PONTIFÍCIA UNIVERSIDADE CATÓLICA DE MINAS GERAIS

Bacharelado em Sistemas de Informação

Dayana Thalita Santos Viana

PROCESSO DE KDD APLICADO AOS MICRODADOS DO CENSO DA EDUCAÇÃO SUPERIOR DO INEP

Belo Horizonte

2012

Dayana Thalita Santos Viana

PROCESSO DE KDD APLICADO AOS MICRODADOS DO CENSO DA EDUCAÇÃO SUPERIOR DO INEP

Monografia apresentada ao Curso de Sistemas de Informação da Pontifícia Universidade Católica de Minas Gerais, como requisito parcial para obtenção do título de Bacharel Sistemas de Informação.

Orientador: Hugo Bastos de Paula

Belo Horizonte

2012

Dayana Thalita Santos Viana

PROCESSO DE KDD APLICADO AOS MICRODADOS DO CENSO DA EDUCAÇÃO SUPERIOR DO INEP

Monografia apresentada ao Curso de Sistemas de Informação da Pontifícia Universidade Católica de Minas Gerais, como requisito parcial para obtenção do título de Bacharel Sistemas de Informação.

Professor 1 (Orientador) – PUC Minas

Professor 2 – PUC Minas

Professor 3 – Universidade

Belo Horizonte, 26 de Novembro de 2012.

A toda minha família e principalmente ao meu pai,
por ter me dado todo carinho e a melhor educação possível,
e por ser um grande exemplo de boa pessoa.

AGRADECIMENTOS

Ao Prof. Hugo Bastos, pela orientação neste trabalho de conclusão de curso.

Aos demais professores, que compartilharam seus conhecimentos e experiências.

Aos colegas do curso de Sistemas de Informação da PUC Minas.

E a minha família pelo apoio e compreensão durante todo esse período.

“Suba o primeiro degrau com fé.
Não é necessário que você veja toda a escada.
Apenas dê o primeiro passo.”

Martin Luther King

RESUMO

O *Knowledge Discovery in Databases* (KDD) é um processo composto de várias etapas para compreensão de padrões nos dados. Dada a divulgação pública dos dados do Censo da Educação Superior realizada anualmente pelo Instituto Nacional de Estudos e Pesquisas Educacionais Anísio Teixeira (Inep) temos uma base de dados para desenvolver o processo. Foi utilizada Mineração de Dados, com o auxílio de ferramentas como o SQL Server e Excel para descoberta de conhecimento nessa base de dados. Visto que um dos maiores desafios que o ensino superior enfrenta hoje é prever as decisões dos alunos, a utilização desse processo e ferramentas pode ajudar a tomada de decisões da Universidade PUC Minas. Os resultados trouxeram informações e previsões sobre ingressos e evasões; análises sobre a quantidade de candidatos vaga; a importância do curso de Sistemas de Informação dentro e fora da PUC Minas; influenciadores da taxa de ocupação, principais cursos que aparecem juntos com grande ocupação e recomendações.

Palavras-chave: Processo KDD. SQL Server. Excel. ETL. Mineração de Dados.
Censo da Educação Superior.

LISTA DE FIGURAS

LISTA DE TABELAS

LISTA DE ABREVIATURAS E SIGLAS

BI *Business Inteligence*

DCBD Descoberta de Conhecimento em Banco de Dados

DW *Data Warehouse*

ETL *Extract Transform Load*

GTI Gerência de Tecnologia de Informação

IES Instituições de Ensino Superior

Inep Instituto Nacional de Estudos e Pesquisas Educacionais Anísio Teixeira

KDD *Knowledge Discovery in Databases*

OLAP *On-Line Analytical Processing*

PUC Minas Pontifícia Universidade Católica de Minas Gerais

SGBD Sistema Gerenciador de Banco de Dados

S2B *Students to Business*

SUMÁRIO

1 INTRODUÇÃO

A questão do acesso ao ensino superior brasileiro vêm sendo discutida como uma questão política brasileira. A partir dos anos 90 vêm se expandindo a quantidade de estudantes que concluem o ensino médio. Esse crescimento deve-se à rede privada e as políticas implementadas no setor público pelo estado de Minas Gerais visando maior eficiência no ensino fundamental. Acrescido a isso, ações de âmbito universitário, como criação de novos cursos, aumento de vagas e facilidades nas inscrições ou realização das provas de vestibular estimulam a procura pela educação superior (MENDES, 1997). Como consequência, percebe-se um aumento na quantidade de ingressantes no ensino superior, conforme podemos ver na Tabela 1.

Tabela 1: Evolução do Número de Ingressos por Categoria Administrativa – Brasil (2001-2010)

Ano	Total	Pública								Privada	
		Total	%	Federal	%	Estadual	%	Municipal	%	Privada	%
2001	1.043.308	251.239	24,1	125.701	12,0	99.214	9,5	26.324	2,5	792.069	75,9
2002	1.431.893	334.070	23,3	148.843	10,4	149.017	10,4	36.210	2,5	1.097.823	76,7
2003	1.554.664	325.405	20,9	153.393	9,9	128.323	8,3	43.689	2,8	1.229.259	79,1
2004	1.646.414	364.647	22,1	165.685	10,1	153.889	9,3	45.073	2,7	1.281.767	77,9
2005	1.805.102	362.217	20,1	148.206	8,2	166.660	9,2	47.351	2,6	1.442.885	79,9
2006	1.965.314	368.394	18,7	177.232	9,0	143.636	7,3	47.526	2,4	1.596.920	81,3
2007	2.138.241	416.178	19,5	193.919	9,1	176.047	8,2	46.212	2,2	1.722.063	80,5
2008	2.336.899	538.474	23,0	211.183	9,0	282.950	12,1	44.341	1,9	1.798.425	77,0
2009	2.065.082	422.320	20,5	253.642	12,3	133.425	6,5	35.253	1,7	1.642.762	79,5
2010	2.182.229	475.884	21,8	302.359	13,9	141.413	6,5	32.112	1,5	1.706.345	78,2

Fonte: (INEP, 2012).

Um dos maiores desafios que o ensino superior enfrenta hoje é prever as decisões dos alunos. Instituições gostariam de saber, por exemplo, quais alunos irão se inscrever em cursos particulares, ou se existem alunos mais propensos realizar transferências do que outros. Além disso, a questão da gestão de inscrição continua a motivar as instituições de ensino superior procurar melhores soluções (LUAN, 2002).

Afim de oferecer informações detalhadas e tendências do setor, o Inep realiza regularmente a coleta dos dados sobre a educação superior. Dentre os dados coletados podemos encontrar informações sobre as instituições de ensino superior, seus cursos, vagas ofertadas, número de inscrições, matrículas, ingressantes e concluintes. Esses dados são coletados através

de questionários respondidos pelas Instituições de Ensino Superior (IES). Então, são publicados apenas como informações estatísticas mostrando, como exemplo, percentuais de crescimento do número de matrículas (INEP, 2011).

Somente a coleta de dados não ajuda nas decisões das instituições. Para que exista melhoria no processo é necessário analisar os dados coletados e estabelecer indicadores, para então descobrir padrões que estavam "escondidos" entre os dados. Dada a quantidade e frequência de dados coletados é necessário despender um alto custo para se realizar toda análise em tempo hábil, sendo necessária a busca por ferramentas que automatizem esse processo. Afim de solucionar este problema, é oportuno utilizar-se a metodologia de Descoberta de Conhecimento em Banco de Dados (DCBD) ou do termo mais conhecido em inglês KDD. Com a utilização de técnicas como a mineração de dados, é possível explicitar o conhecimento antes oculto em grandes quantidades de informações. Através dessas técnicas podemos realizar análises dos dados, permitindo a previsão de tendências e comportamentos. Assim gerentes estratégicos podem tomar suas decisões baseadas nesses fatos descobertos e não mais em premissas (CARDOSO, 2008).

O KDD possui várias etapas como: seleção de dados, limpeza e preparação dos dados, identificação de dados relevantes, *data mining*, avaliação de padrões e apresentação de resultados. A mineração de dados é apenas uma etapa do processo de descoberta, que por sua vez é dividido em tarefas como: análise de regras de associação, classificação e predição, análise de padrões sequenciais, análise de agrupamentos e análise de exceções. Essas tarefas consistem respectivamente em encontrar itens que determinem a presença de outros, definir classes para objetos que ainda não foram analisados, encontrar comportamentos que ocorrem em sequência, identificar grupos com característica iguais e determinar itens que fogem do comportamento padrão da maioria (CARDOSO, 2008).

A demanda por cursos e a evasão estudantil são problemas que atingem instituições de ensino superior em geral (MENDES, 1997; FILHO, 2007). Apesar da abundância de dados fornecida pelo Inep, não é possível conhecer imediatamente as razões que geram esses problemas, para assim aplicar uma solução satisfatória de gestão. Tal quantidade de informações precisam passar por um processo de descoberta de conhecimento para trazerem à tona relações atualmente desconhecidas.

Diversas universidades já realizaram mineração de dados em seus dados educacionais. Mendes (1997) elaborou um artigo analisando a demanda de vagas nos vestibulares da UFMG nos anos 90. Nesse estudo ele observou aspectos socioeconômicos dos candidatos, área de conhecimento mais aquecida no mercado e ações de âmbito universitário. Apesar de analisar os dados relativos às quantidade de alunos inscritos no vestibular, Mendes não analisa informa-

ções pós vestibular, como a quantidade de alunos efetivamente matriculados e valores relativos à alunos que conseguem concluir os cursos. Já Beatriz (2007) avança no ponto criticado anteriormente e publica seu trabalho sobre evasão brasileira no ensino superior. Ela correlaciona evasão e demanda, candidatos por vaga, em diversas áreas de conhecimento, regiões do país e categorias administrativas (púbico/privado). Apesar da relevância das informações, não é incluído no estudo uma proposta ou mesmo uma solução para o problema apresentado. Ramos (1996) teve como objeto de estudo a evasão dos cursos de graduação em IES públicas. Classificou as evasões em nível de curso, instituição e evasão a nível de sistema superior. Ele indica as possíveis causas das evasões classificado-as em três ordens: as que se relacionam ao estudante, ao curso e à instituição ou a fatores sócio-culturais e econômicos externos. Apesar de oferecer disgnósticos rigorosos não apresenta relatórios dimensionando as causa. Christine (2009), semelhante a proposta anterior, analisa os dados referente aos alunos de uma turma e conclui apresentando os motivos das evasões e as soluções cabíveis.

Ao observar as soluções existentes para análise de dados educacionais percebemos que não existe um método automatizado para isso. São organizadas tabelas e gráficos como técnicas de descoberta de dados. Atualmente é possível a aplicação de ferramentas automáticas para extração de informações relevantes, como por exemplo, a extração de dados da plataforma *Lattes* realizada por Cardoso (2008), que utilizou técnicas de *data mining*. Juntamente à necessidade de automatização da descoberta de dados na área escolar percebemos que a PUC Minas ainda não possui um sistema para análise de demandas e evasões. Vê-se aí um ótimo cenário para suplantar as soluções existentes acrescentando o uso de uma metodologia automatizada visando um aumento eficaz de produtividade.

1.1 Objetivo

O objetivo desse trabalho é aplicar as diversas etapas do processo de KDD em um banco com dados recebido pelo Inep. Como foco desse banco teremos os alunos de sistema de informação da Pontifícia Universidade Católica de Minas Gerais (PUC Minas). Assim será possível extrair conhecimento referente ao processo decisório da universidade quanto a esse curso. Será possível até mesmo estabelecer um modelo de gestão à instituição mencionada.

Esse trabalho auxiliará no processo de descoberta do conhecimento, que pode servir de apoio à tomada de decisão, possibilitando aperfeiçoamento do sistema de ensino superior da instituição. Frequentemente más decisões são tomadas pela indisponibilidade do conhecimento para se escolher a melhor decisão (CARDOSO, 2008). Obter uma reflexão sobre demanda e eva-

são nos últimos anos torna-se extremamente importante, permitindo a avaliação e possivelmente reformulação dos processos de seleção. Como também poderá ser possível dar mais suporte aos alunos, afim de que eles não abandonem o curso. Os estudantes serão qualificados garantindo bons resultados com a maior quantidade de diplomados. Portanto, as capacidades do *data mining* aplicadas ao dados do ensino superior economizarão recursos, maximizarão a eficiência e aumentarão a produtividade sem aumentar os custos da instituição (LUAN, 2002).

2 KDD

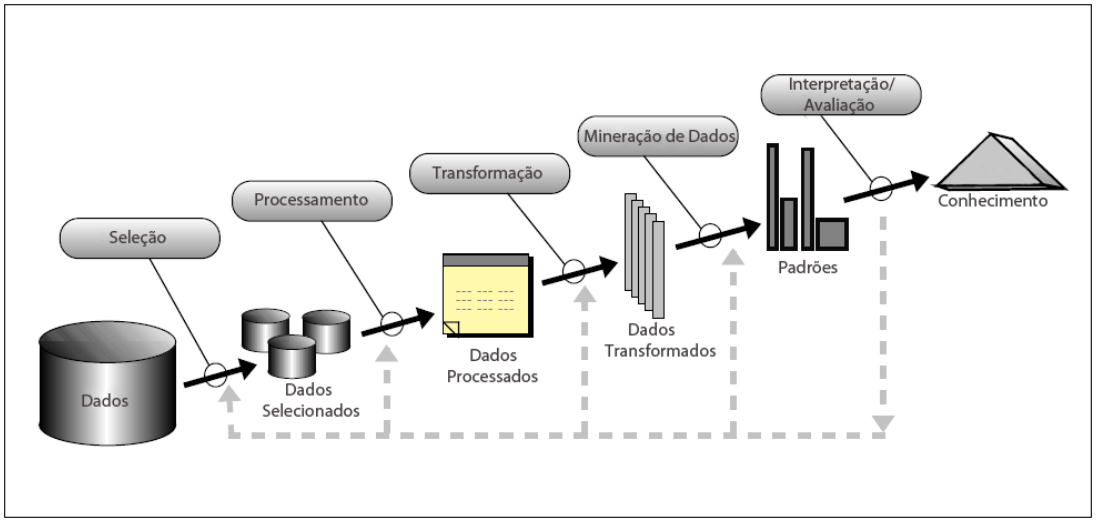
Diversas notações para encontrar padrões úteis nos dados já foram usadas. Entre elas, o termo *data mining* foi o mais comum. A expressão *Knowledge discovery in database* (KDD) apenas começou a ser usada em um workshop em 1989 para enfatizar que o conhecimento (*knowledge*) era o produto final da procura. KDD representa todo processo de descoberta de conhecimento. Inclui como os dados serão armazenados, acessados, como os algoritmos serão aplicados, como os resultados serão interpretados e visualizados. Porém a ênfase maior se dá ao entendimento dos padrões que podem ser interpretados como conhecimento útil. Já *data mining* é a aplicação de algoritmos aos dados para obtenção de regras. É a modelagem de algoritmos para uma grande quantidade de dados inconsistentes (FAYYAD, 1996).

KDD é um processo não trivial de identificação válida, ótima, útil e de fácil compreensão dos padrões nos dados. O termo processo implica que o KDD possui diversos passos como a preparação dos dados, busca de padrões, avaliação do conhecimento e refinamento em múltiplas iterações em que podem conter revisões a cada dois passos. Não trivial significa que são necessárias pesquisas em cima dos dados e não somente computação com valores predefinidos. Útil induz dizer que trará algum benefício ao usuário ou suas tarefas (FAYYAD, 1996).

O KDD é um processo interativo e iterativo que envolve diversos passos envolvendo decisões feitas pelo usuário. Primeiramente é feito um estudo do domínio da aplicação identificando qual é o conhecimento relevante para se atingir o objetivo. Em seguida, os dados coletados são selecionados focando em um subconjunto em que a descoberta será focada. O terceiro passo trabalha com a limpeza e processamento dos dados. Nesse passo as informações erradas, inconsistentes e até mesmo inexistentes são manipuladas. A redução e projeção fazem parte do quarto passo, onde características que representam os dados de acordo com o objetivo são encontradas. O quinto passo consiste em casar os objetivos do processo de KDD a um processo de *data mining*, como por exemplo, clusterização, classificação, sumarização, regressão e etc. O sexto passo consiste na análise, modelagem e hipótese, onde são analisados os modelos e parâmetros mais apropriados. Os resultados são interpretados possibilitando o retorno aos passos 1 a 6 para mais iterações. Finalmente o sétimo passo consiste na busca por padrões de interesse representado em tabelas ou outros tipos de exibições. Como resultado podemos ter uma ação usando o conhecimento adquirido, ou simplesmente produção de uma documentação a ser mostrada às partes interessadas. Mesmo considerando todos os passos muito importantes,

a parte mais trabalhosa do KDD está no passo 5, o *data mining* (FAYYAD, 1996).

Figura 1: Processo KDD (Tradução por Dayana Viana)



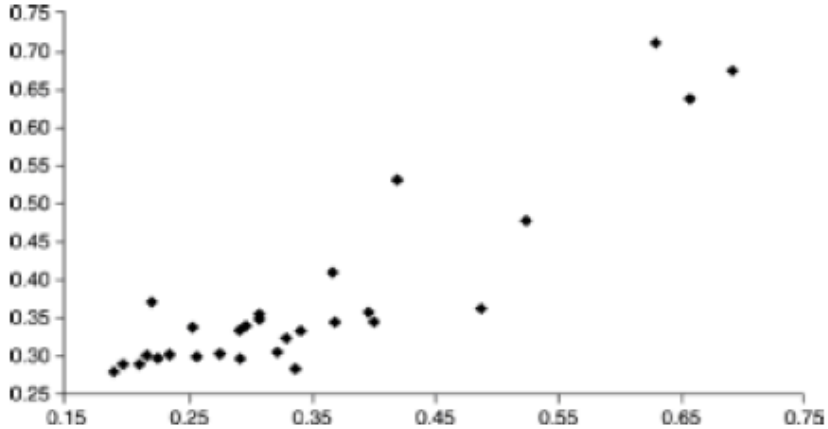
Fonte: (FAYYAD, 1996)

3 MINERAÇÃO DE DADOS

Assim como na mineração geológica (carvão, ouro, etc), não há a garantia da obtenção de resultados significativos pela simples aplicação das ferramentas ao terreno. Uma enorme preparação é necessária. Primeiramente, os dados devem estar preparados. A partir daí é possível fazer a modelagem a fim de transformá-los em informações capazes de serem interpretadas pelos seres humanos. Modelar significa encontrar relações, fazer previsões dos dados para descrever a situação atual. Os fundamentos dos métodos utilizados para mineração são fáceis de entender, porém sua implementação já requer poderosos e sofisticados algoritmos para fazer com que esses métodos funcionem na prática (PYLE, 1999).

Através da observação da Figura 2 é possível percebermos grupos formados pelos pontos. As ferramentas de modelagem tem como tarefa separar e agrupar os dados, nesse caso representado como pontos, de maneira com que tenham significado. Cada algoritmo realiza essa tarefa utilizando abordagens ligeiramente diferentes (PYLE, 1999).

Figura 2: Mineração de Dados



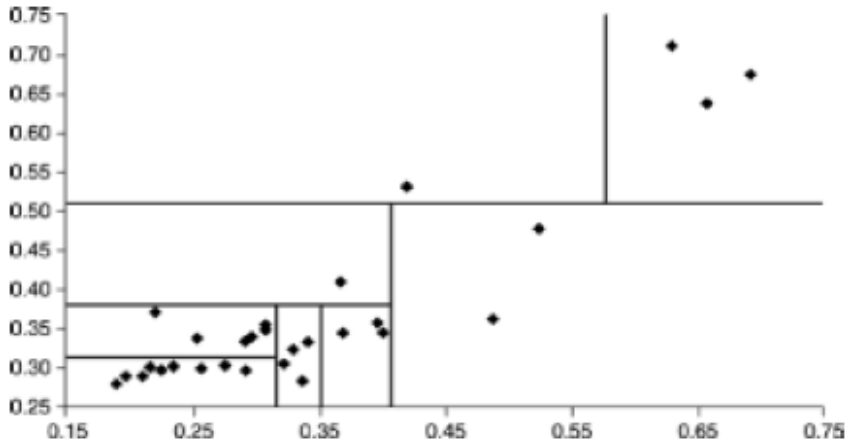
Fonte: (PYLE, 1999).

A mineração de dados, componente do processo KDD, envolve aplicação iterativa e repetida de um método particular. Ajustando os modelos obtêm-se padrões a partir dos dados observados. A maioria dos métodos de *data mining* é baseada em experiências e técnicas de testes das máquinas de aprendizado, reconhecimento de padrões e estatísticas. Algumas técnicas de mineração de dados são: árvores de decisão, clusterização, vizinho mais próximo e

regressão (FAYYAD, 1996).

Árvores de Decisão: Algoritmo baseado no processo de partição. As partições visando a separação dos pontos são feitas através de pontos de decisões até algum critério de parada ou até não ser mais possível realizar separações (Figura 3) (PYLE, 1999).

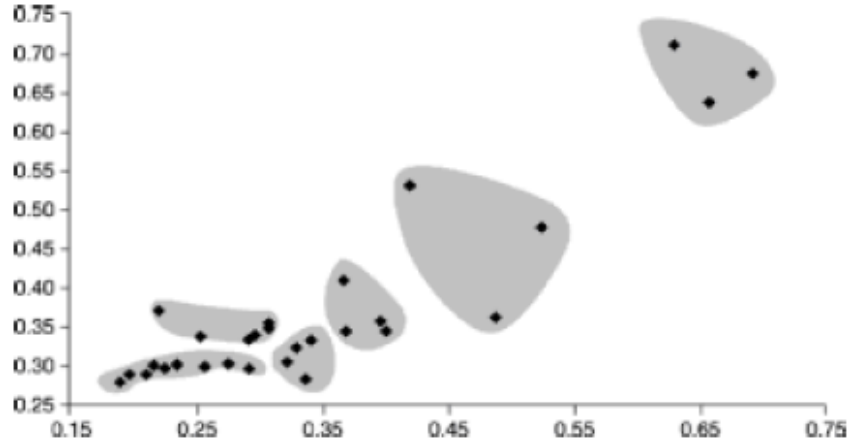
Figura 3: Árvores de Decisão



Fonte: (PYLE, 1999).

Clusterização: Também particionam os espaços, porém agrupando pontos que compartilham as mesmas características. Existem diferentes métodos de clusterização, mas todos produzem esse tipo de arranjo. Uma grande diferença desse método é que ele não separa os grupos linearmente, o que facilita o encontro de similaridades (Figura 4) (PYLE, 1999).

Figura 4: Clusterização

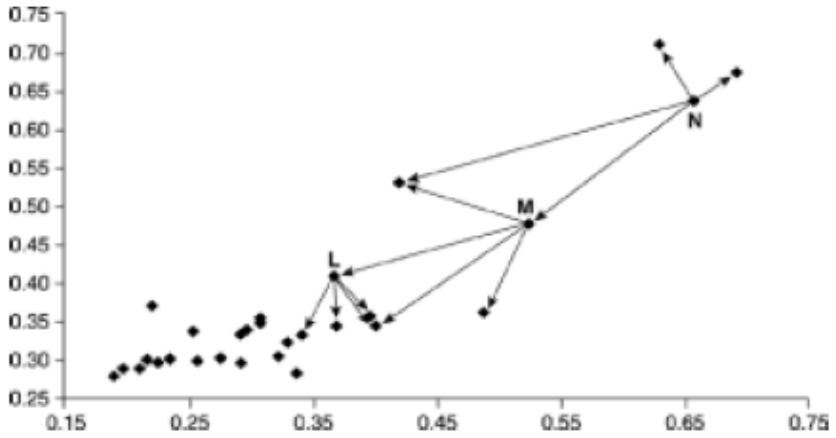


Fonte: (PYLE, 1999).

Vizinho mais próximo: Um tipo de classificação utilizado para descrever interações.

Esse método seleciona um número específico de vizinhos e para cada ponto calcula a vizinhança. A Figura 5 ilustra como os vizinhos podem ser selecionados. Para cada ponto foi calculado os quatro vizinhos mais próximos (PYLE, 1999).

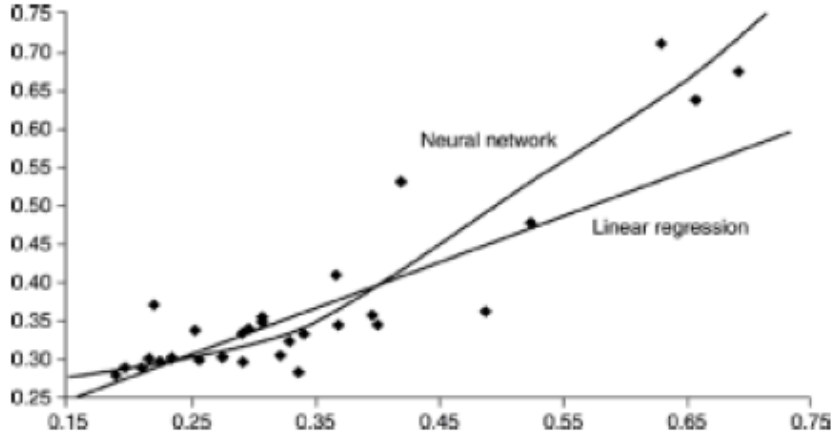
Figura 5: Vizinho mais próximo



Fonte: (PYLE, 1999).

Redes Neurais e Regressão: Esses métodos funcionam através da criação de uma expressão matemática representando uma linha ajustada aos pontos. No caso da regressão linear, para a predição é usado o ponto mais próximo da inferência para o ponto a ser previsto (Figura 6) (PYLE, 1999).

Figura 6: Redes Neurais e Regressão



Fonte: (PYLE, 1999).

4 DATA WAREHOUSE

Data Warehouse (DW), ou armazém de dados, consolidam dados em espaços multidimensionais. Eles podem ser vistos como uma etapa importante para a mineração de dados. Além disso provê integração com ferramentas *On-Line Analytical Processing* (OLAP) para análise interativa dos dados. O DW provê ferramentas e arquitetura para que os responsáveis pelos negócios organizem, entendam e usem seus dados para tomarem decisões estratégicas (HAN, 2005).

O *Data Warehouse* é orientado a um assunto específico, integrado, não volátil e com tempo variante para o suporte do processo de tomada de decisões. Ele é organizado em torno de um objetivo principal, como relações nas vendas, ao invés de se concentrar em operações e transações diárias. Dizemos que o DW é integrado por ser construído através de múltiplas fontes, como banco de dados relacionais diversos, planilhas e outros sistemas. Refere-se a um banco de dados que é mantido separado do banco de dados das operações organizacionais. Então não requer processamento de transações, *backups* contínuos e mecanismos de controle. Basicamente o DW realiza apenas duas operações: carregamento inicial e acesso aos dados. Todas as informações armazenadas dizem respeito a um período de tempo definido normalmente entre 5 a 10 anos (HAN, 2005).

Para permitir modelar e visualizar as múltiplas dimensões do DW utiliza-se os CUBOS. Os Cubos são definidos por dimensões e fato. O fato significa o tema do modelo, representado por uma tabela principal. Já as dimensões são as entidades que dizem respeito aquilo que a organização deseja armazenar, as tabelas ao redor do fato. Apesar de pensarmos no Cubo como uma estrutura 3D, no *data warehousing* ele é n-dimensional. É possível ver no cubo, por exemplo, dados de acordo com o tempo, item, localização e fornecedor. Ou seja, uma visualização 4D (HAN, 2005).

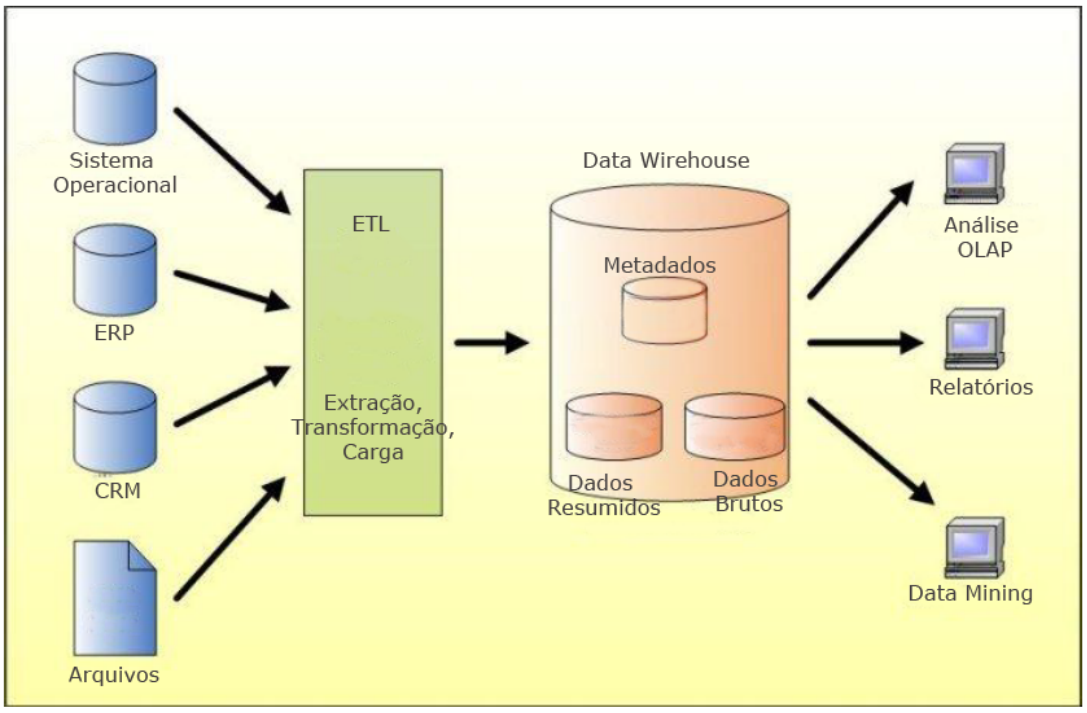
Assim como o fato e as dimensões, a hierarquia é uma característica do DW. O conceito de hierarquia define a sequência do mapeamento dos mais baixos aos mais altos conceitos. Um exemplo dessa hierarquia pode ser visto na dimensão Tempo, onde tem-se as horas como conceitos mais baixos e os anos como conceitos mais altos. Essa hierarquia provê ao usuários a flexibilidade de acordo com suas necessidades (HAN, 2005).

Para o modelo entidade-relacionamento é desenhado um modelo de relações entre as en-

tidades. Entretanto, para o DW é utilizado um modelo multidimensional como o estrela, floco de neve ou mesmo constelação. O esquema estrela contém uma grande tabela central, o fato, com uma série de tabelas menores em volta, as dimensões. O esquema floco de neve é uma variação do esquema estrela, porém as tabelas de dimensões são normalizadas. Então as tabelas existentes são divididas resultando em uma forma final similar a um floco de neve. A maior diferença entre esses dois esquemas é que o segundo modelo reduz as redundâncias no banco, reduzindo também o espaço de armazenamento. Porém apesar dessa redução, esse esquema perde performance por ter que executar mais *joins* em suas consultas. O último esquema, constelação, especifica duas tabelas fatos. Assim é permitido às dimensões serem compartilhadas entre os fatos (HAN, 2005).

A arquitetura de um DW pode ser representada de acordo com a Figura 7. No centro da imagem está o repositório, composto pelos dados e metadados. Para alimentar esse banco são usadas fontes externas, ferramentas de *back-end* e utilitárias. Essas ferramentas executam a extração dos dados das diferentes fontes, assim como sua limpeza e transformação. Essa camada é conhecida como Extração, Transformação e Carga, do inglês *Extract Transform Load* (ETL). A Camada OLAP mapeia as operações nos dados multidimensionais. No topo da arquitetura a camada do cliente, *front-end*. Essa camada contém as ferramentas de consultas, relatórios, análises e mineração de dados (HAN, 2005).

Figura 7: Arquitetura de um Data Warehouse (Tradução por Dayana Viana)



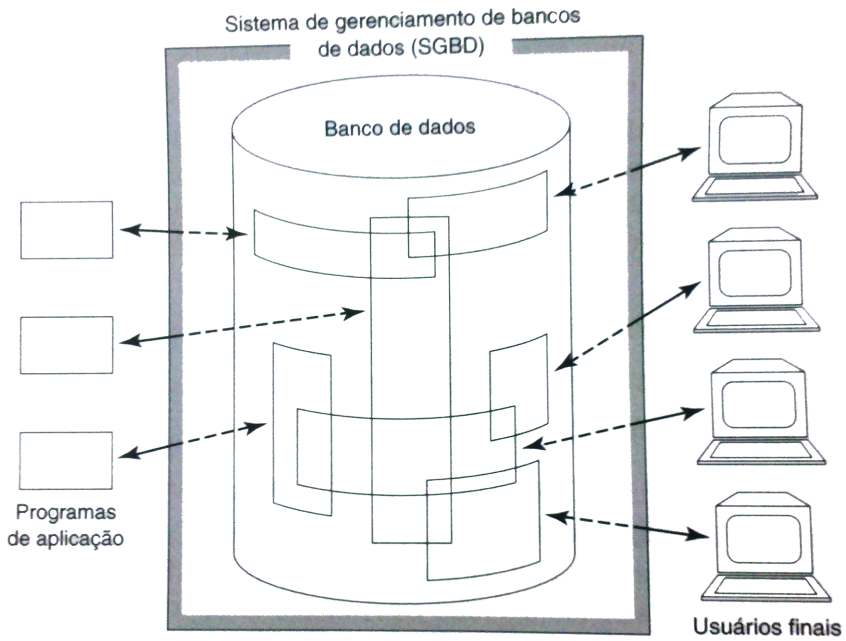
Fonte: (REBOUçAS, 2010).

As informações processadas são baseadas em consultas. Apesar de retornarem informações úteis, refletem diretamente as informações armazenadas. Ou seja, não refletem os padrões do banco de dados. Uma vez que a mineração de dados envolve uma análise mais profunda do que a OLAP, a utilização da mineração permitirá aplicações mais amplas do conhecimento obtido.

5 SQL SERVER

Um banco de dados é um sistema computacional para armazenamento de registros. Ou seja, é um repositório de dados que pode até mesmo ser comparado a um armário de arquivos. Os dados armazenados representam qualquer coisa que tenha sentido à organização. São tudo aquilo que é necessário para auxiliar a tomada de decisões. Intermediando o banco de dados e seus usuários existe uma camada conhecida como Sistema Gerenciador de Banco de Dados (SGBD). Todas as alterações solicitadas ao banco de dados são realizadas pelo SGBD (Figura 8). Uma grande vantagem desse ambiente é que o sistema de banco de dados proporciona um controle centralizado dos dados (DATE, 2000). Basicamente podemos aplicar o banco de dados em qualquer cenário que necessite armazenar informações como, por exemplo, em softwares de gestão e *Data Warehouse*.

Figura 8: Composição do Banco de Dados

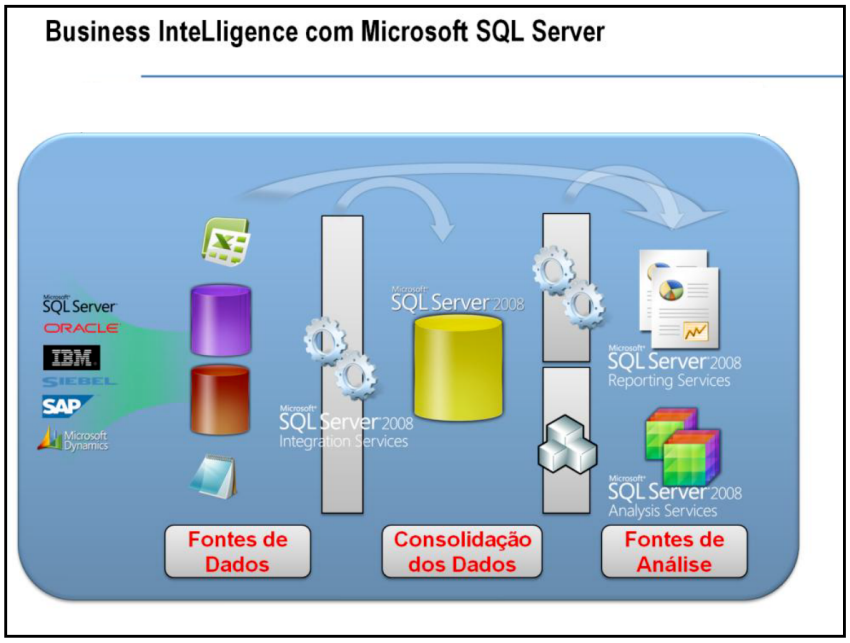


Fonte: (DATE, 2000).

O SQL Server é mais que um banco de dados, ele é uma plataforma de dados. Além de persistir os dados ele também possui todas as ferramentas necessárias para preparação de um Sistema de *Business Inteligence* (BI). Esse tipo de sistema facilita a transformação dos dados em informações para auxiliar as tomadas de decisões. Os componentes do SQL Server

são o *SQL Server Management Studio* e *SQL Server Bussiness Inteligence Development Studio* incluindo *Reporting Services*, *Analysis Services* e *Integration Services*. A interface mais utilizada da plataforma SQL Server é o *SQL Server Management Studio*, um software com foco na administração do banco de dados. A outra interface do produto, usada com foco no desenvolvimento, é o*SQL Server Bussiness Inteligence Development Studio*. Essa interface inclui outras ferramentas (Figura 9) como por exemplo geradores de relatórios (*Reporting Services*), operador de banco de dados multidimensionais (*Analysis Services*) e ferramenta ETL (*Integration Services*).

Figura 9: Estruturas do SQL Server



Fonte: Tutorial *Students to Business* (S2B) - Componentes do Banco de Dados.

O *Analysis Services* é uma ferramenta de *Data Mining* para apoiar as estratégias. A ferramenta possibilita obtenção de informações importantes que podem auxiliar no processo decisório da instituição. O *Analysis Services* oferece diversas soluções para implantar banco de dados analíticos usados para apoio à decisão em aplicativos de BI e até mesmo Excel. A partir de dados históricos já coletados são criados metadados que permitem medir, manipular e comparar esses dados. A partir da criação de um modelo dos dados, ele é então implantado em um servidor do *Analysis Services* como um banco de dados e disponibilizado para conexões externas como Excel ou outras ferramentas (MSDN, 2012).

Uma opção de ferramenta de apresentação para analisar os dados persistentes no *Analysis Services* é o Microsoft Office Excel. O Excel além de criar tabelas e realizar cálculos o é também um software de análise de dados. Para isso necessita do *Data Mining Add-in*, uma

extenção da ferramenta que é instalada separadamente. Após a instalação deve-se conectar a uma fonte de dados de Processamento Analítico Online (OLAP), disponibilizada pelo SQL Server. Através dessa conexão é possível exibir os dados como relatório de tabelas ou gráficos dinâmicos (MICROSOFT, 2012).

6 MICRODADOS DO CENSO DA EDUCAÇÃO SUPERIOR

Desde 1988, nossa Constituição da República Federativa dispôs a necessidade de armazenar dados estatísticos. As informações obtidas através desses dados contribuem para nortear políticas públicas e educacionais. Essa necessidade foi reforçada pelo art. 9º da Lei nº 9.394 em 1996. Surgiram então decretos que culminaram na criação do Decreto no 6.425 em 2008. Esse decreto prevê a obrigatoriedade de Instituições de Ensino Superior (IES) para responderem ao Censo (INEP, 2012).

Anualmente é realizado pelo Inep uma coleta dos dados sobre a educação superior. Um Questionário é enviado para as IES responderem perguntas sobre seus cursos, alunos e sua própria estrutura (Decreto no 6.425). Os dados coletados nos questionários reúnem informações sobre os diversos cursos oferecidos, vagas, inscrições, evasões, etc. Esses dados são então disponibilizados à sociedade em geral para manipulações estatísticas, porém mantendo sigilo quanto as informações dos alunos e instituições. Com os dados podemos obter informações como a situação atual e as tendências das IES e da comunidade (INEP, 2011).

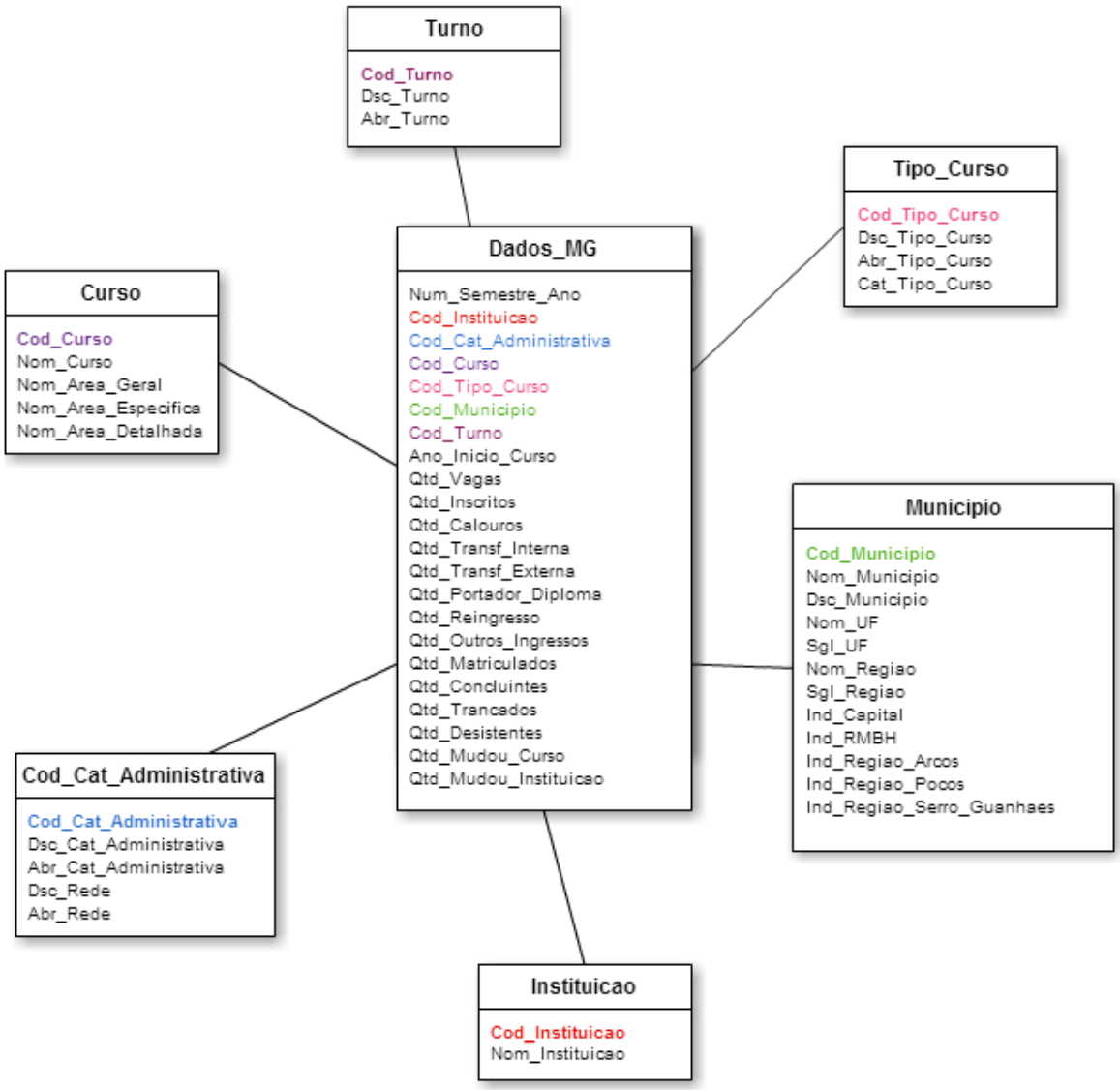
Os microdados coletados ficam disponíveis no portal do Inep: <http://portal.inep.gov-.br/basica-levantamentos-acessar> e são organizados em arquivos separados por ano. Os formatos para download são Texto ASCII, que permite a leitura por diversos softwares, e inputs para a leitura utilizando softwares SAS e SPSS. Para esse trabalho a base de dados utilizada foi manipulada e disponibilizada em formato excel, com as informações acumuladas entre o período de 2001 a 2008.

Os dados obtidos estão organizados em 7 planilhas: Turno, Município, Tipo Curso, Instituição, Categoria Administrativa, Curso e Dados MG. A planilha Turno armazena os turnos disponíveis dos cursos, são eles Diurno e Noturno. Em Município temos listados os 853 municípios de Minas Gerais. Para Tipo de Curso, os dados são divididos entre Graduação e cursos Tecnólogos. Em Instituições temos uma lista de 2217 estabelecimentos onde o nome dos mesmos foi preservado em sigilo. Categoria Administrativa classifica as instituições como Públicas ou Privadas. Na planilha Curso, além da listagem de 600 nomes de cursos, temos também informações sobre a área de cada curso. Finalmente em Dados MG é feita referência a todas planilhas citadas anteriormente, ordenadas por ano e semestre, juntamente com mais alguns dados adicionais como Ano de Início do Curso, Quantidade de Vagas, Quantidade de

Inscritos, Quantidade de Calouros, Quantidade de Transferência Interna, Quantidade de Transferência Externa, Quantidade de Portador de Diploma, Quantidade de Reingresso, Quantidade de Outros Ingressos, Quantidade de Matriculados, Quantidade de Concluintes, Quantidade de Matrículas Trancadas, Quantidade de Desistentes, Quantidade que Mudou de Curso e Quantidade que Mudou de Instituição.

Observando essas planilhas é possível abstrair um modelo de dados, representado pela Figura 10.

Figura 10: Modelo de Dados



Fonte: Criação da autora.

7 DESENVOLVIMENTO

7.1 Processo KDD

Este capítulo tem como objetivo apresentar o processo de KDD que foi aplicado sobre os Microdados do Censo da Educação Superior. Será explicado como foi executada cada etapa desse processo.

7.1.1 Seleção de Dados

Os microdados do Censo da Educação Superior apresentam informações coletadas por todo o país desde 1995. Porém nesse trabalho delimitou-se o escopo nos dados sobre Minas Gerais entre o período de 2001 a 2008. O Gerência de Tecnologia de Informação (GTI) já disponibiliza uma base, em formato Excel, com as informações do portal do Inep agrupadas dentro desse intervalo temporal. O que auxilia no processo de seleção, pois originalmente os dados de cada ano são disponibilizados separadamente. Apesar de trabalhar nessa base selecionada pelo GTI, dentro dela tem ainda um foco maior sobre as informações relacionadas à PUC Minas e ao curso de Sistemas de Informação da PUC Minas.

7.1.2 Pré-Processamento

Em relação aos relacionamentos, os dados trabalhados já estavam organizados de forma eficiente. Garantem agilidade e esforço reduzido nas análises das consultas por manter os campos que serão relacionados com o tipo inteiro.

O maior problema encontrado na base de dados foi a ausência de informação na planilha Dados_MG. Aplicando a função CONTAR.VAZIO do Excel, percebe-se que não havia falhas entre as colunas cujo os códigos se relacionam com as outras planilhas. Porém, observando as outras colunas, com os dados relativos às quantidades foi encontrado uma média de 53% dos campos vazios.

Substituir os valores ausentes em um conjunto de dados é muito importante. Os valores ausentes devem ser substituídos de forma que os valores inseridos não modifiquem os padrões já

existentes nos dados (PYLE, 1999). Pensando nisso e observando que o tipo de dados das colunas com valores ausentes eram números inteiros positivos, foi então preenchido estrategicamente os campos com o valor zero. Assim os padrões das quantidades atuais não foram alterados.

Nessa etapa foi identificado o código da Instituição foco do trabalho. Foi alterado o nome de "Instituição 1934" para "PUC Minas". Para identificar a Instituição foram filtrados os dados selecionando o Município de Arcos (código 310420) e o curso de Sistemas de Informação (código 518). Como resultado tivemos apenas o código de instituição 1934, indicando a comprovação do fato de que apenas a PUC Minas tem o curso de Sistemas de Informação no município de Arcos e que seu código nessa base é o 1934.

Foram também criados dois novos campos: Ano e Semestre, Afim de suprir a necessidade de análises anuais. A base de dados original apresenta esses valores juntos limitando assim as análises por semestre.

7.1.3 Transformação

Nessa etapa foi realizado o enriquecimento dos dados. Analisando, pode-se perceber que existem informações ocultas que poderiam ser explicitadas. Foram adicionadas então quatro novas colunas ao documento afim de agregar valor ao trabalho. Essas colunas informam a Quantidades de Ingressantes, Quantidades de Evasão, Relação Candidato Vaga e Taxa de Ocupação.

Quantidade de Ingressantes (Qt_Ing) é obtida a partir do somatórios das colunas de Quantidade de Calouros (Qt_Cal), Quantidade de Transferência Interna ($Qt\_TransfInt$), Quantidade de Transferência Externa ($Qt\_TransfExt$), Quantidade de Reingresso (Qt_Reing) e Quantidade de Outros Ingressos (Qt_Outros), conforme equação a seguir.

$$Qt_{Ing} = Qt_{Cal} + Qt_{TransfInt} + Qt_{TransfExt} + Qt_{Reing} + Qt_{Outros}$$

A Quantidade de Evasão (Qt_Ev) é obtida a partir do somatório das colunas Quantidade de Matrículas Trancadas (Qt_Tranc), Quantidade de Desistentes (Qt_Deist), Quantidade que Mudou de Curso ($Qt\_MudCurso$) e Quantidade que Mudou de Instituição ($Qt\_MudInst$).

$$Qt_{Ev} = Qt_{Tranc} + Qt_{Deist} + Qt_{MudCurso} + Qt_{MudInst}$$

A relação Candidato Vaga ($Cand\_vaga$) é obtida dividindo-se a Quantidade de Inscritos (Qt_Insc) pela Quantidade de Vagas (Qt_Vagas).

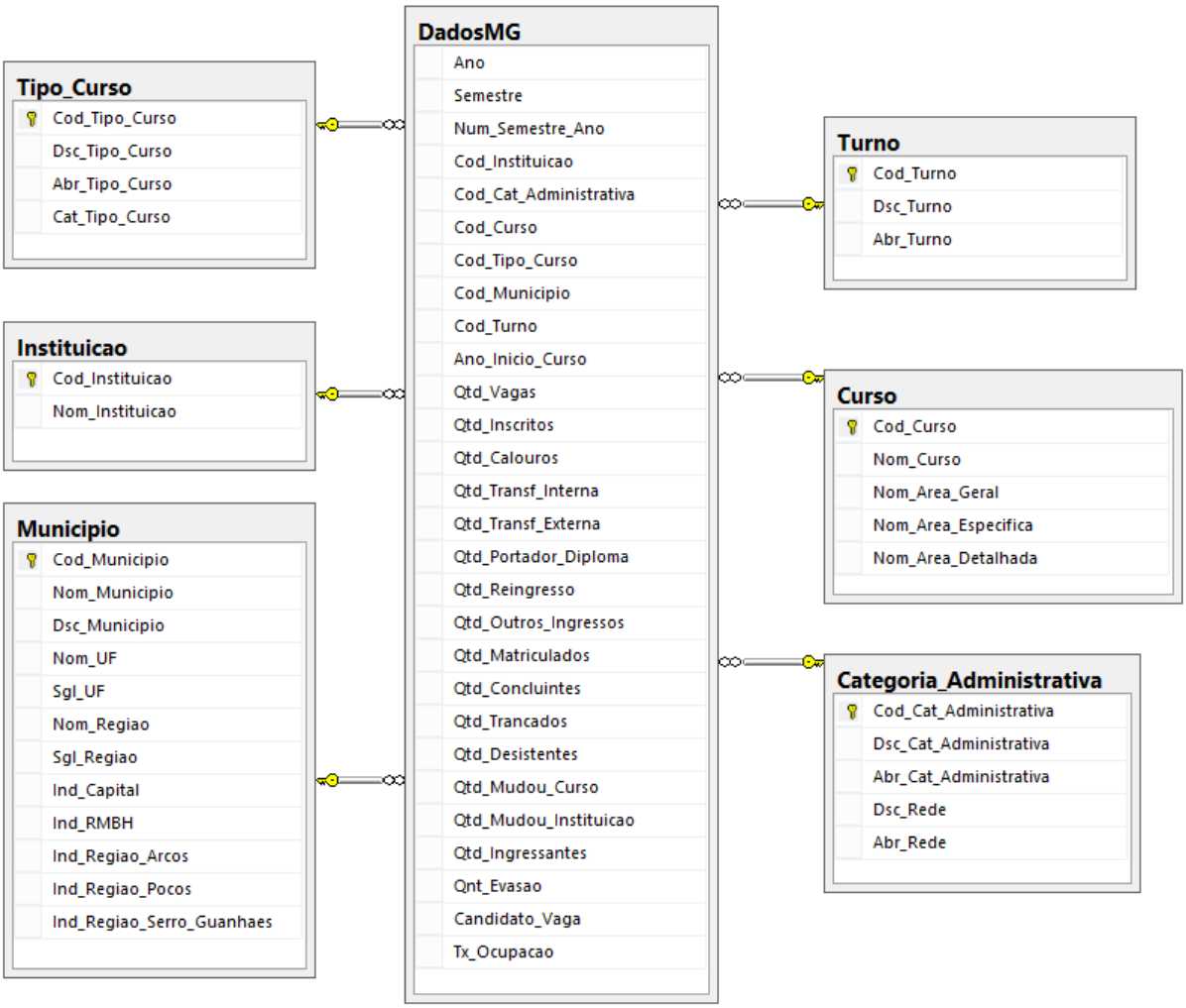
$$Cand\_vaga = \frac{Qt_{Insc}}{Qt_{Vagas}}$$

Por último, a Taxa de Ocupação (Qt_TxOcup) é representada em porcentagem e é o resultado da divisão entre a soma da Quantidade de Calouros (Qt_Cal), Quantidade de Transferência Interna ($Qt\_TransfInt$), Quantidade de Transferência Externa ($Qt\_TransfExt$), Quantidade de Reingresso (Qt_Reing), e Quantidade de Outros Ingressos (Qt_Outros) sobre a Quantidade de Vagas (Qt_Vagas).

$$Qt_{TxOcup} = \frac{Qt_{Cal}+Qt_{TransfInt}+Qt_{TransfExt}+Qt_{Reing}+Qt_{Outros}}{Qt_{Vagas}}$$

Após todo esse processo chegou-se ao modelo de dados representado pela Figura 11. Com o arquivo fonte tratado, foi feita a importação para o SQL Server 2012. Após isso o cubo foi criado utilizando-se o *Analysis Services*.

Figura 11: Modelo de Dados modificado



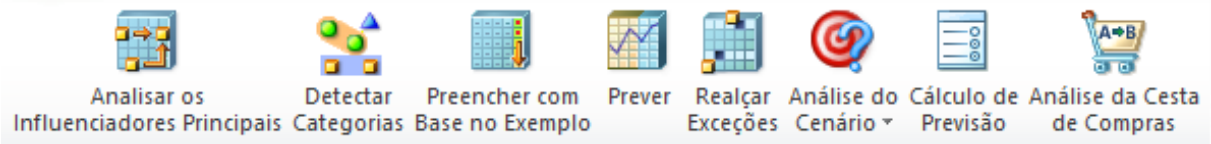
Fonte: Criação da autora.

7.1.4 Mineração de Dados

Na Mineração de Dados foi utilizado o Excel 2010 juntamente com o *Data Mining Add-In* para SQL Server 2012. Com o Excel é possível fazer uma Análise Descritiva dos Dados, ou seja, apresentar o que os dados atuais trazem de informações. O uso do *Add-In* viabiliza a análise de modelos aplicando os algorítimos de Mineração de Dados e visualizando os resultados em forma de gráficos.

Para gerar as Análises Descritivas dos Dados foi realizada uma conexão entre o Excel e o banco de dados. Então cria-se Gráficos Dinâmicos, utilizando essa conexão, selecionando os dados nas quais deseja que a análise seja feita. Nas Análises de Modelo de Dados a conexão é realizada com o *Analysis Services*, assim são aplicados os algoritmos ao cubo criado anteriormente. O *Add-In* possui diversos métodos que podemos utilizar para realizar as análises (Figura 12), porém foram utilizados apenas os métodos de prever, detectar categorias, análise de influências e análise da cesta de compras.

Figura 12: Ferramenta de Análise de Tabela



Fonte: Add-in Excel 2010.

O método Prever executa a previsão dos valores das colunas que forem selecionadas. Como padrão a quantidade de unidade de tempo a ser prevista é 5, porém esse valor pode ser modificado. Os valores gerados são adicionado ao final da tabela que foi utilizada. Também é gerado um gráfico mostrando em tracejados a evolução dos dados atuais para a previsão.

Em Análise de Influências selecionamos uma coluna para análise. Então é detectado quais colunas interferem nos valores da coluna desejada. O resultado é apresentado na forma de relatório, mostrando a porcentagem que cada elemento interfere na coluna destino.

O próximo método pode ser denominado como clusterização devido a sua semelhança nos resultados obtidos. Para Detectar Categorias selecionamos as colunas nas quais desejamos detectar alguma característica semelhante entre seus elementos. É possível também escolher a quantidade de categorias que se deseja criar ou deixar a detecção automática. Como resultado são apresentadas categorias de elementos com características semelhantes.

Na Análise da Cesta de Compras verifica-se itens que costumam aparecer juntos e expõe regras que podem servir em recomendações. Para esse método selecionamos a coluna que

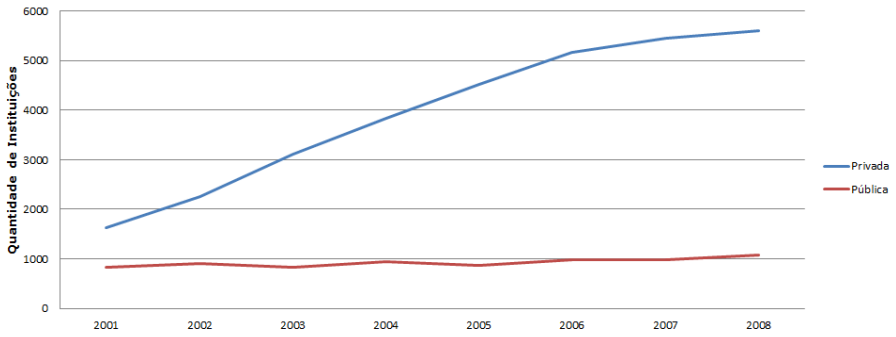
representa o ID da Transação, outra para representar o item e opcionalmente uma coluna para Valor do Item. Em configurações avançadas pode-se ainda definir o suporte mínimo, que é a quantidade mínima de ocorrências da regra no cenário atual, e também pode-se definir a probabilidade de regra mínima, que é a probabilidade daquela regra acontecer.

7.1.5 Interpretação

Após aplicar os diversos métodos citados anteriormente obtém-se os resultados. As primeiras análises foram feitas através de Gráficos Dinâmicos no Excel.

Na Figura 13 conta-se a quantidade de instituições durante o intervalo de anos definido nesse trabalho. Com base nisso, pode-se observar que a quantidade de instituições privadas veio aumentando linearmente, porém a partir de 2007 deu uma desacelerada. Já as instituições públicas mantiveram suas quantidades de instituições basicamente inalterada, com um crescimento irrisório comparado à rede administrativa oposta.

Figura 13: Evolução do Número de Instituições por Rede Administrativa - MG (2001-2008)

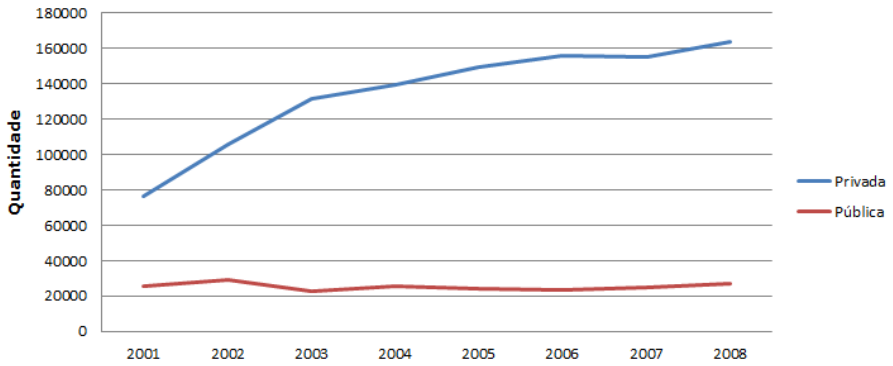


Fonte: Dados da Pesquisa.

Contamos também a quantidade de Ingressantes nas instituições (Figura 14). O Resultado foi bem semelhante ao observado anteriormente. A quantidade de ingressantes aumentou consideravelmente na rede privada e se manteve constante na rede pública. Podemos concluir com isso que devido ao aumento do número de instituições privadas, o número de ingressantes nessas instituições também aumentou. Comparando essa conclusão com os dados da PUC Minas (Figura 15) percebe-se que o mesmo não ocorre nessa instituição. O número de ingressantes se mantém praticamente inalterado durante os anos, aumentando apenas a partir de 2007. Analisando-se também os ingressos no curso de Sistemas de Informação (Figura 16), observamos um grande aumento da procura entre os anos de 2001 e 2003. Após 2003 houve uma desaceleração na procura por esse curso, porém seu crescimento não parou, apenas reduziu. Por

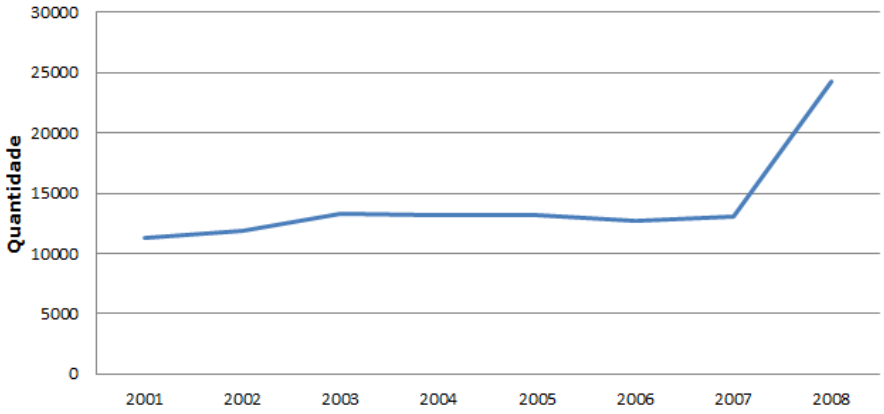
fim, analisamos a evolução dos ingressantes no curso de Sistemas de Informação da PUC Minas (Figura 17). Diferentemente do desempenho geral do curso, nessa instituição a quantidade de ingressantes aumentou consideravelmente até 2005, porém apresentou uma regressão em 2007. Após esse período voltou a crescer novamente.

Figura 14: Evolução de Ingressantes por Rede Administrativa - MG (2001-2008)



Fonte: Dados da Pesquisa.

Figura 15: Evolução de Ingressantes na PUC Minas (2001-2008)

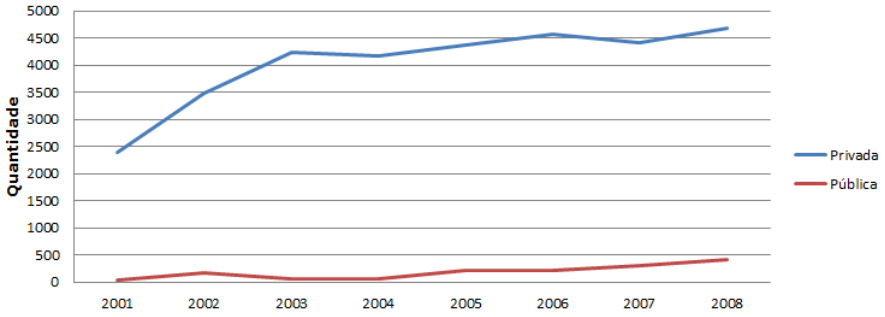


Fonte: Dados da Pesquisa.

Afim de observar o quão representativo é o curso de Sistemas de Informação comparado aos outros, foram geradas as Figuras 18 e Figura 19. Nelas podemos observar que dentre os cursos de todas a instituições de Minas Gerais, Sistemas de Informação está posicionado entre os top 10. Considerando apenas a PUC Minas, o curso sobe para a posição de quarto lugar em número de ingressantes em 2008.

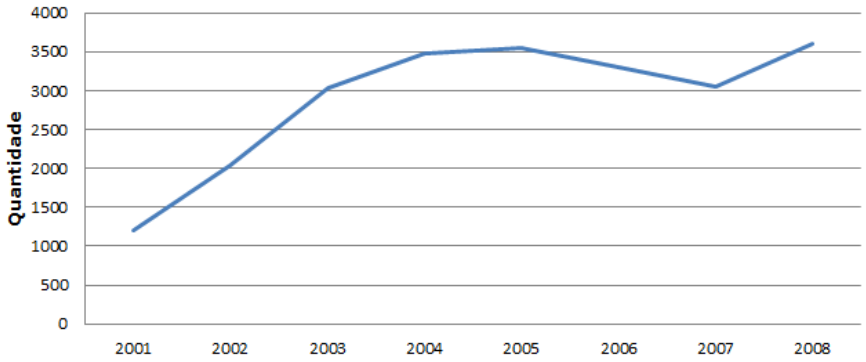
Afim de prever a quantidade de ingressantes para os próximos 6 semestres, foi utilizado o algoritmo de previsão do Add-in no Excel demonstrado na Figura 20. Com isso verifica-se uma queda na quantidade de ingressantes, tanto para os primeiros, quanto para os segundos semestres. Nesse mesmo gráfico aproveita-se para colocar também a representação da Quantidade

Figura 16: Evolução de Ingressantes por Rede Administrativa nos Cursos de Sistemas de Informação - MG (2001-2008)



Fonte: Dados da Pesquisa.

Figura 17: Evolução de Ingressantes na PUC Minas no Curso de Sistemas de Informação (2001-2008)



Fonte: Dados da Pesquisa.

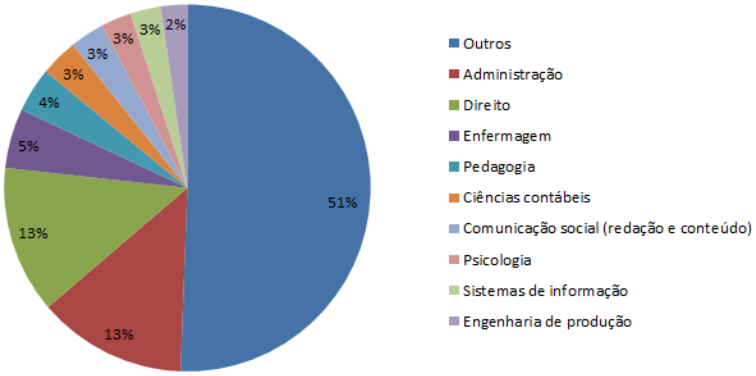
de Evasão. Essa se mantém em constante crescimento. Focando esses resultados no Curso de Sistemas de Informação da PUC Minas Figura (21) verifica-se uma previsão de instabilidade, com variação entre autos e baixos, na quantidade de ingressos e um ligeiro aumento na taxa de evasão.

O próximo algoritmo a ser utilizado é o de Detecção de Categorias. Nesse Algoritmo selecionamos as colunas que possivelmente terão características em comum e então é realizado o agrupamento de todos os seus elementos. Como resultado foram geradas 3 categorias:

- Categoria 1: Categoria com maior quantidade de elementos. Apresenta a quantidade de Candidatos/Vaga muito baixa, menor do que 1,1. A rede administrativa privada, turno noturno, área Educação, semestre 2 e instituição 2098 possuem relevância para que um elemento seja classificado nesse grupo.

- Categoria 2: Nessa categoria a relação candidato/vaga apresenta valores entre 1 e 5. Os

Figura 18: Participação dos 10 maiores Cursos em relação ao total de Ingressantes - MG (2008)



Fonte: Dados da Pesquisa.

fatores que influenciam os itens a pertencerem a essa categoria são: município de Belo Horizonte, rede administrativa privada, área geral em Ciências sociais, negócios e direito, curso de Direito, instituição PUC Minas, e semestre 1.

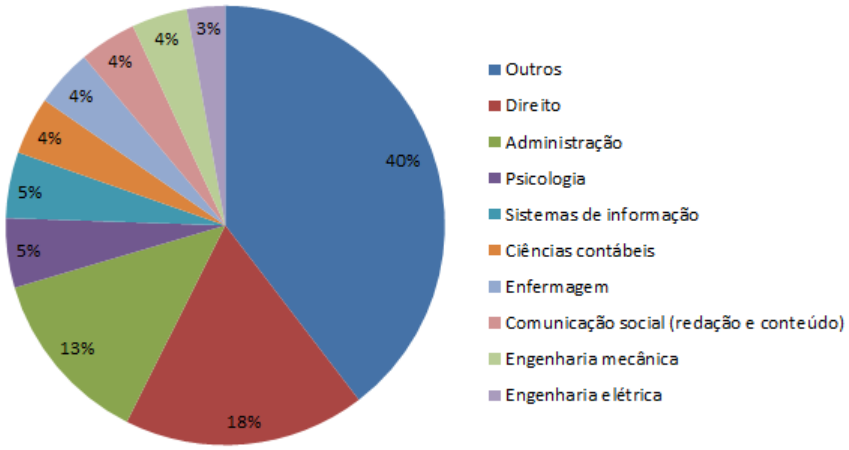
- Categoria 3: Para essa categoria entram os valores maiores que 5 na relação candidato/vaga. Também estão inclusos como influenciadores rede administrativa pública, turno diurno, municípios de viçosa e Ouro Preto, cursos de física e história, instituições 2047, 2058, dentre outros que podem ser visualizados na Figura 22.

Analisando os Gráficos Dinâmicos gerados pelo Excel é possível perceber que a evolução da quantidade de candidatos por vaga em média se mantém entre 1 e 2. Tanto para o curso de Sistemas de Informação da PUC Minas (Figura 24), quanto para os cursos de Sistemas de Informação em geral(Figura 23). Assim conclui-se que os padrões gerais para os cursos de Sistema de Informação podem ser aplicados ao mesmo curso na PUC Minas devido ao seu estreito índice de correlação.

Usando o algoritmo Análise de Influências sobre o Tx_Ocupcao temos como resultado a Figura 25. Nessa figura são apresentadas as colunas que interferem no resultado do campo escolhido. Observando a barra de impacto vemos que o fato de ser o segundo semestre do ano favorece uma ocupação menor que 50%. Já o fato de ser o primeiro semestre, turno noturno e IES privada favorece a ocupação apresentar probabilidades de 50% a 100%. Já a instituição PUC Minas, o município Belo Horizonte e o curso de Direito favorecem para que a ocupação utrapassar seu limite.

Aplicando o algoritmo Análise da Cesta de Compras (Associação) obtemos relação de

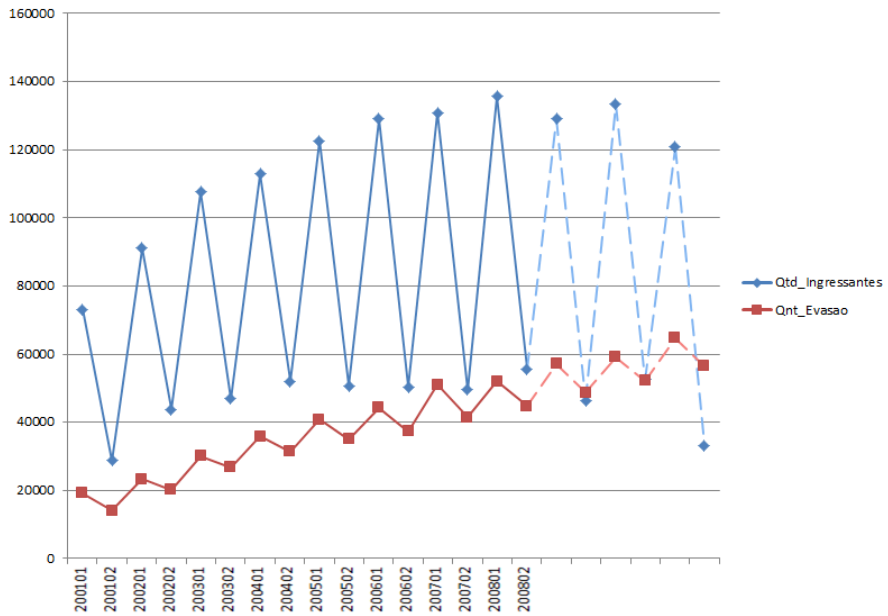
Figura 19: Participação dos 10 maiores Cursos em relação ao total de Ingressantes na PUC Minas (2008)



Fonte: Dados da Pesquisa.

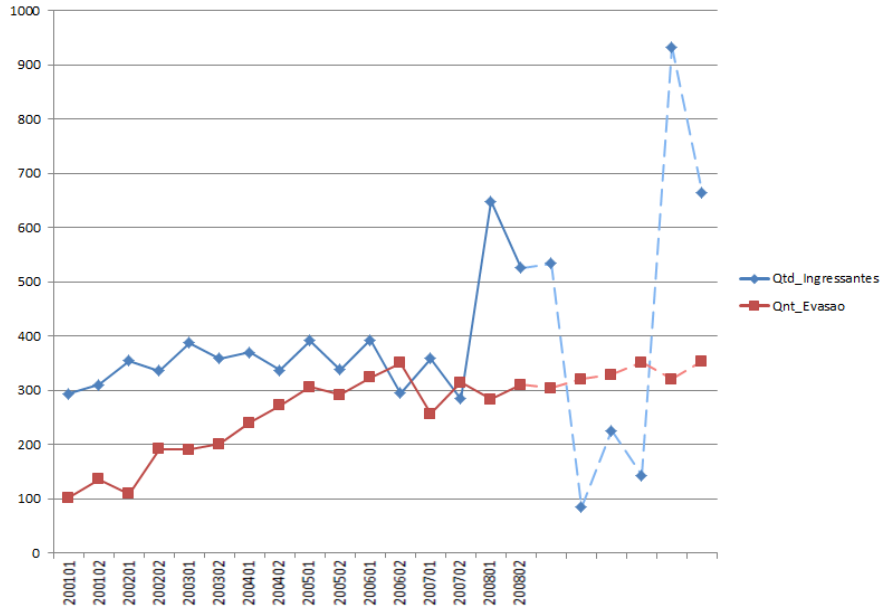
itens que acontecem em conjunto juntamente com recomendações. Para esse trabalho definiu-se como premissa que a taxa de ocupação seja maior que 50%. Como ID foi selecionado a instituição e como item os cursos. Para os resultados foi definido um suporte de 40% e confiança de 80%. O resultado disso é apresentado pela Figura 26 e Figura 27. Nelas observamos, por exemplo, que os cursos de Direito e Administração aparecem constantemente juntos quando a taxa de ocupação é maior que 50% nas suas instituições. O algorítmo também realiza recomendações, ou seja, observando a Figura 27, vemos que ela nos recomenda Enfermagem dado o ocorrência de Fisioterapia com 91% de precisão.

Figura 20: Previsão para Ingressantes e Evasão



Fonte: Dados da Pesquisa.

Figura 21: Previsão para Ingressantes e Evasão no Curso de Sistemas de Informação da PUC Minas



Fonte: Dados da Pesquisa.

Figura 22: Detecção de Categorias

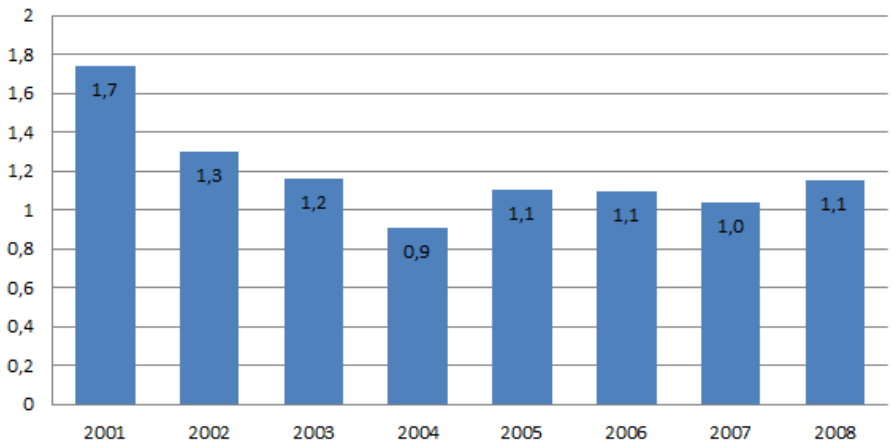
Nome da Categoria	Contagem de Linhas
Categoria 1	22417
Categoria 2	9913
Categoria 3	6660

Características da Categoria

Categoria	Coluna	Valor	Importância Relativa
Categoria 1	Candidato_Vaga	Muito Baixa:< 1.196063721	
Categoria 1	Dsc_Rede	Privada	
Categoria 1	Dsc_Turno	Noturno	
Categoria 1	Nom_Area_Geral	Educação	
Categoria 1	Nom_Instituicao	Instituição 2098	
Categoria 1	Semestre	02	
Categoria 2	Candidato_Vaga	Baixo:1.196063721 - 5.3485163824	
Categoria 2	Dsc_Municipio	Belo Horizonte	
Categoria 2	Dsc_Rede	Privada	
Categoria 2	Nom_Area_Geral	Ciências sociais, negócios e direito	
Categoria 2	Nom_Curso	Direito	
Categoria 2	Nom_Instituicao	PUC Minas	
Categoria 2	Semestre	01	
Categoria 3	Candidato_Vaga	Médio:5.3485163824 - 12.2570298288	
Categoria 3	Candidato_Vaga	Alto:12.2570298288 - 19.3347648736	
Categoria 3	Candidato_Vaga	Muito Alta:>= 19.3347648736	
Categoria 3	Dsc_Municipio	Viçosa	
Categoria 3	Dsc_Municipio	Ouro Preto	
Categoria 3	Dsc_Municipio	São João del Rei	
Categoria 3	Dsc_Rede	Pública	
Categoria 3	Dsc_Turno	Diurno	
Categoria 3	Nom_Area_Geral	Humanidades e artes	
Categoria 3	Nom_Area_Geral	Engenharia, produção e construção	
Categoria 3	Nom_Area_Geral	Agricultura e veterinária	
Categoria 3	Nom_Area_Geral	Ciências, matemática e computação	
Categoria 3	Nom_Curso	Física	
Categoria 3	Nom_Curso	História	
Categoria 3	Nom_Curso	Letras	
Categoria 3	Nom_Curso	Química	
Categoria 3	Nom_Curso	Matemática	
Categoria 3	Nom_Curso	Medicina	
Categoria 3	Nom_Instituicao	Instituição 2047	
Categoria 3	Nom_Instituicao	Instituição 2058	
Categoria 3	Nom_Instituicao	Instituição 2043	
Categoria 3	Nom_Instituicao	Instituição 1637	
Categoria 3	Nom_Instituicao	Instituição 2048	

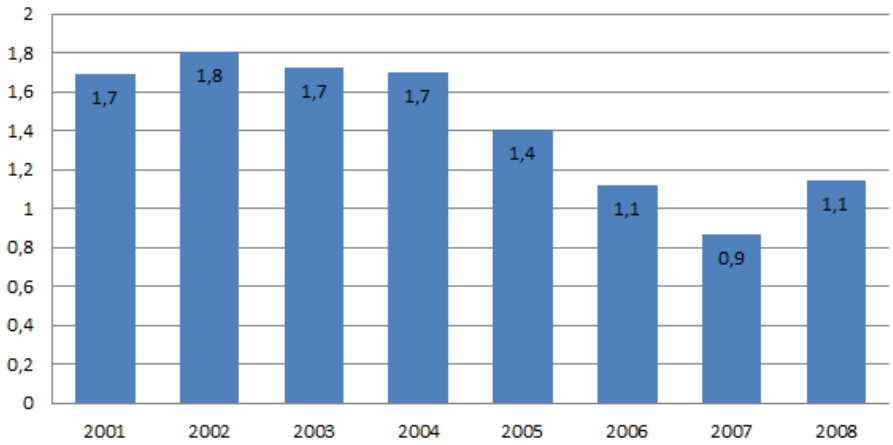
Fonte: Dados da Pesquisa.

Figura 23: Evolução Candidatos/Vaga nos Cursos de Sistemas de Informação (2001-2008)



Fonte: Dados da Pesquisa.

Figura 24: Evolução Candidatos/Vaga no Curso de Sistemas de Informação da PUC Minas(2001-2008)



Fonte: Dados da Pesquisa.

Figura 25: Influenciadores-chave e seu impacto sobre os valores de "Tx_Ocupacao".

Coluna	Valor	Favorece	Impacto Relativo
Semestre	01	50-100	
Dsc_Rede	Privada	50-100	
Dsc_Turno	Noturno	50-100	
Semestre	01	>100	
Nom_Instituicao	PUC Minas	>100	
Dsc_Municipio	Belo Horizonte	>100	
Nom_Curso	Direito	>100	
Semestre	02	<50	
Nom_Curso	Normal superior	<50	
Nom_Instituicao	Instituição 753	<50	
Ano	2008	<50	

Fonte: Dados da Pesquisa.

Figura 26: Associação entre itens

Itens Associados	% de Ocorrências
Direito, Administração	71
Ciências contábeis, Administração	63
Formação de professor de letras, Pedagogia	59
Pedagogia, Administração	57
Ciências contábeis, Direito	48
Enfermagem, Administração	46
Direito, Pedagogia	46
Sistemas de informação, Administração	45
Formação de professor de letras, Administração	44
Ciências contábeis, Direito, Administração	44
Fisioterapia, Enfermagem	43
Normal superior, Pedagogia	42
Formação de professor de matemática, Pedagogia	42
Formação de professor de história, Pedagogia	41
Formação de professor de matemática, Formação de professor de letras	41
Formação de professor de história, Formação de professor de letras	40
Enfermagem, Pedagogia	40
Direito, Pedagogia, Administração	40

Fonte: Dados da Pesquisa.

Figura 27: Recomendações

Item Selecionado	Recomendação	Qnt de Ocorrências	Qnt de Ocorrências Associadas	Precisão
Fisioterapia	Enfermagem	47	43	091%
Formação de professor de história	Formação de professor de letras	45	40	089%
Formação de professor de história	Pedagogia	45	41	091%
Normal superior	Pedagogia	51	42	082%
Ciências contábeis	Administração	73	63	086%

Fonte: Dados da Pesquisa.

8 CONCLUSÃO

De acordo com o trabalho apresentado conclui-se que o principal objetivo foi atingido. Foram aplicadas todas as etapas do processo de KDD em uma base com dados recebidos pelo Inep. Foram comparados resultados gerais com resultados obtidos pela PUC Minas e mais especificamente com o curso de Sistemas de Informação. Assim obtemos informações capazes de interferir no processo decisório da Universidade.

Os resultados apresentados durante o desenvolvimento trouxeram informações sobre como e porque a quantidade de ingressantes está evoluindo; quais fatores classificam em determinado grupo a Instituição, principalmente em relação aos seus candidatos/vaga; análises e evolução da quantidade de candidatos vaga; qual a importância do curso de Sistemas de Informação dentro e fora da PUC Minas, considerando sua procura; previsões para ingressos e evasões gerais e para o curso de Sistemas de Informação na PUC Minas; quais são os influenciadores da taxa de ocupação, quais são os principais cursos que aparecem juntos com grande ocupação nas instituições e finalmente recomendações de um curso dada a presença de outro.

Com esse estudo pudemos observar alguns comportamentos atuais dos estudantes, assim como prever a quantidade de ingressos e evasões para os próximos três anos. Observar essa tendência pode ajudar a PUC Minas a não tomar decisões erradas quanto às sua expectativas. Ou seja, a Universidade pode estar esperando um aumento de alunos, quando na verdades as previsões mostram uma recessão para os próximos dois anos (2009 e 2010) principalmente para o curso de Sistemas de Informação.

Durante a construção desse trabalho percebeu-se dificuldade quanto a ausência de dados, muitos campos como a quantidade de vagas, alunos matriculados e ingressantes por exemplo eram simplesmente apresentadas em branco dificultando assim as análises e fazendo com que essas ausências fossem tratadas manualmente. Outra dificuldade foi encontrada também ao desenvolver perguntas para que então fosse buscada suas respostas nos dados, não se sabia a real necessidade de informações da Universidade, então trazer resultados de forma clara e objetiva para a análise se tornou um grande desafio.

Por não se saber a real necessidade de informações da PUC Minas o resultados apresentados podem deixar a desejar nas suas necessidades no processo decisório. Existe muita informação que ainda pode ser obtida da base de dados utilizada.

8.1 Trabalhos Futuros

Espera-se que esse trabalho não seja apenas o final de uma pesquisa, mas sim o início de um grande projeto. Como proposta para trabalhos futuros propõem-se a utilização de dados diretamente da fonte, o Inep. Assim será possível obter dados mais atualizados, uma vez que não será necessário esperar o GTI trabalhar e distribuir esses dados.

Entrevistas com gestores da Universidade também são indicadas para que se conheça as reais necessidades de conhecimento desejada. Assim pode-se concentrar os esforços em obter apenas as informações necessárias.

Um estudo de caso aprofundado nesses dados poderiam mostrar na íntegra a complexidade das situações reais e apresentar resultados explanatórios e descritivos para as IES de forma geral.

REFERÊNCIAS

CARDOSO, O. N. P. et al. Gestão do conhecimento usando data mining : estudo de caso na Universidade Federal de Lavras *. v. 42, n. 3, p. 495–528, 2008.

DATE, C. J. *Introdução a Sistemas de Banco de Dados*. [S.l.: s.n.], 2000. ISBN 8535205608.

FAYYAD et al. From Data Mining to Knowledge Discovery in Databases. p. 37–54, 1996.

FILHO, R. L. L. e. S. et al. A evasão no ensino superior Brasileiro. *Higher Education*, p. 641–659, 2007.

HAN et al. *Data Mining Concepts and Techniques*. [S.l.: s.n.], 2005. ISBN 9781558609013.

INEP. *Censo da educação superior: 2010 – resumo técnico.* Brasília: [s.n.], 2012. ISBN 9788578630188.

INEP, M. *Censo da Educação Superior*. 2011. Disponível em: <http://portal.inep.gov.br/web/censo-da-educacao-superior>.

LUAN, J. Data mining applications in higher education. *SPSS Executive Report*, 2002. Disponível em: <http://www.insol.lt/media/collateral/modeling/education.pdf>.

MENDES, M. A demanda por vagas no ensino superior: análise dos vestibulares da ufmg na década de 90. p. 1–26, 1997.

MICROSOFT. *Visão geral do OLAP*. 2012. Disponível em: <http://office.microsoft.com/pt-br/excel-help/visao-geral-do-olap-processamento-analitico-online-HP010177437.aspx>.

MSDN, M. *Analysis Services*. 2012. Disponível em: <http://msdn.microsoft.com/pt-br/library/bb522607.aspx>.

PYLE et al. *Data Preparation for Data Mining*. [S.l.: s.n.], 1999. ISBN 4159822665.

REBOUçAS, F. *Data Warehouse*. 2010. Disponível em: <http://www.infoescola.com/informatica/data-warehouse>.