets L(3): 3

Best rules found:

 1. SITUACAO_PROFISSIONAL=EMPREGADO 131 ==> TRABALHA=SIM 131    conf:(1)
 2. SITUACAO_PROFISSIONAL=EMPREGADO ESTA_ESTUDANDO=NAO 116 ==> TRABALHA=SIM 116    conf:(1)
 3. SITUACAO_PROFISSIONAL=EMPREGADO POLITICA_EX_ALUNO=NAO 115 ==> TRABALHA=SIM 115    conf:(1)
 4. INDICARIA_IES=NAO 126 ==> POLITICA_EX_ALUNO=NAO 117    conf:(0.93)
 5. E_CONTACTADO=NAO 129 ==> POLITICA_EX_ALUNO=NAO 119    conf:(0.92)
 6. ESTA_ESTUDANDO=NAO 146 ==> TRABALHA=SIM 134    conf:(0.92)
 7. POLITICA_EX_ALUNO=NAO 142 ==> TRABALHA=SIM 130    conf:(0.92)
 8. E_CONTACTADO=NAO 129 ==> TRABALHA=SIM 118    conf:(0.91)
 9. INDICARIA_IES=NAO 126 ==> TRABALHA=SIM 115    conf:(0.91)
10. ESTA_ESTUDANDO=NAO POLITICA_EX_ALUNO=NAO 126 ==> TRABALHA=SIM 115    conf:(0.91)
```

Figura 29 - Regra de associação na base dos egressos
Fonte: Da pesquisa (2010)

A regra mais expressiva foi a de número 4, Se indicaria a IES = Não, com um total de 126 registros, uma vez que a mesma regra demonstrou que 117 egressos Não conhecem a política de Ex-aluno, com 93% de confiança. Assim como a regra 5, Se É contactado = Não, com 129 registros, sendo que destes 119 Não conhecem a política de Ex-aluno.

```
Apriori
=======

Minimum support: 0.1 (17 instances)
Minimum metric <confidence>: 0.8
Number of cycles performed: 18

Generated sets of large itemsets:

Size of set of large itemsets L(1): 15

Size of set of large itemsets L(2): 13

Size of set of large itemsets L(3): 3

Best rules found:

1. RENDA_BRUTA_MENSAL=ATE_2_SM 28 ==> AVALIACAO_QUALIDADE=BOM 23    conf:(0.82)
2. CURSO=ADM RENDA_BRUTA_MENSAL=DE_2_A_5_SM 21 ==> AVALIACAO_QUALIDADE=BOM 17    conf:(0.81)
```

Figura 30 - Análise Egressos: Curso X Renda Bruta, Avaliação Qualidade e Contribuição
Fonte: Da pesquisa (2010)

A qualidade dos cursos de graduação oferecidos pela IES é considerada boa por 23 egressos que possuem renda bruta mensal de 2 SM até 5 SM, o que representa apenas 14% dos registros, podendo ser compravado quando a análise é feita com os alunos do curso de Administração que se enquadram nesta situação, sendo que dos 33 egressos 25 consideram como boa, o que representa 76% da amostra dos egressos do curso de Administração, porém apenas 15% do universo total da amostra.

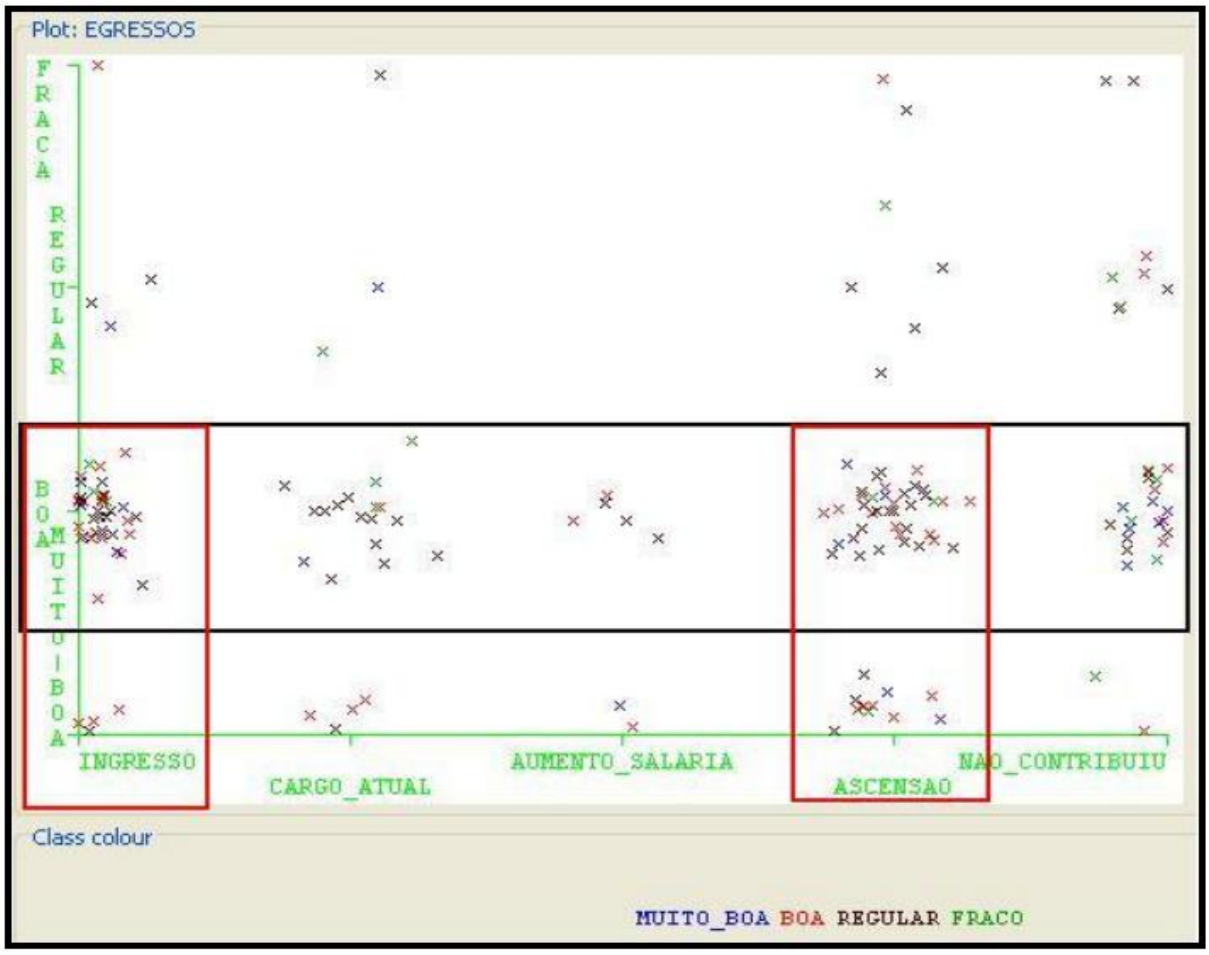


Figura 31 - Análise Contribuição X Qualidade Matriz Curricular
Fonte: Da pesquisa (2010)

A Matriz curricular da IES procura estar em atenção com o mercado, de forma que pode ser verificado com a análise das considerações registradas peloa egressos. O que se observa é que a Qualidade da Matriz Curricular contribui de alguma forma no ingresso da profissão e principlamente na ascensão de cargos por parte dos egressos.

Estas regras refletem a não interação da IES com seus egressos, o que acarreta em mais custos na angariação de alunos para os cursos de extensão e pós graduação que são oferecidos pela IES. Uma política de benefícios para ex-alunos pode ser implantada com

maior clareza se for divulgada nos semestres finais dos cursos de graduação ou nos demais níveis, fato este que serveria de base para a prospecção dos serviços prestados pela IES.

Um forte elemento que pode ser utilizado é a qualidade da Matriz Curricular, que está alinhada com as exigências do mercado no qual a IES está inserida e que tem grande influência na vida profissional dos egressos.

4.1.4.2 Análise de Componentes Principais

A ACP é uma técnica estatística que tem por objetivo a redução do número de variáveis afim de fornecer uma nova visão estatística de um determinado conjunto de dados. Esta técnica fornece ferramentas que possibilitam a identificação de variáveis, consideradas mais importantes, no espaço das componentes principais.

Em função de que apenas os atributos Idade, Renda_Mensal e Conclusão_Ensino_Medio serem numéricos, a ACP foi realizada somente nestes três atributos, assim obtendo os seguintes resultados:

a) Atributo: Idade, resultado: 9,57%;

b) Atributo: Renda_Mensal, resultado: 33,53% e

c) Atributo: Conclusao_Ensino_Medio, resultado: 56,88%.

Assim sendo, o atributo que melhor expressa o conjunto de dados para posterior análise é o atributo Conclusão_Ensino_Medio. Quanto aos egressos a técnica de ACP foi aplicada nos seguintes atributos:

a) Inicio_Atividade_Profissional, resultado: 20,62%;

b) Renda_Bruta_Mensal, resultado: 22,65%;

c) Avaliação_Qualidade, resultado: 30,87%;

d) Matriz_Curricular, resultado: 13,95% e

e) Valores, resultado: 11,88%.

Tendo como atributo mais significativo para posterior análise o atributo Avaliação_Qualidade.

4.1.4.3 Clusterização

A Clusterização é exercida sobre dados nos quais as classes não se encontram definidas. Esta técnica consiste na identificação de novos grupos, que contenham características semelhantes e segmentar os registros com tais características.

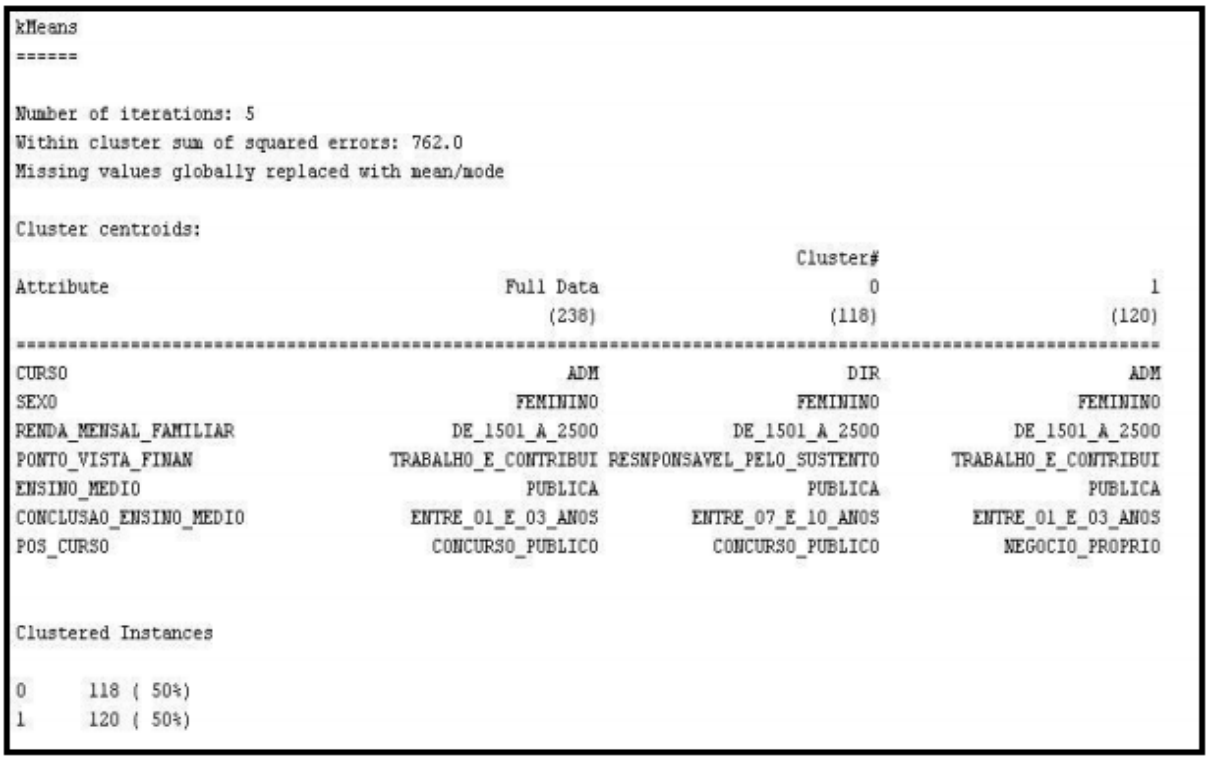
Kampff (2009, p.69) define que:

> A clusterização busca descobrir conhecimento de forma indireta, a partir da identificação de grupos de dados com características semelhantes. O objetivo desta técnica consiste em identificar agrupamentos de dados que podem ser classificados em uma classe comum, descoberta no processo de clusterização.

Em determinadas situações, tornas-se imprescindível que se faça a verificação de como os registros de uma base de dados se agrupam em função de determinadas características intrínsecas de seus atributos. Estes registros podem ser agrupados em clusters com características semelhantes.

Tendo por base uma medida de similaridade, os dados são agrupados, resultando em informações que possibilitam o encontro de relações interessantes entre as instâncias. Assim sendo o usuário pode aplicar uma nova ação em um novo subconjunto de dados, buscando o conhecimento novo sobre os mesmos.

Diferente da tarefa de classificação, onde há classes pré-definidas, a tarefa de clusterização é uma das primeiras técnicas a ser realizada em Mineração de Dados. Nesta pesquisa o algoritmo utilizado para a clusterização foi o algoritmo K Means em conjunto com a medida Euclidiana para medir a similaridade entre os objetos.



```
kMeans
======

Number of iterations: 5
Within cluster sum of squared errors: 762.0
Missing values globally replaced with mean/mode

Cluster centroids:
                                                    Cluster#
Attribute                               Full Data          0                   1
                                          (238)        (118)               (120)
=============================================================================================
CURSO                                        ADM          DIR                 ADM
SEXO                                    FEMININO     FEMININO            FEMININO
RENDA_MENSAL_FAMILIAR                DE_1501_A_2500 DE_1501_A_2500      DE_1501_A_2500
PONTO_VISTA_FINAN       TRABALHO_E_CONTRIBUI RESNPONSAVEL_PELO_SUSTENTO TRABALHO_E_CONTRIBUI
ENSINO_MEDIO                          PUBLICA      PUBLICA             PUBLICA
CONCLUSAO_ENSINO_MEDIO        ENTRE_01_E_03_ANOS ENTRE_07_E_10_ANOS  ENTRE_01_E_03_ANOS
POS_CURSO                     CONCURSO_PUBLICO   CONCURSO_PUBLICO    NEGOCIO_PROPRIO


Clustered Instances

0      118 ( 50%)
1      120 ( 50%)
```

Figura 32 - Tarefa de clusterização – Ingressantes
Fonte: Da pesquisa (2010)

Num primeiro experimento foram selecionados os atributos: Curso, Sexo, Renda_Mensal_Familiar, Ponto_Vista_Finan, Ensino_Medio, Conclusao_Ensino_Medio, Pos_Curso, com os 238 registro contios na base de dados, o que acarretou na geração de dois clusters. O resultado foi equivalente para ambos os clusters, uma vez que a quantidade de registros próximos ao Cluster 0 foi de 118 instancias, enquanto para o Cluster 1 foi de 120 instâncias.

No Cluster 0 (zero), as características para a similaridade foram:

Tabela 12 - Cluster 0 sobre os ingressantes

Atributo	Característica
Curso	Direito
Sexo	Feminino
Renda_Mensal_Familiar	De 1500 a 2500
Ponto_Vista_Finan	Responsavel_Pelo_Sustento
Ensino_Medio	Publica
Conclusao_Ensino_Medio	Entre 07 e 10 Anos
Pos_Curso	Concurso_Publico

Fonte: Da pesquisa (2010)

No Cluster 1 (um), as características para a similaridade foram:

Tabela 13 - Cluster 1 sobre os ingressantes

Atributo	Característica
Curso	Administração
Sexo	Feminino
Renda_Mensal_Familiar	De 1500 a 2500
Ponto_Vista_Finan	Trabalha e Contribui
Ensino_Medio	Trabalha e Contribui
Conclusao_Ensino_Medio	Entre 01 e 03 Anos
Pos_Curso	Negocio_Proprio

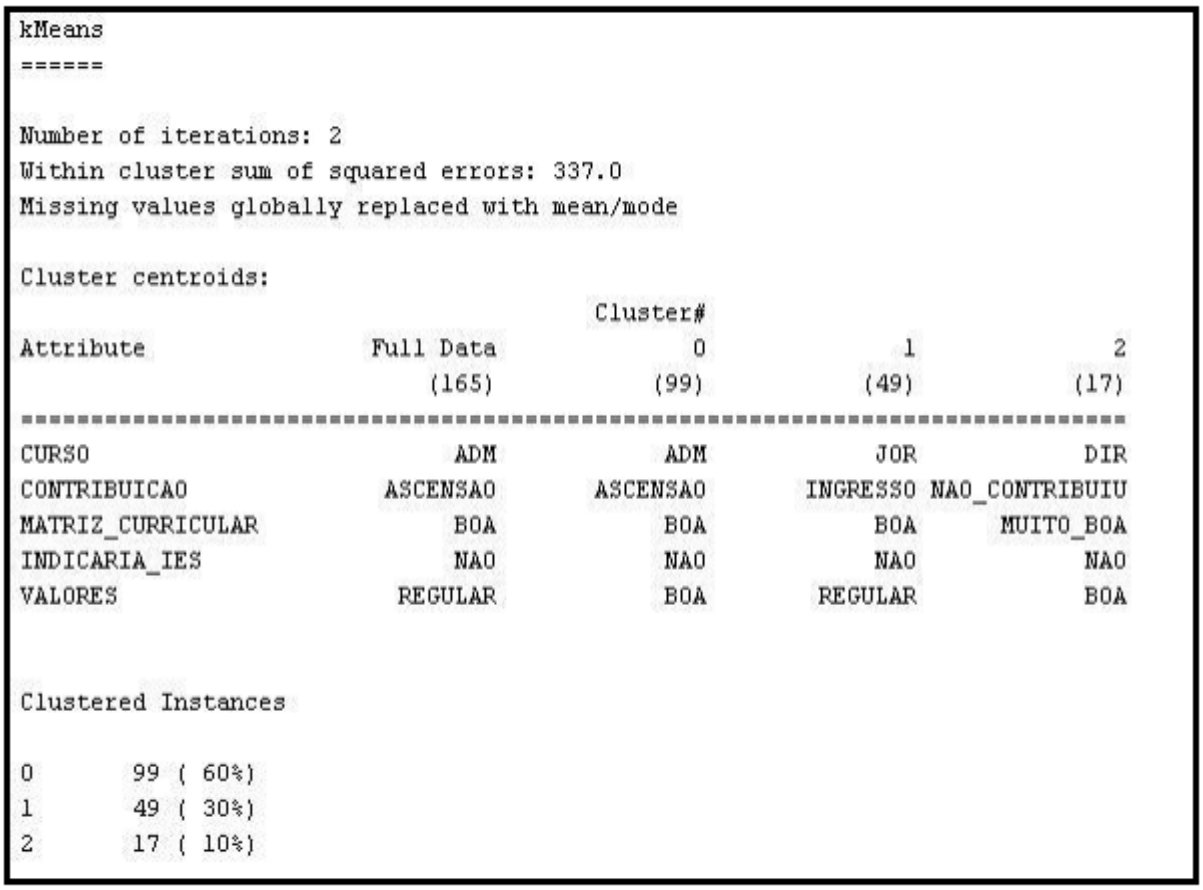
Fonte: Da pesquisa (2010)

Ao aumentar o número de clusters, a divisão foi mais significativa no grupo que estava próximo ao cluster 1, uma vez que a divisão ficou sendo da seguinte maneira:

```
Number of iterations: 7
Within cluster sum of squared errors: 702.0
Missing values globally replaced with mean/mode

Cluster centroids:
                                                Cluster#
Attribute                          Full Data           0           1           2
                                     (238)       (121)        (76)        (41)
=============================================================================================================
CURSO                                 ADM         DIR         ADM         PSI
SEXO                                FEMININO    FEMININO    MASCULINO   FEMININO
RENDA_MENSAL_FAMILIAR          DE_1501_A_2500  DE_1501_A_2500  DE_1501_A_2500  DE_1501_A_2500
PONTO_VISTA_FINAN      TRABALHO_E_CONTRIBUI RESNPONSAVEL_PELO_SUSTENTO  TRABALHO_E_CONTRIBUI  TRABALHO_E_CONTRIBUI
ENSINO_MEDIO                       PUBLICA     PUBLICA     PUBLICA     PUB_PART
CONCLUSAO_ENSINO_MEDIO      ENTRE_01_E_03_ANOS  ENTRE_07_E_10_ANOS  ENTRE_01_E_03_ANOS  MAIS_DE_10_ANOS
POS_CURSO                    CONCURSO_PUBLICO  CONCURSO_PUBLICO  EMPREGADO  NEGOCIO_PROPRIO


Clustered Instances

0      121 ( 51%)
1       76 ( 32%)
2       41 ( 17%)
```

Figura 33 - Criação do terceiro cluster Ingressantes
Fonte: Da pesquisa (2010)

O que vem a reforçar os resultados apresentados na tarefa de Associação e posteriormente serão comprovados com a tarefa de classificação, onde se percebe que os ingressantes no curso de Direito tendem ao finalizar o curso em questão prestar algum concurso público, enquanto os ingressantes no curso de Administração preferem a abertura de um negócio próprio.



```
kMeans
======

Number of iterations: 2
Within cluster sum of squared errors: 337.0
Missing values globally replaced with mean/mode

Cluster centroids:
                                        Cluster#
Attribute               Full Data             0             1             2
                            (165)          (99)          (49)          (17)
=============================================================================
CURSO                        ADM           ADM           JOR           DIR
CONTRIBUICAO           ASCENSAO      ASCENSAO      INGRESSO NAO_CONTRIBUIU
MATRIZ_CURRICULAR           BOA           BOA           BOA      MUITO_BOA
INDICARIA_IES               NAO           NAO           NAO           NAO
VALORES                 REGULAR           BOA       REGULAR           BOA


Clustered Instances

0       99 ( 60%)
1       49 ( 30%)
2       17 ( 10%)
```

Figura 34 - Cluster gerado para os egressos
Fonte: Da pesquisa (2010)

Quanto aos egressos foram criados inicialmente 03 clusters, o que mostrou uma característica muito boa, pois 60% dos registros foram agrupados no cluster 0, que retrata a contribuição do curso na questão de ascensão na vida profissional mas apesar disto não indicaria a realização de um curso na IES para outras pessoas.

4.1.5 Aprendizagem Supervisionada

A Aprendizagem Supervisionada (AS), trabalha com algoritmos preditivos, haja vista que suas tarefas de mineração fazem inferências nos dados com objetivo de fornecer previsões ou tendências, tendo como base informações não disponíveis dos dados a serem minerados.

A AS faz uso de uma classe especificada, isto é, determinada instância contém um atributo classe que determina à qual classe ela está inserida. Diversas técnicas de mineração utilizam este tipo de aprendizado, dentre elas a classificação, que foi uma das técnicas utilizadas nesta pesquisa.

4.1.5.1 Classificação

A tarefa de classificação tem por objetivo encontrar características comuns entre um conjunto de objetos de uma base de dados e classificá-los em classes diferentes. Para chegar a estas classes é necessário seguir alguns passos: 1) definir um conjunto de exemplos (previamente conhecido) para treinamento; 2) aplicar o treinamento sobre este conjunto conhecido e por fim gerar as regras de classificação.

Para Martinhago (2005, p. 20), "Nessa tarefa cada tupla (registro), pertence a uma classe entre um conjunto pré-definido de classes". Pode-se por exemplo classificar os ingressantes em relação o que pretende fazer quando concluírem o curso, atributo Pos_Curso, em: ser empregado de empresa privada, participar de concurso público, gerenciar negócio familiar, gerenciar negócio próprio ou outras atividades.

```
ZeroR predicts class value: CONCURSO_PUBLICO

Time taken to build model: 0 seconds

=== Stratified cross-validation ===
=== Summary ===

Correctly Classified Instances          93               39.0756 %
Incorrectly Classified Instances       145               60.9244 %
Kappa statistic                          0
Mean absolute error                      0.2953
Root mean squared error                  0.384
Relative absolute error                100      %
Root relative squared error            100      %
Total Number of Instances              238

=== Detailed Accuracy By Class ===

               TP Rate   FP Rate   Precision   Recall  F-Measure   ROC Area  Class
                 0         0          0         0         0          0.474    EMPREGADO
                 1         1          0.391     1         0.562      0.48     CONCURSO_PUBLICO
                 0         0          0         0         0          0.412    NEGOCIO_FAMILIAR
                 0         0          0         0         0          0.496    NEGOCIO_PROPRIO
                 0         0          0         0         0          0.452    OUTRO
Weighted Avg.    0.391     0.391      0.153     0.391     0.22       0.475

=== Confusion Matrix ===

  a  b  c  d  e   <-- classified as
  0 53  0  0  0 |  a = EMPREGADO
  0 93  0  0  0 |  b = CONCURSO_PUBLICO
  0 15  0  0  0 |  c = NEGOCIO_FAMILIAR
  0 50  0  0  0 |  d = NEGOCIO_PROPRIO
  0 27  0  0  0 |  e = OUTRO
```

Figura 35 - Tarefa de classificação – Ingressantes
Fonte: Da pesquisa (2010)

Uma forma de validar o desempenho da tarefa de classificação é calcular por meio de uma medida de precisão os resultados do classificador, o que acarreta na atribuição de um nível de confiança ao exemplo classificado.

Uma forma bastante utilizada para validação da classificação é feita com o uso da chamada “matriz de confusão”, que é uma matriz quadrada de dimensões N x N, onde N é o número de classes que se encontram sob investigação. As linhas desta matriz representam as classes desejadas enquanto as colunas são as associações definidas pelo algoritmo classificador. Mori (2008, p. 85) define que “Os elementos da matriz diagonal representam o número de exemplos corretamente classificados (coincidências ou concordâncias). Os elementos acima da diagonal representam os erros de omissão e os abaixo da diagonal os de inclusão”.

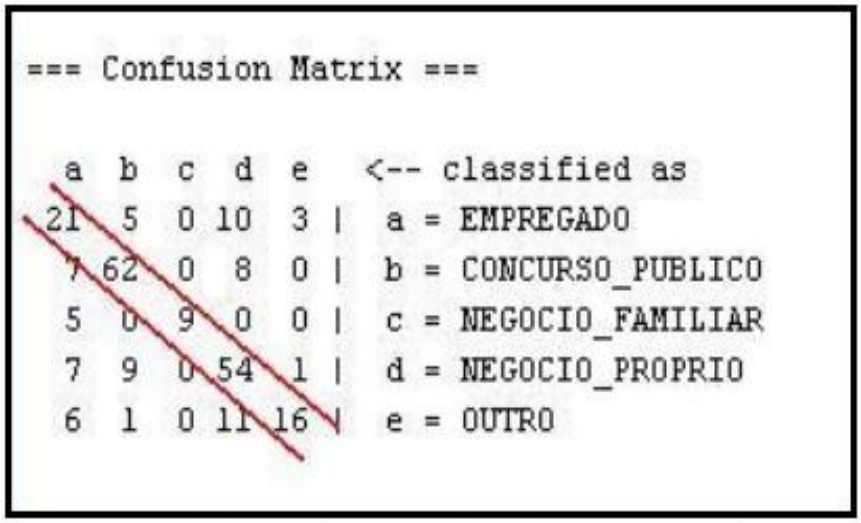


Figura 36 - Matriz de confusão gerada pelo WEKA para ingressantes
Fonte: Da pesquisa (2010)

A matriz gerada apresenta a classificação dos ingressantes em relação ao atributo Pos_Curso, que retrata duas situações interessantes: 1) a classificação de 62 registros para a opção de após a conclusão do curso em questão o ingressante pretende fazer um concurso público, seguida por, 2) a classificação de 54 registros para a abertura do negócio próprio, independente de curso.

Isto tende a estar em acordo com o curso realizado pelos ingressantes, pois dos 71 ingressantes no curso de Direito, 53 optaram pela opção de concurso público, o que representa 75% dos ingressantes neste curso, enquanto no curso de Administração 37% dos ingressantes optaram por abrirem o negócio próprio, sendo 31 num total de 83 ingressantes no curso de Administração.

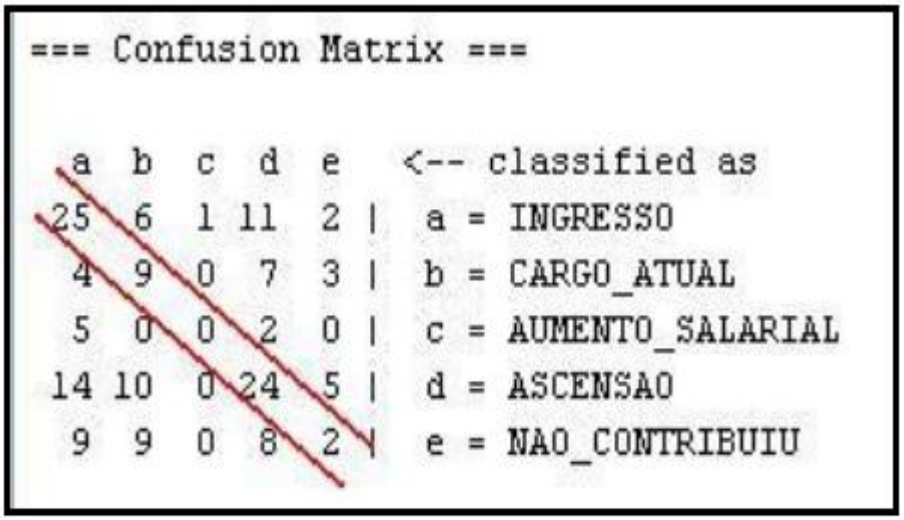


Figura 37 - Matriz de confusão gerada pelo WEKA para os egressos
Fonte: Da pesquisa (2010)

A matriz de confusão gerada para o atributo Contribuição mostra que a matriz curricular aplicada pela IES, tem forte influencia em dois dos quesitos apresentados. A classificação correta para o ingresso na profissão foi de 25 casos e 24 classificações para os egressos que obtiveram ascensão no cargo. Isto vem a comprovar que a matriz curricular oferecida pela IES procura estar adequada ao mercado e oferece a capacidade aos egressos de fazer carreira dentro das organizações.

5 CONCLUSÕES

A aplicação das técnicas de Mineração de Dados com o intuito de descobrir novos conhecimentos auxilia no processo de exploração de uma base de dados, o que permite gerar informações úteis para os gestores, auxiliando-os nas tomadas de decisões.

O objetivo desta pesquisa foi o de mostrar a aplicabilidade das técnicas de Mineração de Dados em um ambiente de gestão educacional de uma IES, apresentando à mesma o perfil de seus ingressantes e egressos, desta forma contribuindo para a gestão e organização de campanhas dirigidas a estes diferentes tipos de perfis de seus futuros e ex-alunos.

A pesquisa apresentou importantes análise sobre o perfil dos ingressantes e egressos da IES, por meioda aplicação das técnicas de Mineração de Dados implementadas na ferramenta WEKA, Associação, Classificação e Clusterização. Os experimentos apresentam uma importante contribuição em termos de quais aspectos são característicos para os ingressantes e também para os egressos.

Uma das principais características que pode ser observada é que a maioria dos ingressantes são oriundos de escolas públicas, escolheram a IES por sua localização, tendo concluido o ensino médio num período relativamente curto de no máximo três anos e pretendem aplicar o seu lado empreendedor, ou seja, abrir seu negócio próprio.

Quanto aos egressos observou-se que a principal característica é que, a matriz curricular aplicada pela IES, influenciou em muito na escolha da mesma e do curso e ainda que a matriz curricular está alinhada com os interesses do mercado, haja vista que muitos egressos conseguiram se promover dentro de suas organizações.

Outra característica importante que pode ser extraída e que deve preocupar os gestores é que a grande maioria dos egressos nào recomendaria a realização de um curso na IES. Cabe aos gestores uma investigação mais aprofundada da situação afim de sanar este descontentamento.

A utilização das técnicas de Mineração de Dados mostrou-se útil para o descobrimento do conhecimento que se encontrava escondido na base de dados do ambiente de gestão da IES. A consistência e eficácia das tarefas de associação, classificação e clusterização

geradas pela ferramenta WEKA, foram analisadas e comprovadas pelo gestor da instituição, o qual irá incorporar este novo conhecimento na tomada de suas decisões.

O trabalho apresentado teve por objetivo contribuir para a análise do perfil dos ingressantes e dos egressos da IES. Acredita-se com este trabalho possa ser utilizado como complemento das técnicas de gestão utilizadaspelos gestores da IES para a melhoria nos procesos de ingresso e também como no trabalho que possa a ser desenvolvido com os egressos.

5.1 CONTRIBUIÇÕES

Esta dissertação teve também como objetivo contribuir para a área de Inteligência Aplicada através do uso das técnicas e ferramentas de Mineração de Dados em conjunto com a metodologia CRISP-DM e a aplicação das tarefas de Associação, Classificação e de Clusterização com o intuito de auxiliar na tomada de decisões. Desta forma, as principais contribuições são:

O uso da metodologia CRISP-DM possibilita a resolução de problemas de extração de informações de uma forma organizada e progressiva, tendo como início uma análise de alto nível, a qual busca a compreensão das regras do negócio, direcionando-se para a definição e implantação de modelos que permitem a obtenção efetiva dos objetivos da mineração.

A utlização da metodologia no ambiente proposto, permitiu a viabilidade e a utilidade prática da metodologia em um estudo de caso real, sendo que os resultados poderão auxiliar os gestores à elucidar características relevantes em relação a diversas situações observadas neste estudo. As conclusões perimitiram mostrar a relevência da metodologia CRISP-DM na obtenção dos resultados da mineração de dados.

O resultado da utilização desta metodologia tende a proporcionar uma melhor interpretação das atividade inerentes ao uso das técnicas de mineração de dados pelos gestores da IES, haja vista que os mesmos não estão familiarizados com tais técnicas e terão mais um recurso a sua disposição para auxiliar nas tomadas de decisões.

A análise de dados feita pelo uso de técnicas de mineração de dados é ainda um pouco difundido em IES, apesar de ser ensino em várias delas, assim sendo este estudo e as sugestões para trabalhos futuros visam contibuir para que o uso das técnicas e metodologas de

mineração de dados seja utilizados como um diferencial competitivo também no setor educacional.

Nesta pesquisa foi demonstrada a relevância do processo de mineração de dados a obtenção de informações no que se refere a análise das informações constantes nos questionários sócio-econômicos aplicados aos ingressante e egressos da IES. Assim, teve-se o objetivo de analisar os motivos que levam aos acadêmicos ingressarem na IES e as considerações a respeito da IES por parte dos egressos por meio da aplicação das tarefas de associação, classificação e de clusterização.

Quando bem aplicada, a Mineração de Dados, através das técnicas de Associação, Classificação e Clusterização, traz muitos benefícios as organizações, auxiliando na tarefa das tomadas de decisões, que são utlizadas para a obtenção de vantagens competitivas. No segmento de Ensino Superior, que está cada vez mais acirrado, a utilização das técnicas de Mineração de dados estão se tornando obrigatórias.

Alguns desafios e dificuldades foram encontrados durante o desenvolver desta pesquisa, entre eles:

a) A dificuldade inicial neste trabalho foi a liberação a base de dados da instituição, haja vista tratar-se de dados pessoais dos acadêmicos ali inscritos e matriculados;

b) A dificuldade de definição dos atributos, por parte dos gestores, que compuseram a base de dados para a mineração;

c) A limitação das informações contidas na base de dados, que não contempla informações a respeito da efetivação de matrículas por candidatos aprovados nos processos seletivos, uma vez que após a aplicação das tarefas de poderia se ter regras relacionadas a condição social do candidato, exemplo se o candidato não efetivou a matrícula pelo fato de não estar trabalhando.

5.2 SUGESTÕES PARA TRABALHOS FUTUROS

Após o estudo abordado nesta dissertação, estabelecem-se algumas recomendações para pesquisas de mesmo cunho. Alguns assuntos merecem aprofundamento em pesquisas ou trabalhos futuros. Os principais são:

a) Utilização de outras técnicas de Mineração de Dados não contempladas neste estudo, como por exemplo, Redes Neurais e Algoritmos Genéticos;

b) Implementação de algoritmos de Mineração de Dados junto a ferramenta de gestão acadêmica da IES, oportunizando ao prórpio gestor elaborar sua mineração;

c) Implementação de um ambiente para a armazenagem dos dados, possibilitando a geração dos arquivos no formato apropriado para a Mineração de Dados e a visualização dos resultados da mineração, acomplado ao ambiente de gestão da IES pesquisada;

d) Implementação do algoritmo fuzzy-means para realização clusterização ao invés do algoritmo K-means disponível no software WEKA.

REFERÊNCIAS BIBLIOGRÁFICAS

ABAR, Celina Aparecida Almeida Pereira. O uso de objetos de aprendizagem no ambiente TELEDUC como apoio ao ensino presencial no contexto da matemática. In: **CONGRESSO INTERNACIONAL DE EDUCAÇÃO A DISTÂNCIA**, 11., 2004, Salvador. Anais... Salvador: ABED, 2004. p. 01 – 07. Disponível em: < http://www.abed.org.br/congresso2004/>. Acesso em: 13 out. 2008.

AGRAWAL, R.; IMIELINSKI, T.; SWAMI, A. Mining associations between sets of items in massive databases. **In: ACM-SIGMOD, 1993**. Proceedings… Int'l Conference on Management of Data, Washington D.C., May 1993..

ALMEIDA, F. J.; ALMEIDA, M. E. B. (Coord.) **Liderança, gestão e tecnologias**: para a melhoria da educação no Brasil. São Paulo: [s.n.], 2006.

ALMEIDA, Felipe S. de. **Otimização de Estruturas de Materiais Compósitos Laminados utilizando Algoritmos Genéticos**. 2006. 146 f. Dissertação (Mestrado em Engenharia na modalidade Acadêmico) - Universidade Federal do Rio Grande do Sul – UFRG, 2006.

ALMEIDA, Manoel V. de.. **Aplicação de Técnicas de Redes Neurais Artificiais na Previsão de Curtíssimo Prazo da Visibilidade e Teto para Aeroporto de Guarulhos – SP**. 2009. 182 f. Tese (Doutorado em Ciências em Engenharia Civil) – Universidade Federal do Rio de Janeiro – UFRJ, Rio de Janeiro, 2009.

ALVARENGA, JÚLIO C. S. de. **Parâmetros de gestão da informação do Centro Universitário São Camilo – Espírito Santo com ênfase na inteligência competitiva**. 2006. 104 f. Dissertação (Mestrado em Ciência da Informação) - Pontifícia Universidade Católica de Campinas, Campinas, São Paulo, 2006.

ALVES, Claudia F. M.. **Gestão da tecnologia da informação nas instituições de ensino superior**. 2005. 151 f. Dissertação (Mestrado em Administração Estratégica) - Universidade Salvador – UNIFACS. Salvador, 2005.

AMORIM, M. ; MANSUR, A. F. U. ; BARONE, D. . **Técnicas de Aprendizado de Máquina Aplicadas na Previsão de Evasão Acadêmica.** In: SBIE 2008 - Simpósio Brasileiro de informática na Educação, 2008, Fortaleza - CE. Anais do SBIE 2008. Ceará : Sociedade Brasileira de Computação, 2008.

BARBOSA, Rommel Melgaço (Org.). **Ambientes virtuais de aprendizagem**. Porto Alegre: Artmed, 2005.

BARION, Eliana C. N. e LAGO, Decio. **Mineração de Textos**. Revista de Ciências Exatas e Tecnologia. São Paulo, Vol. III, Nº. 3, p. 123-140. Dez, 2008.

BARIONI, MARIA C. N.. **Visualização de Operações de Junção em Sistemas de Bases de Dados para Mineração de Dados**. 2002. 65 f. Dissertação (Mestrado em Ciências - Computação e Matemática Computacional) – USP, São Carlos, 2002.

BARRETO _____. As estruturas de suporte da informação no processo do conhecimento: o papel da fluência digital. **DataGramaZero Revista de Ciência da Informação**, v. 7, n. 4, ago. 2006.

BARRETO _____. Os agregados de informação: memórias, esquecimento e estoques de informação. **DataGramaZero: Revista de Ciência da Informação**, Rio de Janeiro, v.1, n.3, p.1-14, jun. 2000. Disponível em: <http://datagramazero.org.br>. Acesso em: 04 ago. 2008.

BARRETO, Aldo de Albuquerque. **A condição da informação**. São Paulo em Perspectiva, v. 16, n. 3, p.67-74, 2002.

BATISTA P., SILVA M.J. "**Mining Web Access Logs of an On-line Newspape**r", Departamento de Informática, Faculdade de Ciências – Universidade de Lisboa. Disponível em:<http://xldb.fc.ul.pt/data/Publications_attach/rpec02.pdf>. Acesso em: 01 ago. 2008.

BERNARDES, José Francisco; ABREU, Aline França de. A contribuição dos sistemas de informações na gestão universitária. Florianópolis, 2004. **Anais** do IV Colóquio Internacional sobre Gestão Universitária na América do Sul.

BERNERS-LEE, T.; HENDLER, J.; LASSILA, O., **The Semantic Web**, Scientific American, May 2001.

BEUREN, I. M. **Gerenciamento da informação**. 2. ed. São Paulo: Atlas, 2000.

BISPO, Carlos A. F.. **Uma Análise da Nova Geração de Sistemas de Apoio à Decisão**. 1998. 174 f. Dissertação (Mestrado em Engenharia da Produção) - Universidade de São Paulo – USP, São Carlos, 1998.

BOENTE, A. N. P. ; OLIVEIRA, F. S. G. ; ROSA, J. L. A.. **Utilização de Ferramenta de KDD para Integração de Aprendizagem e Tecnologia em Busca da Gestão Estratégica do Conhecimento na Empresa**. Anais do Simpósio de Excelência em Gestão e Tecnologia, v. 1, p. 123-132, 2007.

BRAGA, R.; MONTEIRO, C. A**. Planejamento estratégico sistêmico para instituições de ensino**. São Paulo: Hoper, 2005.

BRASIL. **Lei nº 9.394,** de 20 de dezembro de 1996. Dispõe sobre as diretrizes e bases da educação nacional. Brasília, p.10. 1996.

_____. **Decreto nº 3.860,** de 9 de julho de 2001. Dispõe sobre a organização do ensino superior, a avaliação de cursos e instituições, e dá outras providências. Brasília, p. 01. 2001.

BRAY, T.; PAOLI, J.; SPERBERG-MCQUEEN, C. M. **Extensible Markup Language (XML) 1.0** — W3C Recommendation 10-February-1998. [S.l.], fev. 1998.

BUCKLAND, M. K. Information as thing. **Journal** of the American Society for Information Science (JASIS), v. 45, n. 5, p. 351-360, 1991.

BUKOWITZ, Wendi R.; WILLIANS, Ruth. **Manual de gestão do conhecimento**: ferramentas e técnicas que criam valor para a empresa. Porto Alegre: Bookman, 2002.

CARDOSO, Olinda N. P., MACHADO, Rosa T. M. **Gestão do conhecimento usando data mining:** estudo de caso na Universidade Federal de Lavras. Revista de Administração Pública. Rio de Janeiro 42(3) : 495-528, Maio/Jun. 2008.

CARVALHO, Luís A. V.de. **Data mining: a mineração de dados no marketing, medicina, economia, engenharia e administração**. São Paulo: Érica, 2001.

CASTRO, Edna M. M. V.. **Tecnologia da Informação: Fatores relevantes para o sucesso da sua implantação dentro das organizações**. 2002. Dissertação (Mestrado) – Universidade Federal de Santa Catarina- UFSC. Florianópolis. 2002.

CELLA, Antonio S.. **Sistemas de Informações para a Gestão Estratégica das IES-Privadas**. 2006. 204 f. Dissertação (Mestrado em Ciência da Informação) - Pontifícia Universidade Católica de Campinas. São Paulo, 2006.

CHIARA, Ramon. **Aplicação de Técnicas de Data Mining** em Logs de Servidores Web. 2003. Dissertação (Mestrado). Instituto de Ciências Matemáticas e de Computação - ICMC-USP. 2003.

CHIZZOTTI, Antonio. **Pesquisa em ciências humanas e sociais**. 2. ed. São Paulo: Cortez, 1995.

CITELLI. **A comunicação e educação**: A linguagem em movimento. São Paulo: Editora SENAC, 2000.

COLENCI JUNIOR, A. ; GODOY, M. A. ; SAES, Maria Elizete Luz ; SPIGOLON, A. L. . A Gestão Estratégica das Instituições de Ensino Superior: uma contribuição ao melhor desempenho no caso brasileiro. **In:** III Workshop de Pós-Graduação e Pesquisa do Centro Paula Souza, 2008, São Paulo. Anais do IIII Workshop de Pós-Graduação e Pesquisa do Centro Paula Souza, 2008.

CORDEIRO, J.P.C., **Extracção de Elementos Relevantes em Texto/Páginas daWorld Wide Web**. Dissertação(Mestrado). Faculdade de Ciências da Universidade do Porto. 2003.

COSTA, Cezar H. V.. **Posicionamento Geográfico de Dispositivos Móveis em Ambientes Externos Utilizando a Tecnologia Wi–Fi e Redes Neurais Artificiais**. 2010. 100 f. Dissertação (Mestrado em Ciências em Engenharia Civil) – Universidade Federal do Rio de Janeiro – UFRJ, Rio de Janeiro, 2010.

CRISP-DM. **Cross Industry Standart Process for Data Mining**. Disponível em:< http://www.crisp-dm.org/>. Acesso em: out. 2010.

CRISP-DM. **Cross Industry Standart Process for Data Mining**. Disponível em: <http://www.crisp-dm.org/>. Acesso em: 11 nov. 2010.

CRUZ, Armando J. R. da. **Data Mining via Redes Neuronais Artificiais e Máquinas de Vectores de Suporte.** 2007. 123 f. Dissertação (Mestrado em Sistemas de Informação) – Universidade do Minho, Lisboa, 2007.

DALFOVO, Oscar. **Desenho de um Modelo de Sistema de Informação Estratégico para a Tomada de Decisão nas Pequenas e Médias Empresas do Setor Têxtil de Blumenau**. 1998. Dissertação (Mestrado em Administração) – Universidade Regional de Blumenau – FURB. Blumenau. 1998.

_____. **Modelo de Integração de Um Sistema de Inteligência Competitiva com um Sistema de Gestão da Informação e de Conhecimento**. 2007. Tese (Doutorado em Engenharia e Gestão do Conhecimento) – UFSC, Universidade Federal de Santa Catarina. 2007.

DAVENPORT, Thomas H. **Ecologia da Informação**: por que só tecnologia não basta para o sucesso na era da informação. São Paulo: Futura, 1998.

_____.; PRUSAK, Laurence. **Conhecimento empresarial**: como as organizações gerenciam o seu capital intelectual. 4. ed. Rio de Janeiro: Campus, 1998.

_____; MARCHAND, Donald A.; DICKSON, Tim. **Dominando a Gestão da Informação**. Porto Alegre: Bookmann, 2004.

_____; PRUSAK, Laurence. **Conhecimento empresarial**: como as organizações gerenciam o seu capital intelectual. 4.ed. Rio de Janeiro: Campus, 1998.

DE MORI, Luci M. **Sistema de Informação Gerencial para Previsão de Produtividade do Trabalho na Alvenaria de Elevação**. 2008. 232 f. Tese (Doutorado em Engenharia Civil) – Universidade Federal de Santa Catarina – UFSC, Florianópolis, 2008.

DIAS, Maria M. **Um Modelo de Formalização do Processo de Desenvolvimento de Sistemas de Descoberta de Conhecimento em Banco de Dados**. 2001. 212 f. Tese (Doutorado em Engenharia da Produção) – Universidade Federal de Santa Catarina – UFSC, Florianópolis, 2001.

DIAS, Maxwel M.; et all. **Aplicação de Técnicas de Mineração de Dados no Processo de Aprendizagem na Educação a Distância**. XIX Simpósio Brasileiro de Informática na Educação. Florianópolis, 2008.

DINGSOYR, Torgeir. **Knowledge Management in Medium-Sized Software Consulting Companies**. 2002. 256p. (Tese, Doutorado em Ciência da Computação). Trondheim: Norwegian University of Science and Technology. 2002.

FAYYAD _____; PIATETSKY-SHAPIRO, G.; SMYTH, P. **From Data Mining to Knowledge Discovery:** An Overview. Menlo Park, CA: AAAI Press/The MIT Press, 1996

FAYYAD, U. **Advances in knowledge discovery and data mining**. Cambridge: MIT Press, 1996.

FERNEDA , Edberto. **Recuperação de Informação: Análise sobre a contribuição da Ciência da Computação para a Ciência da Informação**.2003. Tese (Doutorado em Ciências da Comunicação) – USP, Universidade de São Paulo, 2003.

FERREIRA, José G. H. de M.. **Tratamento de Dados Geotécnicos para Predição de Módulos De Resiliência de Solos e Britas Utilizando Ferramentas de Data Mining**. 2008. 264 f. Tese (Doutorado em Ciências em Engenharia Civil) - Universidade Federal do Rio de Janeiro- UFRJ, Rio de Janeiro, 2008.

FIALHO, Regina C. N.. **Tecnologia de informação como vantagem competitiva na cadeia de suprimento da FIAT automóveis**. 2001. 170 f. Dissertação (Mestrado em Administração) - Programa de Pós-Graduação em Engenharia de Produção, Universidade Federal de Santa Catarina, Florianópolis, 2001.

FIGUEIRA, Rafael. **Mineração de dados e bancos de dados orientados a objetos.** 1998. Dissertação (Mestrado em Ciências da Computação) – Universidade Federal do Rio de Janeiro, Rio de Janeiro, 1998.

FONSECA, Oswaldo L. H.. **Análise de crédito utilizando inteligência artificial - validação com dados do cartão BNDES.** 2008. 143 f. Tese (Doutorado) - Universidade do Estado do Rio de Janeiro, Instituto Politécnico, Nova Friburgo, 2008.

FUNDAÇÃO NACIONAL DA QUALIDADE. **Cadernos de Excelência: Informações e Conhecimento**. São Paulo. Fundação Nacional da Qualidade, 2007. - (Série Cadernos de Excelência, n. 5.)

FURTADO, M. I. V. **Inteligência competitiva para o ensino superior privado: Uma abordagem através da mineração de textos**. 2004. Tese (Doutorado). COPPE/UFRJ. Universidade Federal do Rio de Janeiro, Rio de Janeiro. 2004.

GALLUCCI, Laura. **Gestão do conhecimento em instituições privadas de ensino superior**: Bases para a construção de um modelo de compartilhamento de conhecimento entre os membros do corpo docente. 2007. Dissertação (Mestrado) – Pontifícia Universidade Católica de São Paulo. 2007.

GIL, Antônio Carlos. **Como elaborar projetos de pesquisa**. São Paulo: Atlas, 2002.

GODOI, Christiane Kelinübing, BANDEIRA-DE-MELO, Rodrigo, DA SILVA, Arielson Barbosa (Organizadores). **Pesquisa Qualitativa em Estudos Organizacionais:** Paradigmas, Estratégias e Métodos. São Paulo. Saraiva, 2006.

GOLDSCHMIDT, R.R.; PASSOS, E. **Data Mining**: Um guia prático, conceitos, técnicas, ferramentas, orientações e aplicações. São Paulo: Elselvier 2005.

GONÇALVES, Caio Márcio, COLAUTO, Romualdo Douglas. BEUREN, Ilse Maria. **Proposta de Indicadores para um Sistema de Inteligência Competitiva em Instituição de Ensino Superior**. Disponível em:< http://www.inpeau.ufsc.br/wp/wp-content/BD_documentos/1301.doc.> Acesso em: 02 nov. 2009.

GOUVEIA, Luis Borges e RANITO, João. **Sistemas de Informação de Apoio à Gestão.** Sociedade Portuguesa de Inovação. Porto, 2004. Disponível em: < https://bdigital.ufp.pt/dspace/bitstream/10284/264/1/Manual_VII.pdf>. Acesso em: 12 Maio 2010.

GOUVEIA, Roberta M. M.. **Mineração de Dados em Data Warehouse para Sistema de Abastecimento de Água.** 2009. 147 f. Dissertação (Mestrado em Informática) – Universidade Federal da Paraíba – UFPB. Paraíba, 2009.

HADDAD, Claudia M. S.. **Sistemas de Informação e a Tomada de Decisão Executiva: Um Estudo Exploratório na Indústria Química Nacional**. 2007. 147 f. Dissertação (Mestrado em Gestão de Negócios) – Universidade Católica de Santos – UCS. Santos, São Paulo, 2007.

HARJINDER, G; RAO, P.C. **The official design the data warehousing**. Que Corporation, 1996.

HIRAGI, GILBERTO de O.. **Mineração de Dados em Base de Germoplasma**. 2008, 108 f. Dissertação (Mestrado em Informática) - Universidade de Brasília- UnB, Brasília, 2008.

HOMMERDING, Nádia M. dos S.. **O profissional da informação e a Gestão do conhecimento nas empresas: um novo espaço de atuação com ênfase no processo de mapeamento do conhecimento e disponibilização por meio da Intranet**. 2001. 221 f. Dissertação (Mestrado) Escola de Comunicações e Artes da Universidade de São Paulo - ECA/ USP. São Paulo, 2001.

INEP – Instituto Nacional de Estudos e Pesquisas Educacionais Anísio Teixeira. **Censo Superior 2008.** Disponível em : <http://www.inep.gov.br/superior/censosuperior/sinopse/default.asp>. Acesso em: 02 Jan 2010.

INMON, Willian H. **Como construir o data warehouse**. Rio de Janeiro: Campus, 1997.

JANISSEK-MUNIZ, R.; FREITAS, H.; LESCA, H. **A Inteligência Estratégica Antecipativa e Coletiva como apoio ao desenvolvimento da capacidade de adaptação das organizações**. Revista Gestão das Organizações. 2008.

KAMPFF , Adriana J. C.. **Mineração de Dados Educacionais para Geração de Alertas em Ambientes Virtuais de Aprendizagem como Apoio à Prática Docente**. 2009. 189 f. Tese (Doutorado em Informática na Educação) – UFRGS, Universidade Federal do Rio Grande do Sul, 2009.

KAMPFF, Adriana Justin Cerveira ; REATEGUI, Eliseo ; LIMA, José Valdeni de . 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, 2008.

KANASHIRO, Augusto. **Um data warehouse de publicações científicas: indexação automática da dimensão tópicos de pesquisa dos datamarts**. 2007. 109 f. Dissertação (Mestrado em Ciência de Computação e Matemática Computacional) – USP, São Carlos, 2007.

KIMBALL, Ralph; MERZ, Richard. **Data Webhouse: construindo o Data Warehouse para a Web**,. Tradução: Edson Furmankiewicz, Joana Figueiredo. Rio de Janeiro, Campus, 2000.

KOBS, Fabio F.; REIS, Dálcio R. dos.. Gestão nas Instituições de Ensino Superior Privado. **Revista Científica de Administração,** v. 10, n. 10, jan./jun. 2008.

KSHIRSAGAR, Sumedha; MAGNENAT - THALMANN, Nadia. **Multimedia communication with virtual humans**. Disponível em: < http://www.miralab.unige.ch//repository/papers/11.pdf>. Acesso em: 02 out. 2008.

LACERDA, Rafael de Alencar. Um modelo pedagógico de atividades colaborativas na web para desenvolvimento de equipes de alto desempenho. **In: CONGRESSO INTERNACIONAL DE EDUCAÇÃO A DISTÂNCIA**, 12., 2005, Florianópolis. Anais... Florianópolis: ABED, 2005. p. 01 – 10. Disponível em: <http://www.abed.org.br/congresso2005/>. Acesso em: 14 out. 2008.

LASSILA, O.; SWICK, R. R. **Resource Description Framework (RDF) Model and Syntax Specification**. 1999. W3C Recommendation. http://www.w3.org/TR/REC-rdf-syntax/.

LAUDON, Kenneth C. e LAUDON, Jane Price. Management information systems: organization and technology. 4th ed. New Jersey - Ed. Prentice-Hall, Inc.1996.

_____. **Gerenciamento de sistemas de informação.** 3. Ed. Rio de Janeiro: LTC. 2001.

LE COADIC, Y. F. **A Ciência da Informação**. 2. ed. Brasília: Briquet de Lemos, 2004.

LOYOLLA, Waldomiro; PRATES, Maurício. Ferramental pedagógico da educação a distância mediada por computador. **In: CONGRESSO INTERNACIONAL DE EDUCAÇÃO A DISTÂNCIA,** 8., 2001, Brasília. Anais... Brasília: ABED, 2001. p. 01 – 10. Disponível em: <http://www.abed.org.br/congresso2001/>. Acesso em: 02. nov. 2008.

LUCAS, Anelise de Macedo. **Utilização de Técnicas de Mineração de Dados considerando os Aspectos Temporais**. 2002. Dissertação (Mestrado). Porto Alegre: PPGC da UFRGS, 2002.

MACCARI, Emerson Antonio, SAUAIA, Antonio Carlos Aidar. **Aderência dos Sistemas de Informação na Tomada de Decisão em Jogos de Empresa. In:** Revista de Gestão da

Tecnologia e Sistemas de Informação. Vol. 3, No.3, 2006, p. 371-388. Disponível em: < www.revistasusp.sibi.usp.br/pdf/jistem/v3n3/07.pdf> . Acesso em: 20 Jul 2010.

MARTINHAGO, Sergio. **Descoberta de Conhecimento sobre o Processo Seletivo da UFPR**. 2005. 125 f. Dissertação (Mestrado em Ciências) - Universidade Federal do Paraná – UFP, Curitiba, 2005.

MORATE, Diego G. **Manual de WEKA**. Valladolid, 2010. Disponível em: < http://www.metaemotion.com/diego.garcia.morate/>. Acesso em: 12 jan 2010.

MOTTA, Custódio Gouvêa Lopes da. **Metodologia para Mineração de Regras de Associação Multiníveis Incluindo Pré e Pós-Processamento.** 2010. Tese (Doutorado em Engenharia Civil) - UFRJ/ COPPE. Rio de Janeiro. 2010.

MURRAY, Peter J.; MASON, Robin. C**omputer-mediated communication (CMC): state of the art**. Revista Brasileira de Aprendizagem Aberta a Distância, Brasília, v. 1, n. 2, jan. 2003.

NEVES, José Luiz. **Pesquisa qualitativa** - Características, usos e possibilidades. Caderno de Pesquisas em Administração, São Paulo, V.1, Nº 3, 2º Sem. 1996.

NOBREGA, Clemente. **A ciência da gestão: marketing, inovação, estratégia:** um físico explica a gestão - a maior inovação do século XX - como uma ciência. 2 ed. Rio de Janeiro. Ed. Senac Rio, 2004.

NOGUEIRA, Mário Lúcio de Lima**. A educação a distância como ferramenta de inclusão.** In: Congresso Internacional de Qualidade em EAD, 5, 2005. São Leopoldo. Anais...São Leopoldo: UNISINOS, 2005.

NONAKA, Ikujiro e TAKEUCHI, Hirotaka. **Criação de Conhecimento na empresa**. Rio de Janeiro: Campus, 1997.

OLIVEIRA, Djalma de P. R. de. **Sistemas de informações gerenciais**. 7. ed. – São Paulo: Atlas, 2001.

PERRENOUD, Phillipe (2000). **Entrevista**. Disponível em:<http:// www.unige.ch / fapse/SSE /teachers / perrenoud/ php main / php_ 2000/2000_31.html>. Acesso em 15 Mar 2008.

PICCOLI, G.; AHMAD, R.; IVES, B. **Web-based virtual learning environments**: a research framework and a preliminary assessment of effectiveness in basic IT skill training. Mis Quarterly, v. 25, n. 4, p. 410 – 426, dec. 2001.

POE, Vidette, KLAUER, Patricia, BROBST, Stephen. **Building a data warehouse for decision support**. New Jersey, Prentice Hall PTR. 1998.

PRESCOTT, J; MILLER, S. **Inteligência Competitiva na prática.** Rio de Janeiro: Campus, 2002.

PRETTI, Orestes. Autonomia do Aprendiz na Educação a Distância: Significados e Dimensões IN O. Pretti (Org). **Educação a Distância: Construindo Significados**, p. 126-145. Cuiabá: Plano, 2000.

QUONIAM, L., et al. **Inteligência obtida pela aplicação de data mining em base de teses francesas sobre o Brasil,** Revista Ciência e Informação, Brasília, v. 30, n. 2, p. 20-28, maio/ago. 2001.

RABELO, Emerson. **Avaliação de Técnicas de Visualização para Mineração de Dados**. 2007. 103 f. Dissertação (Mestrado em Ciência da Computação) - Universidade Estadual de Maringá. Maringá, 2007.

RAMASWAMI, M.; e BHASKARAN, R. A Study on Feature Selection Techniques in Educational Data Mining. **In: Journal of Computing**, volume 1, ISSUE 1, december, 2009. Disponível em: < http://arxiv.org/abs/0912.3924> . Acesso em: 23 out 2009.

REGO, Tereza Cristina. Vygotsky**: uma perspectiva histórico-cultural da educação**. Petrópolis, RJ: Vozes, 1995.

REYES, Sady. C. Fuentes. LOBAINA, Marina. Ruiz. **Minería Web: un recurso insoslayable para el profesional de la información.** Acimed. Cuba. N 16,4 out-2007. Disponível em: < http://scielo.sld.cu/pdf/aci/v16n4/aci111007.pdf> Acesso em: 20 set 2008.

REZENDE, Denis A. e ABREU, Aline F.. **Tecnologia da Informação Aplicada a Sistemas de Informação Empresariais.** São Paulo: Atlas, 2000.

REZENDE, Denis A., **Tecnologia da informação aplicada a sistemas de informação empresariais**, 2. ed. São Paulo: Atlas, 2001.

ROBREDO, J. da. **Ciência da Informação revisitada aos sistemas humanos de informação**. Brasília: Thesaurus, 2003.

RODRIGUES, Leonel C., MACARRI. Emerson. A. **Gestão do conhecimento em instituições de ensino superior**. Revista de Negócios, Vol. 8, No 2 (2003). Disponível em: <http://proxy.furb.br/ojs/index.php/rn/article/viewArticle/318>. Acesso em: 20 out 2009.

RODRIGUES, Carlos Rangel et all. **Ambiente virtual: ainda uma proposta para o ensino**. Ciências & Cognição 2008; Vol 13 (2): 71-83. Disponível em:<http://www.cienciasecognicao.org>, Acesso em: 16 de out. 2008.

ROESCH, S. M. A. **Projetos de estágios e de pesquisa em administração: guias de estágios, trabalhos de conclusão, dissertações e estudo de casos**. 2. ed. São Paulo: Atlas, 1999.

ROSSETTI, Adroaldo Guimarães ; PACHECO, Ana Paula R ; SALLES, Bertholdo W ; GARCIA, Marcos Antonio ; SANTOS, Neri dos . **A organização baseada no**

conhecimento: novas estruturas, estratégias e redes de relacionamento. Ciência da Informação, v. 37, p. 61-72, 2008.

SANTOS, George França dos. **Uma avaliação dos níveis de aceitação de curso de preparação de monitores para educação à distância da UVB** – Universidade Virtual Brasileira. 2002. 90 f. Dissertação (Mestrado em Engenharia de Produção) – Programa de Pós-Graduação em Engenharia de Produção, Universidade Federal de Santa Catarina, Florianópolis, Santa Catarina, 2002.

SANTOS, Geraldo de O. **Redes Complexas em Mineração de Dados: Aplicação no Segmento De Segurança, Meio Ambiente e Saúde**. 2008. 174 f. Tese (Doutorado em Ciências em Engenharia Civil) - Universidade Federal do Rio de Janeiro – UFRJ, Rio de Janeiro, 2008.

SANTOS, N. **Estado da arte em espaços virtuais de ensino e aprendizagem**. Revista Brasileira de Informática na Educação, n.4, abril 1999. p 75-94.

SCARINCI, Rui G.. **SES : sistema de extração semântica de informações**. 1997, 165 f. Dissertação (Mestrado em Ciência da Computação) – Universidade Federal do Rio Grande do Sul – UFRG, Porto Alegre, 1997.

SCOSS, Anne M.. **A Clusterização e Classificação no Processo De Data Mining para Análise do Desempenho Docente no Ensino de Graduação**. 2006. 86 f. Trabalho de Conclusão de Curso (Especialização) - Universidade do Extremo Sul Catarinense - UNESC, Criciúma, 2006.

SHIBA, Sonia Kaoru. M**odelagem de processo de extração de conhecimento em banco de dados para sistemas de suporte à decisão**. 2008. Dissertação (Mestrado) - Escola Politécnica da Universidade de São Paulo. 2008

SILVA F.; CÂNDIDO G., Aplicação da Tecnologia da Informação como Ferramentade apoio para Inteligência Competitiva e a Gestão do Conhecimento: Um Estudo de Caso no Setor Varejista, 2003.

SILVA, E. L. da; MENEZES, E. M. **Metodologia da pesquisa e elaboração de dissertação**. 4. ed. Florianópolis: UFSC, 2005. 138 p. Disponível em: <www.posarq.ufsc.br/download/metPesq.pdf>. Acesso em: 04 set. 2010.

SILVEIRA, Murilo A. A. Da. **Rede de Textos Científicos: um estudo sob à ótica da institucionalização da Ciência da Informação no Brasil**. 2008. 133 p. Dissertação (Mestrado em Ciência da Informação) - Pontifícia Universidade Católica de Campinas. Campinas, 2008.

SMITH, M.; WELTY,C.; MCGUINNESS D. **OWL Web Ontology Language Guide HomePage.** 2004. Disponível em: <http://www.w3.org/TR/owl-guide/>. Acesso em: 08 set 2008.

SOUZA, Renato R.. **Uma proposta de metodologia para escolha automática de descritores utilizando sintagmas nominais**. 2005. 197 f. Tese (Doutorado em Ciência da Informação) - Universidade Federal de Minas Gerais – UFMG, Belo Horizonte, 2005.

SRIVASTAVA, Jaideep. DESIKAN, Prasanna. KUMAR, Vipin .**Web Mining – Accomplishments & Future Directions.** Disponível em:< http://www.cs.umbc.edu/~kolari1/Mining/webmining.html>. Acesso em : 23 out. 2008.

STAIR, Ralph M. **Princípios de sistema de informação: uma abordagem gerencial**. 2. ed. Rio de Janeiro: CTC, 2002.

STAIR, Ralph M; REYNOLDS, George W. **Princípios de sistemas de informação: uma abordagem gerencial.** Trad. Flávio Soares Corrêa da Silva (coord.) Giuliano Mega, Igor Ribeiro Sucupira. 6ª ed. São Paulo: Cengage Learning, 2008.

STAREC, Cláudio. **A dinâmica da informação**: a gestão estratégica da informação para a tomada de decisão nas organizações , in STAREC,C.; GOMES E.; BEZERRA J.(Org). **Gestão estratégica da Informação e inteligência competitiva** . São Paulo : Saraiva, 2005. p. 47-64.

STATA, R. **Aprendizagem Organizacional: a chave da inovação gerencial**. In:STARKEY, K. (Ed.). Como as Organizações Aprendem: relatos do sucesso de grandes empresas. São Paulo: Futura, 1997. cap. XVII, p. 376-96.

SVEIBY, Karl Erik. **A nova riqueza das organizações**: gerenciando e avaliando patrimônios de conhecimento. Rio de Janeiro: Campus, 1998.

TACHIZAWA, Takeshy; ANDRADE, Rui Otávio Bernardes de. **Gestão de instituições de ensino.** 3. ed. Rio de Janeiro : FGV, 2002.

TEIVE, Raimundo C. G. **Raciocínio Baseado em Casos**. Material da Disciplina RBC no Programa de Pós-Graduação em Computação Aplicada, UNIVALI, 2008.

TERRA, José Claudio Cyrineu e GORDON, Cindy. **Portais Corporativos: A revolução na Gestão do Conhecimento**. São Paulo: Negócio, 2002.

TESTA, M. G. **Efetividade dos ambientes virtuais de aprendizagem na internet**: A influência da autodisciplina e da necessidade de contato social do estudante. Disponível em: <http://professores.ea.ufrgs.br/hfreitas/orientacoes/dout_arq/pdf/proposta_gregianin.pdf>. Acesso em 02 nov. 2008.

UNIVERSITY OF WAIKATO**. Weka 3 – Machine Learning Software in Java**. Disponível em:< http://www.cs.waikato.ac.nz/ml/weka>. Acesso em: 20 Jan 2009.

VEDOVELLI, Alexandre S**.. Desenvolvimento de um Sistema de Informação para o Processo de Implantação do Planejamento Estratégico: O Caso de uma IES**. 2005. 133 f.

Dissertação (Mestrado em Administração de Negócios) – Pontifícia Universidade Católica do Rio Grande do Sul – PUCRS, Porto Alegre, 2005.

VERGARA, S. C. **Projetos e relatórios de pesquisa em administração**. 4. ed. São Paulo. Atlas, 2003

VICTORINO, Ana Lúcia Quental et al. Utilização de ambiente colaborativo na internet como suporte para o ensino de graduação e pós-graduação. **In: CONGRESSO INTERNACIONAL DE EDUCAÇÃO A DISTÂNCIA**, 10., 2003, Porto Alegre. Anais... Porto Alegre: ABED, 2003. p. 01 – 09. Disponível em: < http://www.abed.org.br/congresso2003/>. Acesso em: 13 out. 2008.

VIEIRA, Alexandre Thomaz; ALMEIDA, Maria Elizabeth Bianconcini de; ALONSO, Myrtes. **Gestão educacional e tecnologia**. São Paulo : Avercamp, 2003.

WANGENHEIM, Christiane Gresse von. WANGENHEIM, Aldo von. **Raciocínio Baseado em Casos.** Barueri, São Paulo: Manole, 2003. 293p.

WAZLAWICK, Raul Sidney. **Metodologia de Pesquisa para Ciência da Computação**. Rio de Janeiro: Elsevier, 2008.

WILEY, D. A. Connecting learning objects to instructional design theory: A definition, a metaphor, and a taxonomy. **In D. A. Wiley (Ed.) The instructional use of learning objects.** 2001. Disponivel em <http://reusability.org/read/chapters/wiley.doc>. Acesso em 25 nov 2008.

WOODY JR, Thomaz. **Quo vadis, Pindorama?.** Disponível em:< http://cartacapital.com.br/edicoes/2006/11/421/quo-vadis-pindorama/>. Acesso em 28 nov. 2008.

ZAMBENEDETTI, Christian**. Extração de informação sobre bases de dados textuais**. 2002. 142 f. Dissertação (Mestrado em Ciência da Computação) - Universidade Federal do Rio Grande do Sul, UFRGS, Porto Alegre, 2002.

ZWIEREWICZ, Marlene; MOTTA, Neide de Oliveira; VALLEJO, Antonio Pantoja. Inclusión de la diversidad em ambientes virtuales de aprendizaje. **In: CONGRESSO INTERNACIONAL DE EDUCAÇÃO A DISTÂNCIA,** 12., 2005, Florianópolis. Anais... Florianópolis: ABED, 2005. p. 01 – 10. Disponível em: <http://www.abed.org.br/congresso2005/>. Acesso em: 14 out. 2008.

ANEXO A – QUESTIONÁRIO APLICADO AOS INGRESSANTES

Os itens abaixo relacionados são provenientes do questionário sócio-educacional aplicado aos ingressantes pela IES no ato da inscrição. Os itens foram aproveitados na integra como atributos para base de dados desta pesquisa.

Com o objetivo de traçar um perfil dos ingressantes no semestre ANO/1 e avaliar a eficácia das campanhas de marketing, solicitamos a sua atenção para responder as questões que seguem.

1. Curso
(1) Administração
(2) Ciências Contábeis
(3) Publicidade e Propaganda
(4) Jornalismo
(5) Direito
(6) Psicologia

2. Sexo
(1) Masculino
(2) Feminino

3. Idade
(1) Até 20 anos
(2) 21 a 25 anos
(3) 26 a 30 anos
(4) 31 a 40 anos
(5) Mais de 40 anos

4. Bairro: _____

5. Ocupação
(1) Empregado de empresa privada.
(2) Funcionário público.
(3) Tem negócio próprio.
(4) Administra negócios da família.
(5) Não trabalha.
(6) Outra (descreva). _____

6. Estado civil
(1) Solteiro(a)
(2) Casado(a)
(3) Separado(a)//divorciado(a)
(4) Viúvo(a)
(5) Outro

7. Com quem você mora atualmente?
(1) Com os pais
(2) Com esposo(a) e/ou filho(s)
(3) Com amigos
(4) Sozinho(a)
(5) Outro

8. Qual a faixa de renda mensal da sua família?
(1) Até R$ 2.325,00
(2) De R$ 2.325,01 a R$ 4.650,00
(3) De R$ 4.650,01 a R$ 6.975,00
(4) De R$ 6.975,01 a R$ 9.300,00
(5) Mais de R$ R$ 9.300,00

9. Qual meio de transporte utiliza para vir à faculdade?
(1) Veículo próprio
(2) Ônibus
(3) Carona
(4) Outro (descreva) _____

10. Assinale a situação que melhor descreve seu caso do ponto de vista financeiro.
(1) Não trabalho e meus gastos são financiados pela família.
(2) Trabalho e recebo ajuda da família.
(3) Trabalho e me sustento.
(4) Trabalho e contribuo com o sustento da família.
(5) Trabalho e sou o principal responsável pelo sustento da família.

11. Em que tipo de escola você cursou o ensino médio?
(1) Todo em escola pública.
(2) Todo em escola particular.
(3) A maior parte do tempo em escola pública.
(4) A maior parte do tempo em escola particular.
(5) Metade em escola pública e metade em escola particular.

12. Que meio você mais utiliza para se manter atualizado(a)? (resposta ÚNICA)
(1) Jornais
(2) Revistas
(3) TV
(4) Rádio
(5) Internet

13. Há quanto tempo concluiu o Ensino Médio?
(1) Menos de 1 ano
(2) Entre 1 e 3 anos
(3) Entre 4 e 6 anos
(4) Entre 7 e 10 anos
(5) Mais de 10 anos

14. Por que razão você escolheu o seu curso?
(1) Adequação às minhas aptidões pessoais.
(2) Prestígio da profissão.
(3) Bom mercado de trabalho.
(4) Perspectiva de boa remuneração.
(5) Outra (descreva). _____

15. Porque escolheu o IBES?
(1) Localização
(2) Credibilidade/Qualidade
(3) Preço
(4) Parceria com FGV
(5) Outro (descreva). _____

16. Quem tomou a decisão de você estudar no IBES? (UNICA resposta)
(1) Eu mesmo(a)
(2) Meus pais
(3) Companheiro(a)
(4) Eu e meus pais
(5) Eu e meu (minha) companheiro(a)
(6) Outro

17. Quem influenciou a decisão de você estudar no IBES? (MÚLTIPLA resposta)
(1) Amigos
(2) Familiares
(3) Companheiro(a)
(4) Colegas de trabalho
(5) Empregador/chefe
(6) Outro

18. Por quais meios você obteve informações sobre o IBES e seu processo seletivo? (MÚLTIPLA resposta)
(1) Panfleto
(2) Rádio
(3) Jornal
(4) Televisão
(5) Internet
(6) Blitz do Vestibular
(7) Outdoor
(8) Display em Relógios
(9) Boca-a-boca
(10) Convênio com minha empresa
(11) Outro (descreva) _____

19. Qual deles mais atingiu você? (resposta ÚNICA)
(1) Panfleto
(2) Rádio
(3) Jornal
(4) Televisão
(5) Internet

(6) Blitz do Vestibular
(7) Outdoor
(8) Display em Relógios
(9) Boca-a-boca
(10) Convênio com minha empresa
(11) Outro (descreva) _____

20. De que forma você se sentiu tocado pelas ações de divulgação promovidas pelo IBES?
(1) Não me senti tocado.
(2) Fui tocado, mas não o suficiente para me convencer.
(3) As ações de marketing foram responsáveis pela minha decisão.

21. Por qual meio você buscou mais informações sobre o Processo Seletivo do IBES?
(1) Amigos e/ou familiares
(2) Na internet
(3) Por telefone
(4) Vindo pessoalmente ao IBES
(5) Outro (descreva) _____

22. Se acessou o site do IBES para buscar informações, qual sua avaliação?
(1) Encontrei facilmente as informações que precisava.
(2) Encontrei com dificuldade as informações que precisava.
(3) Encontrei uma parte das informações que precisava.
(4) Não encontrei as informações que precisava.
(5) Não acessei o site.

23. Se buscou informações por telefone, qual sua avaliação?
(1) Obtive facilmente as informações que precisava.
(2) Obtive com dificuldade as informações que precisava.
(3) Obtive uma parte das informações que precisava.
(4) Não obtive as informações que precisava.
(5) Não busquei informações por telefone.

24. Qual sua intenção após concluir o curso?
(1) Atuar como empregado de empresa privada.
(2) Realizar concurso público.
(3) Administrar negócios da família.
(4) Criar negócio próprio.
(5) Outra (descreva) _____

25. Comentários finais (OPCIONAL).

_____
_____
_____

Muito obrigado por sua atenção!

ANEXO B – QUESTIONÁRIO APLICADO AOS EGRESSOS

PESQUISA DE EGRESSO
GRADUAÇÃO

Nome: _____e-mail: _____

Graduação:_____ Ano de Conclusão:_____

Telefone:_____ Idade:_____Sexo: _____

Cidade que cursou a graduação_____ Polo (para EAD) _____

DADOS PROFISSIONAIS

1- Você está trabalhando atualmente?

□ Sim □ Não

2 - Trabalha na área da sua formação acadêmica?

□ Sim - Especifique a área:
_____

□ Não - Motivo:
_____

3 - Situação profissional:

□ ***Empregado*** Área: □ Administrativa □ Produção □ Comercial

□ Outro – Qual:
_____

□ ***Autônomo*** Área: □ Prestação de serviços Especializados □ Vendas

□ Outro – Qual:
_____

4 - Dados da empresa em que trabalha:

Nome:_____Cidade:_____

Nº de funcionários: _____ Tempo de atividade: _____

Segmento de Atuação:
□ Indústria □ Serviços □ Pública □ Comércio □ 3º Setor

□ Outro – Qual:
_____

5 - A sua atividade profissional atual teve início:
□ antes da graduação □ durante a graduação □ após formação acadêmica

6 - Classifique sua Renda Bruta Mensal atual:

	□ Até 2 salários mínimos □ De 2 a 5 salários mínimos □ De 5 a 10 salários mínimos □ Acima de 10 salários mínimos
AVALIAÇÃO DA IES	**7 - Como você avalia a qualidade da graduação realizada na Instituição de Ensino?** □ Muito Bom □ Bom □ Regular □ Fraco
	8 - A sua formação contribuiu para: □ Ingressar no trabalho atual □ Ocupar o cargo atual □ Aumento de Salário □ Ascensão profissional □ Não contribuiu
	9 - As matrizes curriculares de todos os cursos de graduação da Sociesc são elaboradas para preparar profissionais de alto desempenho, neste sentido você classifica a matriz curricular do seu curso como: □ Muito Bom □ Bom □ Regular □ Fraco
	10 - Quando você comenta sobre a graduação realizada na nossa Instituição de Ensino a reação das pessoas é: □ Reconhecimento □ Respeito □ Simpatia □ Indiferença
INFORMAÇÕES IMPORTANTES	**11 – No momento, você está estudando?** □ Sim □ SOCIESC □ Outra Instituição Qual: _____ □ Não Por quê? _____
	12 - Você entra em contato com a Sociesc? □ Sim □ Não **Meios de contato:** □ E-mail □ Telefone □ Site □ Pessoalmente Motivo do contato: _____
	13 - Você é contactado pela nossa Instituição de Ensino? □ Sim □ Não Meios de contato: □ E-mail □ Telefone □ Correio **Motivo pelo qual a nossa Instituição de Ensino entra em contato com você:** □ Notícias e Informações □ Divulgação de Novos Cursos □ Convite para eventos, encontros ou seminários □ Outros Qual:_____
	14 - Você tem conhecimento da política de descontos para ex-alunos? □ Sim □ Não **Qual nível de aprimoramento lhe interessa:** □ Outra Graduação □ Pós-graduação □ Mestrado □ Idiomas □ Cursos de Extensão
	15 - Você indicaria os serviços de Educação da nossa Instituição de Ensino para alguém?

	□ Sim □ Não Para quem?:_____ Por quê?_____
	16 – Os valores fundamentais da nossa Instituição de Ensino são: **1 – Crescer com reconhecimento;** **2 – Ser responsável socialmente;** **3 – Valorizar as pessoas.** **Na sua avaliação, estes valores influenciaram a sua vida profissional de forma:** □ Muito Boa □ Boa □ Regular □ Fraco
	17 - Sugira ações que a nossa Instituição de Ensino poderia adotar para estreitar o relacionamento com seus egressos: _____ _____ _____

A sua participação nesta pesquisa foi primordial para que possamos validar os Valores da Nossa Instituição de Ensino.

Obrigado.