APLICAÇÃO DE TÉCNICAS DE DATA MINING NA CARACTERIZAÇÃO DE TURNOVER INTERNO PARA O SUPORTE À GESTÃO DE PESSOAS

Alessandro de Souza Mendes

DISSERTAÇÃO DE MESTRADO EM ENGENHARIA ELÉTRICA
DEPARTAMENTO DE ENGENHARIA ELÉTRICA

FACULDADE DE TECNOLOGIA

UNIVERSIDADE DE BRASÍLIA

UNIVERSIDADE DE BRASÍLIA
FACULDADE DE TECNOLOGIA
DEPARTAMENTO DE ENGENHARIA ELÉTRICA

APLICAÇÃO DE TÉCNICAS DE DATA MINING NA CARACTERIZAÇÃO DE TURNOVER INTERNO PARA O SUPORTE À GESTÃO DE PESSOAS

ALESSANDRO DE SOUZA MENDES

ORIENTADOR: RAFAEL TIMÓTEO DE SOUSA JÚNIOR

DISSERTAÇÃO DE MESTRADO EM ENGENHARIA ELÉTRICA

PUBLICAÇÃO: PPGENE.DM-550/2013

BRASÍLIA / DF, 30 de abril de 2013.

UNIVERSIDADE DE BRASÍLIA
FACULDADE DE TECNOLOGIA
DEPARTAMENTO DE ENGENHARIA ELÉTRICA

APLICAÇÃO DE TÉCNICAS DE DATA MINING NA CARACTERIZAÇÃO DE TURNOVER INTERNO PARA O SUPORTE À GESTÃO DE PESSOAS

ALESSANDRO DE SOUZA MENDES

DISSERTAÇÃO DE MESTRADO ACADÊMICO SUBMETIDA AO DEPARTAMENTO DE ENGENHARIA ELÉTRICA DA FACULDADE DE TECNOLOGIA DA UNIVERSIDADE DE BRASÍLIA, COMO PARTE DOS REQUISITOS NECESSÁRIOS PARA OBTENÇÃO DO GRAU DE MESTRE.

APROVADA POR:

_____
RAFAEL TIMÓTEO DE SOUSA JÚNIOR, Dr., ENE/UNB
(ORIENTADOR)

_____
FLÁVIO ELIAS GOMES DE DEUS, Dr., ENE/UNB
(EXAMINADOR INTERNO)

_____
ROBSON DE OLIVEIRA ALBUQUERQUE, Dr., ABIN
(EXAMINADOR EXTERNO)

BRASÍLIA, 30 DE ABRIL DE 2013.

FICHA CATALOGRÁFICA

MENDES, ALESSANDRO DE SOUZA.

Aplicação de técnicas de Data Mining na caracterização de turnover interno para o suporte à Gestão de Pessoas. [Distrito Federal] 2013.

2013, xiii, 113p, 297 mm (ENE/FT/UnB, MESTRE, Engenharia Elétrica, 2013).

Dissertação de Mestrado - Universidade de Brasília, Faculdade de Tecnologia. Departamento de Engenharia Elétrica.

1. Mineração de Dados
2. Gestão de Pessoas
3. Rotatividade de Pessoal.
4. Data Warehouse
5. Clustering

I. ENE/FT/UnB.
II. Título (Série)

REFERÊNCIA BIBLIOGRÁFICA

MENDES, Alessandro de Souza (2013). Aplicação de técnicas de Data Mining na caracterização da turnover interno para suporte à Gestão de Pessoas. Dissertação de Mestrado em Engenharia Elétrica, Publicação PPGENE.DM-550/2013, Departamento de Engenharia Elétrica, Universidade de Brasília, Brasília, DF, 113p.

CESSÃO DE DIREITOS

AUTOR: ALESSANDRO DE SOUZA MENDES

TÍTULO: Aplicação de técnicas de Data Mining na caracterização da turnover interno para suporte à Gestão de Pessoas.

GRAU: Mestre ANO: 2013

_____

Alessandro de Souza Mendes
Campus Universitário Darcy Ribeiro, Gleba A, Faculdade de Tecnologia.
CEP: 70790-120

AGRADECIMENTOS

A Deus pelo dom da vida e pelas oportunidades concedidas em minha vida, permitindo-me enveredar pelo caminho da ciência e do saber, e dando-me o alento necessário para prosseguir. Nossa aliança é eterna!

A Nossa Senhora, pelo seu grande exemplo de vida, mostrando-me o caminho da fé, superação, esperança, tolerância, doação e principalmente, seu exemplo de amor.

Agradeço a todos que me ajudaram nesta jornada promovendo suporte e encorajamento necessário para realização desta árdua tarefa: professores, amigos, familiares e colegas de trabalho.

Ao orientador e amigo professor Dr. Rafael Timóteo Sousa Júnior, que me orientou de forma profissional nas horas mais complicadas durante a criação deste trabalho e que suportou pacientemente tantas dúvidas e problemas relativos ao assunto e outros pequenos detalhes pertinentes a esta dissertação.

Aos professores e amigos Fábio Lúcio Lopes de Mendonça e Valério Aymoré Martins pelas grandes dicas e ajudas e pelo constante apoio, incentivo, dedicação e amizade, essenciais para o desenvolvimento deste trabalho.

Ao Centro de Apoio ao Desenvolvimento Tecnológico - CDT/UnB, do qual sou bolsista e tive grande auxílio e incentivo no decorrer desse trabalho.

À instituição que me proporcionou a oportunidade de estudar a rotatividade interna de empregados através da mineração de mais de 20 anos de coleta de registros funcionais. O investimento de tempo e acesso às informações fornecidas me promoveram a oportunidade de avançar na pesquisa sobre um dos tópicos mais críticos enfrentados pelo mundo empresarial contemporâneo. Eu realmente aprecio a confiança que depositaram em mim e minha agenda de pesquisa.

Aos colegas do Laboratório de Tecnologias da Tomada de Decisão - LATITUDE/UnB e ao corpo administrativo da Faculdade de Tecnologia que incentivaram para o desenvolvimento deste trabalho.

À minha mãe Teresinha, meu pai Galdino e minha irmã Aline que não perderam a fé na minha dedicação aos meus estudos.

Aos queridos Márcia Keila e Rafael Jorge pelo apoio e incentivo que foi dado durante todo o tempo em que estive envolvido neste trabalho.

DEDICATÓRIA

Para:

José Galdino de Souza Mendes e Teresinha Maria de Jesus Mendes,

Queridos Pais,

Aline de Souza Mendes e Bárbara Luiza,

Minhas queridas irmã e sobrinha.

RESUMO

APLICAÇÃO DE TÉCNICAS DE DATA MINING NA CARACTERIZAÇÃO DE TURNOVER INTERNO PARA O SUPORTE À GESTÃO DE PESSOAS.

Esta dissertação encontra-se no campo da Mineração de Dados e suas aplicações em bases de dados de Gestão de Pessoas, na hipótese de que tais técnicas podem ser agregadas a um modelo multidimensional que leve à descoberta de fenômenos e ao entendimento dos dados relativos à rotatividade interna de pessoal e seus impactos. Foram utilizadas as abordagens de modelagem descritiva e preditiva, a fim de descobrir informações ocultas no histórico de transferências dos empregados entre as unidades de uma organização. Entre as técnicas descritivas, foram aplicados métodos de agrupamento e regras de associação, para descrever os dados. Para as análises preditivas, foi utilizada a técnica de Árvores de Decisão, um método de indução que mostra graficamente o processo de classificação. Para validar a hipótese de que tais proposições levam à descoberta de conhecimento acerca de rotatividade de pessoas, foi desenvolvido um módulo de suporte à decisão no domínio do problema, aplicando as técnicas de Mineração de Dados propostas, além da criação de um novo tipo de dimensão voltada para a descoberta de conhecimento. Para validar as contribuições e atingir o objetivo proposto neste trabalho, foram utilizados, como estudo de caso, dados oriundos de uma instituição financeira de grande porte e com um longo histórico de rotatividade de pessoas. Os resultados obtidos são apresentados e discutidos.

ABSTRACT

APPLICATION OF DATA MINING TECHNIQUES TO THE CHARACTERIZATION OF INTERNAL TURNOVER TO SUPPORT PERSONNEL MANAGEMENT

This dissertation is in the field of Data Mining and its applications in databases of Personnel Management, considering the assumption that such technique can be aggregated to a multidimensional model that leads to the discovery of phenomena and the understanding of data relating to internal employee turnover and its impacts. Descriptive and predictive modeling approaches were used in order to discover hidden information on the history of transfers of employees between departments of an organization. Among the descriptive techniques, cluster analysis and association rules have been applied to describe the data. For predictive analysis, the technique of Decision Trees was used, comprising a method of induction that graphically shows the classification process. For validating the hypothesis that such proposals lead to the knowledge discovery about employee turnover, a decision support module was developed in the problem domain, applying the proposed techniques of Data Mining. This decision support module introduces a new type of dimension focused on knowledge discovery. For validating the contributions and evaluating the achievements of this work, a case study was performed using data from a large financial institution with a long history of employee turnover. The results are presented and discussed.

SUMÁRIO

LISTA DE TABELAS

LISTA DE FIGURAS

ACRÔNIMOS

ARFF	Attribute-relation file format
BD	Banco de Dados
BI	Business Intelligence
CRISP-DM	Cross Industry Standard Process for Data Mining
DM	Data Mining
DW	Data Warehouse
DWB	Data Warehouse Bus
DC	Decision Tree
DASD	Direct Access Storage Device
DSS	Decision Support System
ETL	Extract, Transformation and Load
ID3	Iterative Dichotomiser 3
FEBRABAN	Federação Brasileira de Bancos
KDD	Knowledge discovery in databases
OLAP	Online Analytical Processing
OLAM	On-Line Analytical Mining
OLTP	Online Transaction Processing
PDI	Pentaho Data Integration
PCA	Principal Component Analysis
RH	Recursos Humanos
SGBD	Sistema de Gerenciamento de Banco de Dados
WEKA	Waikato Environment for Knowledge Analysis
WWW	World Wide Web

1 - INTRODUÇÃO

Esta dissertação se insere na temática da Mineração de Dados e suas aplicações em bases de dados de Gestão de Pessoas, particularmente no entendimento dos dados relativos à rotatividade interna de pessoal e seus impactos. São utilizadas abordagens de mineração com base em modelagem descritiva e preditiva, a fim de descobrir informações ocultas no histórico de transferências dos empregados entre as unidades de uma organização. As técnicas descritivas de agrupamento e regras de associação são aplicadas para descrever o conhecimento representado pelos dados. Para análises preditivas, é utilizada a técnica de Árvores de Decisão, um método de indução que mostra graficamente o processo de classificação.

O objetivo da utilização destas técnicas consiste em obter conhecimento para apoiar as decisões dos gestores na área de Gestão de Pessoas, bem como os processos de seleção e políticas de promoção interna. De fato, na atual conjuntura da economia mundial, com o advento de novas tecnologias desenvolvidas principalmente nas áreas de comunicação e informática, os setores produtivos e administrativos das empresas foram submetidos a acentuadas adaptações nas estruturas organizacionais e na forma de lidar com os empregados. A área de Gestão de Pessoas recebeu uma nova abordagem, migrando de uma função vista como burocrática para uma função estratégica, já que as pessoas são o recurso primário para que as organizações alcancem seus objetivos estratégicos de negócios. Assim, pessoas já não são vistas apenas como uma força de trabalho, mas são reconhecidas como um ativo valioso dentro e fora de seus ambientes de trabalho (Chiavenato, 2001). Esse reconhecimento vem da compreensão de que as pessoas são a chave para as estratégias corporativas, sendo a atuação dessas pessoas considerada como um fator determinante para o alcance das metas e objetivos organizacionais.

O suporte tecnológico tem sido um elemento fundamental desta transformação, no entanto, a maioria das organizações não percebe os potenciais benefícios que a tecnologia oferece (Patterson e Lindsey, 2003). Uma imensa quantidade de dados está disponível dentro das organizações, mas muitas vezes não são aproveitados para identificar potenciais áreas em que as empresas podem ganhar vantagem competitiva. Patterson e Lindsey (2003) afirmam

que uma análise efetiva dos dados de RH pode trazer vantagem competitiva para a organização.

No contexto atual, a competitividade está cada vez mais dependente da capacidade de geração de conhecimentos que uma organização possui (Chiavenato, 2004). Portanto, as pessoas assumem papel primordial. As políticas de gestão de pessoas tornam-se de grande importância para as organizações. Conforme aponta Huselid (1995), verifica-se um crescimento de práticas de trabalho voltadas para obtenção de alto desempenho, especificamente porque analisam os procedimentos de recrutamento e seleção, os sistemas de incentivos e compensações, e a agenda de treinamento e ações voltadas para o desenvolvimento do empregado de forma a melhorar seus conhecimentos, habilidade e atitudes no ambiente organizacional. Porém é fundamental a existência de mecanismos de retenção desses empregados na organização ao longo do tempo (Chiavenato, 2008).

A dinâmica de admissão e desligamento de pessoal impacta qualquer organização, uma vez que é um processo contínuo. Uma alta taxa de perda de empregados pode implicar em problemas e desafios organizacionais (Chiavenato, 2008). Quando um empregado deixa a empresa, ou simplesmente muda de um escritório para outro, provavelmente há uma perda de conhecimento, capital intelectual, inteligência de negócios e domínio de processos. E quando isso acontece, o reflexo é sentido visivelmente nos profissionais, uma vez que esse estado instável impacta diretamente, seja na motivação de quem permanece no quadro funcional, seja na capacidade do setor em realizar suas atividades. Neste contexto, a motivação e satisfação no ambiente de trabalho são temas importantes para melhorar o desempenho organizacional.

O termo rotatividade de pessoal é aplicado para caracterizar a dinâmica de entrada e saída de empregados de uma empresa em um determinado período (Vandeber, 1999; Chang, 1999). O estudo da rotatividade de pessoal tem atraído a atenção de muitos pesquisadores em busca de uma maior compreensão sobre o comportamento das relações entre as organizações e seus empregados (Chiavenato, 2008). Além disso, a capacidade de gerenciar os custos decorrentes da rotatividade de pessoal, visando à manutenção de seus talentos e maior competitividade, é um tema que atraiu a atenção da Governança Corporativa (Chiavenato, 2008).

A alta taxa de rotatividade nas organizações pode ser o resultado de muitos fatores. As razões para demissões variam: um indivíduo pode renunciar por não concordar com a política da empresa, falta de motivação, ou busca de uma melhor colocação profissional. Em contrapartida, a empresa tem também o direito de buscar profissionais mais qualificados para fortalecer seu quadro funcional, agindo com base em avaliações de desempenho de sua força de trabalho.

Atualmente, dado um maior investimento das organizações na área de Gestão de Pessoas, é cada vez mais frequente a avaliação das principais causas que levam os empregados a saírem de uma empresa e também dos fatores que levam a organização a demiti-los (Lacombe, 2005).

Porém, rotatividade de pessoal não se refere somente ao desligamento do empregado. De acordo com Bluedorn (1982), a rotatividade de pessoal também engloba as transferências de um indivíduo de uma função ou uma área para outra função ou área dentro da mesma organização. Um empregado, que é transferido de área, será substituído por outro que necessitará de treinamento e tempo para absorver as atividades do antigo, o que afetará temporariamente a produtividade da área. Projetos podem sofrer descontinuidades, podendo ser replanejados, visto que será necessário que o novo empregado entenda o que é e em ponto parou o projeto. O empregado promovido também deverá ser capacitado na nova área e levará um tempo para entender e realizar suas novas atividades. Portanto, é importante compreender esta rotatividade interna, uma vez que gera os mesmos problemas e desafios da rotatividade externa, só que para as diversas áreas componentes da empresa.

Sendo assim, entender a rotatividade interna de pessoal pode prover informação chave para o gerenciamento proativo de pessoal e dos custos associados. Porém, este não é um trabalho trivial devido à grande quantidade de dados existente nos bancos de dados corporativos. Utilizar ferramentas automatizadas e eficientes se torna essencial para realizar esta tarefa. Atualmente a tecnologia que mais chama atenção para a realização desta tarefa é a Mineração de Dados. Com efeito, esta dissertação procura mostrar que a análise dos dados de gestão de pessoas existentes em bases de dados permite, não apenas compreender as características das transições das pessoas no trabalho, como também projetar no futuro os impactos dessas transições. Para tanto, a contribuição desta

dissertação consiste em desenvolver modelos de mineração e propor uma nova estrutura de análise multidimensional especificamente para a compreensão da rotatividade interna de pessoas.

Para validar as contribuições e atingir o objetivo proposto neste trabalho, foram utilizados, como estudo de caso, dados oriundos de uma instituição financeira de economia mista, com unidades espalhadas por todo o território nacional, com um histórico de transferências de pessoas de mais de 20 anos, e atualmente com mais de 89 mil empregados. É pertinente destacar que por questões de ética, o nome da organização será mantido no anonimato.

1.1 OBJETIVOS

Este trabalho propõe-se a utilizar técnicas de agrupamento, sumarização e classificação, a fim de descrever o comportamento da rotatividade de empregados, avaliando as proposições com dados provenientes da empresa em estudo. Também se propõe utilizar consultas OLAP para subsidiar a exploração de dados e a avaliação do processo de mineração.

1.1.1 Objetivo Geral

O objetivo do presente trabalho é estruturar as técnicas de Mineração de Dados para detectar padrões de comportamento na rotatividade de pessoas entre áreas de uma empresa. Além disso, propõe-se a construir um módulo de suporte à decisão nesse domínio validando a proposição por um estudo de caso.

1.1.2 Objetivos Específicos

Dentro dos objetivos específicos cabem as seguintes metas:

- Realizar a coleta do referencial teórico relacionadas ao tema;
- Realizar uma cópia dos dados a serem utilizados junto à organização em estudo;
- Descrever de forma quantitativa e qualitativa a rotatividade interna de pessoal;
- Criar modelos (árvores de decisão, agrupamento, regras de associação) que descrevam o perfil dos empregados que se movimentam na empresa;

- Desenvolver e implementar um protótipo do modelo dimensional de dados sobre rotatividade de pessoal;
- Realizar estudos de métodos e técnicas de algoritmos para classificação e predição de dados;
- Apresentar resultados em soluções sistêmicas que permitam o suporte ao entendimento do assunto.

1.1.3 Justificativa

Entende-se que rotatividade de pessoal é um fenômeno que ocorre em todas as organizações, sendo então de interesse global. Conceitualmente, o fenômeno é de fácil entendimento, porém, quando se vai analisá-lo em nível de sistemas, onde realmente é registrado, o fenômeno se torna complexo. A complexidade se deve ao volume de dados registrados nos sistemas, à enorme quantidade de variáveis e à complexidade dos relacionamentos, o que dificulta as análises do pessoal da área de Gestão de Pessoas.

Diante disto, existe a necessidade de uma solução científica e tecnológica para tratar do assunto. Além disto, não foi encontrada na literatura pesquisada nenhuma solução para este problema.

Neste sentido, os trabalhos realizados e apresentados nesta dissertação têm como objetivo contribuir com uma parcela de conhecimento prático da gestão, com a utilização da tecnologia, permitindo avanços técnico-científicos na área do estudo em questão.

1.1.4 Premissas da Proposta

Como limitações de escopo, algumas pressuposições serão consideradas, entre essas:

- Foi utilizada como fonte de dados a coleta do histórico de lotação entre 2008 e 2012 da empresa em análise;

- Somente foram consideras as movimentações entre as áreas relacionadas à designação efetiva de função gratificada, cujo motivo de transferência foi de interesse da administração ou por interesse pessoal. Esta restrição se deve ao fato

de que existem registros relacionados à reestruturação da empresa ou extinção de área, por exemplo;

- Este trabalho não se destina a evidenciar as causas que levam as pessoas a saírem de uma empresa ou a trocarem de área, nem procura dar suporte à descoberta de talentos críticos;

- Apesar de fatores externos influenciarem a vida funcional dos empregados, somente foram utilizados dados internos da organização.

1.2 ORGANIZAÇÃO DO TRABALHO

Tendo estabelecido a finalidade do estudo, juntamente com as definições que servem como a base da pesquisa, os seguintes capítulos focam na revisão da literatura, descrição do procedimento de investigação e os resultados da pesquisa, e culminam em um resumo que inclui as conclusões e recomendações para um estudo mais aprofundado. Para um melhor entendimento e organização, este trabalho é dividido em Capítulos conforme relacionado a seguir.

O Capítulo 2, revisão da literatura, inclui um sumário de conceitos e pesquisas recentes sobre Gestão de Pessoas - focado nos impactos do alto índice de rotatividade de pessoal -, bem como trata de técnicas e processos focados em Mineração de Dados e *Data Warehouse*.

No Capítulo 3, descrevem-se as soluções propostas para o problema da obtenção de conhecimento sobre rotatividade interna de pessoal.

Já no capítulo 4, são apresentadas as análises e os resultados obtidos com aplicação das técnicas de Mineração de Dados na extração de informações referentes à rotatividade interna de pessoal, especificamente para validação das soluções propostas com base nos dados da empresa objeto do estudo de caso.

No Capítulo 5 é apresentada a conclusão deste trabalho e a exposição de trabalhos futuros.

2 - REVISÃO DA LITERATURA

Este capítulo tem como objetivo apresentar a fundamentação teórica, cujo propósito é abordar aspectos relevantes e um contexto para o entendimento dos principais conceitos aplicados neste trabalho.

2.1 CONCEITOS BÁSICOS DE GESTÃO DE PESSOAS

Este tópico destaca conceitos importantes envolvidos de Gestão de Pessoas, com foco na avaliação do impacto da rotatividade de pessoal e retenção de talentos.

O contexto da gestão de pessoas é formado justamente pelas pessoas e suas relações organizacionais, onde umas dependem das outras para atingir seus objetivos e cumprir suas missões, havendo sempre benefícios ou prejuízos recíprocos. As organizações constituem para as pessoas o meio pela qual elas irão conquistar seus objetivos pessoais, e, por outro lado as organizações usufruem dos esforços de várias pessoas trabalhando em conjunto.

Fisher e Fleury (1998) conceituam a gestão de pessoas como sendo o conjunto de políticas e práticas definidas de uma organização para orientar o comportamento humano e as relações interpessoais no ambiente de trabalho.

Chiavaneto (2005) caracteriza a gestão de pessoas como contingencial e situacional, já que depende de alguns aspectos como, por exemplo, da cultura ou da estrutura organizacional, das características dos conceitos ambientais, do negócio da organização, da tecnologia utilizada ou dos processos internos.

Além disso, a expressão gestão de pessoas pode referir-se ao departamento que, dentro de uma determinada empresa, é responsável por administrar e gerir o capital humano.

Marques (2012) cita alguns dos principais assuntos tratados na Gestão de Pessoas:

- Análise e descrição de cargos;
- Planejamento e administração de cargos e salários;
- Recrutamento, seleção e admissão;
- Orientação e integração de novos empregados;

- Criação de incentivos e benefícios;
- Avaliação de desempenho;
- Comunicação aos empregados;
- Treinamento e desenvolvimento (T&D).

Nesse contexto, as organizações administram seus empregados como recursos organizacionais ou parceiros da organização. Do uso da primeira visão, os mesmos precisam ser bem administrados, uma vez que são considerados parte do patrimônio físico da empresa. Caso sejam observados como parceiros da organização, os profissionais são conduzidos como parte integrante do capital intelectual da organização.

Atualmente, há grandes empresas que vêm mudando o seu conceito sobre gestão e alterando suas práticas gerenciais. Como exemplo, há o caso de empresas que ao invés de investirem diretamente em produtos e serviços, estão investindo nas pessoas que entendem sobre como manusear os mesmos, ou seja, a pessoa passa a ser vista como um pilar de sustentação do sucesso de uma organização.

Para Chiavaneto (2005), tratar as pessoas como recursos organizacionais é um desperdício de talentos. O foco atual é a Gestão de Pessoas como parte de uma estratégia organizacional, e não mais o tratamento de pessoas como recursos humanos, em que as pessoas são vistas apenas como meros empregados remunerados em função do tempo disponibilizado em uma determinada organização.

Conforme Perillo (2009), a Tecnologia da Informação vem aliando-se à Gestão do Conhecimento para que a informação seja transformada em conhecimento com o intuito de ser compartilhada para diversos profissionais para que possam tratá-las conforme a necessidade.

Para determinada organização obter sucesso na Gestão de Pessoas, de acordo com o Sebrae Nacional (2013), é fundamental que a missão dessa organizaçãoesteja transparente aos seus empregados, além da visão de um organograma de funções com vinculação hierárquica bem definida, a fim de se garantir uma correta distribuição de tarefas. Além do mais, a criação de regulamentos internos é de grande importância para que cada pessoa conheça seus direitos e deveres.

Ainda de acordo com o Sebrae Nacional (2013), a tendência é que haja uma evolução com relação à contratação, treinamento e manutenção de empregados motivados, com o intuito de reduzir a rotatividade de pessoal (demissões que geram novas admissões), já que esta gera diversos custos adicionais para a organização.

2.1.1 Rotatividade de Pessoal

Em cenário cada vez mais competitivo dos negócios, aonde o capital intelectual é cada vez mais valorizado, é natural que pessoas mudem de emprego, num movimento natural de mudança, de oxigenação e de transformação das empresas. A rotatividade faz parte da vida e do mundo dos negócios. A expressão rotatividade de pessoal, ou do inglês *turnover*, é um termo utilizado para caracterizar o movimento de entradas e saídas, admissões e desligamentos, de profissionais empregados de uma empresa, em um determinado período.

O *turnover* é a relação entre a entrada e a saída de empregados de uma empresa, podendo ocorrer por iniciativa pessoal ou da empresa. Para Chiavenato (2005), a rotatividade de pessoal é o resultado da saída de alguns empregados e a entrada de outros para substituí-los no trabalho. Na visão do autor, rotatividade é o fluxo de pessoal na organização. Para cada saída de empregado, provavelmente ocorrerá uma reposição.

No que se refere a desligamento, há dois tipos: por iniciativa do empregado ou por iniciativa da organização. Chiavenato (2005) salienta que o desligamento por iniciativa do empregado acontece por razões pessoais ou profissionais, levando-o a encerrar o contrato de trabalho com a organização. O desligamento por iniciativa da organização surge quando a organização demite um empregado, o que pode acontecer por diversos motivos tais como: empregado mal selecionado, substituição ou redução do quadro funcional.

Em resumo, o *Turnover* ou Rotatividade de Pessoal é um conceito proveniente da área de Gestão de Pessoas, e tem como objetivo mensurar as entradas e saídas de empregados por um período de tempo específico e, consequentemente, analisar a capacidade da empresa em manter seus empregados (Bispo, 2005).

Por exemplo, se o percentual de *turnover* estiver muito grande pode significar que esteja ocorrendo um baixo comprometimento dos profissionais perante a empresa, havendo

então, necessidade de avaliação das causas de incapacidade de retenção de pessoal. Por isso, tal parâmetro é muito utilizado como indicador da saúde/estabilidade de uma empresa (Claro, 2009).

Chiavenato (1997) relata que a rotatividade de pessoal entre uma organização e seu ambiente pode ser motivada por diversos fatores:

- Incentivos;
- Recrutamento e seleção com problemas;
- Baixo comprometimento organizacional;
- Remuneração inadequada;
- Reconhecimento profissional;
- Problemas disciplinares;
- Sobrecarga de trabalho.

O mesmo autor destaca que ainda existe um controle estatístico para que novas contratações de empregados acarretem em menos custo possível, ou seja, faz uma previsão de tempo e custo para que seja reestabelecida determinada rotina em uma função.

Para Bispo (2005), o *turnover* não gera apenas perda de capital intelectual, gera também perda de conhecimento, de inteligência e de entendimento. Por isso, tal rotatividade deve ser gerenciada corretamente a fim de causar o menor impacto possível já que tais fatores impactam em perda de produtividade e lucratividade na empresa.

A perda de talentos gera desequilíbrio em uma organização, uma vez que pode gerar descontentamento de seus clientes, e ainda gerar enriquecimentos em seus concorrentes, ou seja, o *turnover* gera perdas de difícil reparação que vão além de simples admissões ou desligamentos (Bispo, 2005). Há ainda mais perdas a serem tratadas, segundo o autor:

- Sobrecarrega os antigos empregados;
- Leva tempo para integrar e orientar o novo profissional;
- Tempo do profissional de RH, desde o recrutamento até a capacitação do novo empregado;
- Menor produtividade, enquanto o novo profissional está em tempo de aprendizado;

- Aumento de acidentes e doenças, processos trabalhistas, entre outros.

Por isso, uma boa gestão de *turnover* preserva o capital intelectual, o ambiente e a imagem da empresa, fatores importantíssimos para que a mesma continue realizando suas atividades no mercado de trabalho.

Segundo Claro (2009), é possível calcular o *turnover* através da seguinte equação (2.1), onde efetivo médio do período é a média da soma do efetivo no início e final do período:

$$Turnover = \{ [(ingressos + desligamentos) /2] / (Efetivo\ médio\ do\ período) \} * 100 \quad (2.1)$$

Por exemplo, considere-se que uma empresa tem 100 empregados. No mês anterior, 10 deles foram demitidos e 6 foram contratados. Aplicando a fórmula, tem-se:

1) Efetivo Médio do Período = (Efetivo no início – Efetivo no final) /2 = (100 + 96) / 2 = 98
2) *Turnover* = { [(06 + 10)] / 2] / (98) } *100 = (8 /98) * 100 = 8,16%

Outro variável relativa ao fenômeno é expressa pela equação (2.2), que define a taxa de desligamento:

$$Taxa\ de\ Desligamento = [(Desligamentos)/(Efetivo\ médio\ do\ período)] * 100 \quad (2.2)$$

Por exemplo, considere-se que uma empresa tem 100 profissionais. No mês anterior, 10 profissionais foram desligados. Logo, aplicando a fórmula, tem-se:

1) Efetivo médio do período = (efetivo no início+ efetivo no final)/ 2 = (100+90)/2 = 95
2) Taxa de Desligamento (análise das perdas) = [(10)/[(95)] * 100 = 10,53%

Peconick (2009) afirma que a equação 1 deve ser utilizada quando houver substituições no quadro de pessoal, ou seja, novas demissões ou admissões não são consideradas. Para o autor, um resultado elevado no índice de *turnover* é fato expressivo determinante para requerer ações preventivas ou até mesmo implantar mudanças em uma organização.

Porém, tal equação, segundo Claro (2009), não é aplicada a todos os casos, e, por isto, deve ser usada com cautela. O autor relata que a equação 1 é questionada na realidade dos *call centers*, por exemplo, ou em organizações que apresentam crescimentos ou diminuições

com grande expressividade. Nessas organizações, a taxa de desligamento é que é utilizada, já que este indicador representa não só perdas de pessoas, mas principalmente, perda de conhecimento, de capital intelectual, de inteligência, de entendimento, de domínio dos processos e de conexões com os clientes.

Robbins (1999) considera que a rotatividade pode ser positiva para a organização, por exemplo, quando um trabalhador que tenha um baixo desempenho desliga-se, sendo substituído por alguém que esteja motivado e que tenha melhores habilidades. Já Fernandez (2009) considera que qualquer saída é traumática por mais argumentos que existam a seu favor. Sempre se deve considerar o quanto foi gasto com treinamento, burocracias, benefícios, encargos dentre outros gastos operacionais. O mesmo autor completa ao dizer que se o percentual do *turnover* se tornar excessivo é sinal de que algo pode estar errado na organização, então novas medidas preventivas devem ser tomadas para conter a situação. A saída de um membro da equipe significa que a organização será afetada de qualquer maneira.

Para Chiavenato (2000), o índice de rotatividade é considerado ideal quando a organização consegue reter seus profissionais bem qualificados e substituir aqueles que apresentam alguma deficiência no desempenho.

Chiavaneto (1997) destaca alguns dos prováveis impactos causados pelo alto índice de *turnover* nas empresas, que muitas vezes passam despercebidos por muitas delas:

1. Recrutamento de empregados substitutos, incluindo despesas administrativas, seleção e entrevistas, e serviços associados com a seleção, como análises de informações, processamento de referências e, possivelmente, testes psicológicos;
2. Custos administrativos de contratação;
3. Perda de produtividade associada com o período de integração do novo empregado, antes que ele exercer sua função na empresa;
4. Perda de produtividade devido ao tempo requerido para o novo empregado ter a produtividade que um empregado experiente;

5. Perda de produtividade associada com o tempo que empregados antigos têm que gastar para ajudar o novo empregado;
6. Custos de treinamento, incluindo tempo de colegas e supervisores em treinamento formal, assim como o tempo que o empregado em treinamento deve gastar fora do trabalho;
7. Custos associados com o período que antecede a demissão voluntária, quando os empregados ficam menos produtivos;
8. Em alguns casos, custos associados com a comunicação de segredos organizacionais, procedimentos, e habilidades a empresas concorrentes;
9. Custos de relações públicas associados com o grande número boatos que surgem sobre a imagem da companhia, devido ao alto número de demissões voluntárias e involuntárias;
10. Aumento dos custos de seguro-desemprego;

De acordo com Chiavenato (2000), as empresas precisam realizar constantes diagnósticos para identificar os fatores que levam ao desligamento de empregados. O mesmo autor relata que, no mundo contemporâneo, é preciso criar políticas que vão além do reajuste salarial. É preciso pensar na satisfação do empregado como um todo.

Dentre essas exigências, estão melhora no clima organizacional, políticas de promoção e plano de carreira, capacitação, o reconhecimento profissional e concessão de benefícios, sendo estes, a estratégia intencional utilizada como forma de vencer a desmotivação dos empregados frente ao mercado concorrente, bem como agregar valores institucionais, profissionais e pessoais. Por fim, o autor questiona se estes fatores são suficientes para reter talentos. Por isso, para evitar futuras perdas, é sugerido que seja questionado o porquê da saída de certo empregado para saber o que está ocorrendo na empresa.

Vale notar que a presente dissertação estende a gestão de *turnover* focada em empregados que são desligados da empresa, para a rotatividade interna dos empregados que são transferidos dentro da empresa, seja por interesse pessoal, da administração ou por promoção. De fato, a rotatividade de pessoal não se refere somente ao desligamento do empregado. De acordo com Bluedorn (1982), a rotatividade de pessoal também engloba as transferências de um indivíduo de uma função ou uma área para outra função ou área

dentro da mesma organização. Portanto, é importante compreender esta rotatividade interna, uma vez que gera os mesmos problemas e desafios só que para as diversas unidades da empresa.

2.2 DATA WAREHOUSE

No início dos anos 1970, o surgimento de uma nova tecnologia de armazenamento e acesso em disco, ou *direct access storage* device (DASD), associou-se a um novo tipo de *software* conhecido como Sistema de Gerenciamento de Banco de Dados (SGBD). Com o conceito de SGBD, surgiu à ideia de um banco de dados definido como uma única fonte de dados para todo o processamento em uma organização.

O conceito de banco de dados promoveu uma visão de uma organização "baseada em dados", em que o computador poderia atuar como coordenador central para atividades de toda a empresa. Nesta visão, o banco de dados tornou-se um recurso corporativo básico, pois passou a permitir o registro de transações diversas da organização à medida da realização dessas transações. O banco de dados passou também a permitir a consulta de tais registros para outras operações da organização, ou para consolidações, comparações, comprovações, etc. Os diversos sistemas registradores e processadores dessas transações foram justamente denominados sistemas transacionais e a própria forma de tratamento da informação ganhou a denominação de *on line transaction processing* (OLTP). Entretanto, muito voltados ao registro e recuperação de transações, tais sistemas passaram a apresentar deficiências no que se refere à análise de fenômenos nos dados. Além disso, a substituição de sistemas por sistemas mais novos passou a requerer a gestão dos sistemas e dados legados.

Em atendimento às solicitações dos gestores em relação à deficiência da análise de informação nos sistemas legados, surgiram no mercado os chamados "programas extratores". Esses programas extraem informações dos sistemas transacionais com o intuito de trabalhá-las em outros ambientes.

Muitas vezes essas extrações ocorriam em arquivos intermediários, onde as informações sofriam novos tratamentos. Isso provocava uma falha na integridade das informações

acarretando, muitas vezes, uma falta de credibilidade dos dados, uma queda da produtividade e a informação sendo publicada com valores diferentes.

Para resolver este problema, começou-se a estudar uma forma de se armazenar a informação contida nos sistemas transacionais numa base de dados central, para que houvesse integração total dos dados. Além disso, era necessário manter o histórico das informações e fazer com que ela fosse disposta por dimensões, ou seja, o analista de negócios poderia visualizar um mesmo fato através de diversas dimensões diferentes. O nome dado a essa modalidade de sistema de apoio à decisão foi o *Data Warehouse* (DW), ou em português, armazém de dados.

O termo DW surgiu como conceito acadêmico na década de 80, correspondendo basicamente a um grande repositório de dados com o objetivo de fornecer informações para tomada de decisão na esfera estratégica.

Em 1990, Bill Inmon ganhou o apelido "*pai do Data Warehouse*" apresentando o termo *Data Warehouse* na publicação *Building the Data Warehouse*. As empresas começaram, desde então a implantar a visão de Inmon, com graus variados de sucesso.

Segundo Taurion (1997), ao reunir informações dispersas nos diversos bancos de dados operacionais da empresa que podem estar em plataformas distintas, o DW permite que sejam feitas consultas e análises bastante eficazes, transformando dados esparsos em informações antes inacessíveis ou subaproveitadas. Essas informações podem ser convertidas em estratégias para os negócios.

Inmon (1994, 1997) apresenta a sua visão sobre a metodologia a adotar no desenvolvimento de DWs. Na terceira edição do seu trabalho, Inmon (2002) descreve uma arquitetura lógica para extrair os dados de BDs operacionais dispersos. Os dados são transformados e organizados temporalmente em um único BD.

A Figura 2.1 apresenta a visão geral, onde partes destes dados são então extraídos para BDs menores, criando BDs departamentais denominadas *Data Mart (DM*), de onde os utilizadores finais exploram os dados e criam relatórios. Para criar o DW e os Data Marts,

Inmon propõe uma metodologia *top-down*, partindo do geral para a pormenorização dos vários sistemas que o compõem.

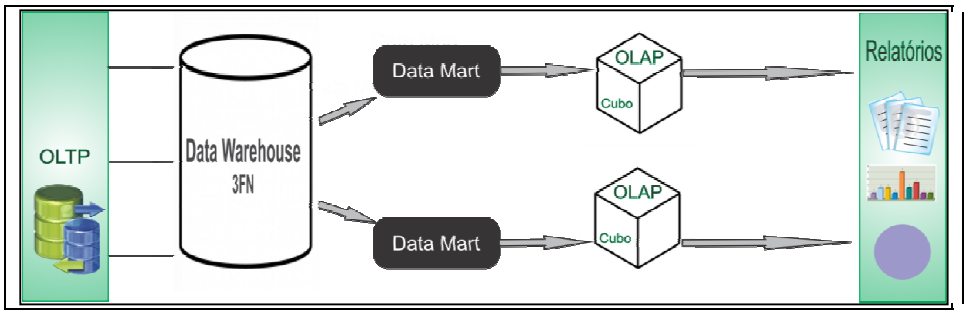


Figura 2.1 Visão de modelo segundo Inmon
(Adaptado de Kimball, 1998)

Depois da publicação do livro de Inmon, outros especialistas de BD começaram a criar DWs. A experiência de Ralph Kimball conduziu-o ao desenvolvimento de uma metodologia própria tendo, em 1998, publicado *The Data Warehouse Toolkit*.

Depois de vários anos de experiência, Kimball (2002) publicou uma segunda edição da sua obra, recomendando nesta versão uma arquitetura de múltiplos BDs e Data Marts, organizadas por áreas de negócio, em que os Data Marts têm que aderir a um canal de comunicação comum denominado Data Warehouse Bus (DWB).

Nesta versão, o DW é definido como sendo a soma dos vários Data Marts. Para o desenvolvimento é recomendada uma metodologia inversa à de Inmon, uma aproximação bottom-up, que parte da análise dos vários sistemas individuais terminando com a agregação dos mesmos num grande DW. Assim, os dados mantidos por uma empresa são chamados de "operacionais" ou "primitivos" conforme apresentado na Figura 2.2.

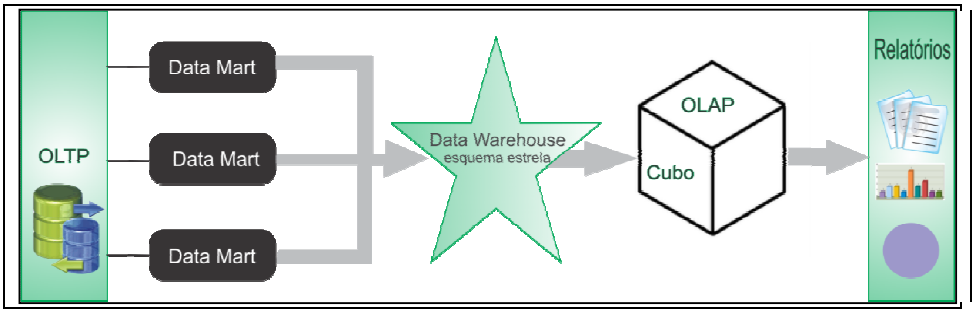


Figura 2.2 Visão do modelo segundo Kimball.
(Adaptado de Kimball, 1998)

Essa diferença de abordagem é mais relativa à terminologia utilizada do que propriamente conceitual. Observa-se que diversas discussões são a respeito de semântica dos dados.

Neste contexto, a capacidade das organizações em identificar, capturar e explorar os seus repositórios de conhecimento de forma a criar valor para o seu negócio é um fator crítico para garantir a competitividade, pois - de acordo com a semiologia - a palavra é um receptáculo de significados conferidos individualmente por cada pessoa. Mesmo que as significações possam ser plurais - e um caos de comunicação ocorra -, há significados compartilhados e comuns a todos.

Discussões são travadas em torno de assuntos recorrentes e ações são executadas sem chegar a resultados - a última ocorre muitas vezes quando as organizações gastam muito dinheiro para manter um banco de dados que não transfere qualquer tipo de informação relevante. Podemos enfim, apontar que os conceitos finais organizados hierarquicamente podem ser aplicados na construção do modelo de dados dimensional sobre uma visão conceitual analítica.

Nas duas visões apresentadas, percebe-se o termo *Online Analytical Processing* (OLAP), que foi citado pela primeira vez por E.F.Codd (2006), quando ele definiu regras que estas aplicações deveriam atender. A visão conceitual multidimensional dos negócios de uma empresa foi umas das regras citadas, a qual se tornou a característica fundamental no desenvolvimento destas aplicações.

Observa-se que as aplicações OLAP diferem das aplicações operacionais chamadas de *Online Transaction Processing* (OLTP) no que se refere aos requisitos funcionais e de desempenho, conforme apresentado na Tabela 3:

Tabela 2.1 Características que diferem as aplicações em OLAP e OLTP.

Características	**OLTP**	**OLAP**
Operação típica	Atualização (update)	Consulta - Análise
Interfase	Imutável	Redefinida
Nível de dados	Atomizado	Altamente sumarizado
Idade dos dados	Presente	Histórico, atual e projetado
Recuperação	Poucos registros	Muitos registros
Orientação	Registros	Arrays
Modelagem	Processo	Assunto

Portando, uma modelagem OLAP é mais do que uma aplicação, é uma solução de ambiente, integração e modelagem de dados. A maioria dos dados de uma aplicação OLAP, é originária de outros sistemas OLTP e armazém de dados transacionais.

2.2.1 Modelagem multidimensional

A modelagem multidimensional representa a principal técnica para atender às necessidades exigidas em ambientes convencionais de BI. Os elementos básicos dessas estruturas são os "cubos multidimensionais" (ou cubo de dados), que são fisicamente *arrays* multidimensionais usados para facilitar o processamento das operações de matemática nas medidas que estão contidos dentro deles.

Tanto Kimball (1998, 2001) como Inmon (1992, 1998, 2005) observam um cubo multidimensional como uma forma materializada de dados que apresenta em suas arestas as “dimensões”, e onde as métricas unitárias estão representadas em cada elemento deste cubo, alocando os valores unitários de métricas em cada elemento desse cubo. A Figura 2.3 ilustra uma representação visual de um cubo multidimensional, formado por diversos cuboides, tendo como dimensões os conceitos de “Região”, “Produto” e “Mês”:

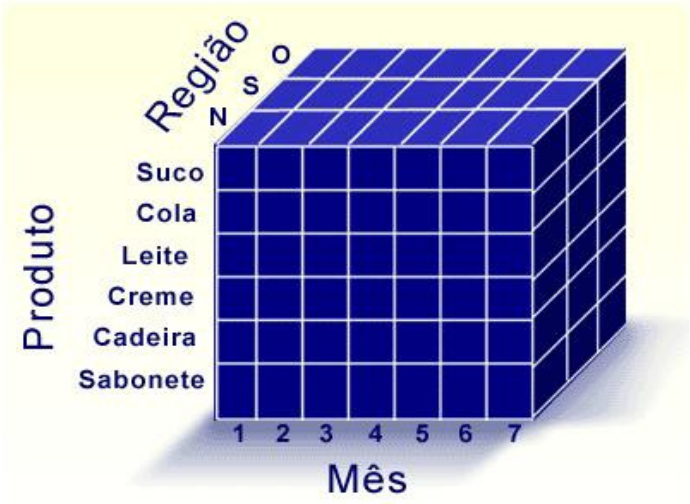


Figura 2.3 Modelo de um Cubo Muldidimensional.
(Siciliano, 2012)

No sentido de se compreender os princípios de modelagem multidimensional é necessária a definição de alguns dos conceitos envolvidos (Anzanello, 2005):

- Cubo é uma estrutura que armazena os dados em formato multidimensional.
- Dimensão é uma unidade de análise que agrupa dados de negócio relacionados. As dimensões se tornam cabeçalho de colunas e linhas.
- Hierarquia é composta por todos os níveis de uma dimensão, podendo ser balanceada (os número de níveis são equivalentes) ou não.
- Membro é um subconjunto de uma dimensão. Cada nível hierárquico tem membros apropriados aquele nível.
- Medidas (ou métricas) são os valores que são fatorados e apresentados.

Aguns desses conceitos são demonstrados na Figura 2.4 exemplificando os elementos do cubo:

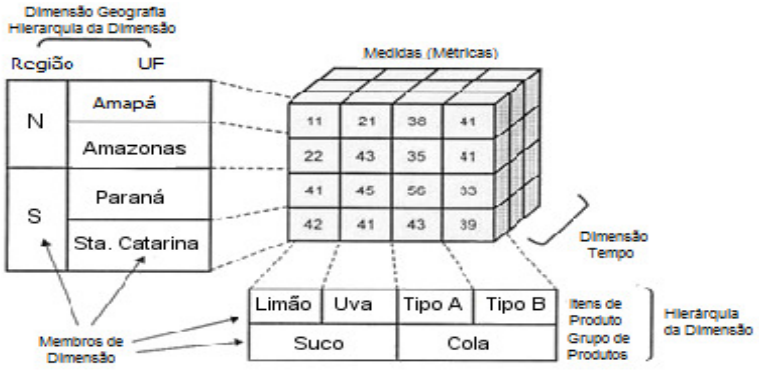


Figura 2.4 Elementos do Modelo Multidimensional.
(Adaptado de Kimball, 1998)

A cada uma das possíveis combinações de dimensões dá se a designação de "cubóide". A computação dos cubóides pode ser total - com todos os cuboides - ou parcial - apenas alguns -, e se concretizar através de uma função de agregação nas medidas (Delis, 1999). Cabe ao tratamento das dimensões filtrar, agrupar e organizar as informações desejadas, segundo as questões gerenciais apresentadas.

As medidas do modelo multidimensional são agregadas conforme são realizadas funções sob as dimensões, funções essas denominadas como operações multidimensionais. Segundo Kimball (1998) e Inmon (1998), as aplicações de BI devem permitir que modelos multidimensionais realizem algumas operações multidimensionais específicas, tais como:

- *Slice* - Extração de informação sumariada (agregada) segundo um valor de dimensão a partir de um cubo de dados.
- *Dice* - Extração de um cubóide ou interseção de vários *slices*. Esta extração verifica as restrições de valor ao longo de várias dimensões.
- *Pivot* - Troca de linhas e colunas numa tabela (*crosstab*) para ajustar a forma como é apresentado o resultado.
- *Drill-up* - Apresentação de dados num nível de abstração superior.
- *Drill-down* - Apresentação de dados num nível de abstração mais específico.
- *Drill-across* - Detalha vários cubóides com dimensões compartilhas, por

desagregação ao longo de um nível específico.

- *Drill-through* - Detalha os valores, ao longo de uma dimensão dada, além do nível mais baixo do cubo, por consultas SQL diretamente na fonte relacional.

- *Ranking* (ou *Rank*) - Ordenação dos membros de uma dimensão de acordo com a ordem de uma das medidas.

O modelo multidimensional possui dois elementos básicos: dimensões e fato. “Fato” é uma coleção de dados implementados sobre tabelas que representam um assunto, sendo composto por dados de medida (quantificadores), e informações do contexto aos quais os dados estão associados (qualificadores) que são discriminados dentro das “dimensões”.

Segundo Kimball (2001) e Inmon (2005), existem dois esquemas lógicos para a implementação dos esquemas lógicos do "fato" e das "dimensões" no suporte às representações de modelos multidimensionais, que são:

- Esquema Estrela (*Star Schema*), criado por Kimball (2001), que propõe uma visão cuja principal característica é a presença de dados altamente redundantes. É chamado de estrela porque a tabela de fatos fica ao centro com várias tabelas de dimensões que não tem outro relacionamento nas suas pontas. Neste modelo, as tabelas de dimensão não são normalizadas visando garantir melhores performance (Figura 2.5).

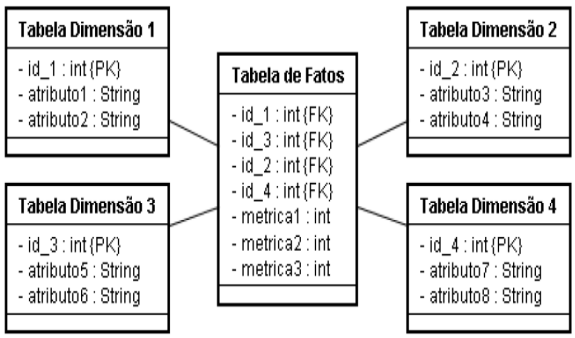


Figura 2.5 Esquema estrela (*star-schema model*)
(Adaptado de Kimball, 1998)

- Esquema Floco de Neve (*Snow Flake*), apresentado na Figura 2.6, cujas tabelas dimensionais relacionam-se com a tabela de "fatos" e com outras tabelas

dimensionais, que são representações da normalização das dimensões principais em diversos níveis de agrupamento. Este esquema tem como objetivo a normalização das tabelas dimensionais para diminuir assim o espaço ocupado por elas (Inmon, 2005).

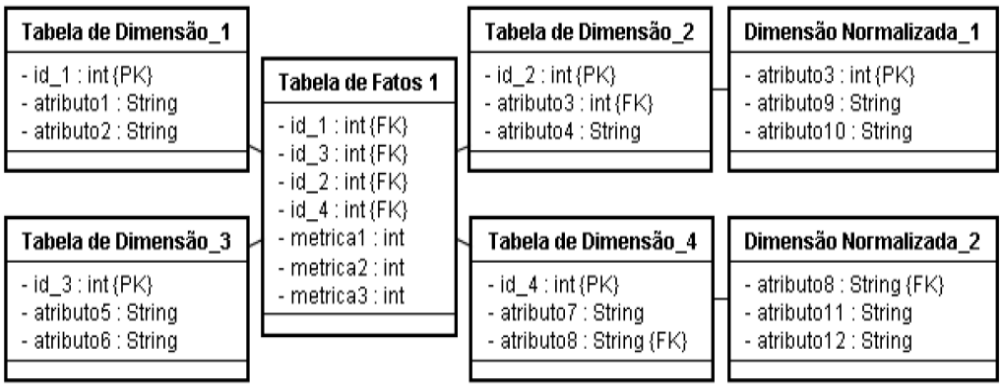


Figura 2.6 Esquema floco-de-neve (snow-flake model).
(Adaptado de Kimball, 1998)

Chaudhuri e Dayal (1997) reportam que os esquemas flocos de neve são um refinamento de esquemas estrela, onde a hierarquia dimensional é explicitamente representado através da normalização das tabelas de dimensão. Basicamente, no modelo estrela, todas as tabelas de dimensões são diretamente relacionadas com a "tabela de fato", enquanto no modelo floco de neve, as tabelas de dimensões formam hierarquias ligadas à "tabela de fato".

Machado (2007) define a modelagem multidimensional como uma técnica de concepção e visualização de modelos de dados de um conjunto de medidas que descrevem aspectos comuns de negócios. E esse modelo é formado por elementos básicos: as dimensões (qualificadores) e as medidas (quantificadores) agrupadas em contextos específicos (tuplas ou registros), chamados de "tabelas fato". Neste sentido:

- As dimensões determinam o contexto do assunto / fato do negócio. Possui uma ou mais hierarquias naturais além de atributos descritivos sem relacionamentos hierárquicos.

- As medidas são atributos numéricos que quantificam um fato e que são tratadas em conjunto ao contexto e as dimensões que participam do "fato". Elas são fatoradas

segundo suas categorias: medidas algébricas a partir de operações algébricas de agregação sobre dados atômicos ou por medidas distributivas / algébricas (média, desvio-padrão, etc.); e por medidas holísticas que armazenam agregados específicos (mediana e ranking).

Neste sentido, tratam especialmente de dois aspectos informacionais: as dimensões, através das quais os conceitos qualificam a informação, e as medidas são representadas por resultados fatorados dos dados através de operações algébricas.

De modo geral, os conceitos informacionais dentro das dimensões são estruturados nos modelos multidimensionais usando árvores enraizadas, que organizam os conceitos através de relações "gênero-espécie", uma representação de conhecimento chamado de taxonomia (Guarino, 1996).

2.3 MINERAÇÃO DE DADOS

O processo de descobrir padrões em dados é conhecido como Mineração de Dados (*Data Mining*, em inglês). Em tese, o processo deve ser semiautomático, isto porque é indispensável a interação com o usuário, que participará do processo desde a definição dos dados a serem analisados, até a análise do conhecimento gerado, de maneira a verificar se este é realmente útil e previamente desconhecido. Ainda assim, o processo semiautomático de mineração de dados visa extrair, de grandes bases de dados, sem nenhuma formulação prévia de hipóteses, informações desconhecidas, válidas e acionáveis, úteis para a tomada de decisão.

2.3.1 Conceitos e Princípios

Conforme descritos, os rápidos avanços na tecnologia de coleta e armazenamento de dados permitiram que as organizações acumulassem vasta quantidade de dados. A extração de informação útil, entretanto, tem provado ser extremamente desafiadora. Muitas vezes, ferramentas e técnicas tradicionais de análise de dados não podem ser usadas devido ao enorme conjunto dos dados. Às vezes a natureza não trivial dos dados significa que abordagens tradicionais não podem ser aplicadas, mesmo se o conjunto de dados for

relativamente pequeno. Em outras situações, questões que precisam ser respondidas não podem ser abordadas usando-se as técnicas existentes para análise de dados e, assim, novos métodos precisam ser desenvolvidos (Tan, Steinbach e Kumar, 2009).

A convergência da informática e da comunicação tem produzido uma sociedade que se alimenta de informações. No entanto, a maior parte da informação está em sua forma bruta: os dados. O dado sozinho não levará a compreender determinada situação, por isto é necessário trabalhá-lo, contextualizá-lo, transformá-lo em informação. Já dizia Naisbitt (1982), nós estamos afogados em dados, mas famintos por informação. Há uma quantidade enorme de informações ocultas em bancos de dados, que são potencialmente importantes, mas que ainda não foram descobertas.

O rápido crescimento de dados, coletados e armazenados das mais diversas formas, tem gerado gigantescos repositórios, que por sua vez excedeu em muito a capacidade humana de compreensão, sem ferramentas adequadas. Como resultado, os grandes bancos de dados se tornam "túmulos de dados" - arquivos de dados que raramente são visitados (Han & Kamer, 2006). Consequentemente, importantes decisões são frequentemente tomadas baseadas apenas na intuição dos gestores, sem o apoio de informações necessárias, simplesmente porque não se tem as ferramentas para extrair valiosos conhecimentos embutidos na vasta quantidade de dados.

Segundo Witten e Frank (2005), Mineração de Dados é um processo de extração de informação implícita, previamente desconhecida, e potencialmente útil a partir de dados brutos. A ideia é construir programas de computadores que vasculham automaticamente gigantescos bancos de dados em busca de correlações e padrões. Padrões este que, se encontrados, provavelmente possibilitarão fazer precisões sobre eventos futuros. Chang & Hsu (2005) acrescentam dizendo que os padrões descobertos devem ser válidos e compreensíveis.

Mineração de Dados é um campo jovem e promissor voltado a descobrir informações e conhecimentos (Han, 2011). Nos últimos anos, Mineração de Dados tem atraído uma grande atenção das organizações e da sociedade como todo, devido à grande disponibilidade de dados e a necessidade iminente de transformar dados em informações

úteis e conhecimento. A informação e o conhecimento adquirido podem ser utilizados nas mais diversas aplicações que vão desde análise de mercado, retenção de clientes, detecção de fraudes, controle de produção e exploração científica.

As técnicas de Mineração de Dados podem ser aplicadas em diversas áreas do conhecimento, dentre elas na Gestão de Pessoas, que por sua vez, é o objeto do estudo de caso deste trabalho. A sua principal característica é a aplicação dos algoritmos aos dados pré-processados, com o objetivo de auxiliar as organizações a gerar indicadores numéricos, indicadores gráficos e relatórios *ad hoc*, i.e., relatórios onde o analista define o que deseja obter no momento da consulta, através de aplicações que possam servir de apoio à tomada de decisão nos diferentes níveis, sejam eles estratégicos, táticos ou operacionais.

2.3.2 Aprendizado Indutivo

A indução é um meio de inferência lógica que permite que conclusões gerais sejam obtidas de exemplos particulares. É caracterizada como o raciocínio que parte do específico para o geral, do particular para o universal, da parte para o todo.

De acordo com (Batista, 2003), um argumento indutivo e correto pode, perfeitamente, admitir uma conclusão falsa, ainda que suas premissas sejam verdadeiras. Se as premissas de um argumento indutivo são verdadeiras, o melhor que pode ser dito é que a sua conclusão é provavelmente verdadeira. Desta forma, esse recurso deve ser utilizado com os devidos cuidados, dado que se o número de observações for insuficiente ou se os dados relevantes forem mal escolhidos, as hipóteses induzidas poderão produzir conclusões errôneas. Apesar disso, a inferência indutiva é um dos principais meios de criar novos conhecimentos e prever eventos futuros.

A Mineração de Dados compreende dois tipos de aprendizado indutivo: Supervisionado e Não Supervisionado. O aprendizado Supervisionado é direcionado a tomada de decisão e é através dele onde se realiza inferências nos dados com o intuito de realizar predições, envolvendo o uso dos atributos para prever o valor futuro. Enquanto que no Aprendizado Não-Supervisionado as atividades são descritivas, o que permite a descoberta de padrões e novos conhecimentos.

2.3.2.1 Aprendizado Supervisionado

O aprendizado supervisionado serve para identificar a classe a que pertence uma nova amostra de dados. Neste tipo de aprendizado é sempre conhecida a classe dos dados que são usados para treino e há um histórico de dados que permite prever sobre dados futuros.

Inicialmente é fornecido ao sistema de aprendizado um conjunto de exemplos $E = \{E_1, E_2,..., E_N\}$, onde cada exemplo $E_i \in E$ possui um rótulo associado. Esse rótulo define a classe a qual o exemplo pertence. Formalmente, cada exemplo $E_i \in E$ corresponde a uma tupla $E_i = (\overrightarrow{x_l}, y_i)$. Sendo $\overrightarrow{x_l}$ um vetor de valores que representam as características (atributos) do exemplo E_i e y_i o valor da classe deste exemplo. O objetivo do aprendizado supervisionado é induzir um mapeamento geral dos vetores $\overrightarrow{x_l}$ para valores y. Portanto, o sistema de aprendizado deve construir um modelo, tal que $y = f(\overrightarrow{x_l})$, onde f é uma função desconhecida (função conceito) que permite predizer valores y.

2.3.2.2 Aprendizado Não Supervisionado

Neste tipo de aprendizado o rótulo da classe de cada amostra de treino não é conhecido e o número de classes a ser treinada pode não ser conhecido a priori. É fornecido ao sistema de aprendizado um conjunto de exemplos E, no qual cada exemplo consiste somente de vetores $\overrightarrow{x_l}$, não incluindo a informação sobre a classe y. O objetivo é construir um modelo que procura por regularidades nos exemplos, formando agrupamentos ou clusters de exemplos com características similares.

O aprendizado não supervisionado utiliza-se de algoritmos descritivos. As atividades descritivas trabalham com conjuntos de dados que não possuem uma classe determinada e têm o objetivo de identificar padrões de comportamento semelhantes nestes dados. As tarefas descritivas podem ser divididas em: Associação, Agrupamento e Generalização.

2.3.3 Principais Tarefas de Mineração de Dados

As tarefas de Mineração de Dados são geralmente divididas em duas categorias principais de acordo com sua natureza (Dunham, 2003): tarefas de previsão e tarefas descritivas (Figura 2.7).

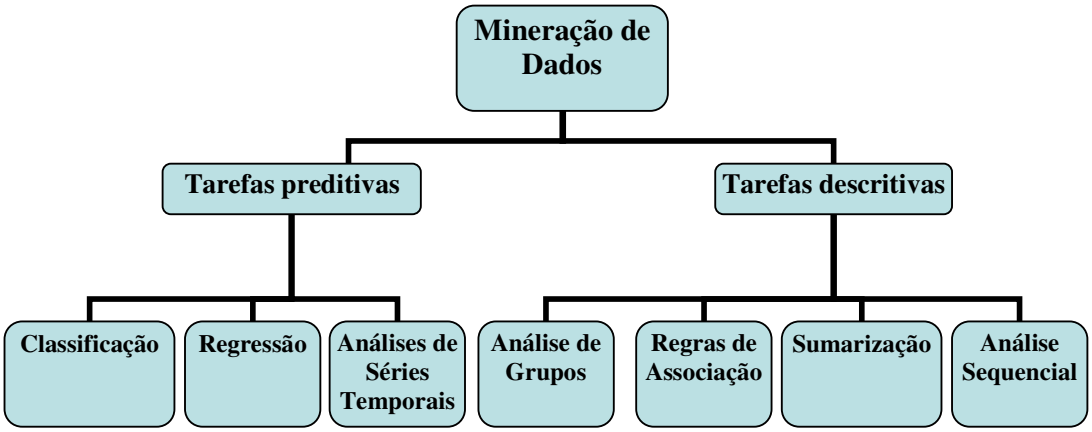


Figura 2.7 Tarefas e modelos de Data Mining

Tarefas de previsão tem o objetivo de prever o valor de um determinado atributo baseado nos valores de outros atributos. Já as tarefas de descrição objetivam identificar padrões ou relacionamentos nos dados. Ao contrário da modelo preditivo, o modelo descritivo se presta a explorar as propriedades dos dados examinados, sem previsão de novas propriedades. As tarefas descritivas analisam eventos passados em buscas de *insight* para tratar eventos futuros, enquanto que tarefas preditivas analisam os dados para determinar o provável resultado de eventos futuros ou a probabilidade de uma situação ocorrem.

Nas tarefas de previsão, o atributo a ser previsto é comumente conhecido como a **variável dependente** ou **alvo**, enquanto que os atributos usados para fazer a previsão são conhecidos como as **variáveis independentes** ou **explicativas**. Sendo assim, a modelagem de previsão se refere à tarefa de construir um modelo para a variável alvo como uma função das variáveis explicativas (Tan, Steinbach e Kumar, 2009).

Tarefas de previsão include os métodos de classificação, regressão, análises de série temporal, enquanto que as tarefas de descrição envolvem os métodos de agrupamento, sumarização, regras de associação e análises sequenciais.

Entre as tarefas preditivas, Classificação é provavelmente a abordagem melhor entendida e mais utilizada. Tarefas de classificação possuem três características em comum:

- Aprendizagem supervisionada;

- A variável dependente é discreta;
- E o modelo construído é capaz de atribuir a novos dados uma das classes pré-definidas.

Classificação é o processo de definir um modelo (ou função) que descreve e distingue classe ou conceitos de dados, com o propósito de ser capaz de usar este modelo para prever classe de objetos cuja classificação é desconhecida. O modelo é construído a partir de um conjunto de dados (*data training*) cuja classificação é conhecida (Han e Kamber, 2006).

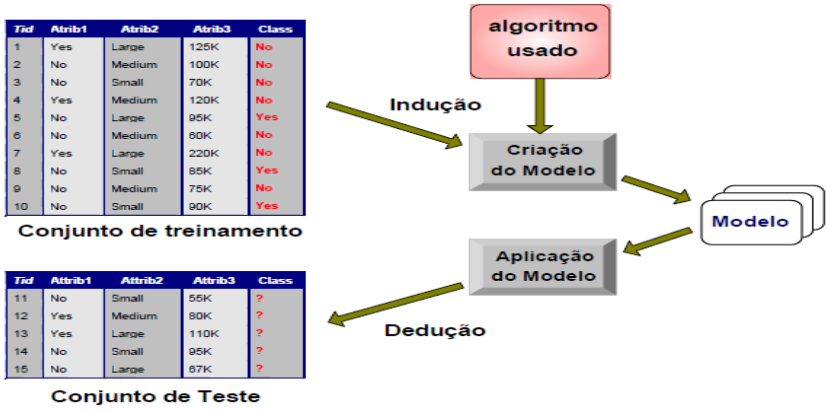


Figura 2.8 Abordagem geral para construção de um modelo de classificação.
(Tan, P., Steinbach, M., Kumar, V., 2009.)

A Figura 2.8 mostra uma abordagem geral para resolver problemas de classificação. Primeiro, um conjunto de treinamento consistindo de registros rotulados devem ser fornecidos. Este conjunto é usado para construir um modelo de classificação, que é subsequentemente aplicado ao conjunto de teste, que consiste de registros com rótulos de classes desconhecidos.

O modelo construído pode ser representado de várias formas, tais como regras de classificação (IF-THEN), árvores de decisão, fórmulas matemáticas ou redes neurais.

Por exemplo, dadas as classes de pacientes que correspondem a um determinado tratamento, identificar o melhor tratamento para um novo paciente (Stephens e Pablo, 2003). Outro exemplo seria prever se um usuário Web fará uma compra em uma livraria online, onde a variável alvo é de valor binário.

Diferente da classificação, que prevê rótulos discretos e não ordenados, a regressão é uma técnica estatística supervisionada usada para prever variáveis alvo contínuas (numéricas). Por exemplo, prever quanto um usuário Web irá comprar numa loja virtual.

De acordo com Han e Kamber (2006), classificação e predição podem ser precedidas de análise de relevância, que tenta identificar atributos que não contribuam para no processo. Estes atributos podem então ser excluídos. Chang (2009) descreve alguns métodos de seleção de atributos a fim de analisar os fatores para encontrar o melhor classificador para *turnover* de empregados.

Segundo Tan, Steinbach e Kumar (2009), o objetivo de ambas as tarefas (classificação e regressão) é aprender um modelo que minimize o erro entre os valores previsto e real da variável alvo. Em outras palavras, o objetivo é fazer um bom, mas não perfeito, trabalho de previsão.

A análise de séries temporais é outra técnica preditiva geralmente utilizada para prever resultados numéricos dependentes do tempo (Roiger e Geatz, 2003). Uma série temporal pode ser definida como um conjunto de observações de uma variável dispostas sequencialmente no tempo (Shumway e Stoffer, 2011). Podemos enumerar os seguintes exemplos de séries temporais: temperaturas máximas e mínimas diárias em uma cidade, vendas mensais de uma empresa, valores mensais do IPC-A, valores de fechamento diários do IBOVESPA, resultado de um eletroencefalograma, gráfico de controle de um processo produtivo. O objetivo da análise de séries temporais é identificar padrões não aleatórios na série temporal de uma variável de interesse, e a observação deste comportamento passado pode permitir fazer previsões sobre o futuro, orientando a tomada de decisões.

As tarefas descritivas são normalmente utilizadas na geração de frequências, análise cruzada e correlação. Métodos descritivos podem ser definidos para descobrir relações interessantes entre os dados, encontrar padrões e agrupamentos interessantes na massa de dados (Marco e Gianluca, 2005).

Segundo Dunham (2005), Sumarização é um método descritivo que mapeia dados em subconjuntos com associações descritivas simples. Esta abordagem usa técnicas básicas de

estatística, tais como, média, moda, mediana, desvio padrão e variância para resumir os dados.

De acordo com Tan, Steinbach e Kumar (2009), Agrupamento ou Análise de grupos ou clusterização (*clustering*) consiste de uma abordagem descritiva que agrupa objetos baseado apenas em informações encontradas nos dados que descrevem os objetos e seus relacionamentos. O objetivo é que os objetos dentro de um grupo, ou *cluster*, sejam semelhantes (ou relacionados) entre si e diferentes de (ou não relacionados aos) outros objetos de outros grupos. Tan, Steinbach e Kumar (2009) descrevem a Análise de grupos como sendo uma classificação não supervisionada. Ao contrário da classificação que possui classes pré-definidas, a análise de grupos cria uma rotulagem de objetos baseado apenas nos dados, ou seja, as classes são extraídas dos próprios dados. Han e Kamber (2006) descrevem que os objetos são agrupados com o princípio de maximizar a similaridade intraclasse e minimizar a semelhança interclasse. Estes *clusters* descobertos podem ser usados para explicar as características da distribuição dos dados subjacentes e assim servir como base para várias técnicas de análise e mineração de dados. As aplicações de clusterização incluem caracterização de diferentes grupos de clientes baseado nos padrões de compra, categorização de documentos na World Wide Web, agrupamento de genes e proteínas que possuem funcionalidades similares, agrupamento de localizações geográficas propensas a terremotos através de dados sismológicos. Farajian e Mohammadi (2011) descrevem a aplicação deste método para descrever padrões no comportamento de clientes de um banco.

Outro método descritivo são as Regras de Associação. Este método é usado para descobrir relacionamentos frequentes entre atributos e itens, isto é, encontrar conjuntos de itens que aparecem frequentemente juntos em uma transação. Tan, Steinbach e Kumar (2009) definem regra de associação como sendo uma expressão de implicação no formato $X \rightarrow Y$, onde X e Y são conjuntos disjuntos de itens ($X \cap Y = \varnothing$). O uso deste método foi empregado por Silva, Stopanovski, Rocha e Cosac (2008) para descobrir fraudes no uso dos cartões de pagamento do Governo Federal.

Por fim, o método de Análise Sequencial é mais um método descritivo que consiste em uma especialização do método de Regras de Associação. Este método é utilizado para

minerar padrões sequenciais nos dados (Dunham, 2003). Nesta abordagem a ordem dos itens é de total importância e o objetivo é encontrar os itens que costumam aparecer na base após o aparecimento de outros.

2.3.4 Mineração de Dados na Gestão de Pessoas

Durante os últimos anos é crescente o número de pesquisas que procuraram adotar, de forma prática, Mineração de Dados (DM) para suportar tomadas de decisões na área de Gestão de Pessoas. Somente na última década a área de RH vem adotando práticas de DM de forma séria (Wilkerson, 2012).

As contribuições abrangem as diversas atividades e processos de RH, tais como: seleção de empregados (Aiolli, Filippo e Sperduti, 2009) ou previsão de rotatividade de pessoal (Chang, 2009); averiguação de competências (Zhu, Goncalves, Uren, Motta e Pacheco, 2005) ou previsão (Thissen-Roe, 2005) e avaliação (Zhao, 2008) de desempenho. Para prover estas funcionalidades, vários abordagens e métodos são empregados, tais como árvores de decisão (Sivaram e Ramar, 2010), análise de grupos (Karahoca, 2008), análise de associação (Danping e Jin, 2011), máquina de vetor de suporte (Li, Xu e Meng, 2009) ou redes neurais (Ning, 2010). A maioria das pesquisas de Mineração de Dados na área de RH é voltada para o quadro de pessoal e especialmente a seleção de empregados é considerada um domínio relevante que deve ser apoiada pela Mineração de Dados (Piazza e Stronmeier, 2011). A justificativa é usualmente baseada na quantidade elevada de dados produzidos, por exemplo, pelo prognóstico de desempenho de empregados (Cho e Ngai, 2003) ou a redução e seleção de atributos relevantes (Wang, Li e Hu, 2009) que são valiosos no suporte de decisões.

Além disso, a modelagem preditiva oferece às organizações uma oportunidade de agir de forma proativa com base no histórico de atividades de seus empregados, antes que eventos aconteçam. Através da Mineração de Dados, as organizações podem, por exemplo, prever com 85% de precisão quais empregados podem se desligar do emprego. Assim, uma organização pode usar esta informação para planejar a alocação de recursos ou capacitação dos empregados que permanecerão na empresa.

O uso de dados não somente aumenta a eficiência das empresas, mas também serve para verificar os efeitos positivos de outros fatores. Kennedy (2003) considerou o uso dos dados como um componente central para que seu modelo de negócio alcançasse os objetivos definidos de maneira mais eficiente.

2.3.5 Processo de descoberta de conhecimento

Segundo Fayyad (1996), o termo *Knowledge Discovery in Databases* ou KDD foi criado em 1989 como referência ao processo amplo de encontrar conhecimento em dados. KDD refere-se a todo processo de descoberta de conhecimento útil de dados, enquanto Mineração de Dados refere-se à aplicação de algoritmos para extrair modelos dos dados.

O processo de KDD é um conjunto de atividades iterativas e contínuas que compartilham o conhecimento descoberto a partir de bases de dados. De acordo com Fayyad (1996), esse conjunto é composto de cinco etapas (Figura 2.9), que são:

- Seleção e Definiç0ão do problema;
- Integração e limpeza dos dados;
- Transformação dos dados;
- *Data Mining* ou Mineração de Dados*;*
- Interpretação e Avaliação dos resultados.

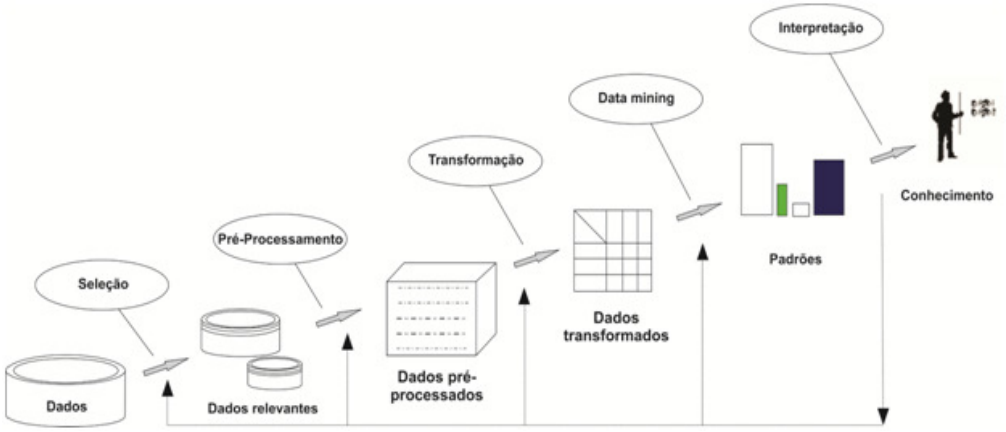


Figura 2.9 Processo de Descoberta do Conhecimento
(Adaptado de Fayyad, 1996)

O processo KDD começa com o entendimento do domínio do problema, dos objetivos finais a serem atingidos e seleção dos dados relevantes para o problema em questão. A etapa da limpeza dos dados e integração (*data cleaning e data integration*) vem a seguir, através de um **pré-processamento** dos dados, fazendo a integração de dados heterogêneos,

eliminação de incompletude dos dados e outras. Segundo Mannila (1996) essa etapa pode tomar 80% do tempo necessário de todo o processo.

A etapa de transformação tem o propósito de adequar de dados para serem utilizado pelo algoritmo utilizado na etapa de mineração de dados.

Tem-se, então, a etapa de DM, que começa com a escolha dos algoritmos a serem utilizados. Essa escolha depende, fundamentalmente, do objetivo do processo de KDD (Witten & Frank, 2005) que pode ser: classificação, regressão, agrupamento, associação ou detecção de outliers. De modo geral, na etapa de DM, os algoritmos utilizados procuram por padrões nos dados.

Por fim tem-se a etapa de interpretação e validação dos resultados, também conhecida como **pós-processamento**. Esta etapa assegura que apenas resultados válidos e úteis sejam incorporados aos sistemas de apoio a decisões (Tan, Steinbach e Kumar, 2009).

2.3.6 Redução de dimensionalidade

Conjuntos de dados podem ter um grande número de características, porém nem todas elas precisam ser consideradas no processo de mineração. Para Witten e Frank (2009), o mundo real usualmente possuem atributos irrelevantes ou redundantes, que degradam a precisão dos algoritmos. Logo, se faz necessário o uso de técnicas para reduzir o número de atributos nos dados – redução de dimensionalidade.

Para Tan, Steinback e Kumar (2009) a redução de dimensionalidade traz diversos benefícios:

- Ajuda a reduzir o número de atributos irrelevantes e remover ruídos;
- Reduz a quantidade de tempo e memória utiliza pelos algoritmos de mineração;
- Facilita a visualização dos dados, uma vez que leva a um modelo mais compreensível.

O termo redução de dimensionalidade é muitas vezes reservado para as técnicas que reduzem a dimensionalidade de um conjunto de dados criando novos atributos que sejam uma combinação dos atributos antigos.

Muitos tipos de análise de dados se tornam significativamente mais difíceis quando a dimensionalidade dos dados aumenta. Segundo Tan, Steinback e Kumar (2009), quando a dimensionalidade aumenta, os dados se tornam cada vez mais dispersos no espaço que eles ocupam; o hipervolume do espaço cresce de forma exponencial com a adição de novos atributos. Os dados ficam muito esparsos o que prejudica o desempenho de algoritmos que operam fundamentalmente com base em medidas de distância. Para a classificação, isto significa que não há objetos de dados suficientes para permitir a criação de um modelo que atribua de forma confiável uma classe a todos os objetos possíveis. Para agrupamento, as definições de densidade e distâncias entre pontos, que são críticas para agrupamento, se tornam menos significativas. Como consequência, muitos algoritmos de agrupamento e classificação têm problemas com dados de alta dimensionalidade – exatidão de classificação e grupos de qualidade inferior.

Algumas das abordagens mais comuns para a redução de dimensionalidade, especialmente para dados contínuos, usam técnicas de álgebra linear para projetar os dados de um espaço de alta dimensionalidade para um de dimensionalidade menor (Tan, Steinback e Kumar, 2009). Uma técnica conhecida é a Análise de Componentes Principais (PCA) que é usada, por exemplo, em (da Costa, de Freitas, David, Amaral & de Sousa Jr, 2012) para a redução do problema da detecção de intrusões em redes de computadores.

Outra forma de reduzir a dimensionalidade é usar apenas um subconjunto das características – seleção de atributos. Embora possa parecer que tal abordagem perca informação, não é o caso se características redundantes e irrelevantes estiverem presentes. Características redundantes duplicam muitas ou todas as informações contidas em um ou mais atributos. Um exemplo encontrado no conjunto de dados deste trabalho foi: UF, Região e Subsistema. Características irrelevantes quase não contêm informações úteis para a tarefa de mineração de dados. Por exemplo, a matrícula do empregado é irrelevante para a tarefa de descrever o comportamento das transferências de unidades.

De acordo com Tan, Steinback e Kumar (2009), há três abordagens padrão para a seleção de características: interna, filtro e envoltório.

- Abordagens Internas: A seleção de características ocorre naturalmente como parte do algoritmo de mineração de dados. Especialmente, durante a operação de algoritmo de mineração, o próprio algoritmo decide quais atributos usar e quais ignorar. Algoritmos para construir classificadores de árvores de decisão muitas vezes operam desta maneira.

- Abordagens de Filtro: Características são selecionadas antes que o algoritmo de mineração seja executado, usando alguma abordagem que seja independente da tarefa de mineração de dados. Por exemplo, podem-se ignorar atributos que são derivados de outro atributo.

- Abordagens de Envoltório: Estes métodos usam o algoritmo de mineração de dados alvo como uma caixa preta para encontrar o melhor subconjunto de atributos, mas geralmente sem enumerar todos os subconjuntos possíveis.

2.3.7 Discretização e Binarização

Alguns algoritmos de Mineração de Dados, especialmente determinados algoritmos de classificação, requerem que os dados estejam na forma de atributos categorizados. Algoritmos que encontram padrões de associação requerem que os dados estejam na forma de atributos binários. Assim, muitas vezes é necessário transformar um atributo contínuo em um categorizado – discretização - e tanto os atributos contínuos quantos os discretos podem precisar ser transformados em um ou mais atributos binários – binarização (Tan, Steinbach e Kumar, 2009). Adicionalmente, se um atributo categorizado possuir um número grande de valores (categorias), ou se algum valor ocorra raramente, então pode ser benéfico para determinadas tarefas de mineração de dados reduzir o número de categorias combinando alguns dos valores.

Assim como a seleção de características, a melhor abordagem de discretização e binarização é a que produz o melhor resultado para o algoritmo que será usado para analisar dados (Tan, Steinbach e Kumar, 2009). A Tabela 1 ilustra um exemplo de binarização.

Tabela 2.2 Conversão de um atributo categorizado em três árvores binárias.
Fonte: Tan, P., Steinbach, M., Kumar, V., 2009.

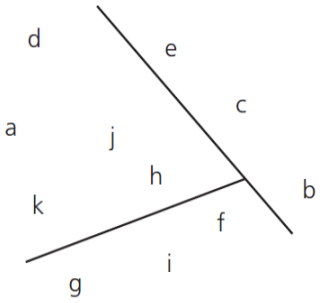
Valor categorizado	*Valor inteiro*	x_1	x_2	x_3
Terrível	0	0	0	0
Fraco	1	0	0	1
Satisfatório	2	0	1	0
Bom	3	0	1	1
Excelente	4	1	0	0

Na discretização de atributos contínuos o resultado pode ser representado como um conjunto de intervalos $\{[x_0,x_1], [x_1, x_2],... [x_{n-1}, x_n]\}$, onde x_0 e x_n podem ser $+ \infty$ ou $- \infty$, respectivamente ou, de forma equivalente, como um série de desigualdades $x_0 < x <= x_1$, ... $x_{n-1} < x < x_n$.

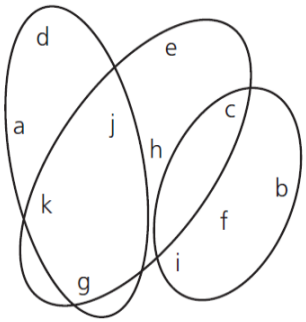
2.3.8 Algoritmos de Agrupamento

Técnicas de Agrupamento ou clusterização são aplicadas quando não há classe a ser predita, quando as instâncias são divididas em grupos naturais. Os grupos gerados refletem características comuns compartilhadas pelos objetos analisados. Existem diferentes maneiras de expressar os resultados objetivos com esta técnica. Os grupos que foram identificados podem ser exclusivos, ou seja, uma instância pertence a somente um grupo, Figura 2.10 (a). Ou pode haver sobreposições, instâncias pertencendo a mais de um grupo, Figura 2.10 (b) - Diagrama de Venn. Ou pode ser por probabilidade, em que uma instância pertence a cada grupo com certa probabilidade, Figura 2.10 (c). Também podem ser hierárquico, de tal forma que existe uma divisão em níveis, com grupos e subgrupos, Figura 2.10 (d). Entretanto, como esses mecanismos são raramente conhecidos, afinal, é algo que se tenta descobrir, a escolha é geralmente ditada pelas ferramentas de agrupamento que estão disponíveis (Tan, Steinbach e Kumar, 2009).

Segundo Witten e Frank (2009), a técnica de agrupamento é geralmente seguida por uma fase em que se inferem árvores de decisão ou regras de associação a fim de alocar cada instância a um determinado grupo. Ou seja, a operação de agrupamente é apenas um passo no caminha para uma descrição estrutural dos dados.



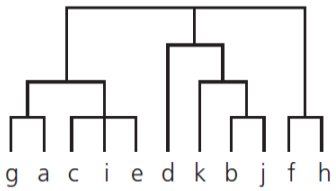
(a)



(b)

	1	2	3
a	0.4	0.1	0.5
b	0.1	0.8	0.1
c	0.3	0.3	0.4
d	0.1	0.1	0.8
e	0.4	0.2	0.4
f	0.1	0.4	0.5
g	0.7	0.2	0.1
h	0.5	0.4	0.1

(c)



(d)

Figura 2.10 Diferente forma de representa grupos.
(Tan, P., Steinbach, M., Kumar, V., 2009)

Existem diversas técnicas de clusterização, e cada uma possui suas vantagens e desvantagens. De acordo com Steinbach (2000), clusterização hierárquica é retratada como a técnica de clusterização de melhor qualidade, sendo limitada pela sua complexidade quadrática, já o K-means e suas variações possuem complexidade de tempo linear, mas produzem clusters de qualidade inferior. O algoritmo de clusterização K-means pode ser também chamado de K-médias. Segundo Jain (1999) o algoritmo K-means é popular devido a sua facilidade de implementação.

De acordo com Fontana e Naldi (2009), K-means utiliza o conceito de centróides como protótipos representativos dos grupos, onde o centróide representa o centro de um grupo, sendo calculado pela média de todos os objetos do grupo. Primeiramente é especificado o número de grupos que serão procurados – este é o parâmetro k. Então k pontos são escolhidos aleatoriamente como os centros dos grupos. Todas as instâncias são atribuídas ao centro mais próximo de acordo com alguma métrica que calcula a distância entre as instâncias. Em seguida, o centróide, ou média, de todas as instâncias em cada grupo é calculado. Estes centróides serão os novos centros dos seus respectivos grupos.

Finalmente, todo o processo é repetido com os novos centros. A iteração continua até que os mesmos centróides sejam atribuídos a cada grupo nas próximas rodadas, estabilizando os centros de cada grupo. A Figura 2.11 é ilustrado o processo K-Means.

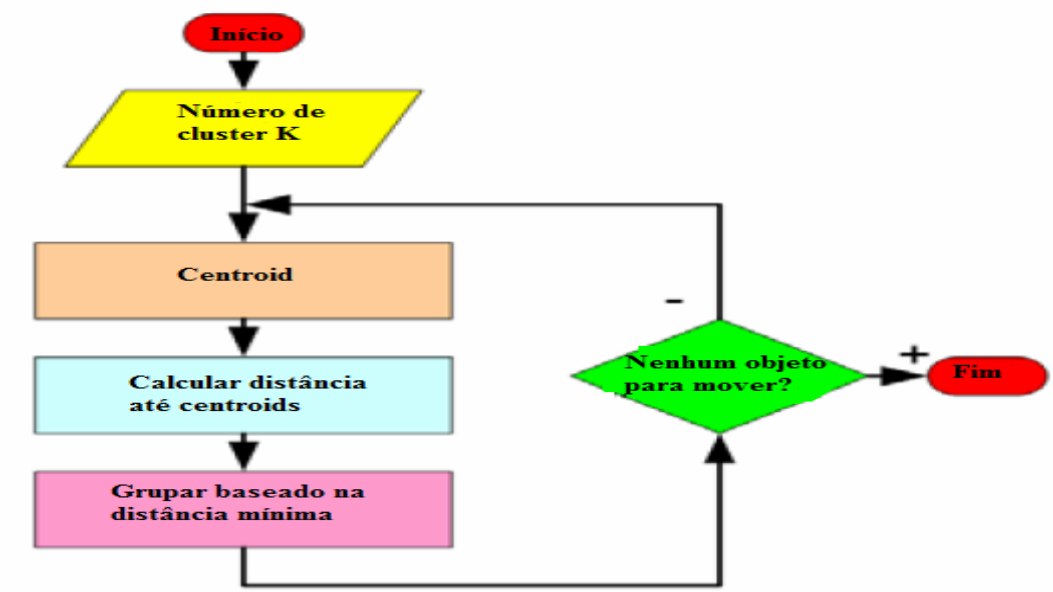


Figura 2.11 Processo K-Means
(Witten e Frank, 2009)

De acordo com Witten e Frank (2009), este processo é simples e eficiente. É fácil provar que a escolha do centro do grupo ser o centróide minimiza o quadrado da distância total (*total squared distance*) de cada um dos pontos do grupo ao seu centro. Uma vez que a iteração se estabiliza, cada ponto é atribuído ao seu centro mais próximo (*cluster*).

As medidas de distância de uma maneira geral podem ser definidas como medidas de similaridade, e dissimilaridade; na qual a primeira é para definir o grau de semelhança entre as instâncias e realizam o agrupamento de acordo com a sua coesão, e a segunda de acordo com as diferenças dos atributos das instâncias. Witten e Frank (2005) realizam uma consideração sobre a utilização das medidas de similaridade: em aprendizado baseado em instância ou exemplo, cada nova instância é comparada a uma instância existente usando métrica de distância, e a instância existente mais próxima é designada classe da nova. Este método é chamado de classificação de vizinho mais próximo.

Distância Euclidiana e distância *Manhattan* são duas métricas de similaridade bastante conhecidas. A distância Euclidiana ou simplesmente distância consiste da raiz quadrada das diferenças entre coordenadas de dois objetos, ou seja, a distâncias entre uma instâncias com valores $a_1^{(1)}$, $a_2^{(1)}$, ... $a_k^{(1)}$ (onde k é o número do atributo) e uma com valores $a_1^{(2)}$, $a_2^{(2)}$, ... $a_k^{(2)}$ é definido como:

$$\textbf{Distância Euclidiana: } \sqrt{((a_1^{(1)} - a_1^{(2)})^2 + \ldots + (a_k^{(1)} - a_k^{(2)})^2)}. \qquad (2.3)$$

Já a Distância *Manhattan* ou *city-block* tem uma definição mais simples na qual é apenas a soma das diferenças entre todos os atributos de dois dados x e y, conforme equação (2.4), não sendo indicada para os casos em que existe uma correlação entre tais atributos (Witten e Frank, 2005).

$$\textbf{Distância } \textit{Manhattan}\textbf{: } |a_1^{(1)} - a_1^{(2)}| + \ldots + |a_k^{(1)} - a_k^{(2)}|. \qquad (2.4)$$

Um dos problemas para a utilização de técnicas de agrupamento é a utilização de dados nominais em seus atributos, os quais por não ter uma métrica implícita dificultam o trabalho dos algoritmos em termos de atribuição de pesos e valores para formação dos clusters. Para este caso, Witten e Frank (2005) apontam a seguinte abordagem: Dado o atributo cor com valores vermelho, amarelo e azul. Usualmente a distância zero é atribuída se os valores são idênticos; caso contrário, a distância é um. Sendo assim, a distância entre vermelho e vermelho é zero, mas a distâncias entre vermelho e azul é um. No entanto, é desejável usar uma representação mais sofisticada dos atributos. Por exemplo, com mais cores pode-se usar uma medida numérica que cria uma escala, tornando amarelo mais para laranja do que é verde e ocre mais perto ainda. Alguns atributos serão mais importantes do que outros, e isso geralmente é refletido na distância métrica por algum tipo de ponderação atributo. Isto, porém, consiste um problema chave na aprendizagem baseada em exemplo, visto que requer atribuição adequada dos pesos de atributos.

2.3.9 Algoritmo de Regras de Associação

Regras de Associação é uma das muitas técnicas de mineração de dados que descrevem eventos que tendem a ocorrer juntos. O conceito de regras de associação pode ser entendido da seguinte forma: Seja I = {i1, i2 ... in} um conjunto de literais, chamados de

itens. Seja T uma transação com um conjunto de itens tal que $T \subseteq I$. Dado um banco de dados de transações D (sobre I), uma regra de associação é uma implicação da forma $X \rightarrow Y$, onde $X \subset I$, $Y \subset I$ e $X \cap Y = \varphi$. Dois conceitos importantes quando se tratar de regras de associação são confiança e suporte. A regra $X \rightarrow Y$ de um conjunto de transações D tem confiança ***c*** se ***c%*** das transações em X também contêm Y, ou seja, a confiança $(X \rightarrow Y)$ = (n º de tuplas contendo X e Y) / (n º de tuplas contendo X) = $P(X \mid Y) = P(X \cup Y) / P(X)$. A regra $X \rightarrow Y$ tem suporte ***s*** na transação do banco de dados D se ***s%*** das transações em D contêm $X \cup Y$, ou seja, o suporte $(X \rightarrow Y)$ = (n º de tuplas contendo ambos, X e Y) / (número total de tuplas) = $P(X \cup Y)$.

Segundo Tan, Steinbach e Kumar (2009), o suporte é uma medida importante porque uma regra que tenha baixo suporte pode acontecer simplesmente por coincidência. Uma regra de baixo suporte também possui grande probabilidade de não ter interesse a partir de uma perspectiva de negócio porque pode não ser lucrativo promover, por exemplo, itens que os clientes raramente compram juntos. A confiança, por outro lado, mede a confiabilidade da inferência feita por uma regra. Para uma determinada regra $X \rightarrow Y$, quanto maior a confiança, maior a probabilidade de que Y esteja presente em transações que contenham Y. A confiança também fornece uma estimativa de probabilidade condicional de Y dado X.

A mineração de regras de associação permite a descoberta de regras da forma $X \rightarrow Y$ e $X \& Y \rightarrow Z$ com suporte e confiança mínima. Segundo Witten e Frank (2005), o desafio é a seleção de algoritmos que podem ser aplicados para extrair regras de associação de um particular conjunto de dados. Para Han e Kamer (2006), outro problema enfrentado por qualquer algoritmo é o problema da dimensionalidade. O número de regras de associação possíveis cresce exponencialmente com o número de atributos. Se existem atributos k (considerando apenas atributos binários, como comprar *SmartPhone* = Sim), há na ordem de $\mathbf{k \cdot 2^{k-1}}$ regras de associação possíveis. Por exemplo, suponha que uma pequena loja tem apenas 100 itens diferentes, e um cliente poderia comprar ou não comprar qualquer combinação desses 100 itens. Depois, há 100×2^{99} possíveis regras de associação que esperam por um algoritmo de busca.

Segundo Tan, Steinbach e Kumar (2009), a medida de suporte auxilia a reduzir o número de conjuntos de itens candidatos explorados durante a geração de conjuntos de itens

frequentes. O uso de suporte para podar conjunto de itens candidatos é guiado pelo princípio a seguir: se um conjunto de itens é frequente, então todos os seus subconjuntos também devem ser frequentes. Este princípio é criado pelo algoritmo Apriori e foi proposto por R. Agrawal e R. Srikant em 1994 para a mineração de conjuntos de itens frequentes na forma de fortes regras de associação booleanas.

Um conjunto de itens frequente é um conjunto de transações que ocorre com um suporte mínimo especificado. Uma regra forte é aquele que satisfaz tanto suporte mínimo e confiança mínima. Algoritmo Apriori usa busca iterativa *level-wise*, onde *k-itemsets* (um conjunto de itens que contém itens k) são usados para explorar *k+1 itemsets*, para mineração de conjuntos de itens frequentes em banco de dados transacional de regras associação booleanas.

Para ilustrar a ideia por trás do princípio Apriori, dado o conjunto {a, b, c, d, e} e suas combinações, a Figura 2.12 demostra o Princípio Apriori. Se {c,d,e} é frequente, então todos os subconjuntos desde conjunto de itens são frequentes. Se o conjunto {c, d, e} for frequente, então todos os subconjuntos de {c, d, e}, isto é, os conjuntos de itens sombreados na Figura 2.12, também devem ser frequentes.

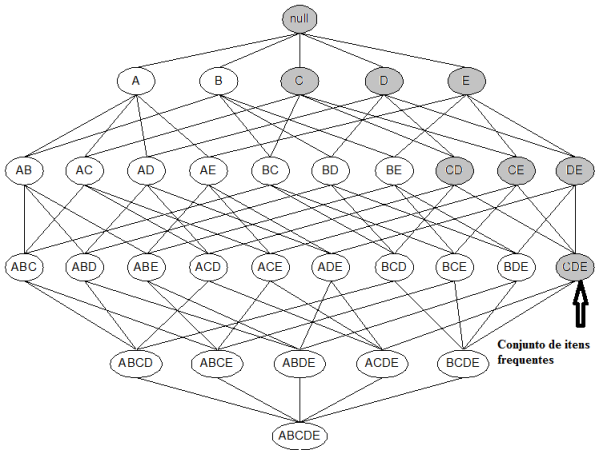


Figura 2.12 Princípio Apriori.
(Tan, P., Steinbach, M., Kumar, V., 2009)

De forma inversa, se um conjunto de itens como {a,b} for infrequente, então todos os seus superconjuntos deve ser infrequentes também. Esta estratégia de se diminuir o espaço de

pesquisa exponencial baseado na medida de suporte é conhecida como poda baseada em suporte. A Figura 2.13 ilustra esta idéia.

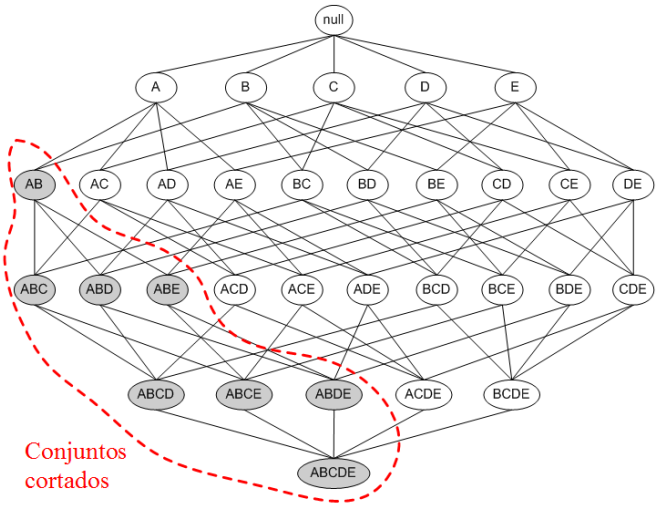


Figura 2.13 Podada baseada em suporte.
(Tan, P., Steinbach, M., Kumar, V., 2009)

O nome do algoritmo é baseado no fato de que o algoritmo usa o conhecimento prévio da frequência do conjunto de itens. A metodologia básica envolvida consiste em primeiro encontrar o conjunto de frequência de conjunto de itens onde k = 1. Este conjunto é chamado L1. L1 é então usada para localizar o conjunto de frequência de conjuntos de itens onde k=2, L2, que é por sua vez é usado para encontrar L3, e assim por diante, até que não haja mais k conjunto de itens frequente que possa ser encontrado. A Figura 2.14 ilustra este processo considderano o suporte mínmo igual a 40%.

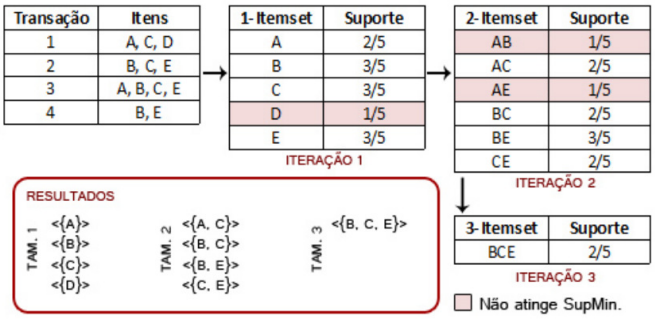


Figura 2.14 Algoritmo Apriori, considerando suporte mínimo igual a 40%.
(Tan, P., Steinbach, M., Kumar, V., 2009)

2.3.10 Algoritmo de Classificação – Árvore de Decisão

Dentre os métodos de classificação, a Árvore de Decisão é um dos mais conhecidos e utilizados (Han e Kamer, 2006). Algoritmos de Árvore de Decisão, tais como ID3, C4.5 e CART, foram originalmente destinado para classificação, no entanto, trata-se de um modelo que é simultaneamente preditivo e descritivo. O seu nome deriva do fato do modelo resultante ser apresentado na forma de uma estrutura de árvore, onde cada nó interno (não folha) corresponde a um teste ou condição, e cada nó externo (folha) denota uma classe prevista. Em cada nó, o algoritmo escolhe o atributo que "melhor" particiona os dados em classes individuais.

Segundo Han e Kamer (2006), a indução de Árvore de Decisão pode ser usada para seleção de subconjunto de atributos, onde os atributos que não aparecem na árvore são considerados irrelevantes. O conjunto de atributos que aparecem na árvore forma o subconjunto reduzido de atributos.

A princípio, há exponencialmente muitas árvores de decisão que podem ser construídas a partir de um determinado conjunto de atributos. Embora algumas árvores sejam mais precisas que outras, encontrar a árvore ótima é computacionalmente inviável por causa do tamanho exponencial do espaço de pesquisa (Tan, Steinbach e Kumar, 2009). Apesar disso, algoritmos eficientes têm sido desenvolvidos para induzir uma árvore de decisão razoavelmente precisa, embora não perfeita, em uma razoável quantidade de tempo. Um desses algoritmos é o algoritmo de Hunt, que é a base de muitos outros algoritmos, incluindo o ID3, C4.5 e CART.

A sigla ID3 significa *Iterative Dichotomizer* 3 e foi um método desenvolvido por Quinlan (1986). O algoritmo ID3 consiste num processo de indução de árvores de decisão. A construção da árvore é realizada de cima para baixo (*top-down*), com o objetivo de escolher sempre o melhor atributo para cada nó de decisão da árvore. É um processo recursivo que após ter escolhido um atributo para um nó, começando pela raiz, aplica o mesmo algoritmo aos descendentes desse nó, até que certos critérios de parada sejam verificados.

A escolha do atributo de partição é concretizada tendo em conta o ganho de informação. O **Ganho de Informação** é uma medida estatística que está na base da construção de árvores de decisão neste algoritmo. Esta medida estatística consiste no seguinte (Quinlan, 1996):

Se tivermos um conjunto de vários exemplos S, e um conjunto de n classes C = {C_1, C_2,...., C_n} , sendo p_i a probabilidade da classe C_i em S, então a entropia do conjunto S, é a homogeneidade deste, traduzida na equação (2.5):

$$Entropia = -\sum_{i=1}^{c} p_i \log_2 p_i \qquad (2.5)$$

A entropia é uma medida aplicável à partição de um espaço de probabilidade, medindo quanto esse espaço é homogéneo, ou por outro lado, quanto maior a entropia maior a desordem. A entropia atinge o seu valor máximo, igual a $\log_2 n$, quando $p_1 = p_2 = ... = p_n = 1/n$, expressando precisamente a existência de um máximo de heterogeneidade. Pelo contrário a homogeneidade máxima corresponderia a $p_1 = p_2 = ... = p_n = 0$ e $p_i = 1$.

De outro modo, pretende-se saber qual o ganho de informação do atributo A, que é dado pela equação (2.6):

$$Ganho(S, A) = Entropia(S) - \sum_{v \in Valores(A)} \frac{|S_v|}{|S|} Entropia(S_v) \qquad (2.6)$$

em que, valores (A) é o conjunto de todos os valores possíveis para o atributo A, e $|S_v|$ é o subconjunto de S para o qual o atributo A tem valor v, confrme equação (2.7):

$$S_v = \{s \in S \mid A = s\} \qquad (2.7)$$

Desta forma, o Ganho de Informação, mede a eficácia de um atributo em classificar os dados de treino, a escolha do atributo mais eficaz – que mais reduz a entropia – faz com que a tendência seja a de gerar árvores, que são, em geral, menos profundas com menos nós e ramificações.

Em suma, o algoritmo ID3 realiza uma procura ávida (*greedy*) no espaço das árvores de decisão, consistentes com os dados, guiada pelo ganho de informação e feita segundo a

estratégia do "subir a colina" (*hill-climbing*). No entanto, no uso desta estratégia corre-se o risco da solução convergir para um óptimo local (Quinlan, 1996).

De acordo com Quinlan (1996), para os atributos cujos domínios sejam valores quantitativos, reordenam-se as instâncias, de acordo com esse atributo e procuram-se pontos extremos nos quais existe uma mudança de valor da classe. Um ponto de mudança de classe marca uma partição binária do conjunto das instâncias, mediante uma condição lógica do tipo A > x, sendo A o atributo numérico em causa e x um valor calculado a partir dos dois valores consecutivos de A nesses pontos. Normalmente, toma-se x igual à média dos valores de A, nos pontos consecutivos. Foi mostrado que, neste tipo de atributos, de todos os possíveis pontos de partição, aqueles que maximizam o ganho de informação correspondem exatamente à separação dos dois exemplos pertencentes a classes diferentes.

Uma das grandes vantagens do ID3 é a sua simplicidade, o seu processo de construção torna relativamente simples a compreensão do seu funcionamento. A maior desvantagem do ID3 é que a árvore de decisão produzida é essencialmente imutável – não se pode eficientemente reutilizar a árvore sem a reconstruir. Usando este algoritmo para atualização, o método tende a produzir uma árvore de decisão que está longe da árvore de decisão óptima, impedindo assim a ideia original de reformular a árvore de decisão a partir da original (Han e Kamber, 2006).

O algoritmo C4.5 (Quinlan, 1993) é um método melhorado relativamente ao ID3 que, entre outras melhorias, combate o problema de *overfitting*, utilizando uma estratégia de poda de árvore. O algoritmo C4.5 adopta a estratégia (pós-poda). Podar uma árvore, neste contexto, significa reduzir algumas sub-árvores a folhas, ou de outra forma, um ramo da árvore, a partir de determinado nó é cortado (transformado em folha). O corte dum ramo da árvore é guiado por um teste estatístico que tem em conta os erros num nó e a soma dos erros nos nós que descendem desse nó. Assim, para cada nó, a poda só se concretiza se o desempenho da árvore não diminuir significativamente. Além do problema do *overfitting*, o C4.5 inclui soluções para problemas concretos e comuns do mundo real como: atributos com valores quantitativos; valores omissos e dados contendo ruído.

Outra possibilidade disponibilizada por este sistema é a capacidade de realizar validação cruzada (*cross-validation*) com dois ou mais grupos (*v-fold* ou validação Jacknife), melhorando assim a estimativa do erro cometido pelo classificador (Tan, Steinbach, Kumar, 2009).

O J48 é um algoritmo baseado na implementação do algoritmo C4.5 *release* 8, e este por sua vez é uma evolução do algoritmo ID-3, ambos foram desenvolvidos por Quinlan (1993). A versão mais recente desta classe de algoritmos é C5.0, contudo, este algoritmo não será discutido neste trabalho por se tratar de uma implementação proprietária e que é disponibilizada apenas comercialmente.

O algoritmos J4.8 surgiu da necessidade de recodificar o algoritmo C4.5, que originalmente é escrito na linguagem C, para a linguagem Java (Witten, 2005). Ele tem a finalidade de gerar uma árvore de decisão baseada em um conjunto de dados de treinamento, sendo este modelo usado para classificar as instâncias no conjunto de teste.

Um dos aspectos para a grande utilização do algoritmo J4.8 pelos especialistas em Mineração de Dados é que o mesmo mostra-se adequado para os procedimentos envolvendo as variáveis (dados) qualitativas e variáveis quantitativas contínuas e discretas presentes nas Bases de Dados.

Um esqueleto de algoritmo de indução de árvore de decisão chamado *CrescimentoDaArvore* é mostrado na Tabela 2.3. A entrada desse algoritmo consiste dos registros de treinamento E e o conjunto de atributos F. O algoritmo funciona selecionando recursivamente o melhor atributo para dividir os dados - passo 7 - e expandir os nodos folha da árvore - passos 11 e 12 - até que o critério de parada seja satisfeito - passo 1- (Tan, Steincbach e Kumar, 2009).

Tabela 2.3 Exemplo de algoritmo de indução de árvore de decisão.
(Tan, P., Steinbach, M., Kumar, V., 2009.)

```
CrescimentoDaArvore(E,F)
1: se cond_parada(E,F) = verdadeiro então
2: folha = criarNodo();
3: folha.rotulo = Classificador(E)
4: retorna folha.
5: senão
6: raiz = criarNodo().
7: raiz.cond_teste = encontrar_melhor_divisao(E,F).
```

```
8: atribuir V = {v|v é um resultado possível de raiz.cond_teste}.
9: para cada v Є V faca
10: Ev = {e | raiz.cond_teste(e) = v e é Є E}.
11: filho = CrescimentoDaArvore(Ev,F).
12: adicionar filho como descendente de raiz e rotule o limite (raiz → filho) como v.
13: fim do para
14: fim se
15: retornar raiz.
```

1. A função criarNodo() estende a árvore de decisão criando um novo nodo. Um nodo na árvore de decisão possui uma condição de teste, denotada como nodo.cond_teste, ou um rótulo de classe denotado como nodo.rotulo;

2. A função encontrar_melhor_divisao() determina qual atributo deve ser selecionado como condição de teste para dividir os registros de treinamento.

3. A função Classifica() determina o rótulo de classe a ser atribuído a um nodo folha. Para cada nodo folha t, p(i|t) denota a fração de registros de treinamento da classe i a associação ao nodo t.

4. A função cond_parada() é usada para determinar o processo de crescimento da árvore testando se todos os registros possuem ou o mesmo rótulo de classe ou os mesmos valores de atributos. Outra forma de terminar a função recursiva é testar se o número de registros está abaixo de algum ponto limite mínimo.

Como forma de exemplificação, na Tabela 2.4 é monstrado cinco registros de treinamento. Todos esses registros de treinamento estão rotulados corretamente e a árvore de decisão correspondente é mostrada na Figura 2.15.

Tabela 2.4 Um exemplo de conjunto de treinamento para classificar mamíferos.
(Tan, P., Steinbach, M., Kumar, V., 2009.)

Nome	Temperatura do	Origina	Quatro patas	Hiberna	Rótulo de Classe
Salamandra	Sangue frio	Não	Sim	Sim	Não
Peixe Guppy	Sangue frio	Sim	Não	Não	Não
Águia	Sangue quente	Não	Não	Não	Não
Poorwill	Sangue quente	Não	Não	Sim	Não
Playpus	Sangue quente	Não	Sim	Sim	Sim

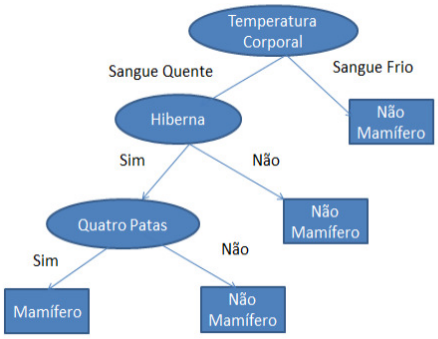


Figura 2.15 Árvore de decisão induzida do conjunto de dados de treinamento.
(Tan, P., Steinbach, M., Kumar, V., 2009)

2.3.11 Relação entre Data Warehouse, OLAP e Mineração de Dados

Os Sistemas de Apoio à Decisão (SAD) ou *Decision Support System* (DSS) agregam importante diferencial competitivo nas organizações, ajudando na tomada de decisão. A implantação dos SAD ocorre principalmente pelo uso de ferramentas *On-line Analytical Processing* (OLAP) e Mineração de Dados, que por sua vez fazem acesso aos dados do *Data Warehouse*. DW e *Data Marts* são utilizados numa grande variedade de aplicações. Os executivos de negócios utilizam os dados em DW e *Data Marts* para realizar a análise de dados e tomar decisões estratégicas.

Tipicamente, quanto mais tempo um DW está em uso, mais ele evoluirá (Inmon, 1996). Esta evolução ocorre ao longo de um número de fases. Inicialmente, o DW é utilizado principalmente para a geração de relatórios e para responder consultas predefinidas. Progressivamente, é usado para analisar dados resumidos e detalhados, onde os resultados são apresentados na forma de relatórios e gráficos. Mais tarde, o DW é utilizado para fins estratégicos, realizando análise multidimensional e sofisticadas operações de S*lice and Dice*. Finalmente, o DW pode ser empregado na descoberta de conhecimento e tomado de decisão estratégica, utilizando ferramentas de Mineração de Dados. Neste contexto, as ferramentas para DW podem ser classificados em ferramentas de acesso e recuperação, ferramentas de relatórios de banco de dados, ferramentas de análise de dados e ferramentas de Mineração de Dados (Han e Kamer, 2006).

Segundo Han e Kamer (2006), Mineração de Dados frequentemente requer limpeza (*data cleaning*) e integração de dados (*data integration*). Eles ainda reportam que, a limpeza de

dados é um importante problema para ambos os processos – DW e Mineração de Dados - visto que dados do mundo real tendem a ser incompletos e inconsistentes. Para Inmon (1996), a existência de um DW provê limpeza, integração e completude dos dados, permitindo que o processo de Mineração de Dados foque na sua principal tarefa: extrair conhecimento compreensível e útil.

A construção de um DW envolve a limpeza, integração e completude dos dados – etapa ETL -, e pode ser visto como uma importante etapa de pré-processamento para DM. Mais ainda, DW prove ferramentas analíticas (OLAP) com análises multidimensionais em diversas granularidades, que podem ser utilizadas nas fases de exploração de dados e validação dos resultados obtidos no processo de mineração.

De acordo com (Sanches, 2003), existe uma relação simbólica entre a atividade de Mineração de Dados e *Data Warehouse*. Os DW organizam os dados para um efetivo processo de mineração, porém, a exploração de dados através da mineração pode ser aplicada onde não exista nenhum DW. O uso do DW aumenta significativamente as chances de sucesso da Mineração de Dados, visto que o DW dispõe de dados integrados; dados detalhados e resumidos; dados históricos e metadados. A utilização desses tipos de dados melhora o desempenho e o resultado do processo de mineração.

Segundo (Kimball, 1997), enquanto OLAP é dedutivo e guiado por especialistas, Mineração de Dados é indutivo e guiado pelos próprios dados. Ambas necessitam de dados limpos e consistentes. E neste caso, o *Data Warehouse* é capaz de fornecer dados para as duas tecnologias, o que o torna a principal fonte de dados para OLAM, cujo termo refere-se à junção de OLAP e Mineração de Dados. A Figura 2.16 ilustra onde o DW se encaixa no processo de DM.

Segundo (Han, 2006), OLAM significa minerar interativamente em diferentes porções dos dados e em diferentes níveis de agregação, utilizando operações OLAP, podendo-se escolher as funções de Mineração de Dados e algoritmos dinamicamente, além de poder navegar pelos resultados da mineração.

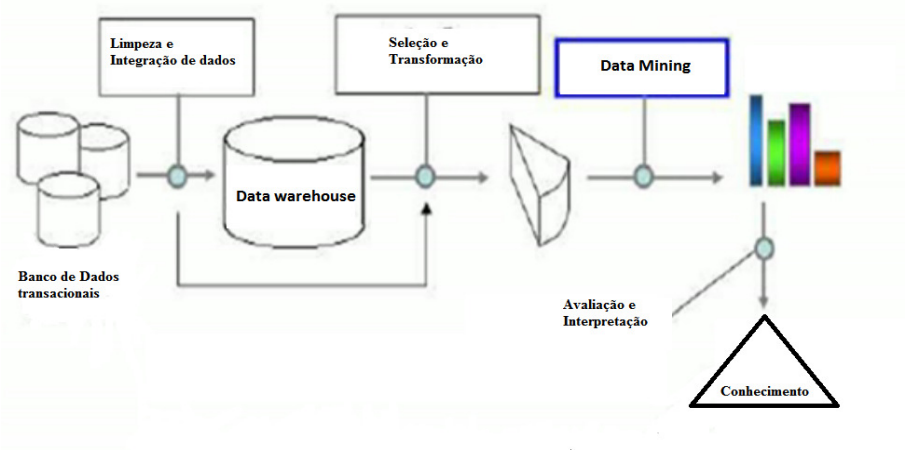


Figura 2.16 Relação entre DW e Mineração de Dados.
(Adaptado de Tan, P., Steinbach, M., Kumar, V., 2009.)

2.4 METODOLOGIA PARA MINERAÇÃO DE DADOS

A pesquisa apresentada nesta dissertação utilizou uma abordagem empírica positivista na análise de registros de empregados durante 14 anos. A organização em estudo forneceu as observações de registro de transferências de empregado entre os anos 2008 e 2012. Esses registros foram analisados por meio de métodos de análise estatística descritiva, bem como técnicas de análise multivariada. Agrupamento, Sumarização, Classificação e Regras de Associação foram utilizados como técnicas de Mineração de Dados, a fim de identificar padrões e modelos descritivos. A mineração de dados foi realizadas usando o WEKA 3.7, ferramenta que reúne uma coleção de algoritmos de aprendizagem de máquina para resolver problemas de DM, implementada em Java e código aberto sob a licença GPL.

Este estudo utiliza a metodologia CRISP-DM sugerida por Chapman (2000). Esta metodologia envolve seis fases: Definição do problema, Exploração dos Dados, Preparação dos Dados, Modelagem, Avaliação e Implementação, conforme Figura 2.17.

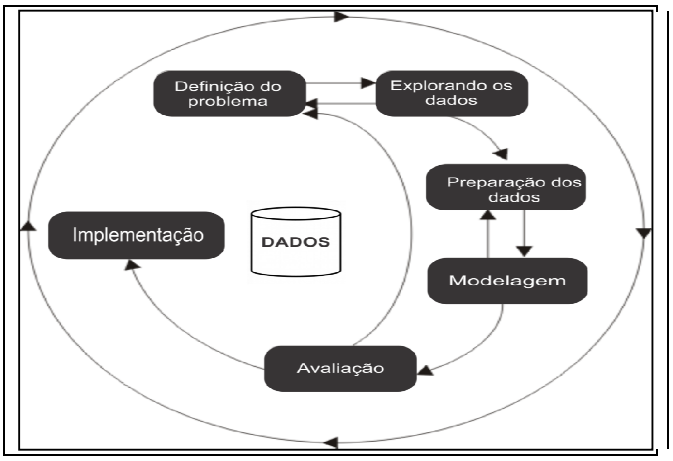


Figura 2.17 Típico processo de mineração de dados
(Adaptado de http://www.crisp-dm.org)

Antes de proceder a mineração propriamente dita, os dados disponíveis foram estudados, aonde os objetivos de negócio foram detalhados.

2.4.1 Definição do problema

A primeira fase da metodologia CRISP-DM é o entendimento do negócio com foco sobre objetivos do projeto e requisitos. O conhecimento obtido nesta fase é transformado em uma definição de problemas de Mineração de Dados, juntamente com a definição preliminar de um plano para alcançar os objetivos. Para identificar os potenciais problemas, um estudo literário foi realizado e trabalhos relevantes e relacionados foram identificados. Neste estudo, pesquisas relacionadas ao tema de Gestão de Pessoas, particularmente rotatividade interna de pessoal, foram levantadas e adequados algoritmos de mineração de dados voltados para modelagem preditiva e descritiva foram selecionados.

2.4.2 Exploração dos dados

A segunda fase consiste no entendimento dos dados. Neste ponto, dados são coletados, sumarizados e entendidos. A fim de se tornar familiarizado com os dados, é necessário identificar problema de qualidade de dados, obter *insights* e selecionar subconjuntos que serão utilizados na fase de Mineração de Dados. Para este estudo, foi construído um *Data Warehouse (DW)* que serviu como fonte de dados para o processo de descoberta de conhecimento. O DW criado auxiliou no processo de limpeza, integração e exploração dos

dados. Foram selecionados mais de 138 mil transferências entre unidades referentes a processo seletivo interno ou interesse da administração entre os anos de 2008 e 2012. Para obter uma visão completa da distribuição dos dados e identificação de desvios (outliers), uma análise descritiva foi realizada para fins exploratórios.

2.4.3 Preparação de dados

Esta fase abrange todas as atividades necessárias para a construção do conjunto de dados final utilizado na fase de modelagem. As tarefas são suscetíveis de serem realizadas várias vezes e não podem estar prescritas, visto que diferentes bancos de dados tendem a expor novos assuntos e desafio. Segundo Siraj e Abdoulha (2011), com o objetivo definido, é importante escolher a ferramenta, algoritmo e métodos de mineração corretos que espera-se dar os melhores resultados com os dados fornecidos. Esta fase foi realizada repetidamente para determinar atributos adequados para serem utilizados pelos algoritmos.

2.4.4 Modelagem

Durante esta fase, técnicas de modelagem são selecionadas e aplicadas ao conjunto de dados usado no estudo. Este fase inclui selecionar uma técnica apropriada, construção do modelo e em seguida avalição dos resultados. De acordo com Siraj e Abdoulha (2011), esta fase envolve a seleção de técnicas adequadas ao problema e o refinamento do modelo sempre que necessário, a fim de atender aos objetivos e restrições definidas. Inicialmente, estatística descritiva foi realizada para investigar a natureza do conjunto de dados e a distribuição de cada atributo. Tabelas de frequência foram geradas e análises de correlação foram conduzidas para determinar relações entre atributos, incluindo análise de tabulação cruzada, através de cubos OLAP. Após exploração dos dados, Análises de grupo (clustering) foram desenvolvidas com o objetivo de agrupar as transferências em grupos distintos. Em seguida foi criado um novo tipo de dimensão no cubo OLAP original que possibilitasse o detalhamento e análise de cada grupo por um especialista de negócio. Por fim, uma técnica de Regras de Associação foi empregada com o objetivo de descrever cada grupo, encontrado na fase anterior, de forma a caracterizar cada grupo.

2.4.5 Avaliação

Nesta fase, os modelos e resultados são avaliados a fim de assegurar que apenas resultados válidos e úteis sejam incorporados ao sistema de apoio a decisões. Para auxiliar a análise dos resultados foi criado um novo modelo dimensional que permitiu aos especialistas de négocio fazerem análises OLAP.

2.4.6 Implementação

Na última fase da metodologia CRISP-DM, o conhecimento adquirido com o modelo é incorporado ao sistema de apoio de decisões. Para isto, um módulo foi criado no domínio do problema através da suíte Pentaho de Inteligência do Negócio.

3 - ESTUDO DE CASO E METODOLOGIA

Este capítulo tem com objetivo apresentar o estudo de caso e a metodologia desenvolvida neste trabalho.

3.1 ESTUDO DE CASO

Um estudo de caso foi realizado em uma organização de economia mista do ramo financeiro, com mais de quatro mil unidades espalhadas em todo território brasileiro e mais de 88 mil empregados.

Essa organização em estudo está estruturada em três subsistemas: Negocial, Logístico e Central. O subsistema Negocial é composto de unidades operacionais, que tem por objetivo atendimento aos clientes e a realização de negócios, e é composta por Superintendências Regionais e Canais de Atendimento. O subsistema Logístico é composto de unidades da rede de sustentação ao negócio, e tem por objetivo garantir o equilíbrio e os meios para realização dos negócios. Já o subsistema Central é composto por unidades da Matriz, possui a representação dos macroprocessos que sustentam as atividades da organização, sendo responsável pela definição de diretrizes e pelo controle dos resultados. Esta diversidade de áreas da organização possibilita a atuação de empregados das mais variadas formações.

A contratação de novos empregados nessa organização se dá por concursos públicos e todos eles, ou grande maioria, iniciam suas atividades em unidades operacionais (subsistema Negocial). Após o estágio probatório, estes novos empregados podem participar de processos seletivos internos para outras áreas da empresa visando sua ascensão profissional ou atuação em área relacionada com sua formação acadêmica.

Esta organização implantou formalmente, um processo seletivo interno (PSI) a partir de 2008, com o objetivo de normatizar, padronizar e criar uma meritocracia baseada na análise das trajetórias profissional e educacional de cada candidato. Segundo o manual normativo da organização, o PSI tem por objetivo identificar empregado com as competências necessárias ao exercício da Função Gratificada, visando à composição e manutenção de equipes qualificadas para alcance dos resultados da instituição.

A sistemática geral do processo seletivo interna envolve as seguintes etapas:

- Abertura do PSI: consiste da publicação do PSI para determinada função gratificada, dos critérios objetivos para a seleção dos empregados que participarão da Avaliação de Competências assim como da produção temática.

- Manifestação de interesse: candidato se inscreve no processo e envia a sua produção temática.

- Apuração dos critérios objetivos: nesta etapa os candidatos são classificados conforme critérios previamente definidos. Os candidatos são avaliados conforme sua trajetória profissional e educacional.

- Avaliação de Competências: nesta etapa os seis candidatos melhor classificados são avaliados por uma banca avaliadora, que selecionará o mais adequado para exercer a função gratificada.

Assim, para este estudo de caso, foram analisados os históricos de lotação com a expectativa de que a aplicação descreva o fluxo de empregados entre os subsistemas, conforme Figura 3.1. A hipótese é que isso contribuirá para uma tomada de decisão focada nos fatos, apoiando na criação e avaliação das políticas de processos seletivos internos.

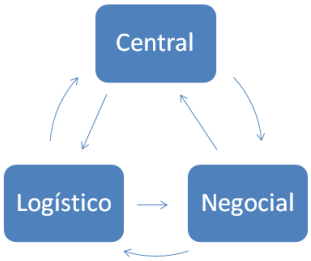


Figura 3.1 Fluxo de empregados entre subsistemas.

Após procedimento de correlação aos desafios e motivação do presente estudo, iniciou-se a etapa de experimentos. Foi criado um módulo de Sistema de Apoio à Decisão (SAD) que consiste em um ambiente projetado para apoiar, contribuir e influenciar no processo de tomada de decisão. Conforme Figura 3.2, o SAD utilizado e implementado nesta pesquisa

é formado por tês componentes: os Dados (dispostos no *Data Warehouse*), o SGBD, e as Ferramentas de Apoio à Decisão.

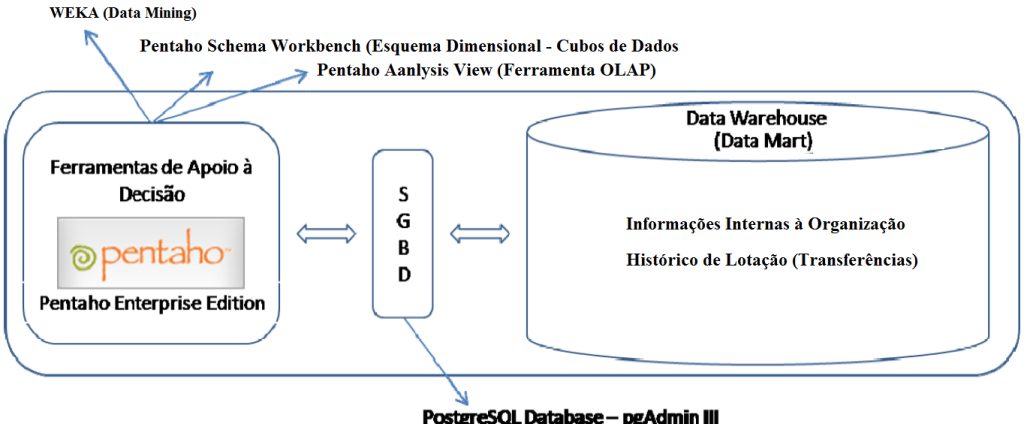


Figura 3.2 Componentes do sistema de apoio à decisão.

O papel do SGBD em um ambiente de apoio à decisão é permitir que os usuários definam, construam e manipulem o Banco de Dados com dados integrados e compartilhados. Um SGBD pode representar a unificação de diversos arquivos, que, de outra forma, seriam distintos, eliminando-se total ou parcialmente a redundância entre os mesmos. Já o compartilhamento não significa apenas que as aplicações existentes podem compartilhar dados do Banco de Dados, mas também que novas aplicações podem ser desenvolvidas para operar sobre os mesmos dados armazenados.

O *Data Warehouse* deste trabalho foi implementado no SGBD *PostgreSQL* 9.1, utilizando a modelagem dimensional. Este DW corresponde aos dados internos à organização em estudo, constituído principalmente pelo histórico de lotação de empregados.

As Ferramenas de Apoio à Decisão são softwares utilizados para manipular os dados extraídos do *Data Warehouse* através da estrutura de cubos de dados, de funções de agregações (sumarização, médias, mínimos, máximos, *count*, etc.), de funções estatísticas ou de funções gráficas. Elas auxiliam na simulação e análise dos dados, proporcionando a descoberta de novos conhecimentos. As ferramentas de apoio à decisão utilizadas neste trabalho foram:

- *Pentaho Schema Workbench*: ferramenta responsável pela criação dos cubos de dados (tabelas de fatos), dimensões (tabelas de dimensões) e métricas do esquema dimensional. No Apêndice B é apresentado o arquivo XML do esquema dimensional criado neste trabalho;

- *Pentaho Analysis View*: ferramenta OLAP que executa operações *Slice and Dice* sobre o arquivo XML do esquema dimensional. Este ferrameno foi utilizada nas etapas de exploração de dados e análise de resultado como será mostrado adiante.

- *WEKA (Waikato Environment for Knowledge Analysis)*: ferramenta que implementa os principais algoritmos de Mineração de Dados. A Figura 3.3 mostra a tela incial de pré-mineração dos dados.

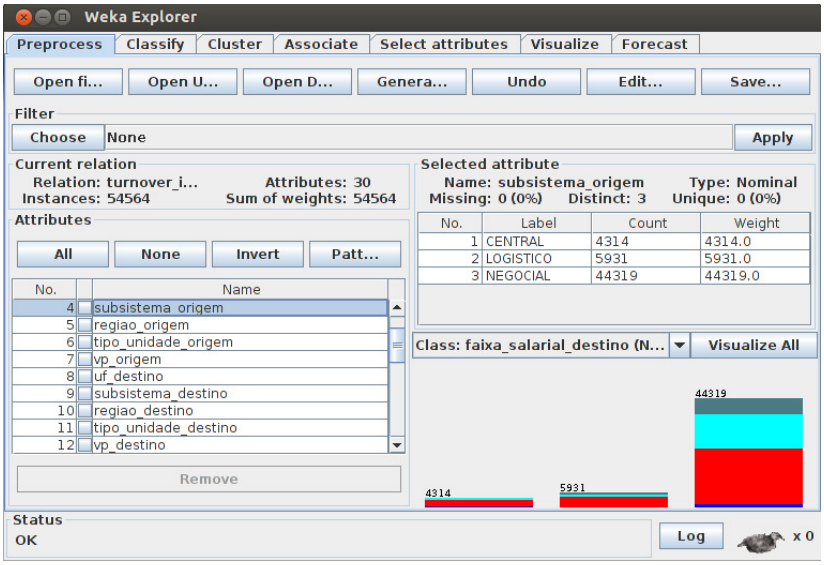


Figura 3.3 Mineração de dados pela ferramenta WEKA.

Foi utilizada uma abordagem em cascata envolvendo os algoritmos K-*means*, Apriori e C4.5 com o objetivo de extrair informações úteis relativas à rotatividade interna de pessoal. O K-*means* foi empregado para segmentar as transferências, e Apriori e C4.5 foram usados para caracterizar os grupos criando perfis de transferências (Figura 3.4).

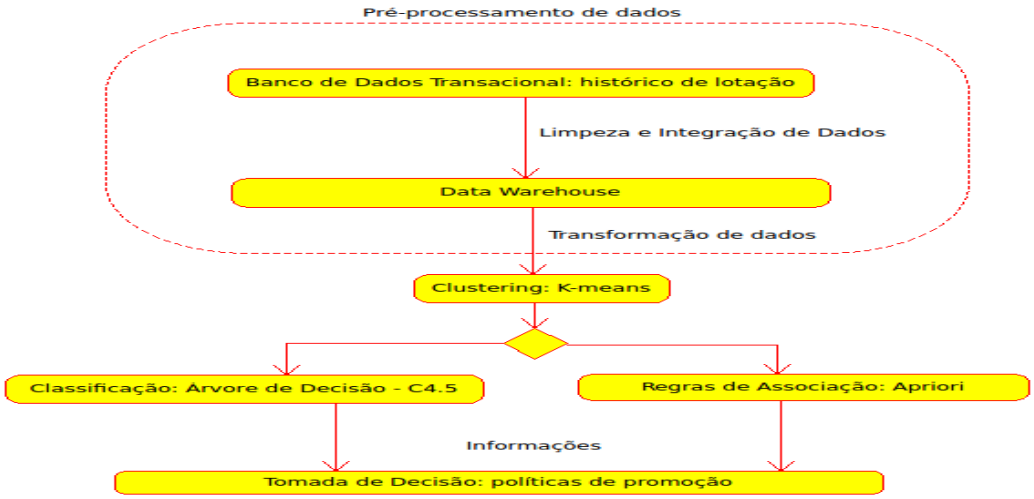


Figura 3.4 Abordagem em cascata para descrever transferências.

A Figura 3.4 ilustra as atividades realizadas durante os experimentos. Em resumo, primeiramente foi construído um *Data Warehouse* para integrar e tratar os dados extraído do sistemas de Gestão de Pessoas da organização em estudo. Em seguida foram realizadas análises OLAP através do *Pentaho Analysi View*. Algoritmos de Mineração de Dados são aplicados e os resultados interpretados com o auxílio, novamente, de consultas OLAP. Por fim as informações encontradas são utilizadas como insumos pelos tomadores de decisão.

3.2 IMPLEMENTAÇÃO DO DATA WAREHOUSE

O *Data Warehouse* foi criado a partir dos dados cadastrais de histórico de lotação (matrícula empregado, unidade origem, unidade destino, função origem, função destino, data início unidade origem, data fim unidade origem, código de ocorrência, motivo transferências), de empregados (sexo, idade, estado civil, geração, escolaridade, formação), de funções gratificadas (nome, tipo função) e unidades (subsistema, UF, região). Os dados foram orientados de modo a permitir os agrupamentos principalmente por sexo, tipo função e subsistema, visando às informações referentes às transferências de empregados.

Foram coletados e analisados mais de 138 mil registros de transferências entre os anos 2008 e 2012, cujas características são listadas na Tabela 3.1:

Tabela 3.1 Variáveis de entrada utilizadas na mineração de transferências

Atributo	**Descrição e valores no momento da transferência**
Idade	Idade em anos até data fim de lotação.
Geração	Geração do empregado: Veteranos, Boomers, Geração X e Geração Y.
Sexo	Sexo: F ou M.
Casado?	Se o empregado estava casado: S ou N.
Escolaridade	Escolaridade: Ensino Médio, Graduação ou Pós-Graduação.
Formação	Informa qual a área de formação do empregado: Administração, Direito, etc.
Número de dependentes	Quantidade de dependentes informados para o Imposto de Renda.
Possui experiência externa?	Informa se empregado já trabalhou fora da empresa: S ou N.
Rendimento	Salário do empregado
Tipo de Função Gratificada	Categoria da função exercida: Sem função, Chefia (gerencial) ou Técnico.
UF, Região, Subsistema e Tipo de Unidade	Dados das unidades de origem e destino. Subsistema: Central, Logístico ou Negocial. Tipo de Unidade: Agência, Centralizadora, Centro Administrativo, Diretoria e Presidência.
Tempo de empresa	Quanto tempo (em anos) o empregado está na empresa
Tempo de unidade	Quanto tempo o empregado ficou na unidade. Atributo a ser previsto: <=2 anos ou > 2 anos.
Tipo de transferência	Promoção, Transferência por lateralidade (Sem Função Gratificada ou mesma Função Gratificada) ou Decesso.
Horas de Treinamento	Quantidade de horas treinadas até a data fim da lotação.

O motivo principal que levou ao desenvolvimento de um ambiente de *Data Warehouse* ao invés de um ambiente de Banco de Dados tradicional reside no fato dos ambientes de suporte a decisão e extração do conhecimento em bases de dados serem caracterizados pela não-volatilidade dos dados e pela complexidade das consultas ad hoc.

A modelagem dimensional do DW desenvolvida neste trabalho foi implementada fisicamente no SGBD Relacional *PostgreSQL* 9.1, conforme ilustra a Figura 3.5.

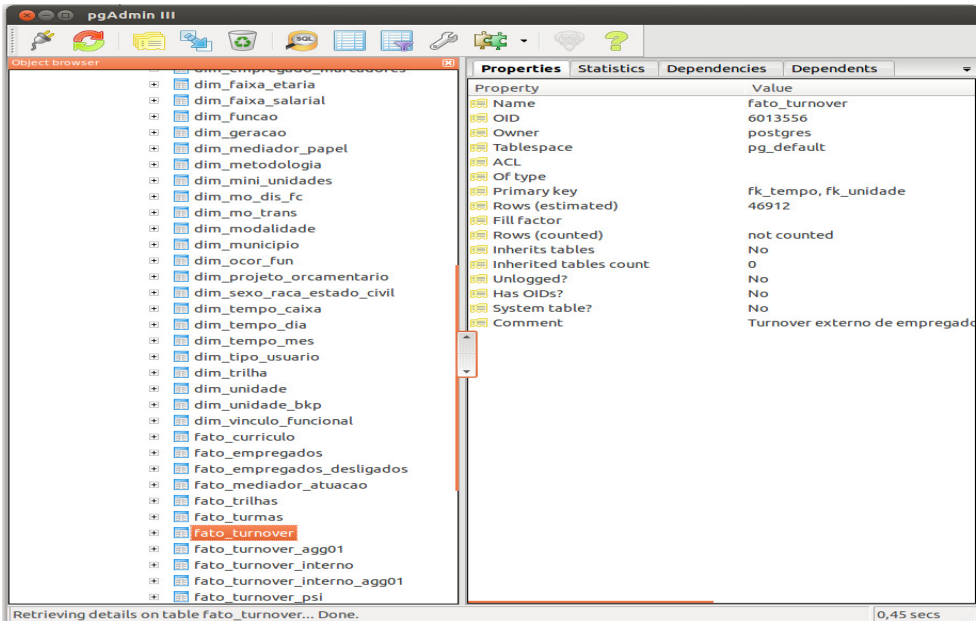


Figura 3.5 Desenvolvimento da modelagem dimensional no SGBD *PostgreSQL*.

A modelagem lógica do *Data Warehouse* possui cinco tabelas de fatos (Transferências, Transferências segmentada, Turnover Interno, Turnover Externo, Empregados) e 15 tabelas de dimensões (dentre elas: Dimensão Unidade, Dimensão Faixa Etária, Dimensão Função Gratificada, Dimensão Geração, Dimensão Motivo Transferência, etc.). A Figura 3.6 apresenta a tabela fato Transferências e suas tabelas dimensões. No Apêndice A é apresentado esquema dimensional completo utilizado neste trabalho.

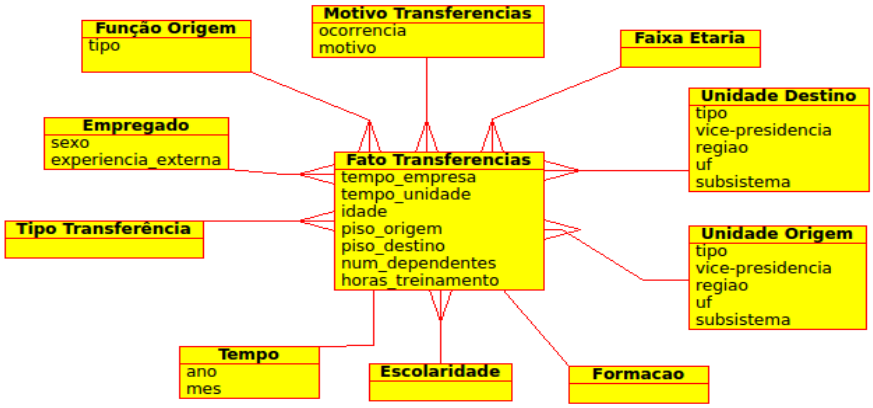


Figura 3.6 Tabela fato de transferências utilizado na exploração de dados.

Seguem as definições das tabelas fato:

- Fato Tunover Externo (fato_tunvover_externo): é responsável pelo reconhecimento e análise dos empregados que se desligaram da organização entre os anos de 2008 e 2012. Possui 2 tabelas dimenssão, são elas: Tempo e Unidade de Lotação. Possui 3 métricas: quantidades de empregados admitidos, desligados e total de empregados ativos por ano e mês.

- Fato Tunover Interno (fato_tunvover_interno): é responsável pelo reconhecimento e análise das transferências ocorridas entre os anos de 2008 e 2012. Posui 2 dimenssões, são elas: Tempo e Unidade de Lotação. Possui as métricas: quantidades de empregados admitidos, desligados e total de empregados ativos por ano e mês.

- Fato Transferências (fato_transferencias): é a tabela fato responsável pelas análises de transferências ocorridas na etapa de exploração dos dados descrita na seção 3.3. Possui 11 tabelas dimensão, são elas: Tempo, Sexo, Faixa Etária, Geração, Escolaridade, Formação, Função de Origem, Função de Destino, Unidade de Origem, Unidade de Destino e Motivo da Transferência. As métricas de cubos são: Idade, Tempo de Unidade, Tempo de Empresa, Horas de Treinamento e Piso Salarial;

- Fato Transferências Segmentadas (fato_transferencias_predicted): esta tabela possui as mesmas dimensões e métricas da fato anterior, acrescentando uma nova dimensão chamada de *cluster*. Esta tabela fato foi utilizada na análise das caracteríscias de cada grupo gerado pelo algoritmo K-means, conforme descrito na seção 3.4;

- Fato Empregados (fato_empregados): esta tabela fato é responsável pelas análises do quadro de empregados existentes na organização estudada. Foram coletados o histórico de empregados dos anos 2011 e 2012.

Para atender as nessidades de análise das informações, o SAD utiliza o *Data Warehouse* para dar suporte às operações OLAP do tipo *Slice and Dice* e também para dar suporte às técnicas de Mineração de Dados.

O *Data Warehouse* se apresentou de forma satisfatória para realização das consultas OLAP e para aplicação das técnicas de Mineração de Dados, conforme será discutido mais adiante.

3.2.1 Extração, Transformação e Carga (ETL) do DW

A etapa de ETL serve para detectar os erros de cadastros e inconsistências dos dados extraídos do ambiente opracional, ou seja, tratar questões de qualidade de dados. É realizada a limpeza dos dados a fim de adequar e carregar apenas os dados necessários no *Data Warehouse*. Esta adequação dos dados se dá através da integração de dados heterogêneos, remoção de dados incompletos, eliminação de repetição dos dados e dos problemas de tipagem.

Houve limpeza e transformação dos dados com as datas de fim de lotação que se encontravam nulas (em branco) e que foram atualizadas com a data de início de lotação posterior mais um dia no histórico de lotação do empregado em questão. Desta forma, somente a lotação atual do empregado possui a data fim nula.

Alguns registros da base de dados foram excluídos por não apresentarem informações concisas ou por não serem de interesse ao estudo desta dissertação. Nesta situação se encontram os registros cujo motivo de transferência não se tratava de promoção ou transferência por interesse da administração, como por exemplo, extinção de unidade ou reestruração.

Na construção do DW utilizou-se a ferramenta Pentaho Data Integrator (PDI 4.1), também conhecido como Kettle, ferramenta que tem como objetivo realizar o processo de ETL em sistemas de DW. A Figura 3.7 ilustra o processo implementado para carregar a tabela Fato Turnover Interno.

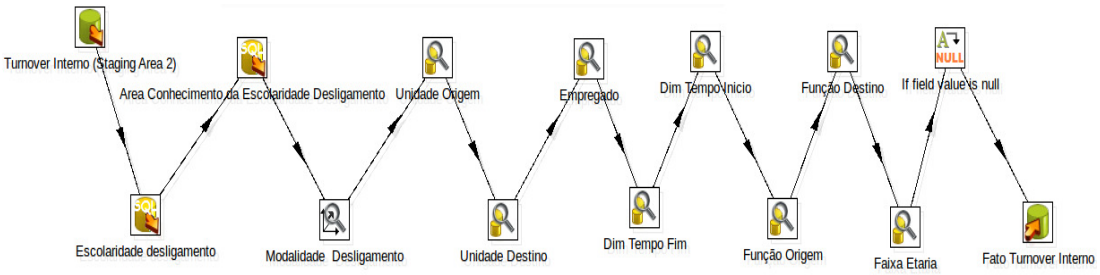


Figura 3.7 Processo ETL implementado com o PDI – Carga da tabela Fato *Turnover* Interno.

Em resumo, a transformação apresentada na Figura 3.7 executa os seguintes passos:

1. Turnover Interno (*Staging Area 2*): Passo que recuperar o histórico de lotação do empregados;
2. Escolaridade desligamento: Passo que recupera a escolaridade do empregado na data de desligamento (Dimensão Escolaridade);
3. Área Conhecimento da Escolaridade desligamento: Passo que recupera a área de conhecimento (Administração, Direito, Tecnologia da Informação, etc) referente a formação do empregado (*lookup* na Dimensão área de conhecimento);
4. Modalidade Desligamento: Passo que recupera o motivo de desligamento (*lookup* na Dimensão Motivo);
5. Unidade origem: Passo que recupera a unidade de lotação de origem (*lookup* na Dimensão Unidade);
6. Unidade destino: Passo que recupera a unidade de lotação de destino (*lookup* na Dimensão Unidade);
7. Empregado: Passo que recupera o empregado que realiza a transferência (*lookup* na Dimensão Empregado);
8. Dim Tempo Fim: Passo que vincula a data fim de lotação à Dimensão Tempo;
9. Dim Tempo Início: Passo que vincula a data de início de lotação à Dimensão Tempo;
10. Função Origem: Passo que recupera a função comissionada do empregado na unidade de origem (*lookup* na Dimensão Função);
11. Função Destino: Passo que recupera a função comissionada do empregado na unidade de destino (*lookup* na Dimensão Função);
12. Faixa Etária: Passo que recupera a faixa etária do empregado na data de desligamento (*lookup* na Dimensão Faixa Etária);
13. *If field values is null*: Passo que define valores *default* para campos nulos;
14. Fator Turnover Interno: Passo que atualiza a tabela fato.

3.2.2 Pentaho Schema Workbench – Modelagem Dimensional

O esquema dimensional foi modelado utilizando o módulo Schema Workbench da plataforma de código aberto de *Business Intelligence*, Pentaho, conforme ilustra a Figura 3.8.

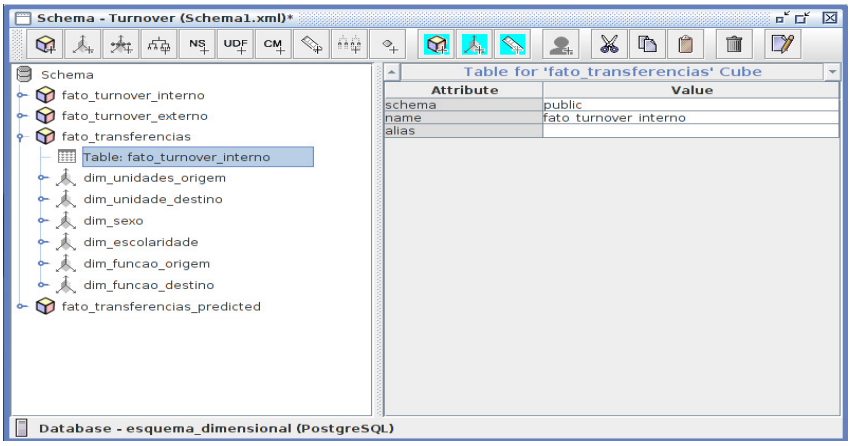


Figura 3.8 Criação do esquema dimensional através da ferramenta *schema workbench*.

A ferramenta *Schema Workbench* está incorporada na plataforma do Pentaho e proporciona a geração dos cubos de dados OLAP. Ela tem uma interface visual para navegar entre definições do cubo, permitindo criar métricas, dimensões e hierarquias, que proporcionam a correta utilização e exploração do cubo de dados OLAP.

Foram implementados quatros cubos de dados que representam de forma clara e concisa o setor estudado. Estes cubos consistem em uma camada lógica implementada acima do modelo físico do *PostgreSQL*.

Os cubos de dados (fato_turnover_interno, fato_tunvover_externo, fato_transferencias e fato_transferencias_predicted) implementados pela ferramenta *Schema Workbench* são salvos no formato XML e precisam ser publicados para que consultas analíticas sejam realizadas pela ferramenta OLAP Pentaho Analysis View. No Apêndice B é apresentado o arquivo *mondrian* gerado pela ferramenta.

3.3 EXPLORAÇÃO DE DADOS

A partir dos dados fornecidos pela organização em estudo, foi possível levantar o histórico de admissões e desligamentos da empresa nos últimos anos, conforme ilustra Figura 3.8. Segundo pesquisa divulgada em 2010 pela FEBRABAN (Federação Brasileira de Bancos), o percentual de rotatividade médio das instituições financeiras é de 10%, enquanto que no mercado formal é de 33%, o que demonstra o elevado tempo de permanência da maioria

dos bancários no setor. Na Figura 3.9 é possível observar que a Taxa de Desligamento média é de 3.4 estando portando abaixo da média de mercado.

	Ano				
Métricas	2008	2009	2010	2011	2012
Qtd. Empregados	75.171	78.823	81.934	83.260	88.620
Qtd. Desligados	2.614	1.939	4.029	2.440	3.750
Taxa Desligamento	3	2	5	3	4
Qtd. Admitidos	5.820	5.062	5.903	4.893	11.031
Turnover	6	4	6	4	8

Figura 3.9 Taxa de *Turnover* Externo

A Figura 3.10 ilustra as movimentações que ocorreram entre 2008 e 2012 cujo motivo de transferência se deu através de processo seletivo interno (por promoção) ou interesse da administração (movimentação por lateralidade). Neste caso, observou-se que o índice de rotatividade interna de pessoal é maior principalmente nos subsistemas Central e Negocial.

		Ano				
Unidades por Subsistema	Métricas	2008	2009	2010	2011	2012
⊞ CENTRAL	Qtd. Empregados	4.648	4.463	4.408	4.873	4.925
	Qtd. Admitidos	1.088	794	1.151	1.687	1.375
	Qtd. Desligados	781	470	872	1.404	949
	Taxa Desligamento	17	11	20	29	19
	Turnover	20	14	23	32	24
⊞ LOGISTICO	Qtd. Empregados	27.153	27.393	24.543	19.599	26.212
	Qtd. Admitidos	1.003	899	1.432	1.572	2.118
	Qtd. Desligados	1.097	948	5.363	773	913
	Taxa Desligamento	4	3	22	4	3
	Turnover	4	3	14	6	6
⊞ NEGOCIAL	Qtd. Empregados	43.370	46.967	52.983	58.788	57.483
	Qtd. Admitidos	5.429	5.660	6.714	9.289	14.453
	Qtd. Desligados	4.812	5.000	6.090	7.572	11.410
	Taxa Desligamento	11	11	11	13	20
	Turnover	12	11	12	14	22

Figura 3.10 Taxa de *Turnover* Interno por subsistema.

Através de consultas OLAP sobre a tabela fato empregados, foi possível verificar que a instituição possui em seu quadro de pessoal por volta de 54% de homens e 46% de mulheres (Figura 3.11).

	Ano/Mês	
Sexo	⊞ 2011	⊞ 2012
Feminino	42.770	44.033
Masculino	51.759	52.510

Slicer: [Measure=Qtd. Empregados]

Figura 3.11 Quantidade de empregados por sexo.

	Ano/Mês			
	⊞ 2011		⊞ 2012	
	Sexo		Sexo	
Unidades por Subsistema	Feminino	Masculino	Feminino	Masculino
⊞ CENTRAL	2.674	2.655	2.780	2.757
⊞ LOGISTICO	12.651	15.449	13.545	16.407
⊞ NEGOCIAL	27.627	33.847	29.529	35.597

Slicer: [Measure=Qtd. Empregados]

Figura 3.12 Quantidade de empregados por subsistema.

Conforme a Figura 3.12, existe um equilíbrio no número de homens e mulheres no subsistema Central. Já nos demais susbsistema existe mais homens que mulheres.

	Ano/Mês.Mensal				
Sexos	⊞ 2008	⊞ 2009	⊞ 2010	⊞ 2011	⊞ 2012
Feminino	2.942	2.761	3.624	4.727	6.475
Masculino	3.883	3.889	4.855	6.479	9.127

Slicer: [Measure=Qtd. Empregados] [Ocorrencia=TRANSFERENCIA]

Figura 3.13 Transferências por sexo.

Na Figura 3.13 é possível demonstrar um número maior de transferências realizadas por homens durantes os anos 2008 e 2012. Esta proporção é reflexo do quadro de empregado ser maioria de homens, conforme ilustrado na Figura 3.11.

A Figura 3.14 mostra a correlação entre as transferências entre os subsistemas. É possível observar que o número de transferências vem aumentado a cada ano, muito em virtude do crescimento do número de empregados e de novas unidades operacionais. Observa-se que a maioria das transferências dos subsistemas Central e Negocial, 60% e 95% respectivamente, ocorrem dentro do próprio subsistema. Já no subsistema Logístico, este

percentual diminui, visto que ocorrem muitas transferências vindas do subsistema Negocial.

		Unidades (Origem).Unidades por Subsistema		
Ano/Mês.Mensal	Unidades (Destino).Unidades por Subsistema	CENTRAL	LOGISTICO	NEGOCIAL
2008	CENTRAL	566	284	168
	LOGISTICO	73	504	391
	NEGOCIAL	14	201	4.734
2009	CENTRAL	434	213	93
	LOGISTICO	94	390	381
	NEGOCIAL	13	333	4.776
2010	CENTRAL	684	269	131
	LOGISTICO	152	768	474
	NEGOCIAL	15	162	5.939
2011	CENTRAL	923	347	275
	LOGISTICO	125	693	720
	NEGOCIAL	27	140	8.131
2012	CENTRAL	715	378	204
	LOGISTICO	118	678	1.292
	NEGOCIAL	31	463	11.937

Slicer: [Measure=Qtd. Empregados] [Ocorrencia=TRANSFERENCIA]

Figura 3.14 Tabulação Cruzada: Subsistema Origem x Subsistema Destino.

Também na tabulação cruzada entre Tipo de Função Gratificada Origem x Destino e ilustrada na Figura 3.15, é possível observar que os números aumentam a cada ano. Um comportamento que se destacar é o número de movimentações de empregados sem função gratificada, provavelmente porque estes não estão vinculados às unidades depois do estágio probatório, podendo ser transferidos em busca de promoções. Outro comportamento evidenciado é a quantidade de transferências cujo tipo de função é gerencial (chefia), indicando rotatividade entre gestores.

		Cargos Comissionados (Destino).Tipo Cargo Comissionado		
Ano/Mês.Mensal	Cargos Comissionados (Origem).Tipo Cargo Comissionado	Chefia	Sem Função Gratificada	Técnico
⊞ 2008	Chefia	2.241	357	211
	Sem Função Gratificada	176	1.402	690
	Técnico	488	575	978
⊞ 2009	Chefia	2.404	256	221
	Sem Função Gratificada	142	1.135	659
	Técnico	467	437	1.232
⊞ 2010	Chefia	2.890	229	307
	Sem Função Gratificada	211	971	1.065
	Técnico	837	384	1.993
⊞ 2011	Chefia	3.544	353	425
	Sem Função Gratificada	349	1.386	1.475
	Técnico	1.465	475	2.434
⊞ 2012	Chefia	5.248	421	566
	Sem Função Gratificada	411	2.311	2.123
	Técnico	1.877	571	3.187

Slicer: [Measure=Qtd. Empregados] [Ocorrencia=TRANSFERENCIA]

Figura 3.15 Tabulação Cruzada: Função Gratificada Origem X Destino.

Conforme Figura 3.16 que correlaciona Sexo com Faixa Etária, as transferências continuam refletindo a proporção de homens e mulheres por faixa etária, sendo que o número de transferências diminui com empregado com mais de 50 anos.

		Faixa Etária			
Ano/Mês.Mensal	Sexos	18-30	31-40	41-50	> 50
⊞ 2008	Feminino	988	694	1.194	73
	Masculino	1.146	878	1.616	255
⊞ 2009	Feminino	897	664	1.096	118
	Masculino	1.116	879	1.597	314
⊞ 2010	Feminino	1.239	951	1.263	181
	Masculino	1.402	1.274	1.788	403
⊞ 2011	Feminino	1.598	1.546	1.329	273
	Masculino	2.044	1.962	1.921	586
⊞ 2012	Feminino	2.207	2.380	1.540	380
	Masculino	3.049	3.117	2.176	836

Slicer: [Measure=Qtd. Empregados] [Ocorrencia=TRANSFERENCIA]

Figura 3.16 Transferências por Faixa Etária.

Analisando o tempo de permanência na unidade, Figura 3.17, foi possível definir que em média a rotatividade de empregados se dá por volta de 2,5 anos.

		Unidades (Origem).Unidades por Subsistema		
Ano/Mês.Mensal	Métricas	⊞CENTRAL	⊞LOGISTICO	⊞NEGOCIAL
⊞2010	Qtd. Empregados	842	1.193	6.520
	Idade Média	36,492	36,486	37,603
	Tempo Média Empresa	10,927	10,103	11,569
	Tempo Média Unidade	2,513	3,149	2,737
⊞2011	Qtd. Empregados	1.067	1.174	9.088
	Idade Média	37,81	36,452	36,648
	Tempo Média Empresa	12,89	10,314	10,018
	Tempo Média Unidade	2,625	2,818	2,505
⊞2012	Qtd. Empregados	861	1.513	13.344
	Idade Média	37,943	35,236	36,154
	Tempo Média Empresa	12,686	8,397	9,275
	Tempo Média Unidade	2,463	2,285	2,214

Slicer: [Ocorrencia=TRANSFERENCIA]

Figura 3.17 Tempo médio nas unidades.

Após esta fase de exploração de dados proporcionada pelas consultas OLAP através módulo *Pentaho Analysis View*, seguiram-se então as etapas de transformação de dados e Mineração de Dados propriamente dita.

3.4 PREPARANDO OS DADOS

Embora a etapa de preparação de dados do processo de descoberta de conhecimento usualmente consumir muito esforço, durante a construção do DW, as tarefas de limpeza e integração de dados já foram realizadas, encurtando assim o tempo que seria necessário para tal tarefa. No entanto, outras tarefas foram necessárias, tais como, a seleção de características, discretização de dados e exportação dos dados em arquivos ARFF (*attribute-relation file format*). O formato ARFF é uma forma padrão de representar conjunto de dados que consistem em instâncias independentes, não ordenadas e que não possuem relacionamentos entre si. A Figura 3.18 ilustra um exemplo de conjunto de dados no formato ARFF.

```
% ARFF file for the weather data with some numeric features
%
@relation weather

@attribute outlook { sunny, overcast, rainy }
@attribute temperature numeric
@attribute humidity numeric
@attribute windy { true, false }
@attribute play? { yes, no }

@data
%
% 14 instances
%
sunny, 85, 85, false, no
sunny, 80, 90, true, no
overcast, 83, 86, false, yes
rainy, 70, 96, false, yes
rainy, 68, 80, false, yes
rainy, 65, 70, true, no
overcast, 64, 65, true, yes
sunny, 72, 95, false, no
sunny, 69, 70, false, yes
rainy, 75, 80, false, yes
sunny, 75, 70, true, yes
overcast, 72, 90, true, yes
overcast, 81, 75, false, yes
rainy, 71, 91, true, no
```

Figura 3.18 Exemplo de conjunto de dados no formato ARFF.

A estrutura do arquivo ARFF é comporta de três partes: Relação, Atributos e Dados. A relação é a primeira linha do arquivo, e deve conter a palavra reservada @*relation* seguida de uma palavra-chave que identifique a tabela/relação ou a tarefa que está sendo analisada. Os atributos formam um conjunto de linhas onde cada inicia com a palavra reservada @*attribute* seguida do nome do atributo e do seu tipo, que pode ser nominal ou numérico. A última parte do arquivo ARFF corresponde ao conjunto de instâncias de dados (@*data*), inseridos logo após a definição dos atributos.

Utilizando o PDI foi possível criar o conjunto de dados de treinamento no formato adequado para a utilização da ferramenta WEKA, conforme Figura 3.19. No Apêndice C é apresentado o arquivo ARFF gerado por esta transformação.

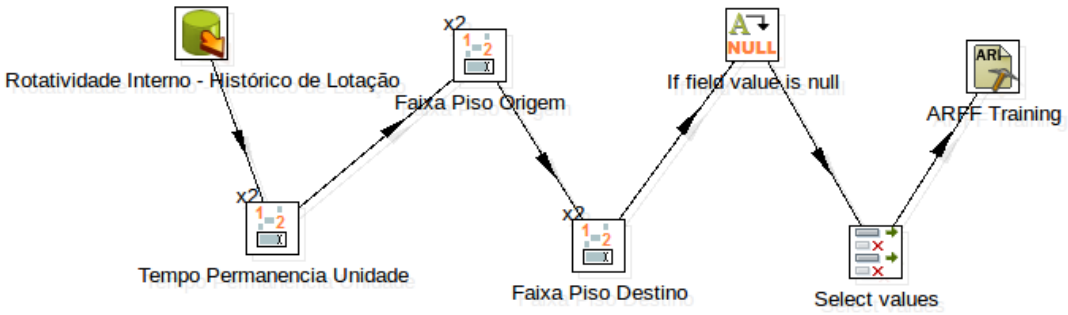


Figura 3.19 Processo ETL responsável por criar conjunto de treinamento no formato ARFF.

3.5 AGRUPANDO TRANSFERÊNCIAS

Esta seção apresenta os experimentos realizados e resultado obtidos com a aplicação do algoritmo K-*means* (método descrito de aprendizado não supervisionado), uma abordagem não hierárquica, com o objetivo de dividir as transferências em grupos com características similares, tendo como métrica de similaridade a distância Euclidiana. O algoritmo é implementado no WEKA com o nome de *Simple K-Means*.

Foram realizados três experimentos de acordo as necessidades da organização em estudo. O primeiro experimento consistiu em descrever as transferências que ocorreram em toda organização, permitido obter visão global. Em seguida procurou-se descrever as transferências cujo destino foi alguma unidade do subsistema Central, permitido avaliar o perfil dos empregados que se deslocam para a matriz. Por último foram descritas as transferências que ocorrem dentro do subsistema Negocial.

Primeiramente, o algoritmo K-*means* foi aplicado sobre todo o conjunto de transferências visando construir um modelo que segmentasse as transferências em cinco grupos (k = 5). A Figura 3.20 lista os cinco grupos ou *clusters* e seus respectivos centróides.

Como evidenciado na Figura 3.20, o algoritmo produziu cinco grupos e realizou quinze iterações até chegar ao resultado. A distorção média (*average within cluster sum of squared errors*) dentro dos grupos foi de 245.967 unidades. Os grupos e seus respectivos centróides para cada atributo são listados na forma de tabela. Os seguintes resultados poderam ser inferidos:

- Na maioria dos grupos as transferências são de homens da geração Y, com graduação e ocorrem dentro do subsistema Negocial.

- O grupo 0 é constituído de transferências de homens sem graduação e com 5 anos de empresa.

- O grupo 3 é constituído de transferências realizadas por mulheres da geração X que são promovidas geralmente para função de chefia.

```
=== Run information ===

Scheme:       weka.clusterers.SimpleKMeans -N 5 -A "weka.core.EuclideanDistance -R first-last" -I 500 -S 10
Relation:     turnover_interno
Instances:    54564
Attributes:   30
              co_sexo
              de_geracao
              subsistema_origem
              regiao_origem
              tipo_unidade_origem
              subsistema_destino
              regiao_destino
              tipo_unidade_destino
              tipo_funcao_origem
              tipo_funcao_destino
              nu_idade_desligamento
              formacao
              escolaridade
              temexperienciaexterna
              num_dep
              qt_tempo_empresa
              qt_tempo_unidade
              horas_treinamento
              piso_origem
              piso_destino
              tipo_transferencia
kMeans
======

Number of iterations: 15
Within cluster sum of squared errors: 245967.90753123735
Missing values globally replaced with mean/mode

Cluster centroids:
                                                    Cluster#
Attribute                               Full Data           0           1           2           3           4
                                          (54564)     (11730)     (10002)      (7297)     (15466)     (10069)
=============================================================================================================
co_sexo                                        M           M           M           M           F           M
de_geracao                           Geração Y   Geração Y      Boomers   Geração Y   Geração X   Geração Y
subsistema_origem                      NEGOCIAL    NEGOCIAL    NEGOCIAL     CENTRAL    NEGOCIAL    NEGOCIAL
regiao_origem                           Sudeste     Sudeste         Sul  Centro-Oeste     Sudeste         Sul
tipo_unidade_origem                    AGENCIA     AGENCIA     AGENCIA GERENCIA NACIONAL  AGENCIA     AGENCIA
subsistema_destino                     NEGOCIAL    NEGOCIAL    NEGOCIAL     CENTRAL    NEGOCIAL    NEGOCIAL
regiao_destino                          Sudeste     Sudeste         Sul  Centro-Oeste     Sudeste         Sul
tipo_unidade_destino                   AGENCIA     AGENCIA     AGENCIA GERENCIA NACIONAL  AGENCIA     AGENCIA
tipo_funcao_origem                      CHEFIA    SEMFUNCAO      CHEFIA      TECNICO      CHEFIA      TECNICO
tipo_funcao_destino                     CHEFIA    SEMFUNCAO      CHEFIA      TECNICO      CHEFIA      CHEFIA
nu_idade_desligamento                   36.922      32.9147     46.9322     35.7462     37.5722     31.5002
formacao                       ADMINISTRACAO    EDUCACAO  ADMINISTRACAO  ADMINISTRACAO  ADMINISTRACAO  ADMINISTRACAO
escolaridade                       GRADUACAO  ENSINO MEDIO    GRADUACAO  POSGRADUACAO    GRADUACAO    GRADUACAO
temexperienciaexterna                    NAO         NAO         SIM         NAO         NAO         NAO
num_dep                               1.0439      0.5622      2.2908      0.8423      0.9606      0.6406
qt_tempo_empresa                     10.6082      5.1934     22.2544     10.2482     10.6034      5.6159
qt_tempo_unidade                      2.5114      1.8579      3.2436      2.5642      2.5514      2.4458
horas_treinamento                    220.6844    191.7348    196.0916    214.239     210.4907    299.167
piso_origem                        5162.2156   2646.7404   7786.0401   5524.5594   6007.8535   3924.7908
piso_destino                       5609.0086   2831.0263   7551.7627   5910.7073   6505.5247   5319.7341
tipo_transferencia                   PROMOCAO  TRANSFSEMFUNCAO  LATERALIDADE     PROMOCAO     PROMOCAO     PROMOCAO

Time taken to build model (full training data) : 3.78 seconds

=== Model and evaluation on training set ===

Clustered Instances

0      13346 ( 24%)
1      10606 ( 19%)
2      11944 ( 22%)
3      11570 ( 21%)
4       7098 ( 13%)
```

Figura 3.20 Modelo gerado a partir de todas as transferências.

3.6 UMA NOVA DIMENSÃO DO CONHECIMENTO

Após análise dos resultados gerados pelo algoritmo *K-means* surgiu a necessidade de melhor entender os grupos encontrados. Para isto foi criado um novo tipo de dimensão que auxiliasse neste processo. Trata-se da dimensão *Cluster* que indica a qual grupo cada transformação pertence. A Figura 3.21 ilustra o novo modelo dimensional criado para detalhar cada grupo de transferências.

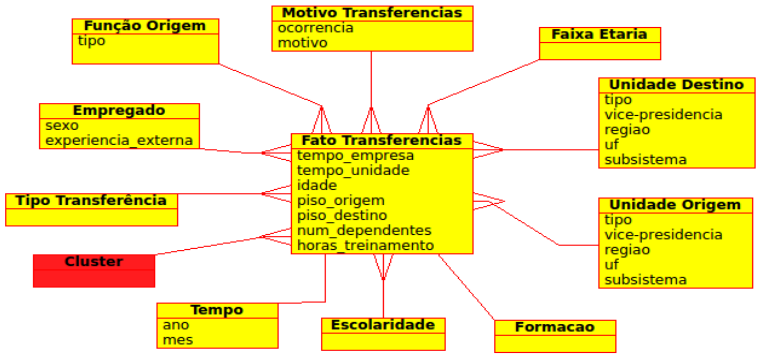


Figura 3.21 Modelo dimensional com um novo tipo dimensão do conhecimento. Tabela Fato Transferências X *Cluters Predicted.*

A criação da dimensão *Cluster*, além de possibilitar uma melhor caracterização de cada grupo de transferências, proporciona uma forma mais amigável e intuitiva para que analistas de negócio interpretem os resultados da mineração de dados. Desta forma, utilizando tecnologias OLAP, o analista poderia visualizar os dados em diversas dimensões.

A Figura 3.22 ilustra o processo ETL, criado na ferramenta *Pentaho Data Integration*, de segmentação das transferências e carga da tabela fato Transferências X *Cluters Predicted.* O passo *KMeans* (*Weka Scoring Step*) da transformação ilustrada na Figura 3.22, consiste na etapa responsável por aplicar o modelo gerado pelo algoritmo K-*means* a cada transferência. Desta forma, cada transformação é classificada em um dos 5 grupos.

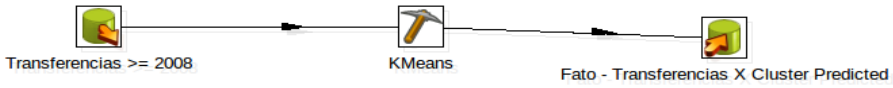


Figura 3.22 Transformações com K-*Means* para classificar as transferências.

4 - ANÁLISES E RESULTADOS

Este capítulo tem como objetivo apresentar os experimentos e análises realizados através da aplicação de técnicas de Mineração de Dados na caracterização da rotatividade interna de pessoal.

4.1.1 Utilizando a dimensão do conhecimento

As Figuras 4.23, 4.24, 4.25, 4.26, 4.27, 4.28, 4.29 e 4.30 ilustram algumas das consultas OLAP utilizadas para descrever cada grupo identificado pelo algoritmo de agrupamento.

	Cluster				
Sexo	0	1	2	3	4
F	887	2.669	2.450	15.237	1.524
M	3.823	8.853	2.839	10.442	5.840

Figura 4.1 Distribuição das transferências por sexo em cada grupo.

De acordo com a Figura 4.23, os homens prevalem nos grupos 0, 1 e 4, o que confirma a leitura do modelo K-*means* (Figura 3.20). Já no grupo 2, apesar do modelo K-*means* indicar que prevalece homens, há um relativo equilíbrio entre homens e mulheres. Ou seja, com a criação da dimensão *Cluster* e uso de tecnologias OLAP foi possível melhorar o entendimento de cada grupo de transferências. Por fim no grupo 3, a maioria é de transferências realizadas por mulheres.

	Cluster				
Geracao	0	1	2	3	4
Boomers	419	9.127	1.172	3.969	154
Geração X	451	2.134	1.421	11.988	693
Geração Y	3.838	258	2.696	9.714	6.517
Veteranos	2	3		8	

Figura 4.2 Distribuição das transferências por geração em cada grupo.

No DW criado para análise da rotatividade de pessoal da organização em estudo, foi definida a dimensão Geração. Esta dimensão, proveniente da área de Gestão de Pessoas, classifica as pessoas de acordo com a data de nascimento. A geração Veteranos são os nascidos entre 1922 e 1945, a geração *Baby Boomers* nascidos entre 1945 e 1965, a geração X nascidos entre 1965 e 1977, e a geração Y são os nascidos entre 1977 e 2000.

Segundo os estudiosos, o comportamento de cada geração depende do momento socioeconômico e histórico em que ela se desenvolve.

De acordo com a Figura 4.24, no grupo 1 prevalece empregados da geração *Boomers*. Nos grupos 0, 2 e 4 a maioria são transferências de pessoas da geração Y. Já no grupo 3 prevalece a geração X.

Realizando a leitura conjunta das Figuras 4.23 e 4.34 foi possível entender que o grupo 1 é formado por transferências realizadas por pessoas do sexo masculino e da geração *Boomers*, por exemplo.

	Cluster				
Tipo Funcao Origem	0	1	2	3	4
⊞ CHEFIA	468	8.452	983	9.775	1.985
⊞ SEMFUNCAO	3.105	1.049	872	7.375	2.534
⊞ TECNICO	1.137	2.021	3.434	8.529	2.845

Figura 4.3 Distribuição das transferências por Tipo de Função Origem em cada grupo.

De acordo com a Figura 4.25, no grupo 1 as transferências são de empregados com função gerencial (chefia). Já no grupo 3 há um relativo equilíbrio entre os tipos de função de origem, ou seja, o tipo de função de origem não é relevante. Novamente é possível observar que esta leitura é mais detalhada e precisa que a leitura do modelo K-*means* (Figura 3.20) que aponta que no grupo 3 prevalece empregados com função gerencial. Já o grupo 4 é composto por transferências de empregados sem função ou com função técnica.

	Cluster				
Métricas	0	1	2	3	4
Qtd. Transferencias	4.710	11.522	5.289	25.679	7.364
Idade	29,899	46,907	35,518	36,068	29,777
Piso Origem	2.667,207	7.322,333	6.025,409	4.785,547	4.067,155
Piso Destino	2.817,113	7.149,463	6.410,502	5.399,698	5.139,407
Tempo Empresa	3,916	21,455	10,48	8,59	5,049
Tempo Unidade	1,665	3,274	2,489	2,45	2,094
Horas Treinamento	199,611	210,248	226,231	244,637	309,359

Figura 4.4 Médias dos atributos números de cada grupo.

Na Figura 4.26 são listadas as médias de alguns atributos numéricos das transferências. De acordo com esta visualização é possível verificar, por exemplo, que a média do tempo de permanência na unidade no grupo 3 é igual a 2,45 anos, ou seja, os empregados deste grupo foram transferidos após 2,45 anos na unidade.

	Cluster														
	0			1			2			3			4		
	Subsistema Destino			Subsistema Destino			Subsistema Destino			Subsistema Destino			Subsistema Destino		
Subsistema Origem	CENTRAL	LOGISTICO	NEGOCIAL	CENTRAL	LOGISTICO	NEGOCIAL	CENTRAL	LOGISTICO	NEGOCIAL	CENTRAL	LOGISTICO	NEGOCIAL	CENTRAL	LOGISTICO	NEGOCIAL
CENTRAL	7	3	6	137	89	34	3.381	382	15	126	87	37		2	8
LOGISTICO	14	179	187	102	655	220	1.127	204	8	241	1.682	685	13	408	206
NEGOCIAL	62	442	3.810	78	407	9.800	159	11	2	427	1.761	20.633	147	647	5.933

Figura 4.5 Análise de Tabulação Cruzada por Subsistema em cada grupo.

De acordo com a Figura 4.27, que ilustra uma análise de tabulação cruzada entre os atributos Subsistema Origem e Destino, no grupo 3 prevalencem as transferências dentro do subsistema Negocial. Já no grupo 2 prevalecem as transferências dentro do subsistema Central.

Fazendo a leitura conjunto das Figuras 4.24, 4.25, 4.26 e 4.27 é possível descrever o grupo 1 como sendo transferências de empregados da geração *Boomers*, com 21 anos de empresa, que são gestores de alguma unidade do subsistema Negocia e que permanecem em média 3 anos na unidade.

	Cluster				
Subsistema Origem	0	1	2	3	4
CENTRAL	16	260	3.778	250	10
LOGISTICO	380	977	1.339	2.608	627
NEGOCIAL	4.314	10.285	172	22.821	6.727

Figura 4.6 Distribuição das transferências por Subsistema Origem em cada grupo.

A Figura 4.28 representa a analise OLAP utilizada para visualizar a distrubuição das transferências por subsistema de origem. De acordo com esta visualização, nos grupos 0, 1, 3 e 4 prevalecem as transferências cujo subsistema de origem é o Negocial. O mesmo pode

ser observado na Figura 4.29, aonde somente no grupo 2 prevalece o subsistema Central como destino.

	Cluster				
Subsistema Destino	0	1	2	3	4
CENTRAL	83	317	4.667	794	160
LOGISTICO	624	1.151	597	3.530	1.057
NEGOCIAL	4.003	10.054	25	21.355	6.147

Figura 4.7 Distribuição das transferências por Subsistema Destino em cada grupo.

	Cluster				
Tipo Transferencia	0	1	2	3	4
DECESSO	207	2.088	883	2.784	438
LATERALIDADE	689	5.183	1.395	5.438	1.210
PERDEUFUNCAO	376	1.081	456	1.950	251
PROMOCAO	776	2.409	2.201	12.423	4.706
TRANSFSEMFUNCAO	2.662	761	354	3.084	759

Figura 4.8 Distribuição das transferências por Tipo Transferências em cada grupo.

De acordo com a Figura 4.30, o grupo 1 se caracteriza por transferência do tipo lateralidade – quando o empregado muda de unidade, mas permance na mesma função. Já o grupo 3 é caracterizado por promoções.

A Tabela 4.1 resume as características de cada grupo obtidas a partir de consultas OLAP sobre o cubo Transferências Seguimentadas (Fato Transferências x *Cluters Predicted).* Desta forma, pode-se inferir que o grupo 0 é caracterizado por transferências de empregado sem função gratificada, do sexo masculino e geração Y, lotados em alguma unidade do subsistema Negocial da região Sudeste, que não possuem experiência externa e com pouco tempo de empresa, que são transferidos sem função em média após 1,5 anos de unidade, por exemplo.

Tabela 4.1 Resumo das características dos grupos de transferências.

Características	**Cluster 0**	**Cluster 1**	**Cluster 2**	**Cluster 3**	**Cluster 4**
Percentual de transferências	8,63%	21,12%	9,69%	47,06%	13,05%
Sexo	Masculino	Masculino	Masculino e Feminino	Feminino	Masculino
Geração	Gestação Y	Boomers	Boomers, Geração X e	Geração X e Y	Geração Y

Características	Cluster 0	Cluster 1	Cluster 2	Cluster 3	Cluster 4
			Y		
Experiência Externa	Não	Sim	Não	Não	Não
Subsistema Origem	Negocial	Negocial	Central e Logístico	Negocial	Negocial
Subsistema Destino	Negocial	Negocial	Central e Logístico	Negocial	Negocial
Tipo Função Origem	Sem função	Chefia	Técnico	Chefia, Técnico e Sem função	Sem função e Técnico
Tipo Função Destino	Sem função	Chefia	Técnico	Chefia, Técnico e Sem função	Chefia e Técnico
Tempo Empresa	4	21	10	8	5
Tempo Unidade	1,5	3	2,5	2,5	2
Tipo Transferência	Transferência Sem Função	Lateralidade	Lateralidade e Promoção	Promoção	Promoção
Região Origem	Sudeste	Sudeste e Sul	Centro-Oeste	Sudeste	Sul
Região Destino	Sudeste	Sudeste e Sul	Centro-Oeste	Sudeste	Sul
Escolaridade	Ensino Médio	Graduação	Pós-Graduação	Graduação	Graduação

Em seguida, os mesmos passos acima foram executados, porém somente com as transferências cujo destino foi o subsistema Central e a origem os subsistemas Negocial ou Logístico. Desta vez o algoritmo foi configurado para extrai três grupos. A Figura 4.31 lista o resultado gerado pelo WEKA após executar o algoritmo K-*means*.

O objetivo deste experimento foi descrever as transferências ou o perfil dos empregados que saíam dos subsistemas Negocial e Logístico e se delocavam para alguma unidade da Matriz da empresa, subsistema Central. A importância deste tipo de análise para a empresa em estudo se deve ao fato de que as unidades do subsistema Central são áreas estratégicas aonde decisões importantes, que afetam toda a empresa, são tomadas.

```
=== Run information ===

Scheme:       weka.clusterers.SimpleKMeans -N 5 -A "weka.core.EuclideanDistance -R first-last" -I 500 -S 10
Relation:     turnover_interno_central
Instances:    54564
Attributes:   30
              co_sexo
              de_geracao
              subsistema_origem
              regiao_origem
              tipo_unidade_origem
              tipo_funcao_origem
              tipo_funcao_destino
              nu_idade_desligamento
              formacao
              escolaridade
              temexperienciaexterna
              num_dep
              qt_tempo_empresa
              qt_tempo_unidade
              horas_treinamento
              piso_origem
              piso_destino
              tipo_transferencia

kMeans
======

Number of iterations: 10
Within cluster sum of squared errors: 13480.323809855228
Missing values globally replaced with mean/mode

Cluster centroids:
                                                              Cluster#
Attribute                           Full Data                        0                        1                        2
                                       (2370)                    (602)                    (827)                    (941)
==========================================================================================================================
co_sexo                                     M                        M                        F                        M
de_geracao                          Geração Y                  Boomers                Geração Y                Geração Y
subsistema_origem                   LOGISTICO                LOGISTICO                LOGISTICO                 NEGOCIAL
regiao_origem                    Centro-Oeste             Centro-Oeste             Centro-Oeste             Centro-Oeste
tipo_unidade_origem        GERENCIA DE FILIAL       GERENCIA DE FILIAL       GERENCIA DE FILIAL                  AGENCIA
tipo_funcao_origem                    TECNICO                  TECNICO                SEMFUNCAO                  TECNICO
tipo_funcao_destino                   TECNICO                  TECNICO                  TECNICO                  TECNICO
nu_idade_desligamento                 33.4785                  42.5066                  29.8585                  30.8842
formacao                        ADMINISTRACAO            ADMINISTRACAO                 EDUCACAO            ADMINISTRACAO
escolaridade                        GRADUACAO             POSGRADUACAO                GRADUACAO                GRADUACAO
temexperienciaexterna                     NAO                      SIM                      NAO                      NAO
num_dep                                0.6865                   1.4551                   0.3652                   0.4772
qt_tempo_empresa                       7.8671                  16.4894                    4.529                   5.2848
horas_treinamento                    260.9452                 253.5364                 224.5244                 297.6935
piso_origem                         4323.4772                6778.3588                3470.3186                3502.7794
piso_destino                        5093.6078                8269.0027                3961.9509                4056.7242
tipo_transferencia                   PROMOCAO                 PROMOCAO                 PROMOCAO                 PROMOCAO




Time taken to build model (full training data) : 2.78 seconds

=== Model and evaluation on training set ===

Clustered Instances

0      602 ( 25%)
1      827 ( 35%)
2      941 ( 40%)
```

Figura 4.9 Modelo de transferência do Sistema Geral

Como evidenciado na Figura 4.31, o algoritmo produziu três grupos e realizou dez iterações até chegar ao resultado. A distorção média (*average within cluster sum of squared errors*) dentro dos grupos foi de 13488 unidades. Os grupos e seus respectivos centróides para cada atributo são listados na forma de tabela e os seguintes resultados podem ser inferidos:

- Na maioria dos grupos as transferências são de homens com graduação e originárias do subsistema Logístico.
- O grupo 0 é constituído de transferências de homens pós-graduados , com idade média de 42 anos, que se exercem alguma função técnica;

- O grupo 1 é constituído de transferências realizadas por mulheres da geração Y que são promovidas;

- O grupo 2 é constituído por transferências de homens da geração Y que são promovidos para algum função técnica.

As Figuras 4.32, 4.33, 4.34, 4.35 e 4.36 ilustram as consultas OLAP utilizadas para melhor caracterizar e interpretar cada grupo.

	Cluster		
Sexo	0	1	2
F	212	505	443
M	450	141	619

Figura 4.10 Distribuição por sexo de transferências para subsistema Central.

De acordo com a visualização ilustrada na Figura 4.32, o grupo 0 é caracterizado por transferências de homens; o grupo 1 é caracterizado por transferências realizadas por mulheres; e no grupo 2 existe uma relativa equiparação de transferências de ambos os sexos.

	Cluster		
Geracao	0	1	2
Boomers	311	10	52
Geração X	250	144	230
Geração Y	101	492	780

Figura 4.11 Distribuição por geração de transferências para subsistema Central.

De acordo com a Figura 4.33, o grupo 0 consiste de transferências de empregados da Geração *Boomers* e Geração X. Já os grupos 1 e 2 consistem de transferências realizadas pela Geação Y. Analisando ambas as Figuras 4.32 e 4.33, infere-se, por exemplo, que o grupo 1 é caracterizado por transferências realizadas por empregados do sexo feminino da geração Y.

	Cluster		
Regiao Origem	0	1	2
⊞ Centro-Oeste	369	494	727
⊞ Nordeste	54	22	20
⊞ Norte	7	7	14
⊞ Sudeste	162	104	256
⊞ Sul	70	19	45

Figura 4.12 Distribuição por região de origem de transferências para subsistema Central.

Na Figura 4.34 é apresentado o resultado de consulta OLAP que distribui as transferências, cujo destino é o subsistema Central, por região. É possível constatar que nos 3 grupos prevalece as transferências provenientes da região Centro-Oeste.

	Cluster		
Subsistema Origem	0	1	2
LOGISTICO	573	646	278
NEGOCIAL	89		784

Figura 4.13 Distribuição por subsistema origem de transferências para subsistema Central.

De acordo com a Figura 4.35, os grupos 0 e 1 são consistituídos de transferências cuja origem é o subsistema Logistico. Já o grupo 2 é composto por transferências cuja origem é o subsistema Negocial.

	Cluster		
Tipo Funcao Origem	0	1	2
⊞ CHEFIA	239	29	118
⊞ SEMFUNCAO	87	268	464
⊞ TECNICO	336	349	480

Figura 4.14 Distribuição por tipo de função origem.

Por fim, de acordo com a Figura 4.36, é possível perceber que o grupo 0 é caracterizado por transferências de empregados que já possuem função gratificadas, seja gerencia ou técnica. Já nos demais grupos prevalecem transferências realizadas por empregados com função técnica e sem função.

Em resumo, as características observadas para cada grupo são as listadas na Tabela 4.2. Desta forma, pode-se inferir, por exemplo, que o grupo 0 é formado por transferências realizadas por empregados com função gratificada do sexo masculino, das gerações *Boomers* e X, e que foram promovidos.

Tabela 4.2 Características de cada grupo de transferências cujo destino foi o subsistema Central e origem os subsistema Negocial e Logístico.

Características	**Cluster 0**	**Cluster 1**	**Cluster 2**
Percentual de transferências	28%	27%	45%
Sexo	Masculino	Masculino	Masculino e Feminino
Geração	Boomers e Geração X	Geração Y	Geração Y
Experiência Externa	Não e Sim	Não	Não
Subsistema Origem	Logístico	100% Logístico	Negocial
Subsistema Destino	Central	Central	Central
Tipo Função Origem	Chefia e Técnico	Técnico e Sem Função	Técnico e Sem Função
Tipo Função Destino	Chefia e Técnico	Chefia	Técnico
Tempo Empresa	15	4,5	5
Tempo Unidade	3,7	2,4	2,4
Tipo Transferência	Promoção	Promoção	Promoção
Região Origem	Centro-Oeste	Centro-Oeste	Centro-Oeste
Região Destino	Centro-Oeste	Centro-Oeste	Centro-Oeste

Por último, foram analisadas somente transferências que ocorreram entre unidades do subsistema Negocial. O algoritmo foi configurado para particionar os dados em cinco grupos, conforme Figura 4.37. O objetivo deste experimento é entender as condições e características das transferências ocorridas dentro do subsistema que possuí maior número de transferências.

```
=== Run information ===

Scheme:       weka.clusterers.SimpleKMeans -N 5 -A "weka.core.EuclideanDistance -R first-last" -I 500 -S 10
Relation:     turnover_interno_negocial
Instances:    54564
Attributes:   30
              co_sexo
              de_geracao
              subsistema_origem
              regiao_origem
              regiao_destino
              tipo_unidade_origem
              tipo_funcao_origem
              tipo_funcao_destino
              nu_idade_desligamento
              formacao
              escolaridade
              temexperienciaexterna
              num_dep
              qt_tempo_empresa
              qt_tempo_unidade
              horas_treinamento
              tipo_transferencia


kMeans
======

Number of iterations: 12
Within cluster sum of squared errors: 142959.75359496163
Missing values globally replaced with mean/mode

Cluster centroids:
                                                 Cluster#
Attribute                               Full Data        0            1            2            3            4
                                          (40178)   (9168)       (7053)      (10794)       (7995)       (5168)
=============================================================================================================
co_sexo                                        M        M            M            F            M            M
de_geracao                           Geração Y  Geração X    Geração Y    Boomers      Geração Y    Geração Y
subsistema_origem                    NEGOCIAL   NEGOCIAL     NEGOCIAL     NEGOCIAL     NEGOCIAL     NEGOCIAL
regiao_origem                          Sudeste    Sudeste          Sul      Sudeste      Sudeste      Nordeste
subsistema_destino                   NEGOCIAL   NEGOCIAL     NEGOCIAL     NEGOCIAL     NEGOCIAL     NEGOCIAL
regiao_destino                         Sudeste    Sudeste          Sul      Sudeste      Sudeste      Nordeste
tipo_funcao_origem                     CHEFIA     TECNICO      CHEFIA       CHEFIA       SEMFUNCAO    SEMFUNCAO
tipo_funcao_destino                    CHEFIA     CHEFIA       CHEFIA       CHEFIA       TECNICO      SEMFUNCAO
nu_idade_desligamento                  37.2335    35.9729      36.4117      45.1638      30.4957      34.4512
formacao                       ADMINISTRACAO  ADMINISTRACAO  ADMINISTRACAO  ADMINISTRACAO  EDUCACAO  ADMINISTRACAO
escolaridade                       GRADUACAO  GRADUACAO    POSGRADUACAO GRADUACAO    ENSINO MEDIO GRADUACAO
temexperienciaexterna                    NAO        NAO          NAO          SIM          NAO          NAO
num_dep                                1.1083     1.0058       1.2393       1.774        0.4443       0.7479
qt_tempo_empresa                      10.8731     8.2586      11.5369      19.6634       4.5195       6.0749
qt_tempo_unidade                       2.4868     2.8062       2.3354       3.0011       1.9533       1.8778
horas_treinamento                    220.6388   244.0711     239.8541     183.4235     238.5511     202.8638
tipo_transferencia                   PROMOCAO   PROMOCAO     PROMOCAO     LATERALIDADE PROMOCAO TRANSFSEMFUNCAO

Time taken to build model (full training data) : 3.78 seconds

=== Model and evaluation on training set ===

Clustered Instances

0      9168 ( 23%)
1      7053 ( 17%)
2     10794 ( 27%)
3      7995 ( 20%)
4      5168 ( 13%)
```

Figura 4.15 Transferências cujo subsistema origem Negocial e Logístico.

Como evidenciado na Figura 4.37, o algoritmo produziu cinco grupos e realizou doze iterações até chegar ao resultado. A distorção média (*average within cluster sum of squared errors*) dentro dos grupos foi de 142959 unidades. Os grupos e seus respectivos centróides para cada atributo são listados na forma de tabela. Os seguintes resultados podem ser inferidos:

- Na maioria dos grupos as transferências são de homens da geração Y e que não possuem experiência externa.

- O grupo 0 é constituído de transferências de homens graduados que são promovidos para funções gerencias.

- O grupo 4 é constituído de transferências realizadas por homens sem função do subsistema Negocial e região Nordeste que foram transferidos por lateralidade.

Novamente as transferências foram segmentadas e utilizando-se a tecnologia OLAP foi possível detalhar cada grupo.

	Cluster				
Sexo	0	1	2	3	4
F	2.944	2.048	5.857	3.099	2.064
M	6.239	5.005	4.936	4.900	3.086

Figura 4.16 Distribuição por sexo de transferências do subsistema Negocial.

De acordo com a Figura 4.38, as transferências realizadas por empregados do sexo masculino são a maioria nos grupos 0, 1, 3 e 4. Sendo que nos grupos 3 e 4 a diferença entre os sexos começa a dimunir.

	Cluster				
Geracao	0	1	2	3	4
Boomers	731	1.652	7.438	605	957
Geração X	5.613	1.825	2.558	1.227	1.225
Geração Y	2.837	3.576	796	6.165	2.962
Veteranos	2		1	2	6

Figura 4.17 Distribuição por geração de transferências do subsistema Negocial.

A Figura 4.39 ilustra a distribuição das transferências ocorridas dentro do subsistema Negocial por geração. É possível constatar que somente no grupo 2 a geração *Boomers* prevalece. Nos demais grupos as gerações X e Y são maioria. Desta forma é possível inferir que empregados mais novos se movimentam com maior frequência entre unidade o subsistema Negocial.

	Cluster				
Escolaridade	0	1	2	3	4
ENSINO MEDIO	638	416	843	5.241	859
GRADUACAO	6.982	2.213	6.921	2.014	3.625
POSGRADUACAO	1.563	4.424	3.029	744	666

Figura 4.18 Distribuição por escolaridade de transferências do subsistema Negocial.

De acordo com a Figura 4.40, com exceção do grupo 3 cuja maioria dos empregados só possui Ensino Médio, todos as transferências são realizadas por pessoas com nível superior.

	Cluster				
Tipo Funcao Origem	0	1	2	3	4
⊞ CHEFIA	3.326	5.101	9.306	824	761
⊞ SEMFUNCAO	889	537	476	4.878	3.552
⊞ TECNICO	4.968	1.415	1.011	2.297	837

Figura 4.19 Distribuição por tipo função origem de transferências do subsistema Negocial.

A Figura 4.41 evidência que os grupos 3 e 4 são caracterizados por transferências de empregados sem função gratificada, e que os grupos 1 e 2 são caracterizados por transferências de empregados com função gerencial.

Em resumo, as características observadas para cada grupo são as listadas na Tabela 4.3. Desta forma é possível inferir que o grupo 0 é caracterizado por transferências realizadas por empregados masculino, com função gratificada, da geração X, com graduação, com 8 anos de empresa e que permanecem em média 2,8 anos na unidade, por exemplo.

Tabela 4.3 Características de cada grupo de transferências cujo destino foi o subsistema Central e origem os subsistema Negocial e Logístico.

Características	**Cluster 0**	**Cluster 1**	**Cluster 2**	**Cluster 3**	**Cluster 4**
Percentual de transferências	23%	18%	27%	20%	13%
Sexo	Masculino	Masculino	Feminino	Masculino e Feminino	Masculino Feminino
Geração	Gestação X	Geração Y	Boomers	Geração Y	Geração Y
Experiência Externa	Não	Não	Sim	Não	Não
Subsistema	Negocial	Negocial	Negocial	Negocial	Negocial

Origem					
Subsistema Destino	Negocial	Negocial	Negocial	Negocial	Negocial
Tipo Função Origem	Chefia e Técnico	Chefia	Chefia	Sem função	Sem função
Tipo Função Destino	Chefia	Chefia	Chefia	Técnico	Sem função
Tempo Empresa	8	11,5	19,6	4	6
Tempo Unidade	2,8	2,3	3	1,9	1,8
Tipo Transferência	Promoção	Promoção	Lateralidade	Promoção	Lateralidade
Região Origem	Sudeste	Sul	Sudeste	Sudeste	Nordeste
Região Destino	Sudeste	Sul	Sudeste	Sudeste	Nordeste
Escolaridade	Graduação	Pós-graduação	Graduação	Ensio Médio	Graduação

4.2 CARACTERIZAÇÃO ATRÁVES DA INDUÇÃO DE REGRAS DE ASSOCIAÇÃO

Na seção anterior, foi utilizado o algoritmo K-*means* a fim de classificar as transferências em grupos com características similares. Nesta seção, dando continuidade a abordagem em cascata, foi empregada a técnica de mineração de regras de associação a fim de identificar os perfis das transferências em cada grupo.

Como forma de demonstração, foram extraídas as regras de associação somente do grupo 3 - *cluster-3* - do primeiro experimento – todas as transferências ocorridas entre 2008 e 2012.

Para esta tarefa somente foram selecionados os atributos: sexo, subsistema (origem e destino), região (origem e destino), número de dependentes, escolaridade, formação, experiência externa, tempo de empresa, tempo na unidade e tipo de função (origem e destino). Os demais atributos foram desconsiderados porque possuíam algum tipo de dependência em relação aos selecionados, o que prejudica a aplicação do método. Por exemplo, UF está associada à Região, logo o algoritmo evidenciaria uma relação que não seria interessante, visto que já é conhecida.

Dado a limitação do algoritmo *Apriori* em trabalhar dados numéricos, foi necessário realizar a discretização dos atributos: idade, tempo de empresa, tempo de unidade e número de dependentes, conforme segue na Tabela 4.4:

Tabela 4.4 Discretização de dados

Atributo	Intervalos	Método
Idade	x <= 31.5 31.5 <= x < 39.5 x >= 39.5	Filtro *Discretize* do próprio WEKA
Número de dependentes	X <= 0.5 0.5 <= x < 1.5 X >= 1.5	Filtro *Discretize* do próprio WEKA
Tempo de Unidade	0 <= x < 6 6 <= x < 10 10 <= x < 15 15 <= x < 20 20 <= x < 25 25 <= x < 30 x >= 30	
Tempo de Empresa	x < 2 x >= 2	

Para identificar as regras de associação, o WEKA foi configurado com um suporte mínimo de 50% e uma confiança mínima de 60%. Ao executar o algoritmo *Apriori* sobre os dados do *cluster-3*, o *software* WEKA gerou as 100 melhores regras de associação. Ao diminuir o valor do suporte mínimo o algoritmo gera mais regras de associações, contudo, a confiança das regras tende a diminuir. O tempo de processamente do algoritmo Apriori não é informado pelo *software*.

A Tabela 4.5 ilustra algumas regras de associação que caracterizam o *cluster*-3, onde cada regra representa um perfil de transferências que foi dominante ou mais fortemente associada com o conjunto de instâncias do grupo.

Tabela 4.5 Resultado da Regra de Associação

	Regra de associação	*Suporte*	*Confiança*
1	co_sexo=F 15237 ==> subsistema_origem=NEGOCIAL 13338	52%	88%
2	subsistema_origem=NEGOCIAL temexperienciaexterna=NAO 18553 ==> subsistema_destino=NEGOCIAL 16640	65%	90%
3	subsistema_origem=NEGOCIAL 22821 ==> escolaridade=GRADUACAO 14171	55%	60%
4	num_dep='(-inf-0.5]' 15360 ==> subsistema_origem=NEGOCIAL 13578	53%	88%

5	qt_tempo_unidade2=<2 14297 ==> subsistema_origem=NEGOCIAL 12845	50%	100%
6	subsistema_origem=NEGOCIAL regiao_origem=Sudeste subsistema_destino=NEGOCIAL 15433 ==> regiao_destino=Sudeste 15373	60%	100%
7	escolaridade=GRADUACAO 15830 ==> temexperienciaexterna=NAO 13016	51%	82%
8	subsistema_origem=NEGOCIAL 22821 ==> temexperienciaexterna=NAO 18553	72%	81%

Através da regra 8 pode-se dizer com 81% de acerto que no *cluster-3* as transferências são de empregado do subsistema Negocial que não possuem experiência externo. Da mesma forma, através da regra 2 pode-se dizer com 90% de acerto que empregados do subsistema Negocial que não possuem experiência externa se transferem para unidades do subsistema Negocial.

4.3 CONSTRUINDO O MODELO DE CLASSIFICAÇÃO

De acordo com a análise realização na fase de exploração de dados, foi construído um modelo que classificasse o tempo de permanência do empregado na unidade em: menor ou igual a dois anos (<=2) ou maior que dois anos (>2).

Para isto utilizou-se o algoritmo J48, implementação do *software* WEKA do algoritmo C4.5 *release* 8, que induz uma árvore de decisão. Para a construção do modelo de classificação foram selecionadas, como conjunto de dados de treinamento, todas as transferências de empregados admitidos a partir de 2008, totalizando mais de 34 mil tuplas.

Através do WEKA foi utilizado o método de seleção de atributos *Ranker* juntamente com o método de avaliação de atributos *InfoGainAttributeEval*, que selecionou os seguintes atributos como relevantes:

- Tempo de empresa;
- Rendimento;
- Tipo Função (da unidade de origem);
- Idade;
- Escolaridade.

Ao executar o algoritmo J48 com as instâncias oriundas do Cubo de Transferências, o *software* WEKA gerou as informações da mineração, conforme mostra Tabela 4.6. Ao todo foram utilizadas 34526 instâncias de treinamento para a classificação, sendo 26142 instâncias classificadas corretamente (taxa de acurácia 75,71%) e 8384 instâncias classificadas incorretamente (taxa de erro 24,29%). O tempo de processamento foi 0,39 segundos.

A taxa de acuraria de 75,71% indica uma boa precisão na classificação. A Matriz de Confusão produzida pelo algoritmo J48 mostra que das 29208 instâncias classificadas como da classe <= 2, 6554 foram classificadas incorretamente e 22654 foram classificadas corretamente. E das 5318 instâncias classificadas com da class > 2, 1830 forma classificadas incorretamente e 3488 classificadas corretamente.

Tabela 4.6 Árvore de Decisão gerada pelo algoritmo J48.

```
=== Run information ===

Scheme:       weka.classifiers.trees.J48 -C 0.25 -M 200
Relation:turnover_interno-turnover_interno-
weka.filters.supervised.attribute.AttributeSelection-
Eweka.attributeSelection.InfoGainAttributeEval-Sweka.attributeSelection.Ranker -
T -1.7976931348623157E308 -N 4
Instances:    34526
Attributes:   6
                tipo_funcao_origem
                qt_tempo_ empresa
                nu_idade_desligamento
                escolaridade
                salario
                tempo_permanencia_unidade
Test mode:    10-fold cross-validation
=== Classifier model (full training set) ===
J48 pruned tree
------------------
tipo_funcao_origem = CHEFIA: <= 2 (7694.0/1736.0)
tipo_funcao_origem = SEMFUNCAO
|   qt_tempo_ empresa <= 6
|   |   qt_tempo_empresa <= 4.6: <= 2 (9706.0/1734.0)
|   |   qt_tempo_ empresa > 4.6
|   |   |   salario <= 5164.666667
|   |   |   |   salario <= 3653.916667
|   |   |   |   |   nu_idade_desligamento <= 41
|   |   |   |   |   |   escolaridade = ENSINOMEDIO
|   |   |   |   |   |   |   qt_tempo_ empresa <= 5.1: >2 (269.16/126.0)
|   |   |   |   |   |   |   qt_tempo_ empresa > 5.1: <= 2 (328.72/140.72)
|   |   |   |   |   |   escolaridade = GRADUACAO: <= 2 (980.16/441.44)
|   |   |   |   |   |   escolaridade = POSGRADUACAO: <= 2 (255.44/106.72)
|   |   |   |   |   nu_idade_desligamento > 41: >2 (315.72/144.72)
|   |   |   |   salario > 3653.916667: >2 (367.23/111.37)
|   |   |   salario > 5164.666667: <= 2 (464.56/179.09)
|   qt_tempo_ empresa > 6: >2 (4291.0/1426.0)
tipo_funcao_origem = TECNICO
|   qt_tempo_ empresa <= 7.1: <= 2 (6632.0/1006.0)
```

```
|   qt_tempo_ empresa > 7.1
|   |   salario <= 8034.571429: <= 2 (2753.51/971.4)
|   |   salario > 8034.571429: >2 (468.49/176.89)

Number of Leaves  :       13

Size of the tree :  23

Time taken to build model: 0.39 seconds

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

Correctly Classified Instances       26142               75.7169 %
Incorrectly Classified Instances      8384               24.2831 %
Kappa statistic                          0.3165
Mean absolute error                      0.3483
Root mean squared error                  0.4177
Relative absolute error                 84.4269 %
Root relative squared error             91.9798 %
Coverage of cases (0.95 level)         100      %
Mean rel. region size (0.95 level)     100      %
Total Number of Instances            34526

=== Detailed Accuracy By Class ===

TP Rate FP Rate Precision Recall F-Measure MCC ROC Area PRC Area Class
0.925    0.653    0.776    0.925   0.844   0.343  0.713 0.825     <= 2
0.347    0.075    0.656    0.347   0.454   0.343  0.713 0.536     >2

Weighted Avg.
0.757    0.485    0.741    0.757   0.731   0.343  0.713 0.741

=== Confusion Matrix ===

        a     b   <-- classified as
    22654  1830 |     a = <= 2
     6554  3488 |     b = >2
```

Partindo-se do nó raiz pode-se descrever o comportamento de todas as transferências. Fazendo uma breve análise da árvore de decisão gerada, pode-se concluir que a maioria das transferências de empregado com função gerencial ocorre com no máximo dois anos na unidade, assim como empregado sem função gratificada e idade maior que 41 anos tendem há permanecer mais tempo nas suas unidade. Pode-se também inferir que empregados com menos tempo de empresa (menos que 4.6) e sem função também tende a permanecer menos de dois anos na sua unidade de lotação, provavelmente porque estes buscam ascensão em outras unidades.

Um das principais vantagens do algoritmo J48 é a árvore de decisão fornecida graficamente pelo *software* WEKA, facilitando o entendimento e melhor análise dos resultados da Mineração de Dados.

5 - CONCLUSÕES

Nesta dissertação, foram propostas técnicas de mineração de dados e uma nova estrutura de análise multidimensional para a descoberta de conhecimentos acerca de rotatitividade de pessoas, partindo dos registros existentes em bases de dados de gestão de pessoal.

Com a revisão bibliográfica, puderam-se conhecer as técnicas de Mineração de Dados necessárias para descrever a rotatividade interna de pessoal, a fim de suportar as tomadas de decisão relativas à definição de políticas de pessoal. Foi definido um modelo para aplicação das técnicas estudadas e criado um novo tipo de dimensão voltada para o processo de descoberta do conhecimento nesse domínio.

A definição de um modelo de mineração para esse domínio constitui outra contribuição da dissertação, pois tal modelo determina os passos que devem ser realizados para obtenção dos resultados com sucesso.

Especificamente, modelo proposto articulou as técnicas de Mineração de Dados possibilitando a classificação de transferências de empregados entre unidades, o que pode ser utilizado no embasamento de políticas de seleção e promoção de pessoas.

Já a criação de um novo tipo de dimensão possibilitou uma análise mais detalhada dos resultados obtidos com as técnicas de Mineração de Dados; e se mostrou uma forma mais interessante e amigável para que os usuários, especialistas de negócios, interpretem os resultados.

Com base em tais contribuições, a criação de um módulo de suporte à decisão mostrou-se funcional quanto ao seu propósito, o que se confirmou após validação por um estudo de caso, facilitando e enriquecendo o processo de descoberta de conhecimento no domínio escolhido.

Quanto aos objetivos específicos, verificou-se que o *Data Warehouse* desenvolvido foi eficiente para aplicação das técnicas de Mineração de Dados, possibilitando também informações para tomada de decisões através da criação de relatórios com ferramentas apropriadas.

Por fim, vale notar que informações relevantes para ententer o fenômeno de rotatividade interna foram descobertas, podendo ser utilizadas como base para decisões estratégicas e melhorias no processo seletivo de pessoal interno das organizações.

5.1 TRABALHOS FUTUROS

Como sugestões para trabalhos futuros, os resultados alcançados permitem apontar nas seguintes direções:

- Identificação de outros indicadores:
    - Preditivos: Identificação de empregados com alto risco de transferência em um período determinado; Classificação de empregados em grupos de alto, médio e baixo risco de desligamento da unidade nos próximos 12 meses. Isto possibilitaria ao gestor tomar ações relacionas a retenção de talento e/ou gestão do conhecimento;
    - Descritivos: Identificação de padrões na trajetória profissional percorrida por empregados com função gerencial. Isto possibilitaria a construção de uma trilha de aprendizado que auxiliaria outros empregados que desejam assumir alguma função gerencial.
- Realização de estudo comparativo entre as técnicas de Mineração de Dados a fim de elencar o algoritmo de aprendizagem que melhor se adeque a esta pesquisa, baseando-se nas funcionalidades e desempenhos apresentados.
- O estudo de mecanismos inteligentes para detecção do tamanho da amostra de dados e dos parâmetros ideais para serem aplicados aos algoritmos de Mineração de Dados.

Desse modo, espera-se que este trabalho possa contribuir significativamente para o aumento da qualidade e eficiência na gestão de informações de apoio à decisão para a área de Gestão de Pessoas, visando o aperfeiçoamento das políticas de promoções e retenção de

talentos, não apenas pelos resultados já alcançados, mas também por abrir novas perspectivas de estudos nesse domínio.

REFERÊNCIAS BIBLIOGRÁFICAS

Agrawal, R, Mannila, H., Srikant, R., Toivonen, H., Verkamo, I. (1996). "Fast Discovery of Association Rules." In *Advances in Knowledge Discovery and Data Mining. Advances in Knowledge Discovery and Data Mining, AAAI/MIT Press.*

Bluedorn, A.C. (1982.) "A unified model of turnover from organizations." Human Relations, 35,p. 135-153.

Bispo, Patrícia. (2005). "A importância da gestão de turnover". In: *http://www.rh.com.br/Portal/Relacao_Trabalhista/Entrevista/3998/aimportancia-da-gestao-do-turnover.html. Acessado em Março de 2013.*

Chang, H. (2009). "Employee Turnover: A Novel Prediction Solution with Effective Feature Selection." , WSEAS Transactions on Information Science and Applications, vol. 6 n.3, p.417-426.

Chiavenato, I., (2001). "Advances and challenges in human resource management in the new millennium." Public Personnel Management, vol. 30, pp.17-26.

Chiavenato, I. (2009). "Gestão de Pessoas." In: *Terceira Edição, Rio de Janeiro: Campus*

Chiavenato, I. (2008). "Planejamento, Recrutamento e Seleção de Pessoal: Como agregar talentos à empresa." *Editora Manole, 7a edição.*

Claro, Roberto (2009). "Como calcular o turnover?" In: *http://www.rellaciona.com.br/blog2009/gestao/como-calcular-o-turnover. Acessado em Março de 2013.*

Creecy, L., Klenz, B. (2008). "Retention Analytics for Human Capital Management." In: *SAS Institute Inc., Cary, NC.*

J. P. C. L. da Costa, E. P. de Freitas, B. M. David, D. Amaral & R. T. de Sousa Jr. (2012). "Improved Blind Automatic Malicious Activity Detection in Honeypot Data," The International Conference on Forensic Computer Science (ICoFCS), Brasília, Brazil.

Fadzilah Siraj and Mansour Ali Abdoulha (2011). "Mining Enrollment Data Using Descriptive and Predictive Approaches." Knowledge-Oriented Applications in Data Mining. In: *http://www.intechopen.com/books/knowledge-oriented-applications-in-datamining/mining-enrollment-data-using-descriptive-and-predictive-approaches. Acessado em Janeiro de 2013.*

Farajian, M. A., Mohammadi, S.(2011). “Mining the Banking Customer Behavior using Clustering and Association Rules Methods.” In: *International Journal of Industrial Engineering& Production Research.*

Fayyad, U.M., G.Piatetsky–Shapiro, P.Smyth (1996). “Knowledge Discovery and Data Mining: Towards a Unifying Framework.” In: *Second International Conference on Knowledge Discovery and Data Mining (KDD-96).*

Fontana, A., Naldi, M. C. (2009). “Estudo de Comparação de Métodos para Estimação de Números de Grupos em Problemas de Agrupamento de Dados.” *In: Universidade de São Paulo. ISSN - 0103-2569*

Han, J.; Kamber, M. (2006). “Data Mining: Concepts and Techniques, 2nd edition.”

Heinrichs, J. H.; Lim, J. S. (2003). “Integrating web-based data mining tolls with business models for knowledge management.” In: *Decision Support Systems, v. 35, n. 1, p. 103-112.*

Inmon, W. H. (1997). “Como construir o Data Warehouse.” In: *2a edição. Rio de Janeiro: Campus.*

Inmon, W.H. (1994). “Using the data warehouse.” In: *John Wiley & Sons, Inc.*

Inmon, W.H.(1997). “Managing the data warehouse.” In: *John Wiley & Sons, Inc.*

Inmon, W.H. (2002). “Building the data warehouse. 3th edition.” In: *John Wiley and Sons, Inc.*

Inmon, W.H. (1996). “The Data Warehouse and Data Mining.” In: *Communcation of the ACM, Vol. 39, No. 11*

Kane-Sellers, M.L. (2006). “Voluntary Employee Turnover in the Industrial Distribution Sales Force: Conceptual Models and Implications”. In: *Review of the Electronic and Industrial Distribution Industries. Vol. 5, No. 1.*

Kimball, Ralph (1997). “Digging into data mining - your data warehouse is your data mining platform. DBMS and Internet System.”

Kimball, Ralph (2002) “The data warehouse toolkit: the complete guide to dimensional modeling.” In: *New York: John Wiley & Sons.*

Lacombe, F. (2005). “Recursos humanos: Princípios e tendências.” In: *Editora Saraiva, SP*

Naisbitt, J. (1982). “Megatrends: Ten new directions transforming our lives.”

Quinlan, J. Ross (1986). “Introduction of decision trees”. Machine Learning, *vol. 1, pp. 81-106.*

Quinlan, J. Ross (1993). “C4.5: Programs for machine learning.” In: *Morgan Kaufmann Publishers: San Mateo, USA. ISBN: 1-55860-238-0.*

Robbins, Stephen Paul (1999). “Comportamento Organizacional. 8ª ed.” In: Rio de Janeiro: Livros Técnicos e Científicos.

Sanches, André Rodrigo (2003). “Uma visão Geral sobre Mineração de Dados.” In: *Relatório de Estudo - Tópicos em Ciência da Computação, Dept. Ciência da Computação, Universidade de São Paulo - USP, São Paulo*.

Sebrae Nacional (2013). “Boa gestão resulta em sucesso no negócio.” In: *http://www.sebrae.com.br/momento/quero-melhorar-minha-empresa/entenda-os-caminhos/gestao-de-pessoas/bia-670-3-a-importancia-de-uma-boa-gestao-de-pessoas/BIA_6703. Acessado em Março de 2013.*

Shumway, Robert H.; Stoffer, David S (2011). “Time Series Analysis and its Aplications. With R Examples. Third Edition”

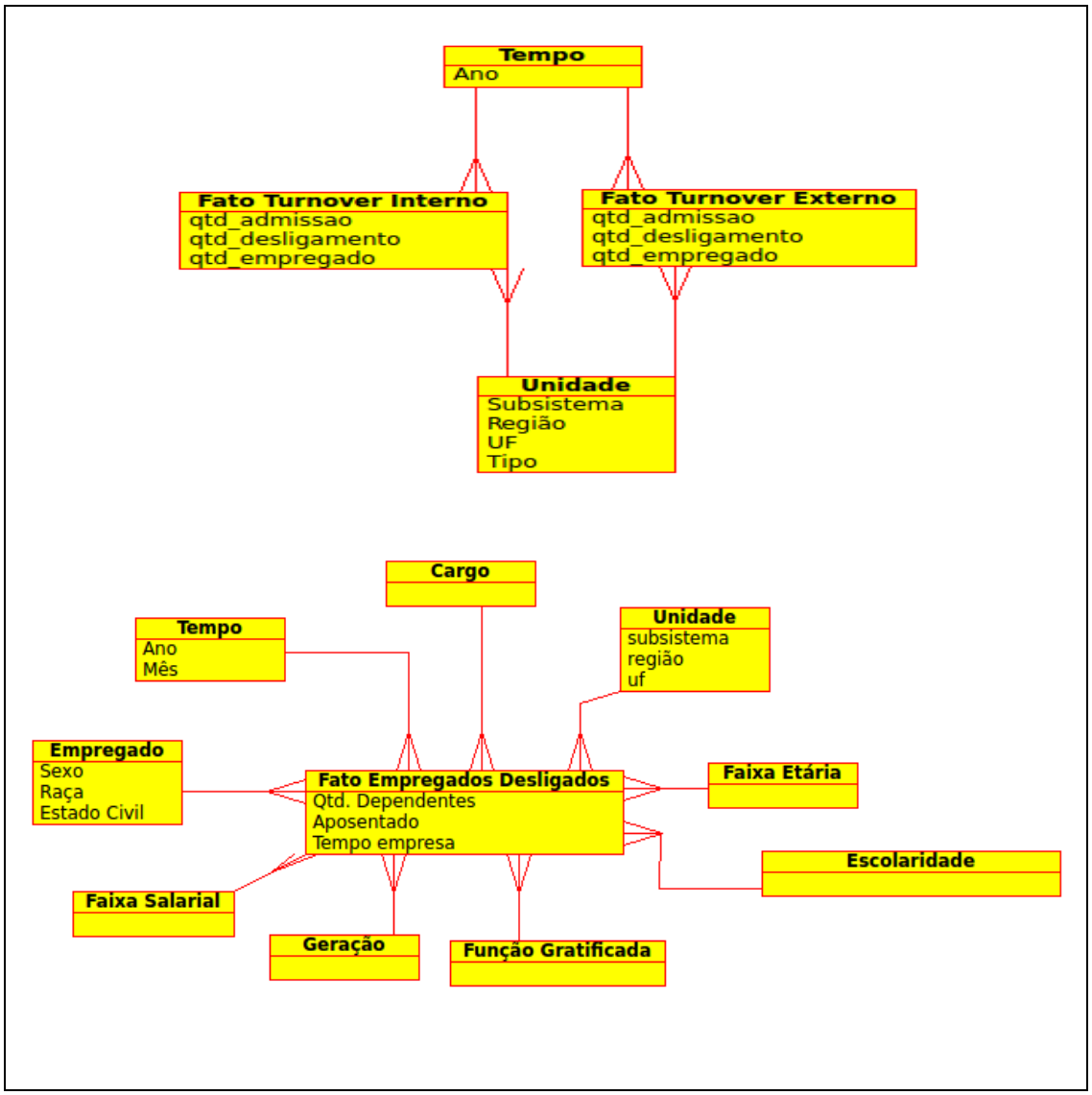
Steinbach, M., Karypis, G., and Kumar, V.(2000). “A comparison of document clustering techniques. KDD workshop on text mining.”

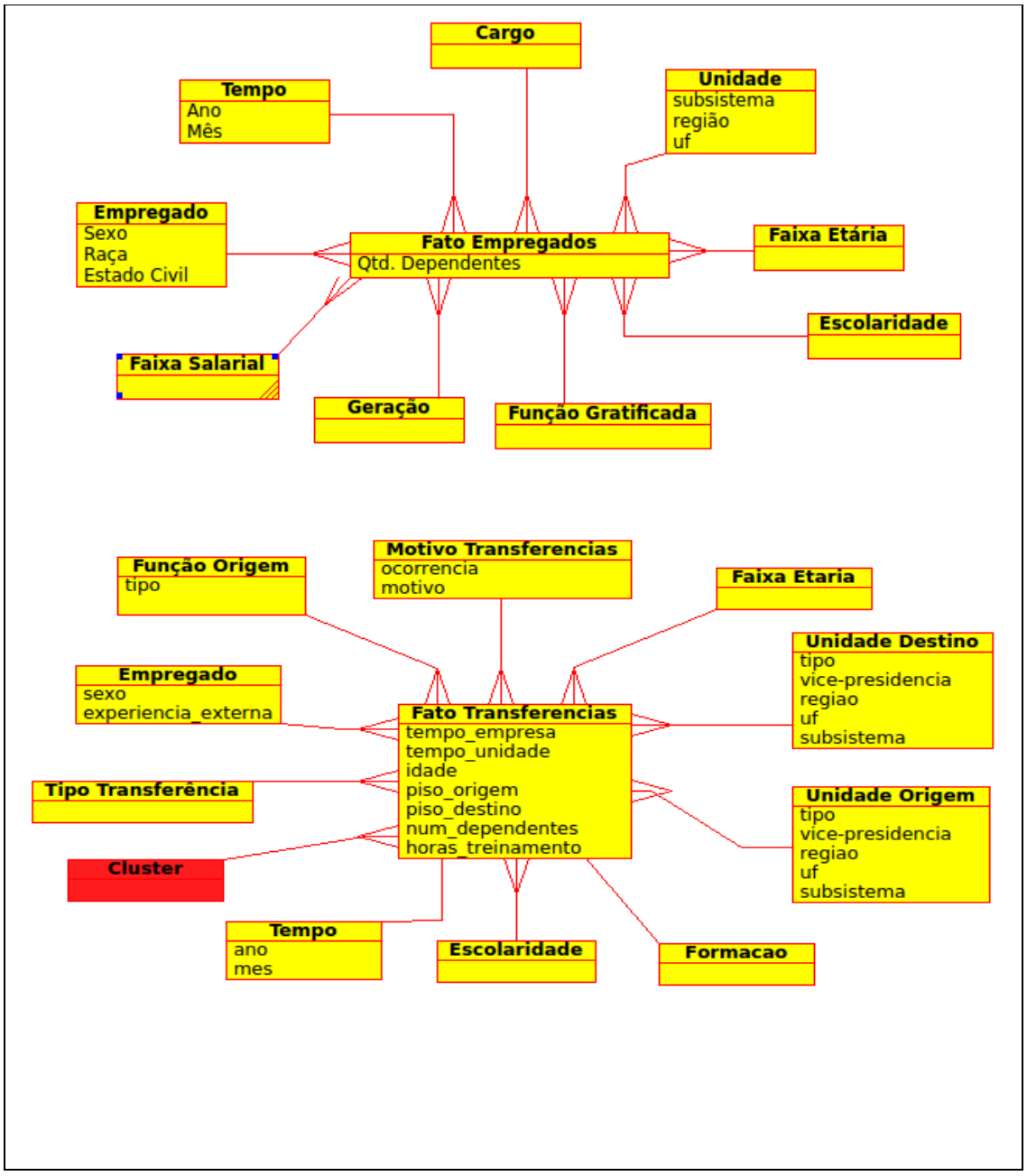
Tan, P., Steinbach, M., Kumar, V.(2009). “Introduction to DATAMINING.” In: *Addison-Wesley.*

Witten, I., Frank, E. (2005). “Data Mining – Pratical Machine Learning Tools and Techniques. 2nd edition.” In: *Elsevier, USA*

APÊNDICES

APÊNDICE A – MODELAGEM DIMENSIONAL DO ESQUEMA CONSTELAÇÃO DE FATOS DO DATA WAREHOUSE




Cargo
Tempo
Ano
Mês
Unidade
subsistema
região
uf
Empregado
Sexo
Raça
Estado Civil
Fato Empregados
Qtd. Dependentes
Faixa Etária
Escolaridade
Faixa Salarial
Geração
Função Gratificada
Função Origem
tipo
Motivo Transferencias
ocorrencia
motivo
Faixa Etaria
Unidade Destino
tipo
vice-presidencia
regiao
uf
subsistema
Empregado
sexo
experiencia_externa
Fato Transferencias
tempo_empresa
tempo_unidade
idade
piso_origem
piso_destino
num_dependentes
horas_treinamento
Tipo Transferência
Unidade Origem
tipo
vice-presidencia
regiao
uf
subsistema
Cluster
Tempo
ano
mes
Escolaridade
Formacao

APÊNDICE B – SCHEMA MONDRIAN GERADO PELO SCHEMA WORKBENCH

```
<Schema name="GENEC - Empregados" measuresCaption="M&#233;tricas">
<Dimension name="dim_sexo" caption="Sexos"  >
  <Hierarchy name="h_sexo" hasAll="true" primaryKey="sk_empregado"
primaryKeyTable="dim_empregado"
allMemberName="todos" allMemberCaption="Total Sexos" caption="Sexos">
    <Join leftKey="fk_sexo_raca_estado_civil"
rightKey="sk_sexo_raca_estado_civil">
      <Table name="dim_empregado"/>
      <Table name="dim_sexo_raca_estado_civil"/>
    </Join>
    <Level name="sexo" table="dim_sexo_raca_estado_civil" captionColumn="no_sexo"
nameColumn="co_sexo" ordinalColumn="no_sexo" column="co_sexo"
uniqueMembers="false" caption="Sexo"/>
  </Hierarchy>
</Dimension>


 <Dimension type="TimeDimension" highCardinality="false" name="dim_tempo_dia"
caption="Data">
    <Hierarchy name="h_default" hasAll="true" allMemberName="todos"
allMemberCaption="Total Anos" primaryKey="sk_tempo_dia" caption="Mensal">
     <Table name="dim_tempo_dia">
      </Table>
      <Level name="ano" column="ano" nameColumn="ano"  ordinalColumn="ano"
type="String" uniqueMembers="true" levelType="TimeYears" hideMemberIf="Never"
caption="Ano">
        <SQL dialect="generic">
            <![CDATA[( ano::varchar(4) )]]>
          </SQL>
      </Level>

      <Level name="mes" column="ds_mes" nameColumn="ds_mes"  ordinalColumn="mes"
type="String" uniqueMembers="false" levelType="TimeMonths" hideMemberIf="Never"
caption="M&#234;s" captionColumn="ds_mes">
      </Level>
    </Hierarchy>
    <Hierarchy name="h_trimestre" hasAll="true" allMemberName="todos"
allMemberCaption="Total Anos/Trimestre" primaryKey="sk_tempo_dia"
caption="Trimestral">
      <Table name="dim_tempo_dia">
      </Table>
      <Level name="ano" column="ano" nameColumn="ano"  ordinalColumn="ano"
type="String" uniqueMembers="true" levelType="TimeYears" hideMemberIf="Never"
caption="Ano">
        <SQL dialect="generic">
            <![CDATA[( ano::varchar(4) )]]>
          </SQL>
      </Level>
      <Level name="trimestre" column="ds_trimestre" ordinalColumn="ds_trimestre"
type="String" uniqueMembers="false"
              levelType="TimeQuarters" hideMemberIf="Never"
captionColumn="ds_trimestre" caption="Trimestre">
          <Annotations>
            <Annotation name="AnalyzerDateFormat">[yyyy].['QTR'q]</Annotation>
          </Annotations>


      </Level>
      <Level name="mes" column="ds_mes" nameColumn="ds_mes"  ordinalColumn="mes"
```

```
type="String" uniqueMembers="false" levelType="TimeMonths" hideMemberIf="Never"
caption="M&#234;s" captionColumn="ds_mes">
      </Level>
   </Hierarchy>
  </Dimension>
  <Dimension highCardinality="false" name="dim_area_conhecimento"
caption="&#193;reas de Conhecimento" >


    <Hierarchy name="h_default" hasAll="true" allMemberName="todos"
allMemberCaption="Total &#193;reas Conhecimentos"
primaryKey="sk_area_conhecimento" caption="&#193;rea Conhecimento">
      <Table name="dim_area_conhecimento">

      </Table>
      <Level name="area_conhecimento" column="no_area_conhecimento"
nameColumn="nu_area_conhecimento" ordinalColumn="no_area_conhecimento"
type="String" uniqueMembers="true" levelType="Regular" hideMemberIf="Never"
caption="&#193;rea Conhecimento" captionColumn="no_area_conhecimento">
      </Level>
      <Level name="area_concentracao" column="no_area_concentracao"
nameColumn="nu_area_concentracao" ordinalColumn="no_area_concentracao"
type="String" uniqueMembers="true" levelType="Regular" hideMemberIf="Never"
caption="&#193;rea Concentra&#231;&#227;o" captionColumn="no_area_concentracao">
      </Level>
    </Hierarchy>
  </Dimension>
  <Dimension visible="true" highCardinality="false" name="dim_raca"
caption="Ra&#231;a">
    <Hierarchy name="h_default" visible="true" hasAll="true"
allMemberName="total_racas" allMemberCaption="Total Ra&#231;as"
primaryKey="sk_sexo_raca_estado_civil" caption="Ra&#231;a">
      <Table name="dim_sexo_raca_estado_civil">
      </Table>
      <Level name="raca" visible="true" column="nu_raca" nameColumn="nu_raca"
type="String" uniqueMembers="true" levelType="Regular" hideMemberIf="Never"
caption="Ra&#231;a" captionColumn="no_raca">
      </Level>
    </Hierarchy>
  </Dimension>
  <Dimension visible="true" highCardinality="false" name="dim_sexo"
caption="Sexo">
    <Hierarchy name="h_default" visible="true" hasAll="true"
allMemberName="total_sexo" allMemberCaption="Total Sexos"
primaryKey="sk_sexo_raca_estado_civil" caption="Sexo">
      <Table name="dim_sexo_raca_estado_civil">
      </Table>
      <Level name="sexo" visible="true" column="co_sexo" nameColumn="co_sexo"
type="String" uniqueMembers="true" levelType="Regular" hideMemberIf="Never"
caption="Sexo" captionColumn="no_sexo">
      </Level>
    </Hierarchy>
  </Dimension>
  <Dimension visible="true" highCardinality="false" name="dim_mo_dis_fc"
caption="Motivo Dispensa">
    <Hierarchy name="h_default" visible="true" hasAll="true"
allMemberName="total_mod_dis_fc" allMemberCaption="Total Motivo"
primaryKey="sk_mo_dis_func" caption="Motivo Dispensa">


      <Table name="dim_mo_dis_fc">
      </Table>
      <Level name="mo_dis_func" visible="true" column="nu_mo_dis_func"
```

```
nameColumn="nu_mo_dis_func" captionColumn="no_mo_dis_func"  type="String"
uniqueMembers="true" levelType="Regular" hideMemberIf="Never" caption="Motivo
Dispensa">
      </Level>
    </Hierarchy>
  </Dimension>
  <Dimension visible="true" highCardinality="false" name="dim_mo_trans"
caption="Motivo Tranferencia">
    <Hierarchy name="h_default" visible="true" hasAll="true"
allMemberName="total_mo_trnas" allMemberCaption="Total Motivo Transferencia"
primaryKey="sk_mo_trans" caption="Motivo Transferencia">
      <Table name="dim_mo_trans">
      </Table>
     <Level name="mo_trans" visible="true" column="nu_mo_trans"
nameColumn="nu_mo_trans" captionColumn="no_mo_trans" type="String"
uniqueMembers="true" levelType="Regular" hideMemberIf="Never" caption="Motivo
Transferencia">
     </Level>
    </Hierarchy>
  </Dimension>
  <Dimension visible="true" highCardinality="false" name="dim_ocor_fun"
caption="Ocorrencia">
    <Hierarchy name="h_default" visible="true" hasAll="true"
allMemberName="total_ocor_fun" allMemberCaption="Total Ocorrencia"
primaryKey="sk_ocor" caption="Ocorrencia">
      <Table name="dim_ocor_fun">
      </Table>
      <Level name="ocor" visible="true" column="nu_ocor" nameColumn="nu_ocor"
captionColumn="no_ocor" type="String" uniqueMembers="true" levelType="Regular"
hideMemberIf="Never" caption="Ocorrencia">
      </Level>
    </Hierarchy>
  </Dimension>
  <Dimension visible="true" highCardinality="false" name="dim_estado_civil"
caption="Estado Civil">
    <Hierarchy name="h_default" visible="true" hasAll="true"
allMemberName="total_sexo" allMemberCaption="Total Sexos"
primaryKey="sk_sexo_raca_estado_civil" caption="Estado Civil">
      <Table name="dim_sexo_raca_estado_civil">
      </Table>
      <Level name="estado_civil" visible="true" column="co_estado_civil"
nameColumn="co_estado_civil" type="String" uniqueMembers="true"
levelType="Regular" hideMemberIf="Never" caption="Estado Civil"
captionColumn="no_estado_civil">
      </Level>
   </Hierarchy>
 </Dimension>

  <Dimension visible="true" highCardinality="false" name="dim_funcao"
caption="Cargos Comissionados">
    <Hierarchy name="h_default" visible="true" hasAll="true"
allMemberName="total_funcoes" allMemberCaption="Total Cargos Comissionados"
primaryKey="sk_funcao" caption="Cargo Comissionado">
      <Table name="dim_funcao">
      </Table>
      <Level name="nu_funcao" visible="true" column="nu_funcao"
nameColumn="nu_funcao" type="String" uniqueMembers="true" levelType="Regular"
hideMemberIf="Never" caption="Cargo Comissionado" captionColumn="no_funcao">


      </Level>
    </Hierarchy>
    <Hierarchy name="h_tipo_funcao" visible="true" hasAll="true"
allMemberName="total_funcoes" allMemberCaption="Total Cargos Comissionados"
```

```
primaryKey="sk_funcao" caption="Tipo de Cargo Comissionado">
      <Table name="dim_funcao">
      </Table>
      <Level name="nu_tipo_funcao" visible="true" column="no_tipo_funcao"
nameColumn="no_tipo_funcao" type="String" uniqueMembers="true"
levelType="Regular" hideMemberIf="Never" caption="Tipo Cargo Comissionado"
captionColumn="no_tipo_funcao">
      </Level>
      <Level name="nu_funcao" visible="true" column="nu_funcao"
nameColumn="nu_funcao" type="String" uniqueMembers="true" levelType="Regular"
hideMemberIf="Never" caption="Cargo Comissionado" captionColumn="no_funcao">


      </Level>
    </Hierarchy>
    <Hierarchy name="h_tipo_funcao2" visible="true" hasAll="true"
allMemberName="total_tipo_funcoes" allMemberCaption="Total tipo cargos
comissionados" primaryKey="sk_funcao" caption="Tipo Cargo Comissionado">
      <Table name="vw_dim_tipo_funcao">
      </Table>
      <Level name="funcao" visible="true" column="no_tipo_funcao"
nameColumn="no_tipo_funcao" type="String" uniqueMembers="true"
levelType="Regular" hideMemberIf="Never" caption="Tipo Cargo Comissionado"
captionColumn="no_tipo_funcao">
      </Level>
    </Hierarchy>
  </Dimension>
  <Dimension type="StandardDimension" visible="true" highCardinality="false"
name="dim_unidade" caption="Unidades Subordinacao">
    <Hierarchy name="h_default" visible="true" hasAll="true"
allMemberName="total_unidades" allMemberCaption="Total Unidades"
primaryKey="sk_unidade" caption="Unidades Subordinacao">
      <Table name="vw_dim_unidades" schema="public">
      </Table>
      <Level name="pai_filho" visible="true" table="vw_dim_unidades"
column="sk_unidade" nameColumn="no_unidade" ordinalColumn="no_unidade"
parentColumn="fk_unidade" type="String" uniqueMembers="true" levelType="Regular"
hideMemberIf="Never">
        <Property name="Tipo Unidade" column="no_tipo_unidade" type="String">
        </Property>
        <Property name="Subsistema" column="co_subsistema" type="String">
        </Property>
        <Property name="Regi&#227;o" column="co_regiao" type="String">
        </Property>
        <Property name="UF" column="co_uf" type="String">
        </Property>
        <Property name="GIPES" column="no_gipes" type="String">
        </Property>
      </Level>
    </Hierarchy>
    <Hierarchy name="h_subsistema" visible="true" hasAll="true"
allMemberName="total_unidades" allMemberCaption="Todas Unidades"
primaryKey="sk_unidade" caption="Unidades por Subsistema">
      <Table name="vw_dim_unidades">
      </Table>
      <Level name="subsistema" visible="true" table="vw_dim_unidades"
column="co_subsistema" nameColumn="co_subsistema" ordinalColumn="co_subsistema"
type="String" uniqueMembers="false" levelType="Regular" hideMemberIf="Never"
caption="Subsistema" captionColumn="co_subsistema">
      </Level>
      <Level name="unidade" visible="true" table="vw_dim_unidades"
column="nu_unidade" nameColumn="nu_unidade" ordinalColumn="no_unidade"
type="String" uniqueMembers="false" levelType="Regular" hideMemberIf="Never"
caption="Unidade" captionColumn="no_unidade">
```

```
      </Level>
    </Hierarchy>
    <Hierarchy name="h_regiao" visible="true" hasAll="true"
allMemberName="total_unidades" allMemberCaption="Total Unidades"
primaryKey="sk_unidade" caption="Unidades por regi&#227;o">
      <Table name="vw_dim_unidades" schema="public">
      </Table>
      <Level name="regiao" visible="true" table="vw_dim_unidades"
column="co_regiao" ordinalColumn="co_regiao" type="String" uniqueMembers="false"
levelType="Regular" hideMemberIf="Never" captionColumn="co_regiao">
      </Level>
      <Level name="uf" visible="true" table="vw_dim_unidades" column="co_uf"
ordinalColumn="co_uf" type="String" uniqueMembers="false" levelType="Regular"
hideMemberIf="Never" caption="UF" captionColumn="co_uf">
      </Level>
      <Level name="unidade" visible="true" table="vw_dim_unidades"
column="nu_unidade" nameColumn="nu_unidade" ordinalColumn="no_unidade"
type="String" uniqueMembers="false" levelType="Regular" hideMemberIf="Never"
caption="Unidade" captionColumn="no_unidade">
      </Level>
    </Hierarchy>
    <Hierarchy name="h_todas" visible="true" hasAll="true"
allMemberName="total_unidades" allMemberCaption="Total Unidades"
primaryKey="sk_unidade" caption="Unidades">
      <Table name="vw_dim_unidades" schema="public">
      </Table>
      <Level name="unidade" visible="true"  column="nu_unidade"
nameColumn="nu_unidade" ordinalColumn="no_unidade"   caption="Unidade"
        type="String" uniqueMembers="false" levelType="Regular"
hideMemberIf="Never" >

        <CaptionExpression>
          <SQL dialect="generic">
            <![CDATA[( no_unidade  || ' (' || nu_unidade || ')')]]>
          </SQL>
        </CaptionExpression>
        <Property name="Tipo Unidade" column="no_tipo_unidade" type="String">
        </Property>
        <Property name="Subsistema" column="co_subsistema" type="String">
        </Property>
        <Property name="Regi&#227;o" column="co_regiao" type="String">
        </Property>
        <Property name="UF" column="co_uf" type="String">
        </Property>
        <Property name="GIPES" column="no_gipes" type="String">
        </Property>
      </Level>
    </Hierarchy>
  </Dimension>

 <Dimension type="StandardDimension" visible="true" highCardinality="false"
name="dim_faixa_salarial" caption="Faixa Salarial">
    <Hierarchy name="h_default" visible="true" hasAll="true"
allMemberName="total_faixa_salarial" allMemberCaption="Total Faixa Salarial"
primaryKey="sk_faixa_salarial" caption="Faixa Salarial">


      <Table name="dim_faixa_salarial">
      </Table>
      <Level name="faixa_salarial" visible="true" column="de_faixa_salarial"
nameColumn="de_faixa_salarial" ordinalColumn="sk_faixa_salarial" type="String"
uniqueMembers="false" levelType="Regular" hideMemberIf="Never" caption="Faixa
Salarial" captionColumn="de_faixa_salarial">
      </Level>
```

```
    </Hierarchy>
  </Dimension>

  <Dimension type="StandardDimension" visible="true" highCardinality="false"
name="dim_faixa_etaria" caption="Faixas Et&#225;ria">
    <Hierarchy name="h_default" visible="true" hasAll="true"
allMemberName="total_faixa_etaria" allMemberCaption="Total Faixa Et&#225;ria"
primaryKey="sk_faixa_etaria" caption="Faixa Et&#225;ria">
      <Table name="dim_faixa_etaria">
      </Table>
      <Level name="faixa_etaria" visible="true" column="de_faixa_etaria"
nameColumn="de_faixa_etaria" ordinalColumn="sk_faixa_etaria" type="String"
uniqueMembers="false" levelType="Regular" hideMemberIf="Never" caption="Faixa
Et&#225;ria" captionColumn="de_faixa_etaria">
     </Level>
    </Hierarchy>
  </Dimension>
  <Dimension type="StandardDimension" visible="true" highCardinality="false"
name="dim_geracao" caption="Gera&#231;&#245;es">
    <Hierarchy name="h_default" visible="true" hasAll="true"
allMemberName="total_geracao" allMemberCaption="Total Gera&#231;&#227;o"
primaryKey="sk_geracao" caption="Gera&#231;&#245;es">
      <Table name="dim_geracao">
     </Table>
      <Level name="geracao" visible="true" column="de_geracao"
nameColumn="de_geracao" ordinalColumn="sk_geracao" type="String"
uniqueMembers="false" levelType="Regular" hideMemberIf="Never"
caption="Gera&#231;&#227;o" captionColumn="de_geracao">
      </Level>
    </Hierarchy>
  </Dimension>
  <Dimension type="TimeDimension" visible="true" highCardinality="false"
name="dim_tempo_mes" caption="Ano/M&#234;s">
    <Hierarchy name="h_default" visible="true" hasAll="true"
primaryKey="sk_tempo_mes"
        caption="Ano/M&#234;s"
        allMemberName="total_tempo_mes" allMemberCaption="Total Ano/M&#234;s" >
      <Table name="dim_tempo_mes">
      </Table>
      <Level name="ano" visible="true" column="ano" nameColumn="ano"
type="Integer" uniqueMembers="false" levelType="TimeYears" hideMemberIf="Never"
captionColumn="ano" caption="Ano">
      </Level>
      <Level name="mes" visible="true" column="mes" nameColumn="mes"
ordinalColumn="mes" type="String" uniqueMembers="false" levelType="TimeMonths"
hideMemberIf="Never" caption="M&#234;s" captionColumn="ds_mes">
      </Level>
    </Hierarchy>
  </Dimension>
  <Dimension type="StandardDimension" visible="true" highCardinality="false"
name="dim_empregados" caption="Empregados">


    <Hierarchy name="h_defaut" visible="true" hasAll="true"
allMemberCaption="Total Empregados" primaryKey="sk_empregado"
caption="Empregados">


      <Table name="dim_empregado">
      </Table>


      <Level name="empregado" visible="true" column="no_empregado"
```

```
nameColumn="nu_matricula" ordinalColumn="no_empregado" type="String"
uniqueMembers="false" levelType="Regular" hideMemberIf="Never"
caption="Empregado">
      </Level>
    </Hierarchy>
  </Dimension>

  <Dimension type="StandardDimension" highCardinality="false"
name="dim_vinculo_funcional" caption="Vinculo Funcional">
    <Hierarchy name="h_default" hasAll="true"
allMemberName="total_vinculo_funcional" allMemberCaption="Total Vinculo
Funcional" primaryKey="sk_vinculo_funcional" caption="Vinculo Funcional">
      <Table name="dim_vinculo_funcional">
      </Table>
      <Level name="vinculo_funcional" column="nu_vinculo_funcional"
nameColumn="nu_vinculo_funcional" ordinalColumn="no_vinculo_funcional"
type="String" uniqueMembers="false" levelType="Regular" hideMemberIf="Never"
caption="Vinculo Funcional" captionColumn="no_vinculo_funcional">
      </Level>
    </Hierarchy>
  </Dimension>
  <Dimension type="StandardDimension" visible="true" name="dim_afast_lep"
caption="Marcadores - Afast LEP">
    <Hierarchy name="h_afast_lep" visible="true" hasAll="true"
allMemberCaption="Total Afast LEP" primaryKey="sk_empregado_marcador"
caption="Marcadores - Afast Lep">
      <Table name="dim_empregado_marcadores" alias="afast_lep">
      </Table>
      <Level name="afast_lep" visible="true" column="ic_afast_lep"
nameColumn="ic_afast_lep" type="String" uniqueMembers="false" caption="Afast Lep"
captionColumn="ds_afast_lep">
      </Level>
    </Hierarchy>
      </Dimension>
      <Dimension type="StandardDimension" visible="true" name="dim_deficiente"
caption="Marcadores - Deficiente">
    <Hierarchy name="h_deficiente" visible="true" hasAll="true"
allMemberCaption="Total Deficiente" caption="Marcadores - Deficiente"
primaryKey="sk_empregado_marcador">
      <Table name="dim_empregado_marcadores" alias="deficiente">
      </Table>
      <Level name="deficiente" visible="true" column="ic_deficiente"
nameColumn="ic_deficiente" type="String" uniqueMembers="false"
caption="Deficiente" captionColumn="ds_deficiente">
      </Level>
    </Hierarchy>
      </Dimension>
      <Dimension type="StandardDimension" visible="true" name="dim_aposentado"
caption="Marcadores - Aposentado">


    <Hierarchy name="h_aposentado" visible="true" hasAll="true"
allMemberCaption="Total Aposentado" primaryKey="sk_empregado_marcador"
caption="Marcadores - Aposentado">
      <Table name="dim_empregado_marcadores" alias="aposentado">


      </Table>
      <Level name="aposentado" visible="true" column="ic_aposentado"
nameColumn="ic_aposentado" type="String" uniqueMembers="false"
caption="Aposentado" captionColumn="ds_aposentado">
      </Level>
    </Hierarchy>
  </Dimension>
```

```
  <Dimension highCardinality="false" name="dim_cargo" caption="Cargos">
    <Hierarchy name="h_default" hasAll="true" allMemberName="total_cargos"
allMemberCaption="Total cargos" primaryKey="sk_cargo" caption="Cargo">
      <Table name="dim_cargo">
      </Table>
      <Level name="co_cargo" column="co_cargo" nameColumn="co_cargo"
ordinalColumn="no_cargo" type="String" uniqueMembers="true" levelType="Regular"
hideMemberIf="Never" caption="Cargo" captionColumn="no_cargo">
      </Level>
    </Hierarchy>
  </Dimension>




  <Dimension highCardinality="false" name="dim_escolaridade"
caption="Escolaridade">


    <Hierarchy name="h_default" hasAll="true" allMemberName="total_escolaridades"
allMemberCaption="Total Escolaridade" primaryKey="sk_modalidade"
caption="Escolaridades">
      <Table name="dim_modalidade">
            <SQL dialect="generic">
           <![CDATA[( nu_tipo_modalidade in (1,2) ) ]]>
          </SQL>
      </Table>
      <Level name="modalidade" column="no_modalidade" nameColumn="nu_modalidade"
ordinalColumn="no_modalidade" type="String" uniqueMembers="true"
levelType="Regular" hideMemberIf="Never" caption="Escolaridade"
captionColumn="no_modalidade">
      </Level>
    </Hierarchy>


  </Dimension>


  <Dimension highCardinality="false" name="dim_tempo_caixa" caption="Tempo
Empresa">
    <Hierarchy name="h_default" hasAll="true" allMemberName="total_tempo_caixa"
allMemberCaption="Total Tempo Empresa" primaryKey="sk_tempo_caixa" caption="Tempo
Empresa">


      <Table name="dim_tempo_caixa">


      </Table>


      <Level name="tempo_caixa" column="sk_tempo_caixa"
nameColumn="sk_tempo_caixa" ordinalColumn="sk_tempo_caixa" type="String"
uniqueMembers="true" levelType="Regular" hideMemberIf="Never"
caption="Escolaridade" captionColumn="de_tempo_caixa">
      </Level>
    </Hierarchy>
  </Dimension>


<Cube name="TurnoverInterno" caption="Rotatividade Interna de Pessoal"
visible="true" cache="true" enabled="true">
```

```
    <Table name="fato_turnover_interno" schema="public">
    </Table>
    <DimensionUsage source="dim_tempo_dia" name="dim_tempo_dia_fim"
caption="Ano/M&#234;s" visible="true" foreignKey="fk_tempo_dia_fim"
highCardinality="false">
    </DimensionUsage>



   <DimensionUsage source="dim_empregados" name="dim_empregados"
caption="Empregados" visible="true" foreignKey="fk_empregado"
highCardinality="false">
    </DimensionUsage>


    <DimensionUsage source="dim_unidade" name="dim_unidade_origem"
caption="Unidades (Origem)" visible="true" foreignKey="fk_unidade_origem"
highCardinality="false">
    </DimensionUsage>


    <DimensionUsage source="dim_unidade" name="dim_unidade_destino"
caption="Unidades (Destino)" visible="true" foreignKey="fk_unidade_destino"
highCardinality="false">
    </DimensionUsage>


    <DimensionUsage source="dim_funcao" name="dim_funcao_origem" caption="Cargos
Comissionados (Origem)" visible="true" foreignKey="fk_funcao_origem"
highCardinality="false">
    </DimensionUsage>

    <DimensionUsage source="dim_funcao" name="dim_funcao_destino" caption="Cargos
Comissionados (Destino)" visible="true" foreignKey="fk_funcao_destino"
highCardinality="false">
    </DimensionUsage>


    <DimensionUsage source="dim_ocor_fun" name="dim_ocor_fun"
caption="Ocorrencia" visible="true" foreignKey="fk_ocor" highCardinality="false">
    </DimensionUsage>
   <DimensionUsage source="dim_mo_trans" name="dim_mo_trans" caption="Motivo
Transferencia" visible="true" foreignKey="fk_mo_trans" highCardinality="false">
    </DimensionUsage>
    <DimensionUsage source="dim_faixa_etaria" name="dim_faixa_etaria"
caption="Faixas Et&#225;rias" visible="true" foreignKey="fk_faixa_etaria"
highCardinality="false">



    </DimensionUsage>
    <DimensionUsage source="dim_sexo" name="dim_sexo" caption="Sexos"
visible="true" foreignKey="fk_empregado" highCardinality="false">
    </DimensionUsage>

  <Measure name="qtd_empregados" column="fk_empregado" datatype="Integer"
aggregator="distinct-count" caption="Qtd. Empregados" visible="true">
  </Measure>
  <Measure name="idade_desligamento" column="nu_idade_desligamento"
datatype="Integer" aggregator="avg" caption="Idade Média" visible="true">
  </Measure>
  <Measure name="qt_tempo_caixa" column="qt_tempo_caixa" datatype="Integer"
aggregator="avg" caption="Tempo Média Empresa" visible="true">
  </Measure>
```

```
  <Measure name="qt_tempo_unidade" column="qt_tempo_unidade" datatype="Integer"
aggregator="avg" caption="Tempo Média Unidade" visible="true">
  </Measure>

</Cube>
<Cube name="TurnoverInternoPSI" caption="Rotatividade Interna de Pessoal via PSI"
visible="true" cache="true" enabled="true">
    <Table name="fato_turnover_psi" schema="public">
    </Table>
    <DimensionUsage source="dim_faixa_etaria" name="dim_faixa_etaria"
caption="Faixas Et&#225;rias" visible="true" foreignKey="fk_faixa_etaria"
highCardinality="false">
    </DimensionUsage>
    <DimensionUsage source="dim_tempo_dia" name="dim_tempo_dia_fim"
caption="Ano/M&#234;s" visible="true" foreignKey="fk_tempo_dia_fim"
highCardinality="false">
    </DimensionUsage>


    <DimensionUsage source="dim_empregados" name="dim_empregados"
caption="Empregados" visible="true" foreignKey="fk_empregado"
highCardinality="false">
    </DimensionUsage>

    <DimensionUsage source="dim_unidade" name="dim_unidade_origem"
caption="Unidades (Origem)" visible="true" foreignKey="fk_unidade_origem"
highCardinality="false">
    </DimensionUsage>
    <DimensionUsage source="dim_unidade" name="dim_unidade_destino"
caption="Unidades (Destino)" visible="true" foreignKey="fk_unidade_destino"
highCardinality="false">
    </DimensionUsage>
    <DimensionUsage source="dim_funcao" name="dim_funcao_origem" caption="Cargos
Comissionados (Origem)" visible="true" foreignKey="fk_funcao_origem"
highCardinality="false">
    </DimensionUsage>
    <DimensionUsage source="dim_funcao" name="dim_funcao_destino" caption="Cargos
Comissionados (Destino)" visible="true" foreignKey="fk_funcao_destino"
highCardinality="false">
    </DimensionUsage>
    <DimensionUsage source="dim_ocor_fun" name="dim_ocor_fun"
caption="Ocorrencia" visible="true" foreignKey="fk_ocor" highCardinality="false">
    </DimensionUsage>


    <DimensionUsage source="dim_mo_dis_fc" name="dim_mo_dis_fc" caption="Motivo
Dispensa" visible="true" foreignKey="fk_mo_dis_func" highCardinality="false">


    </DimensionUsage>
   <DimensionUsage source="dim_sexo" name="dim_sexo" caption="Sexos"
visible="true" foreignKey="fk_empregado" highCardinality="false">
    </DimensionUsage>

<Measure name="qtd_empregados" column="fk_empregado" datatype="Integer"
aggregator="distinct-count" caption="Qtd. Empregados" visible="true">
  </Measure>
  <Measure name="idade_desligamento" column="nu_idade_desligamento"
datatype="Integer" aggregator="avg" caption="Idade Média" visible="true">
  </Measure>
  <Measure name="qt_tempo_caixa" column="qt_tempo_caixa" datatype="Integer"
aggregator="avg" caption="Tempo Média Empresa" visible="true">
  </Measure>
```

```
  <Measure name="qt_tempo_unidade" column="qt_tempo_unidade" datatype="Integer"
aggregator="avg" caption="Tempo Média Unidade" visible="true">
  </Measure>
  <Measure name="qt_tempo_funcao" column="qt_tempo_funcao" datatype="Integer"
aggregator="avg" caption="Tempo Média Função" visible="true">
  </Measure>


</Cube>


<Cube name="EmpregadosDesligados" caption="Empregados Desligados" visible="true"
cache="true" enabled="true">
    <Table name="fato_empregados_desligados" schema="public">
    </Table>
  <Dimension highCardinality="false" name="dim_aposentado" caption="Aposentado"
     foreignKey="ic_aposentado" >
    <Hierarchy name="h_default" hasAll="true" allMemberName="total_aposentado"
allMemberCaption="Total Aposentado" caption="Aposentado">
      <Level name="aposentado" column="ic_aposentado"
        type="String" uniqueMembers="true" levelType="Regular"
hideMemberIf="Never" caption="Aposentado">
      </Level>
    </Hierarchy>
  </Dimension>
    <DimensionUsage source="dim_tempo_mes" name="dim_tempo_mes"
caption="Ano/M&#234;s" visible="true" foreignKey="fk_tempo_mes"
highCardinality="false">
    </DimensionUsage>


    <DimensionUsage source="dim_tempo_mes" name="dim_data_admissao"
caption="Ano/M&#234;s Admissao" visible="true" foreignKey="fk_data_admissao"
highCardinality="false">
    </DimensionUsage>
    <DimensionUsage source="dim_unidade" name="dim_unidade_admissao"
caption="Unidades Admissao" visible="true" foreignKey="fk_unidade_admissao"
highCardinality="false">
    </DimensionUsage>

    <DimensionUsage source="dim_unidade" name="dim_unidade_desligamento"
caption="Unidades Desligamento" visible="true"
foreignKey="fk_unidade_desligamento" highCardinality="false">


    </DimensionUsage>



    <DimensionUsage source="dim_estado_civil" name="dim_estado_civil"
caption="Estado Civil" visible="true"
foreignKey="fk_sexo_raca_estado_civil_desligamento" highCardinality="false">
    </DimensionUsage>
   <DimensionUsage source="dim_raca" name="dim_raca" caption="Ra&#231;as"
visible="true" foreignKey="fk_sexo_raca_estado_civil_desligamento"
highCardinality="false">
    </DimensionUsage>
    <DimensionUsage source="dim_sexo" name="dim_sexo" caption="G&#234;neros"
visible="true" foreignKey="fk_sexo_raca_estado_civil_desligamento"
highCardinality="false">
    </DimensionUsage>
    <DimensionUsage source="dim_cargo" name="dim_cargo" caption="Cargos"
visible="true" foreignKey="fk_cargo_admissao" highCardinality="false">
```

```
    </DimensionUsage>
    <DimensionUsage source="dim_funcao" name="dim_funcao" caption="Cargos
Comissionados" visible="true" foreignKey="fk_funcao_desligamento"
highCardinality="false">
    </DimensionUsage>
    <DimensionUsage source="dim_faixa_etaria" name="dim_faixa_etaria"
caption="Faixas Et&#225;rias" visible="true"
foreignKey="fk_faixa_etaria_desligamento" highCardinality="false">
    </DimensionUsage>

   <DimensionUsage source="dim_geracao" name="dim_geracao"
caption="Gera&#231;&#245;es" visible="true" foreignKey="fk_geracao"
highCardinality="false">
    </DimensionUsage>
    <DimensionUsage source="dim_tempo_caixa" name="dim_tempo_caixa"
caption="Tempo Empresa" visible="true" foreignKey="fk_tempo_caixa"
highCardinality="false">
    </DimensionUsage>
    <DimensionUsage source="dim_afast_lep" name="dim_afast_lep"
caption="Marcadores - Afast. Lep" visible="true"
foreignKey="fk_empregado_marcador_desligamento" highCardinality="false">
    </DimensionUsage>


  <DimensionUsage source="dim_aposentado" name="dim_aposentado"
caption="Marcadores - Aposentado" visible="true"
foreignKey="fk_empregado_marcador_desligamento" highCardinality="false">
    </DimensionUsage>



  <DimensionUsage source="dim_deficiente" name="dim_deficiente"
caption="Marcadores - Deficiente" visible="true"
foreignKey="fk_empregado_marcador_desligamento" highCardinality="false">


    </DimensionUsage>



    <DimensionUsage source="dim_vinculo_funcional" name="dim_vinculo_funcional"
caption="Vinculo Funcional" visible="true" foreignKey="fk_vinculo_funcional"
highCardinality="false">


    </DimensionUsage>







    <DimensionUsage source="dim_escolaridade" name="dim_escolaridade_admissao"
caption="Escolaridade (Admissao)" visible="true"
foreignKey="fk_escolaridade_desligamento" highCardinality="false">


    </DimensionUsage>
        <DimensionUsage source="dim_escolaridade"
name="dim_escolaridade_desligamento" caption="Escolaridade (Desligamento)"
visible="true" foreignKey="fk_escolaridade_admissao" highCardinality="false">
    </DimensionUsage>

        <DimensionUsage source="dim_area_conhecimento"
```

```
name="dim_area_conhecimento_desligamento" caption="Area Formacao (Desligamento)"
visible="true" foreignKey="fk_area_conhecimento_desligamento"
highCardinality="false">
    </DimensionUsage>
   <Measure name="qtd_empregados" column="fk_empregado" datatype="Integer"
aggregator="distinct-count" caption="Qtd. Empregados" visible="true">
    </Measure>
  <Measure name="qt_tempo_serv_cef" column="qt_tempo_serv_cef" datatype="Integer"
aggregator="avg" caption="Tempo Serviço CEF (dias)" visible="false">
    </Measure>
  <Measure name="qt_tempo_serv_priv" column="qt_tempo_serv_priv"
datatype="Integer" aggregator="avg" caption="Tempo Serviço Privado"
visible="true">
    </Measure>
  <Measure name="qt_dependentes" column="qt_dependentes" datatype="Integer"
aggregator="sum" caption="Nr Dependentes" visible="true">
    </Measure>
 <Measure name="idade_desligamento" column="nu_idade_desligamento"
datatype="Integer" aggregator="avg" caption="Idade Média" visible="true">
    </Measure>

  <Measure name="qt_tempo_primeira_funcao" column="qt_tempo_primeira_funcao"
datatype="Integer" aggregator="avg" caption="Tempo Médio Primeira Função"
visible="true">
    </Measure>
    <CalculatedMember name="tempo_medio_caixa" formatString="#,##" caption="Tempo
Serviço CEF" formula="[Measures].[qt_tempo_serv_cef] / 365" dimension="Measures"
visible="true">
    </CalculatedMember>
  </Cube>
<Cube name="TaxaDesligamentoInterno" caption="Taxa Desligamento Interno"
visible="true" cache="true" enabled="true">

    <Table name="fato_turnover_interno_agg01" schema="public">

    </Table>


 <Dimension highCardinality="false" name="dim_tempo" caption="Ano"
foreignKey="fk_tempo" >
    <Hierarchy name="h_default" hasAll="true" allMemberName="total_ano"
allMemberCaption="Total Ano" caption="Ano">


      <Level name="ano" column="fk_tempo"


        type="String" uniqueMembers="true" levelType="Regular"
hideMemberIf="Never" caption="Ano">
      </Level>
    </Hierarchy>
  </Dimension>
  <DimensionUsage source="dim_unidade" name="dim_unidade" caption="Unidades"
visible="true" foreignKey="fk_unidade" highCardinality="false">
    </DimensionUsage>
  <Measure name="qt_empregados" column="qt_empregados" datatype="Integer"
aggregator="sum" caption="Qtd. Empregados" visible="true">
  </Measure>
  <Measure name="qt_admitidos" column="qt_admitidos" datatype="Integer"
aggregator="sum" caption="Qtd. Admitidos" visible="true">
  </Measure>
  <Measure name="qt_desligados" column="qt_desligados" datatype="Integer"
aggregator="sum" caption="Qtd. Desligados" visible="true">
```

```
  </Measure>
  <CalculatedMember name="taxa_desligamento" formatString="#,##" caption="Taxa
Desligamento"
       formula="([Measures].[qt_desligados] / [Measures].[qt_empregados]) * 100"
dimension="Measures" visible="true">
  </CalculatedMember>

  <CalculatedMember name="turnover" formatString="#,##" caption="Turnover"


       formula="((([Measures].[qt_desligados] +  [Measures].[qt_admitidos]) / 2)
/ [Measures].[qt_empregados]) * 100" dimension="Measures" visible="true">


  </CalculatedMember>


</Cube>


<Cube name="TaxaDesligamentoExterno" caption="Taxa Desligamento Externo"
visible="true" cache="true" enabled="true">
    <Table name="fato_turnover_agg01" schema="public">
    </Table>

<Dimension highCardinality="false" name="dim_tempo" caption="Ano"
foreignKey="fk_tempo" >
    <Hierarchy name="h_default" hasAll="true" allMemberName="total_ano"
allMemberCaption="Total Ano" caption="Ano">
      <Level name="ano" column="fk_tempo"
        type="String" uniqueMembers="true" levelType="Regular"
hideMemberIf="Never" caption="Ano">
      </Level>
    </Hierarchy>
  </Dimension>
  <DimensionUsage source="dim_unidade" name="dim_unidade" caption="Unidades"
visible="true" foreignKey="fk_unidade" highCardinality="false">
    </DimensionUsage>
  <Measure name="qt_empregados" column="qt_empregados" datatype="Integer"
aggregator="sum" caption="Qtd. Empregados" visible="true">
  </Measure>
  <Measure name="qt_admitidos" column="qt_admitidos" datatype="Integer"
aggregator="sum" caption="Qtd. Admitidos" visible="true">
  </Measure>
  <Measure name="qt_desligados" column="qt_desligados" datatype="Integer"
aggregator="sum" caption="Qtd. Desligados" visible="true">
  </Measure>
  <CalculatedMember name="taxa_desligamento" formatString="#,##" caption="Taxa
Desligamento"
       formula="([Measures].[qt_desligados] / [Measures].[qt_empregados]) * 100"
dimension="Measures" visible="true">
  </CalculatedMember>
   <CalculatedMember name="turnover" formatString="#,##" caption="Turnover"

       formula="((([Measures].[qt_desligados] +  [Measures].[qt_admitidos]) / 2)
/ [Measures].[qt_empregados]) * 100" dimension="Measures" visible="true">


  </CalculatedMember>

</Cube> </Schema>
```

APÊNDICE C – TRECHO DO ARQUIVO ARFF UTILIZADO PELO WEKA NO PROCESSO DE MINERAÇÃO DE DADOS

```
@relation turnover_interno
@attribute co_sexo {F,M}
@attribute de_geracao {Boomers,'Geração X','Geração Y',Veteranos}
@attribute uf_origem
{AC,AL,AM,AP,BA,CE,DF,ES,GO,MA,MG,MS,MT,PA,PB,PE,PI,PR,RJ,RN,RO,RR,RS,SC,SE,SP,TO}
@attribute subsistema_origem {CENTRAL,LOGISTICO,NEGOCIAL}
@attribute regiao_origem {Centro-Oeste,Nordeste,Norte,Sudeste,Sul}
@attribute subsistema_destino {CENTRAL,LOGISTICO,NEGOCIAL}
@attribute regiao_destino {Centro-Oeste,Nordeste,Norte,Sudeste,Sul}
@attribute tipo_funcao_origem {CHEFIA,SEMFUNCAO,TECNICO}
@attribute tipo_funcao_destino {CHEFIA,SEMFUNCAO,TECNICO}
@attribute nu_idade_desligamento numeric
@attribute escolaridade {'ENSINO MEDIO',GRADUACAO,POSGRADUACAO}
@attribute temexperienciaexterna {NAO,SIM}
@attribute num_dep numeric
@attribute qt_tempo_empresa numeric
@attribute qt_tempo_unidade numeric
@attribute horas_treinamento numeric
@attribute piso_origem numeric
@attribute piso_destino numeric
@attribute tipo_transferencia {DECESSO,LATERALIDADE,PERDEUFUNCAO,PROMOCAO,TRANSFSEMFUNCAO}
@attribute tempo_permanencia_unidade {<=2,>2}
@attribute faixa_salarial_origem {'12 a 17 mil','2 a 7 mil','7 a 12 mil','ate 2 mil','mais 17
mil'}
@attribute faixa_salarial_destino {'12 a 17 mil','2 a 7 mil','7 a 12 mil','ate 2 mil','mais 17
mil'}

@data F,'Geração Y',RS,LOGISTICO,Sul,'GERENCIA DE FILIAL','VP LOGISTICA E
RETAGUARDA',RS,LOGISTICO,Sul,'GERENCIA DE FILIAL','VP LOGISTICA E RETAGUARDA','SEM CARGO
COMISSIONADO',SEMFUNCAO,'SEM CARGO
COMISSIONADO',SEMFUNCAO,27,DIREITO,GRADUACAO,NAO,0,5,1.3,153,1630,1630,TRANSFSEMFUNCAO,<=2,'ate
2 mil','ate 2 mil' F,'Geração Y',RS,NEGOCIAL,Sul,AGENCIA,'VP ATENDIMENTO E DISTRIBUICAO
NEGOCIO',RS,NEGOCIAL,Sul,AGENCIA,'VP ATENDIMENTO E DISTRIBUICAO NEGOCIO','SEM CARGO
COMISSIONADO',SEMFUNCAO,'SEM CARGO COMISSIONADO',SEMFUNCAO,23,EDUCACAO,'ENSINO
MEDIO',NAO,0,2.8,2.8,116,1314,1314,TRANSFSEMFUNCAO,>2,'ate 2 mil','ate 2 mil'
```