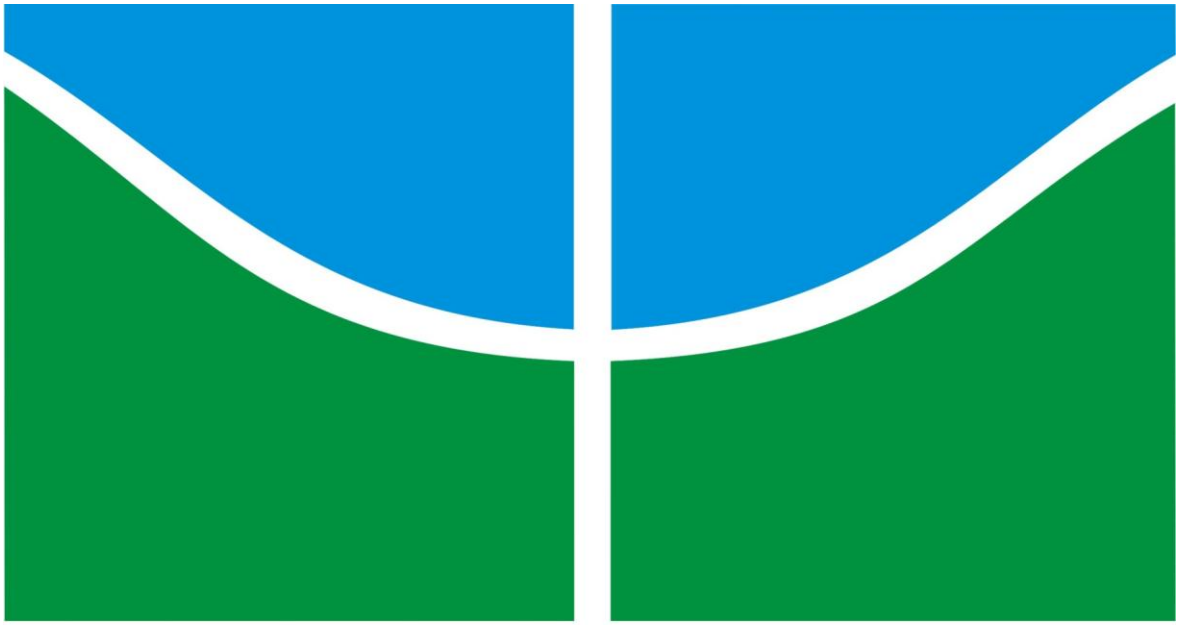


APOIO À TOMADA DE DECISÃO EM TRANSPORTE: APLICAÇÃO DO PROCESSO DE MINERAÇÃO DE DADOS

WANDERLEY GONÇALVES FREITAS

DISSERTAÇÃO DE MESTRADO EM TRANSPORTE

DEPARTAMENTO DE ENGENHARIA CIVIL E AMBIENTAL

FACULDADE DE TECNOLOGIA

UNIVERSIDADE DE BRASÍLIA

UNIVERSIDADE DE BRASÍLIA
FACULDADE DE TECNOLOGIA
DEPARTAMENTO DE ENGENHARIA CIVIL E AMBIENTAL
PROGRAMA DE PÓS-GRADUAÇÃO EM TRANSPORTE

APOIO À TOMADA DE DECISÃO EM TRANSPORTE: APLICAÇÃO DO PROCESSO DE MINERAÇÃO DE DADOS

WANDERLEY GONÇALVES FREITAS

ORIENTADORA: YAEKO YAMASHITA

DISSERTAÇÃO DE MESTRADO EM TRANSPORTE

PUBLICAÇÃO: T.DM-007A/2012

BRASÍLIA/DF: MARÇO DE 2012

UNIVERSIDADE DE BRASÍLIA

FACULDADE DE TECNOLOGIA

DEPARTAMENTO DE ENGENHARIA CIVIL E AMBIENTAL

PROGRAMA DE PÓS-GRADUAÇÃO EM TRANSPORTE

APOIO À TOMADA DE DECISÃO EM TRANSPORTE: APLICAÇÃO DO PROCESSO DE MINERAÇÃO DE DADOS

WANDERLEY GONÇALVES FREITAS

DISSERTAÇÃO DE MESTRADO SUBMETIDO AO DEPARTAMENTO DE ENGENHARIA CIVIL E AMBIENTAL DA FACULDADE DE TECNOLOGIA DA UNIVERSIDADE DE BRASÍLIA, COMO PARTE DOS REQUISITOS NECESSÁRIOS PARA A OBTENÇÃO DO GRAU DE MESTRE EM TRANSPORTE.

APROVADA POR:

_____

PROF[a]. YAEKO YAMASHITA, PhD., (ENC- UnB)
(orientadora)

_____

PROF[o]. JOSÉ MATSUO SHIMOISHI, DR., (ENC- UnB)
(examinador interno)

_____

PROF[o]. WILLER LUCIANO CARVALHO, DR., (ULBRA-TO)
(examinador externo)

BRASÍLIA/DF: 23 DE MARÇO DE 2012

FICHA CATALOGRÁFICA

FREITAS, WANDERLEY GONÇALVES

Apoio à tomada de decisão em transportes: aplicação do processo de mineração de dados [Distrito Federal], 2012.

xvii, 147p., 210x297 mm (ENC/FT/UnB, Mestre, Transportes, 2012).

Dissertação de Mestrado – Universidade de Brasília. Faculdade de Tecnologia.

Departamento de Engenharia Civil e Ambiental.

1. Planejamento em Transporte	2. Processo de Tomada de Decisão
3. Mineração de Dados	4. Regras de Associação
I. ENC/FT/UnB	II. Título (série)

REFERÊNCIA BIBLIOGRÁFICA

FREITAS, Wanderley Gonçalves (2012). Apoio à tomada de decisão em transportes: aplicação do processo de mineração de dados. Dissertação de Mestrado, Publicação T.DM-007A/2012, Departamento de Engenharia Civil e Ambiental, Universidade de Brasília, Brasília, DF, 147p.

CESSÃO DE DIREITOS

AUTOR: Wanderley Gonçalves Freitas

TÍTULO DA DISSERTAÇÃO: Apoio à tomada de decisão em transporte: aplicação do processo de mineração de Dados.

GRAU/ANO: Mestre / 2012.

É concedida à Universidade de Brasília permissão para reproduzir cópias desta dissertação de mestrado e para emprestar ou vender cópias somente para propósitos acadêmicos e científicos. O autor reserva outros direitos de publicação e nenhuma parte desta dissertação de mestrado pode ser reproduzida sem a autorização do autor.

Wanderley Gonçalves Freitas
Email : wanderley.ppgt@gmail.com

DEDICATÓRIA

Deus quer,
O homem sonha,
A obra nasce.
(Fernando Pessoa)

Dedico este projeto
a minha esposa Gláucia,
ao meu filho Luan,
a minha mãe Irenilde,
as minhas irmãs: Vanderléia, Josirene e Jacirene e
aos meus amigos.

AGRADECIMENTOS

A Deus que esteve ao meu lado neste projeto; que deu-me força interior para perseverar, a fim de contribuir para a edificação do saber científico.

A minha mãe Irenilde, por me ensinar seus valores e acreditar que a educação é a base mais sólida para se conquistar objetivos na vida

A minha esposa Gláucia e ao meu filho Luan, pela paciência e compreensão nos longos períodos ausentes e mesmo assim, sempre me deram força para que eu não desistisse desse projeto acadêmico.

As minhas irmãs Vanderléia, Josirene e Jacirene que sempre estiveram ao meu lado nesta empreitada, me incentivando a dar o meu melhor.

A todos os alunos do curso de mestrado com os quais tive a felicidade de conviver e àqueles que me deram suporte em questões burocráticas para o desenvolvimento desse projeto.

A minha orientadora, Prof[a]. Yeako, pela sábia orientação e todas as oportunidades e, principalmente, pelas frases que me fez iniciar o mestrado: "*Acho que vai valer a pena*!", "*mestrado é um grande aprendizado", "Conhecimento ninguém tira de você, mas bens materiais são facilmente subtraídos, retirados e repartidos"* e "*por ter estado certa disso!"* Pelo privilégio de poder conviver, aprender e trabalhar ao seu lado com tantos desafios e de forma tão gratificante. Pelo exemplo, dedicação à Universidade e incentivo ao desenvolvimento dos seus alunos. Palavras não conseguem expressar minha gratidão e admiração.

A todos os professores do mestrado em transportes, por todo o conhecimento repassado e pelo exemplo de amor à vida acadêmica.

A todos vocês, muito obrigado!

RESUMO

APOIO À TOMADA DE DECISÃO EM TRANSPORTE: APLICAÇÃO DO PROCESSO DE MINERAÇÃO DE DADOS

Os sistemas de apoio à tomada de decisão fornecem ferramentas e modelos analíticos para analisar grande quantidade de dados. Especificamente, o sistema de transporte, por suas características, envolve grande quantidade e variedades de informações necessárias ao seu planejamento, gestão e controle, principalmente, devido a sua complexidade e enorme quantidade de informações existentes as quais são armazenadas em bancos de dados. Porém, nem sempre, esses dados são explorados em seu máximo potencial, pois o aumento significativo de informações supera a capacidade humana de interpretar e examinar esses dados em curto espaço de tempo. Dessa forma, para propor soluções para problemas de transportes e apoiar suas atividades, o especialista em transportes, no processo de análise e de tomada de decisão, precisa de informações e dados coerentes e adequados para subsidiar os estudos de transporte. Desse modo, este trabalho buscou o desenvolvimento de uma metodologia que utilizasse o processo de mineração de dados para análise de banco de dados de transporte, a fim de extrair padrões e regras significativas e adequadas às necessidades de informação; para servir de base à tomada de decisão em estudo de transporte durante o processo de planejamento. Para tal, fundamentou-se na metodologia proposta nos princípios e conceitos existentes em sistemas inteligentes de informação, com ênfase em mineração de dados, planejamento de transporte e processo de tomada decisão. Para validação da metodologia, foi realizado um estudo de caso que restringe a caracterização do transporte escolar rural utilizando-se da mineração de dados na análise dos dados publicados na pesquisa web realizada pelo CEFTRU (2007a, 2007b). Dessa forma, pôde-se constatar que a metodologia é apropriada para extrair informações válidas e úteis em base de dados de transporte, para fornecer aos tomadores de decisão subsídios objetivos, para formulação de ações que visem à melhoria desse serviço de transporte e a elaboração do planejamento de transporte. Conclui-se que a metodologia atingiu os objetivos propostos, consistindo em uma ferramenta relevante, para análise de grande quantidade de dados de forma ágil e eficiente, para o estudo do sistema de transporte.

Palavras chaves: Planejamento em transporte, Processo de tomada de decisão, Mineração de dados, Regra de associação, Descoberta de conhecimento em base de dados, Medidas de interesse.

ABSTRACT

SUPPORT FOR DECISION MAKING IN TRANSPORT: APPLICATION OF DATA MINING PROCESS

Decision-making support systems provide tools and analytical models to handle large amounts of data. Specifically the transport system, by nature, involves a large and varied amount of information required for planning, controlling and the management processes, mainly due to the great complexity and the enormous volume of information stored in databases. Nevertheless, these data are not always explored in their full potential, due to the significant increase of information that exceeds the human capacity to interpret and examine these data timely. Thus, in order to propose solutions to transport problems and support his activities, the transport specialist, at the analysis and decision making process, needs consistent and suitable information and data to support his studies of transport. Therefore, this study aimed to develop a methodology that utilizes the process of data mining for database analysis of transport sector in order to extract meaningful patterns and rules and identify appropriate information needs, to serve as a basis for making decision in a study of transport during the planning process. In order to accomplish this task, the proposed methodology has been based on the existing principles and concepts of intelligent information systems, emphasizing data mining, transport planning and decision making process. To validate the methodology, we performed a case study that restricts the characterization of the rural school transportation using data mining analysis on the published web survey conducted by CEFTRU (2007a, 2007b). Thus, it is possible to affirm that the methodology is suitable for extracting valid and useful information from transport databases, to provide objective subsidies for decision makers and to establish actions for the improvement of the transport service and the development of transport planning. The conclusion was drawn that the methodology achieved its objectives, consisting of a relevant instrument, developed to analyze large amounts of data quickly and efficiently, to study the transport system.

Keywords: Transportation planning, Decision-making process, Data mining, Association rules, Knowledge discovery in databases, Measures of interest

SUMÁRIO

1 INTRODUÇÃO 1

1.1 APRESENTAÇÃO 1

1.2 PROBLEMA 2

1.3 HIPÓTESE 3

1.4 OBJETIVO 3

1.5 JUSTIFICATIVA 3

1.6 METODOLOGIA 4

1.7 ESTRUTURA DA DISSERTAÇÃO 7

2 PLANEJAMENTO DE TRANSPORTE E DIAGNÓSTICO 8

2.1 APRESENTAÇÃO 8

2.2 CONCEITOS DE PLANEJAMENTO 8

2.3 CONCEITOS DE PLANEJAMENTO DE TRANSPORTE 9

2.4 PLANAJAMENTO TRADICIONAL 10

2.5 PLANEJAMENTO ESTRATÉGICO 11

2.6 PLANEJAMENTO INTEGRADO 14

2.6.1 Nível estratégico 14

2.6.2 Nível tático 15

2.6.3 Nível operacional 16

2.7 DIAGNÓSTICO 17

2.7.1 Processo de elaboração do diagnóstico 19

2.8 TÓPICOS CONCLUSIVOS EM PLANEJAMENTO DE TRANSPORTE 21

3 SISTEMA INTELIGENTE DE INFORMAÇÃO 23

3.1 APRESENTAÇÃO 23

3.2 TOMADA DE DECISÃO 23

3.2.1 Tipo de decisão 24

3.2.2 Processo de tomada de decisão 25

3.3 CONCEITOS BÁSICOS DE SISTEMA INTELIGENTES 26

3.4 MINERAÇÃO DE DADOS 28

3.4.1 Conceitos básicos de mineração de dados 29
3.4.2 Processo de KDD 32
3.4.3 Atividades e tarefas de mineração de dados 34
3.4.3.1 Tarefa de classificação 36
3.4.3.2 Tarefa de seleção de atributos 37
3.4.3.3 Tarefa de regressão (estimativa) 37
3.4.3.4 Tarefa de agrupamento 37
3.4.3.5 Tarefa de associação 37
3.4.4 Técnica de mineração de dados 38
3.4.4.1 Regra de associação 38
3.4.4.2 Árvore de decisão 40
3.4.4.3 Rede neurais 40
3.4.5 Compreender as medidas de interesse das regras associativas 41
3.4.5.1 Medida de interesse objetiva 41
3.4.5.2 Medida de interesse subjetiva 44
3.4.6 Entendo característica dos dados 45
3.4.7 Validação de resultado 45
3.4.8 Metodologias 46
3.4.8.1 CRISP-DM 46
3.4.8.2 SEMMA 46
3.4.9 Ferramentas 47
3.4.9.1 WEKA 47
3.4.9.2 Transporte mining 48
3.5 TÓPICOS CONCLUSIVOS DE SISTEMAS INTELIGENTES 49

4 METODOLOGIA DE APOIO À TOMADA DE DECISÃO EM TRANSPORTE UTILIZANDO O PROCESSO DE MINERAÇÃO DE DADOS 51
4.1 APRESENTAÇÃO 51
4.2 METODOLOGIA PROPOSTA DE ELABORAÇÃO DE UM SISTEMA DE APOIO À DECISÃO EM TRANSPORTE UTILIZANDO O PROCESSO DE MINERAÇÃO DE DADOS 52
4.3 DESCRIÇÃO DA METODOLOGIA 55
4.3.1 Etapa I – Concepção (requisitos) 55

4.3.2 Etapa II - Elaboração (modelagem) 57
4.3.3 Etapa III – Construção (implementação) 67
4.3.4 Etapa IV – Transição (interpretação) 68
4.4 TÓPICOS CONCLUSIVOS DA METODOLOGIA PROPOSTA 71

5 ESTUDO DE CASO: CARACTERIZAÇÃO DO TRANSPORTE ESCOLAR RURAL UTILIZANDO O PROCESSO DE MINERAÇÃO DE DADOS 73
5.1 APRESENTAÇÃO 73
5.2 CONTEXTUALIZAÇÃO DO OBJETO DE ESTUDO: TRANSPORTE ESCOLAR RURAL BRASILEIRO 74
5.2.1 Representação do conhecimento 75
5.2.1.1 Ontologia 75
5.2.1.2 Redes semânticas 77
5.2.2 Conceituando o sistema de transporte escolar rural 79
5.2.3 Rede semântica do transporte escolar rural 81
5.3 BASE DE DADOS UTILIZADA: TRANSPORTE ESCOLAR RURAL 83
5.3.1 Apresentação resumida da caracterização do TER - relatório web 85
5.3.1.1 Serviço 85
5.3.1.2 Clientela 87
5.3.1.3 Recursos 88
5.4 APLICAÇÃO DA METODOLOGIA PROPOSTA 89
5.4.1 Etapa I – Concepção (requisitos) 90
5.4.2 Etapa II – Elaboração (modelagem) 93
5.4.3 Etapa III – Construção (implementação) 98
5.4.4 Etapa IV – Transição (interpretação) 101
5.5 TÓPICOS CONCLUSIVOS DO ESTUDO DE CASO 118

6 CONCLUSÕES 120
6.1 APRESENTAÇÃO 120
6.2 CONSIDERAÇÕES SOBRE A APLICABILIDADE DA METODOLOGIA 120
6.3 AVALIAÇÃO DA METODOLOGIA PROPOSTA 122
6.4 SUGESTÕES PARA FUTURAS PESQUISAS 124

REFERÊNCIAS BIBLIOGRÁFICAS 125

APÊNDICES **132**

A - MANUAL SIMPLIFICADO DO SOFTWARE MINERAÇÃO DE DADOS **133**

ANEXOS **138**

A – QUESTIONÁRIO WEB SIMPLIFICADO **139**

B – REDE SEMÂNTICA DO SISTEMA DE TRANSPORTE ESCOLAR RURAL SIMPLIFICADA **143**

LISTA DE FIGURAS

Figura 1.1 : Estrutura da metodologia adotada para o desenvolvimento da pesquisa 5

Figura 1.2 : Estrutura da metodologia que utilize o processo de mineração de dados 6

Figura 2.1 : Processo de planejamento contínuo. Fonte: Papacosta e Provedouros (1993) 10

Figura 2.2 : Processo estratégico. Fonte: Guell (1997) 12

Figura 2.3 : Processo de planejamento integrado. Fonte: Magalhães e Yamashita (2008) 14

Figura 2.4 : Visão conexão do diagnóstico. Fonte: adaptação de Tedesco (2008)..... 18

Figura 3.1 :Tipos de decisão. Fonte adaptação de Laudon e Laudon (2007) 24

Figura 3.2: Processo de tomada de decisão. Fonte: adaptação de Simon (1969)..... 25

Figura 3.3 : Pirâmide da informação. Fonte: adaptação de Nonaka e Takeuchi(1997) 26

Figura 3.4: Multidisciplinaridade da MD. Fonte: adaptação de Fayyad *et al* (1996a)..... 28

Figura 3.5 : Processo de KDD. Fonte: adaptação de Fayyad *et al*.(1996a)..... 33

Figura 3.6 : Visão hierárquica do processo de mineração de dados..... 34

Figura 3.7: Tarefas de MD. Fonte: adaptação de Santos e Azevedo (2005) 35

Figura 3.8 : Árvore de classificação. Fonte (WEKA) 36

Figura 3.9 : Objetivos das medidas de interesse . Fonte: Geng e Hamilton (2006) 42

Figura 4.1: Metodologia proposta 52

Figura 4.2 : Fluxograma da metodologia proposta..... 54

Figura 4.3 : Etapa de concepção da metodologia 55

Figura 4.4 : Etapa de elaboração da modelagem..... 58

Figura 4.5: Etapa de construção da metodologia..... 67

Figura 4.6: Etapa de transição da metodologia..... 69

Figura 5.1 : Elementos do STER. Fonte (CEFTRU/FNDE, 2008a, 2008b). 80

Figura 5.2 : Rede semântica do STER. Fonte (CEFTRU/FNDE, 2008a, 2008b) 81

Figura 5.3 : Estrutura semântica do TER. Fonte (CEFTRU/FNDE, 2008a, 2008b)..... 82

Figura 5.4 : Elementos da rede semântica. Fonte (CEFTRU/FNDE, 2008a, 2008b)..... 82

Figura 5.5 : Elementos de Representação. Fonte (CEFTRU/FNDE, 2008a, 2008b) 83

Figura 5.6: Gráficos da frota do TER Fonte: CEFTRU/FNDE (2007b) 86

Figura 5.7: Gráficos da clientela atendida no TER. Fonte: CEFTRU/FNDE (2007b)..... 88

Figura 5.8: Gráficos das fontes de recursos no TER. Fonte: CEFTRU/FNDE (2007b) 88

Figura 5.9 : Conhecendo o TER. Fonte: adaptação de Carvalho (2011)..... 89

Figura 5.10 : Estrutura semântica do TER. Fonte: CEFTRU (2007a, 2007b) 91

Figura 5.11 : Elementos da rede semântica–serviço. Fonte: CEFTRU (2007a, 2007b)..... 91

Figura 5.12 : Elementos da rede semântica–clientela. Fonte: CEFTRU (2007a, 2007b).... 92

Figura 5.13 : Elementos da rede semântica – recursos . Fonte: CEFTU (2007a, 2007b) .. 92

Figura 5.14 : Base de dados na planilha *excel* a partir da consulta *SQL* 95

Figura 5.15 : Arquivo gerado no formato *ARFF* 98

Figura 5.16 : Resultados dos testes na base de dados da pesquisa web 99

Figura 5.17 : Tela de parâmetro de entrada do algoritmo *apriori* 100

Figura 5.18 : Tela de saída do algoritmo *apriori*..... 101

Figura 5.19: Gráficos das variáveis da regra 99 . Fonte: CEFTRU/FNDE (2007b) 104

Figura 5.20: Gráficos das variáveis da regra 30 . Fonte: CEFTRU/FNDE (2007b) 107

Figura 5.21: Gráficos das variáveis da regra 6 . Fonte: CEFTRU/FNDE (2007b) 109

Figura 5.22: Gráficos das variáveis da regra 23. Fonte: CEFTRU/FNDE (2007b) 111

Figura 5.23: Gráficos das variáveis da regra 1. Fonte: CEFTRU/FNDE (2007b) 113

Figura 5.24: Gráficos das variáveis da regra n.9. Fonte: CEFTRU/FNDE (2007b) 115

Figura Apêndice A. 1: Tela principal do software 133

Figura Apêndice A. 2: Tela da tarefa pré-processamento 134

Figura Apêndice A. 3: Tela da tarefa de classificação 135

Figura Apêndice A. 4: Tela da tarefa de agrupamento..... 136

Figura Apêndice A. 5: Tela da tarefa de associação..... 137

Figura Anexo A. 1: Questionário web.– Parte A Fonte (CEFTRU, 2007a, 2007b)..... 139

Figura Anexo A. 2: Questionário web – Parte B. Fonte (CEFTRU, 2007a, 2007b)..... 140

Figura Anexo A. 3: Questionário web – Parte C. Fonte (CEFTRU, 2007a, 2007b)..... 141

Figura Anexo A. 4: Questionário web – Parte D. Fonte (CEFTRU, 2007a, 2007b)..... 142

Figura Anexo B. 1: Rede semântica do STER. Fonte: (CEFTRU, 2008a, 2008b) 143

Figura Anexo B. 2: Elementos físicos. Fonte (CEFTRU, 2008a, 008b)..... 145

Figura Anexo B. 3: Elementos lógicos. Fonte (CEFTRU, 2008a, 008b)..... 146

Figura Anexo B. 4: Atores. Fonte (CEFTRU, 2008a, 008b)..... 147

LISTA DE TABELAS

Tabela 3.1 : Estrutura de uma matriz de confusão 30

Tabela 3.2 : Tarefas de mineração de dados..... 36

Tabela 3.3 : Técnicas de mineração de dados 38

Tabela 3.4 : Base de dados hipotética..... 39

Tabela 3.5 : Característica de dados 45

Tabela 3.6: Estrutura do arquivo com extensão *ARRF*..... 48

Tabela 5.1: Distribuição dos municípios respondentes por estado..... 84

Tabela 5.2 : Variáveis identificas no mapeamento..... 95

Tabela 5.3: Transformação de diversos campos..... 97

Tabela 5.4 : Resultado após execução do algoritmo *apriori* 102

Tabela 5.5: Leitura da regra de associação..... 103

Tabela 5.6 Exemplo de regra de associação hipotética 105

Tabela 5.7 : estrutura da regra 30 108

Tabela 5.8 : estrutura da regra 6 110

Tabela 5.9 : estrutura da regra 23 112

Tabela 5.10 : estrutura da regra 1 114

Tabela 5.11 : estrutura da regra 09 116

ABREVIATURAS

ABNT	-	Associação Brasileira de Normas Técnicas
AI	-	Inteligência artificial
AM	-	Aprendizado de máquina
ARFF	-	Arquivo *ASCII* usado para definir atributos e seus valores
CEFTRU	-	Centro Interdisciplinar de Estudos em Transportes
CRISP-DM	-	*O Cross Industry Standard Process for Data Mining* - Metodologia
DM	-	Mineração de Dados
FNDE	-	Fundo Nacional de Desenvolvimento da Educação
GEIPOT	-	Empresa Brasileira de Planejamento de Transportes
IA	-	Agentes Inteligentes
IDEB	-	Índice do desenvolvimento da educação básica
INEP	-	Instituto Nacional de Estudos e Pesquisas Educacionais Anísio Teixeira
KDD	-	Descoberta de conhecimento em base de dados
MEC	-	Ministério da Educação
MPOG		Ministério do Planejamento, Orçamento e Gestão
PNATE	-	Programa Nacional de Apoio ao Transporte do Escolar
SEMMA	-	Metodologia *SEMMA: metod - Sample, Explore, modify, Model, Assess*
SGBD	-	Sistema de gerenciamento de banco de dados
SQL	-	Linguagem de Consulta Estruturada
STER	-	Sistema de Transporte Escolar Rural
TER	-	Transporte Escolar Rural
UNB	-	Universidade de Brasília
WEKA	-	*Waikato Environment for Knowledge Analysis*

1 INTRODUÇÃO

1.1 APRESENTAÇÃO

O sistema de transporte é um conjunto de elementos, atores, atividades organizadas e inter-relacionadas, que mutuamente se influenciam, e que tem como objetivo permitir o deslocamento indispensável de pessoas e bens (CEFTRU, 2008a, 2008b). Esse deslocamento é propiciado por um sistema complexo de vias, veículos terminais e usuários. Nesse sistema, são identificadas falhas que comprometem o papel do transporte nas diversas sociedades. Os principais problemas de transportes estão nessas falhas e, quanto mais complexos são esses problemas, maior é a necessidade de planejamento que integre os interesses dos diversos atores que participam desse ambiente.

Os estudos de transportes necessitam de informações durante todo o processo de planejamento. Essas informações (dados) precisam ser armazenadas, recuperadas, filtradas, monitoradas e readequadas de acordo com a necessidade (GIACAGLIA, 1998). A obtenção de dados para estudos de transportes, bem como a qualidade destes, são alguns dos principais problemas com os quais o planejador se depara, ocorrendo logo no início do seu planejamento. Uma das primeiras etapas é a coleta primária ou secundária de dados que devem ser conduzida com rigor, de forma a obter dados confiáveis, representativos e adequados em quantidade e qualidade sobre o objeto a ser estudado (TEIXEIRA, 2003).

Especificamente, o sistema de transporte, por suas características, envolve grande quantidade e variedade de informações necessárias ao seu planejamento, gestão e controle; principalmente, devido à sua complexidade e à enorme quantidade de informações existentes, armazenadas em bancos de dados. Porém, nem sempre esses dados são explorados em seu máximo potencial, pois o aumento significativo de informações negociais superam a capacidade de interpretar e examinar estes dados em curto espaço de tempo.

A incapacidade do ser humano de interpretar a enorme quantidade de dados produzidos, certamente, leva ao desperdício de informação e conhecimento, contidos nas bases de dados. Nesse contexto, na era da informação, recebe-se diariamente cada vez mais dados e informações. Muitos desses dados e informações são estruturados e organizados nos

bancos de dados do sistema de informação de diversas organizações, assim há a necessidade de transformação desses dados em conhecimento. As dificuldades de transformar dados em conhecimento estão relacionadas a uma dos seguintes eventos: ausência de conhecimento da existência da mineração de dados; dificuldade na escolha do algoritmo a ser utilizado; falta de software adequado; custo elevado na sua aquisição e pouca referência bibliográfica na seleção da técnica mais apropriada, para um problema específico a ser solucionado (DIAS, 2001, 2002).

A partir dessa visão faz-se necessário a criação de nova geração de métodos e técnicas capazes de auxiliar os gestores e os tomadores de decisão a buscarem conhecimentos úteis nas bases de dados de transporte. A Mineração de Dados - MD surgiu, como uma das principais soluções, para auxiliar no processo de descoberta de conhecimento em bases de dados. Este trabalho visa contribuir no contexto apresentado de desenvolvimento do processo sistematizado, com uma metodologia que utilize técnica de mineração de dados para análise de banco de dados de transporte.

1.2 PROBLEMA

Partindo do princípio de que existem dados estruturados em bancos de dados e sistemas de informação em instituições públicas e/ou privadas, transformar dados em informações e conhecimento úteis, pode servir de base para a tomada de decisão em estudos de transportes. Nesse sentido, embora já existam procedimentos de análise de dados, muitos deles estão baseados em documentos impressos e planilhas digitais, dentre outros, e na maioria das vezes, esses dados não são explorados em seu máximo potencial. Dessa forma, o problema a ser considerado, especificamente, é: Como melhorar o uso das grandes quantidades de dados de transporte, para subsidiar nas tomadas de decisões, nos diversos níveis de planejamento de transporte?

1.3 HIPÓTESE

A aplicação de sistema de apoio à tomada de decisão, que utilize o processo de mineração de dados, para analisar grande quantidade de dados, a fim de descobrir padrões e regras significativas que possam gerar conhecimentos úteis e compreensíveis, para subsidiar o processo de tomada de decisão aos diversos níveis de planejamento de transporte.

1.4 OBJETIVO

O objetivo desta dissertação é o desenvolvimento de uma metodologia de apoio à tomada de decisão em transporte, que utiliza o processo de mineração de dados, a fim de gerar conhecimentos úteis para subsidiar as decisões mais fundamentadas e inteligentes nos diversos níveis de planejamento de transporte.

Os objetivos específicos:

i) Desenvolver procedimentos por meio dos quais seja possível realizar análise de dados, em curto espaço de tempo, para subsidiar os estudos de transporte;

ii) Auxiliar na identificação das variáveis mais relevantes, de um conjunto de observações, que produzam o maior ganho de informação estratégica.

1.5 JUSTIFICATIVA

Os sistemas de transportes, por suas características distintas, abrangem uma enorme quantidade e variedade de informações necessárias ao processo de planejamento. Isso ocorre, pelo fato do sistema de transporte ter uma característica específica, que é a sua operação descentralizada, o que complica atividades de avaliação e controle dos serviços prestados e o entendimento real da oferta e demanda. Desta forma, todos os dados disponíveis são extremamente valiosos, devendo-se extrair destes todas as informações possíveis para dar suporte ao seu planejamento.

Assim, a etapa de elaboração do diagnóstico é, usualmente, a primeira das etapas no processo de planejamento, sem o qual não é possível traçar as metas, objetivos e situação

desejada, ou seja, situação na qual se deseja chegar. No entanto, só é possível identificar os problemas e encontrar as soluções mais adequadas por meio de um diagnóstico, que reflita o estado do objeto. Percebe-se, então, a necessidade de elaboração de uma metodologia que forneça subsídio na fase de diagnóstico, que é uma etapa fundamental no processo de planejamento, pois precede e define as demais etapas; sendo vital à estruturação do processo de planejamento, independente de qual modelo está sendo adotado pelo gestor.

Os sistemas de apoio à tomada de decisão utilizam a técnica de mineração de dados e se apresentam como ferramenta indispensável à eficiência no uso de informações úteis, na concepção e na elaboração do diagnóstico, para dar suporte ao processo de tomada de decisão.

Dessa forma, esses sistemas fornecem modelos analíticos, que analisam grande quantidade de dados, além de consultas interativas de apoio aos planejadores, para enfrentarem situações de tomada de decisão.

A metodologia apresentada neste trabalho, fornece uma ferramenta poderosa, com alta capacidade de análise de dados. Esse instrumento pode ser utilizado como suporte ao processo de planejamento, que tem forte participação dos tomadores de decisão.

1.6 METODOLOGIA

O método de abordagem foi o hipotético-dedutivo, o qual, parte de uma hipótese. A pesquisa é realizada na tentativa de comprová-la. Como método de procedimento foi desenvolvido um estudo de caso. Quanto as técnicas a serem utilizadas, para obtenção do propósito da pesquisa, destacam-se a pesquisa bibliográfica e pesquisa secundária, por documentação indireta, por meio de dados já coletados e publicados pelo CEFTRU (2007a, 2007b, 2007c). O estudo de caso permite verificar a aplicabilidade da metodologia proposta. De forma sistematizada, para cumprir seus objetivos a pesquisa segue as etapas apresentadas na Figura 1.1.

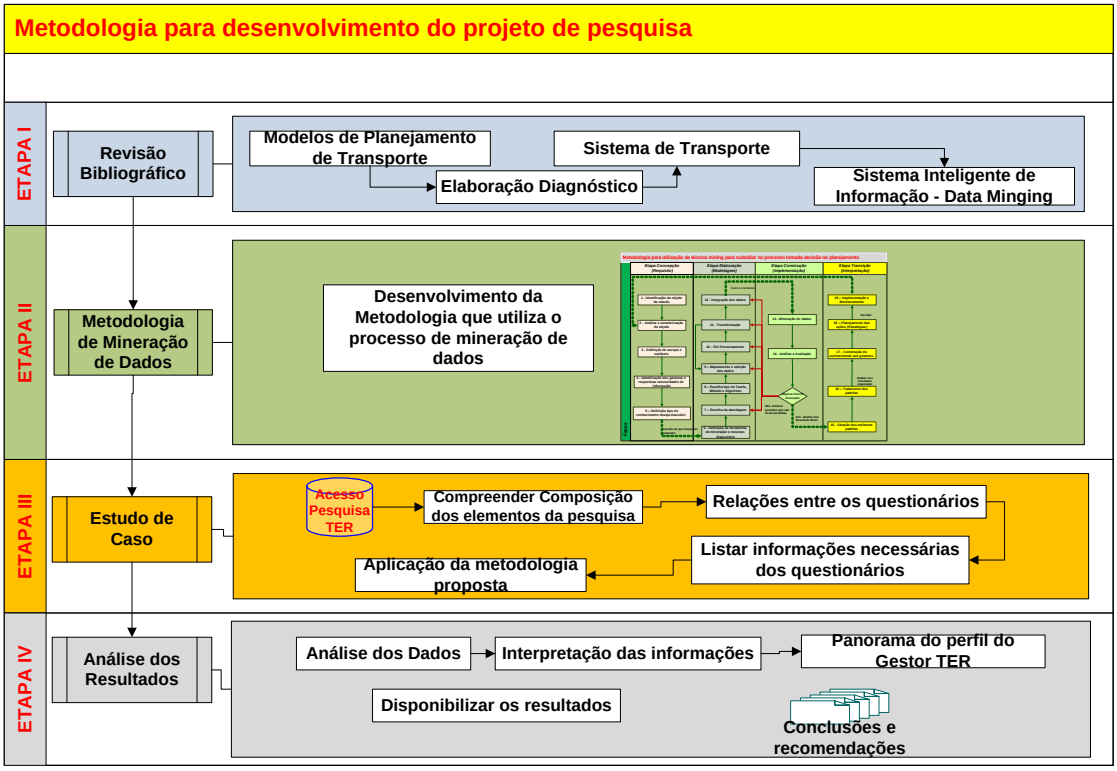


Figura 1.1 : Estrutura da metodologia adotada para o desenvolvimento da pesquisa

Essa proposta contém quatro etapas: revisão bibliográfica, desenvolvimento da metodologia, estudo de caso e análise dos resultados. (MARCONI e LAKATOS, 2009). Cada uma dessas etapas é detalhada a seguir:

Etapa I: Revisão bibliográfica – a revisão bibliográfica teve por objetivo, num primeiro momento, o entendimento das bases teóricas que nortearam o desenvolvimento da pesquisa. Num segundo momento, objetivou-se a consolidação e o amadurecimento desse conhecimento. Nessa etapa, enfim, está incluída a revisão sobre: planejamento de transporte e diagnóstico; sistemas inteligentes com ênfase em mineração de dados e sistema de transporte escolar.

Etapa II: Desenvolvimento da metodologia que utiliza a técnica de mineração de dados – Com base no referencial teórico e na busca para atingir os objetivos previstos para a pesquisa, estruturou-se a proposta metodológica para elaboração de um sistema de apoio à tomada de decisão que utilize a técnica de mineração de dados para estudos de transportes, conforme ilustra a Figura 1.2.

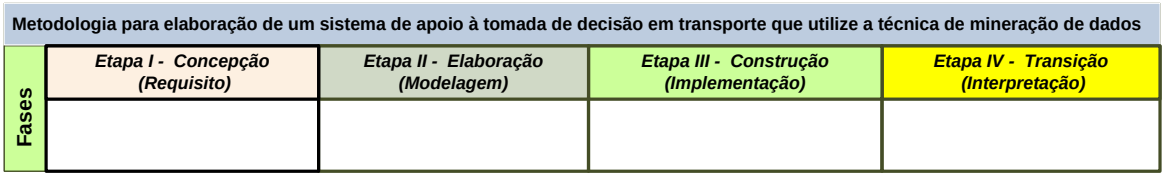


Figura 1.2 : Estrutura da metodologia que utilize o processo de mineração de dados

Cada uma dessas etapas da metodologia proposta é detalha a seguir:

- **Concepção** – nesta fase, incluem-se o entendimento do problema, a delimitação do escopo, a identificação dos usuários finais, com suas necessidades de informação;
- **Elaboração** – que consiste na escolha da ferramenta e do tipo de atividade, que será seguida no processo de mineração dos dados, pré-processamento dos dados, transformação e integração dos dados e geração da base de dados;
- **Construção** – será realizada a implementação da mineração de dados, conforme especificado na etapa de elaboração.
- **Transição** – neste tópico, atividade de pós-processamento realiza-se o tratamento das informações, para construção do novo conhecimento, para subsidiar o planejador na tomada decisão.

Etapa III aplicação da metodologia em estudo de caso – esta etapa teve por objetivo a aplicação da metodologia proposta, com finalidade de validar a aplicação de cada uma de suas etapas.

Etapa IV – Análise dos resultados – esta etapa, teve a finalidade de avaliar os resultados da aplicabilidade da metodologia proposta em relação ao seu objetivo.

1.7 ESTRUTURA DA DISSERTAÇÃO

A dissertação foi estruturada, em sete capítulos, no intuito de alcançar os objetivos propostos e a validação da hipótese apontada, sendo cada um deles descritos a seguir:

O primeiro capítulo introduz o estudo, contextualizando a pesquisa, apresentando seus objetivos e hipóteses.

O segundo capítulo revisa os conceitos de planejamento, o processo planejamento de transporte, os modelos de planejamento tradicional, estratégico, integrado e elaboração do diagnóstico.

O terceiro capítulo apresenta um embasamento teórico sobre sistema inteligente de informação, o processo de tomada de decisão e revisão dos principais conceitos de mineração de dados.

O quarto capítulo detalha o desenvolvimento da metodologia, que utiliza técnica de mineração, com base no referencial teórico e busca atingir o objetivo da pesquisa.

O quinto capítulo apresenta o estudo de caso, com aplicação da metodologia de apoio à tomada de decisão em transporte, que utilize o processo de mineração de dados, na caracterização do transporte escolar rural, com dados da pesquisa efetuada nos gestores do transporte escolar rural, nos 2.277 municípios pelo CETRU (2007a, 2007b).

Por fim, no último capítulo, sistematizam-se os resultados obtidos e as recomendações para trabalhos futuros, dentro do tema desenvolvido.

2 PLANEJAMENTO DE TRANSPORTE E DIAGNÓSTICO

2.1 APRESENTAÇÃO

O planejamento tem ganhado destaque, devido à escassez de recursos para investimento no setor de transporte e as exigências de organismos financiadores e fiscalizadores, por maiores retornos sobre as aplicações de recursos. Assim, os gestores têm sido pressionados, no sentido da busca da efetividade das ações. Essas metas são alcançadas com um planejamento correto e com uma ferramenta de gestão, para dar suporte ao processo de tomada de decisão. Para ter sucesso, o planejamento deve ser um processo contínuo e permanente, em que é fundamental, a identificação do problema, a solução, as causas e as consequências do conhecimento do caminho a ser percorrido, e dos instrumentos para melhor utilização dos recursos.

Neste capítulo, apresenta-se uma visão sistematizada sobre modelos de planejamento, a fim de possibilitar uma melhor compreensão de como a técnica de mineração de dados pode e dever atuar no processo. Este capítulo foi estruturado da seguinte forma: os conceitos de planejamento, os conceitos de planejamento de transporte, planejamento tradicional, planejamento estratégico, planejamento integrado e, por último, o conceito e a peculiaridade do diagnóstico.

2.2 CONCEITOS DE PLANEJAMENTO

Para Papacosta e Provedouros (1993), planejamento é definido como uma atividade responsável pelas tomadas de decisões futuras, baseadas em um plano elaborado antecipadamente, o qual servirá de estrutura para decisões específicas a serem tomadas de maneira racional, seja para atingir um objetivo proposto, ou para evitar algum tipo de problema futuro.

Na visão de Chiavenato e de Sapiro (2004), o planejamento determina antecipadamente os objetivos a serem alcançados e como fazer para alcançá-los. Para os autores, os objetivos vão influenciar todo o funcionamento da instituição no sentido *top-down*, neste caso, o

planejamento tem como missão as questões: "*Onde se pretende chegar*?"; "*O que deve ser feito?"; "Quando?"; "Como?"; e "Em que sequência?"*.

Ferrari (1979) define o planejamento como um método contínuo e permanente destinado a resolver, de forma racional, os problemas que afetam uma sociedade em determinado espaço e determinada época, por meio de uma previsão ordenada, de maneira a antecipar suas consequências posteriores. Este autor complementa sua definição ao colocar que todo planejamento pressupõe uma pesquisa, uma análise e uma síntese.

Na próxima seção, serão abordados os diversos conceitos sobre planejamento de transporte.

2.3 CONCEITOS DE PLANEJAMENTO DE TRANSPORTE

Os sistemas de transportes são estruturas complexas, sujeitas aos impactos e às transformações sociais, cujos problemas estão sujeitos às variáveis quantitativas e qualitativas.

Para Mello (1981), o processo de planejamento de transporte é baseado na coleta, análise e interpretação das condições existentes, tendo como base, os dados que refletem os anseios da comunidade. Seu desenvolvimento, metas e objetivos a serem alcançados dependem da questão da demanda futura pelo transporte.

Vanconcellos (2000) afirma que o objetivo do planejamento de transporte é definir a infraestrutura viária de transportes (vias e terminais), os meios (veículos), os serviços de transporte que irão permitir os deslocamentos de pessoas e mercadorias.

O planejamento de transporte tem as seguintes etapas: (i) definição de metas e objetivos; (ii) pesquisa e análise das condições existentes; (iii) previsões quanto ao uso do solo e padrões de movimentos; (iv) desenvolvimento de alternativas de rede; (v) análise das alternativas; (vi) avaliação; e (vii) implementação (BUCHANAN apud BRUTON, 1979).

Na próxima seção, são apresentados os diversos modelos de planejamento, a fim de possibilitar uma melhor compreensão de como a técnica de mineração de dados pode e

dever atuar no processo. Com esta finalidade, são discutidos os conceitos sobre planejamento tradicional, estratégico, integrado e o processo de diagnóstico.

2.4 PLANAJAMENTO TRADICIONAL

Para Ferrari (1979), o planejamento é um método contínuo utilizado para solução de problemas, os quais afetam uma sociedade e uma antecipação de suas consequências num momento futuro. É, portanto, um processo continuado que segue métodos científicos, para a condução da análise e elaboração de soluções.

Segundo Papacosta e Provedouros (1993), o processo tradicional de planejamento contínuo envolve uma visão sistêmica e é composto de oito etapas do planejamento, que se realimentam por meio da avaliação. A Figura 2.1, apresenta esquema sobre o processo de planejamento, formando um processo cíclico, em que é destacada a necessidade da avaliação contínua do processo.

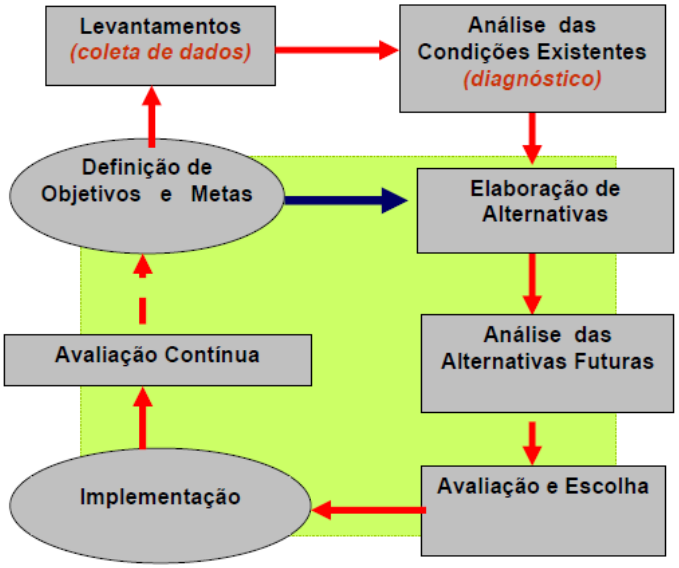


Figura 2.1 : Processo de planejamento contínuo. Fonte: Papacosta e Provedouros (1993)

Tomando como base as definições postas para processo tradicional de planejamento contínuo, pode-se generalizar cada uma das etapas, conforme a Figura 2.1:

1) **Definição de objetivos e metas** – etapa na qual são definidos os resultados finais desejados do processo de planejamento (objetivos) e resultados parciais com prazo definido (metas);

2) **Coleta de dados –** etapa que consiste no levantamento de dados e informações sobre o(s) objeto(s) de análise;

3) **Análise das condições existentes –** etapa de diagnóstica da situação atual, com base nos dados coletados, subsidiando a etapa de elaboração de alternativas;

4) **Elaboração de alternativas –** etapa propositiva, que define possíveis soluções aos problemas encontrados;

5) **Análise de alternativas –** etapa de investigação de cada alternativa particular, quanto a sua eficiência, na solução dos problemas apontados;

6) **Avaliação e escolha** – etapa de seleção das melhores alternativas para implementação;

7) **Implementação** – etapa que consiste na operacionalização do plano;

8) **Avaliação continuada** – etapa que consiste no constante monitoramento das ações implementadas pelo plano, com vistas a adequá-las ao ambiente dinâmico no qual são implementadas.

Conforme a Figura 2.1, para as etapas (i) análise das condições existentes e (ii) elaboração de alternativas. A metodologia proposta seria de fundamental importância para extração de informações úteis e relevantes, para elaboração de um diagnóstico alternativo mais preciso para o planejamento.

2.5 PLANEJAMENTO ESTRATÉGICO

As definições para planejamento estratégico são as mais diversas. Segundo Lucas (2001), é um processo dinâmico que define os caminhos que a empresa deve percorrer, por meio de um ação proativa, levando em conta a análise do seu ambiente interno, integrado com o ambiente externo, com propósito de construir o futuro desejado.

Chiavenato (2003) determina três níveis de planejamento estratégico, tais como: estratégico, tático e operacional. Nesta visão, o nível estratégico é o mais abrangente, de

longo prazo, e envolve toda a organização, preocupando-se sempre com os objetivos e metas a serem obtidos e os horizontes de tempo para estas realizações. O nível tático é o um pouco menos abrangente, a médio prazo, e se incube do desafio de apontar os caminhos para a consecução dos resultados desejados, no nível estratégico. O nível operacional é voltado para cada tarefa ou atividade, sendo de curto prazo, com metas específicas e definidas no nível operacional, para cada tarefa ou atividade.

Conforme a Figura 2.2, define planejamento estratégico como um método sistemático de administrar as mudanças na empresa, com o propósito de concorrer vantajosamente no mercado, adaptar-se ao seu ambiente externo, redefinir os produtos e maximizar os benefícios. Este processo disponibiliza uma série de estratégias, para que a empresa expanda seu crescimento, sua rentabilidade ou sua eficiência, levando em consideração os pontos fortes e fracos presentes. Assim, como as ameaças e oportunidades presentes e futuras (GUELL, 1997).

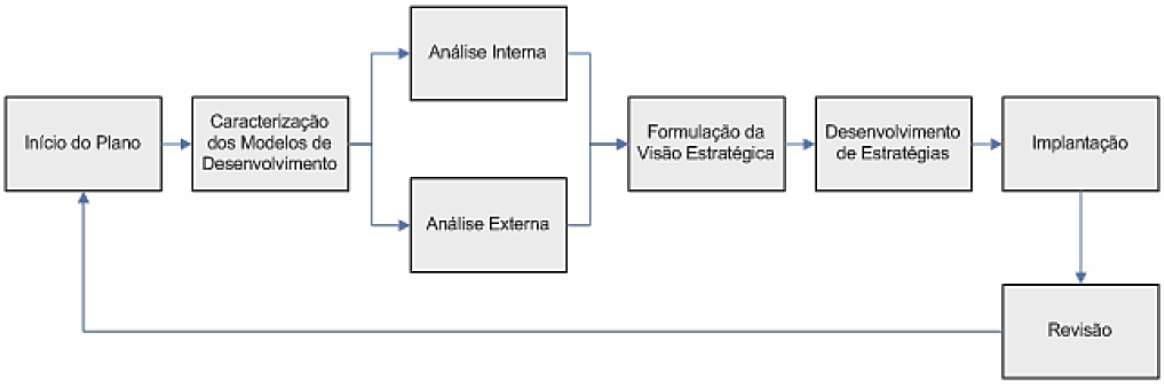


Figura 2.2 : Processo estratégico. Fonte: Guell (1997)

Tomando como base as definições postas para o processo estratégico de planejamento, pode-se generalizar cada uma das etapas, conforme a Figura 2.2 (GUELL, 1997):

1) **Início do Plano**: Nesta etapa são identificados os principais agentes socioeconômicos envolvidos, estabelecidas uma estrutura organizativa e participativa, além de uma política de comunicação, para difundir publicamente os objetivos do plano;

2) **Caracterização dos Modelos de Desenvolvimento**: Estes modelos descrevem os padrões de desenvolvimento físico, econômico e social, que conduziram ao

estabelecimento das condições atuais e estabeleceram o marco de referência para a análise interna e externa;

3) **Análise Externa**: Relaciona as oportunidades, potencialidades e ameaças relacionadas aos eventos externos, mas que estão fora de controle.

4) **Análise Interna**: Diagnosticam os principais elementos da empresa considerados de importância estratégica, identificando pontos fortes e fracos, problemas e restrições;

5) **Formulação da Visão Estratégica**: Corresponde ao modelo de futuro desejado para a comunidade ou os diversos agentes envolvidos. Os ajustes necessários à adequação entre a visão desejada e a realidade existente permitem a seleção daquelas metas mais fundamentais para o desenvolvimento;

6) **Desenvolvimento de Estratégias**: Baseada na formulação da visão desejada, as estratégias são desenvolvidas para possibilitar a consecução dos objetivos definidos, composto de programas de atuação, com seus respectivos planos de ação;

7) **Implantação**: Consiste na aplicação e na operacionalização dos planos de ação;

8) **Revisão**: Consiste na avaliação e no ajuste de programas e de ações, no sentido de obtenção dos objetivos definidos.

Conforme a Figura 2.2, as etapas (i) de caracterização do modelo, (ii) formulação da visão estratégica e (iii) avaliação do processo estratégico, podem utilizar a metodologia proposta para descobrir padrões de comportamento e associações entre os eventos, para extração de informações úteis para um diagnóstico alternativo, como ferramenta de apoio na elaboração do planejamento.

2.6 PLANEJAMENTO INTEGRADO

O planejamento integrado confronta as visões dos modelos apresentados anteriormente, com a finalidade de integrar os enfoques de auditoria e planejamento num quadro conceitual. Nesse contexto, surge o modelo esquemático de planejamento integrado (MAGALHÃES e YAMASHITA, 2008). Esse modelo baseia-se os princípios de planejamento, apresentado pelo Ministério do Planejamento, Orçamento e Gestão – MPOG (2006).

Dessa forma, nesse modelo torna-se possível compreender todo o processo, toda a guia dos planos de ação na implementação, no controle e na avaliação dos esforços, com o objetivo de transformar o objeto planejado. A estrutura é dividida em três níveis de decisão hierárquicos, conforme a Figura 2.3.

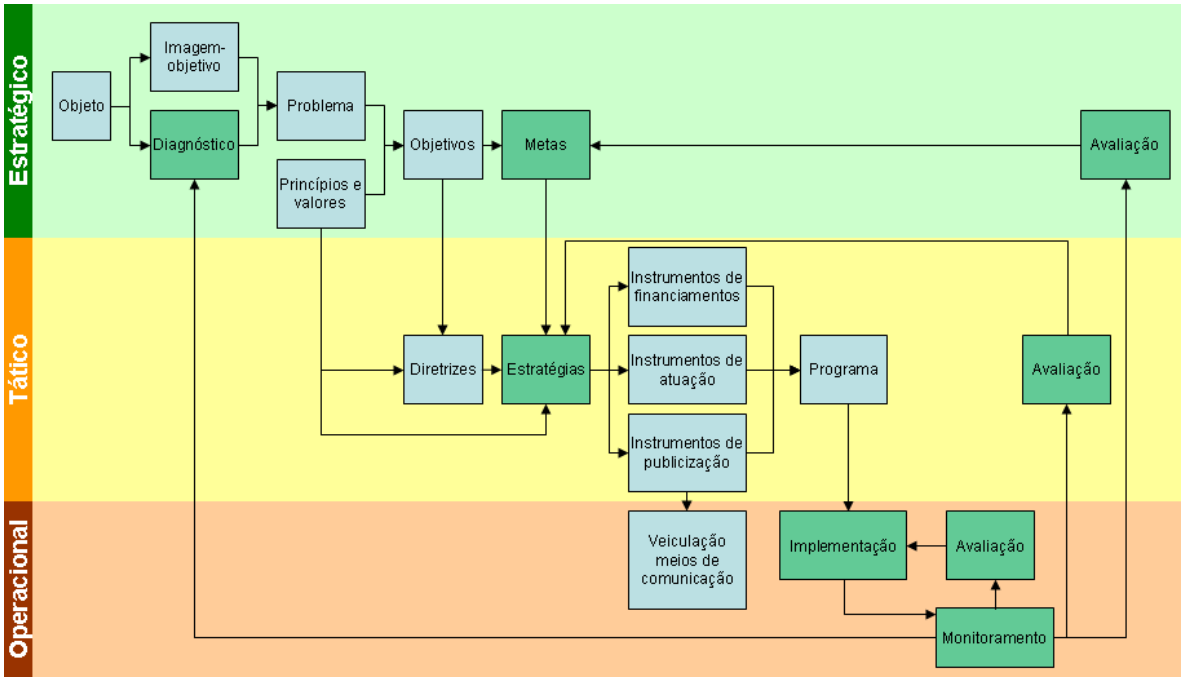


Figura 2.3 : Processo de planejamento integrado. Fonte: Magalhães e Yamashita (2008)

2.6.1 **Nível estratégico**

O **nível estratégico** é responsável pelas expectativas dos resultados a serem alcançados e as perspectivas de tempos para estas realizações. Desse processo, surge uma estrutura analítica do objeto, por meio da qual pode-se inserir, de forma adequada e coerente, todos

os elementos componentes e que intervenham ou influenciem no serviço (MAGALHÃES e YAMASHITA, 2008).

As etapas desse nível são:

- **Objeto***:* definição clara do objeto planejado em que se cria uma estrutura analítica, que possibilita a inserção de forma adequada e lógica de todos os elementos que fazem o serviço. Corresponde àquilo que se pretende planejar, e é obrigatório ter uma definição clara e um bom entendimento;
- **Imagem-objetivo:** é a situação desejada para o objeto que guia o planejado. Esse elemento é referencial, para o qual se deve guiar todo esforço do plano;
- **Diagnóstico:** que é o resultado obtido dos levantamentos de dados, realizados para cada uma dos elementos de representação do objeto;
- **Problema:** é a diferença entre um estado atual do objeto e o estado desejado, dentro dos limites de consentimento;
- **Princípios e Valores:** a escolha dos objetivos deve sempre ser embasada ou ainda limitada por valores e princípios, visando garantir o espaço de aceitabilidade no desenho das ações e na integridade de variáveis, que não devem ou não podem ser afetadas pelas ações previstas no plano;
- **Objetivos:** são resultados a serem alcançados e os elementos que devem orientar o desenvolvimento das ações. São determinados a partir das causas dos problemas identificados;
- **Metas:** são resultados, com prazo definido para execução, refletindo o compromisso político dentro de um horizonte de realização (curto, médio e longo prazo). Nessa etapa, o estudo de viabilidade deve ser realizado no contexto político e técnico.

2.6.2 **Nível tático**

O *nível tático* é responsável em apontar os caminhos para a consecução dos resultados desejados, no nível estratégico, e ainda em preparar o ambiente para implementação do serviço (MAGALHÃES e YAMASHITA, 2008).

As etapas desse nível são:

▪ **Diretrizes:** são atividades, critérios ou valores a serem seguidos de forma geral, valendo para todas ou várias ações definidas. As ações são atividades mais operacionais, com produtos e resultados claros;

▪ **Estratégias:** são conjuntos de projetos e ações escolhidos, para a realização dos diferentes objetivos, sendo limitado pelas diretrizes;

▪ **Organização da estrutura institucional:** etapa da definição das competências, das atribuições e das responsabilidades de cada ator, evitando que as diversas instituições fujam das responsabilidades e transfiram para terceiros, quando conveniente;

▪ **Instrumentos de financiamento**: disponibilidade de recursos financeiros do plano, pois se caracteriza como um elemento que pode impedir a realização do plano;

▪ **Instrumentos de publicação**: responsável pela divulgação dos dados e informações importantes para os diversos atores.

2.6.3 **Nível operacional**

O **nível operacional** é o responsável pela implementação das definições e por garantir a conformidade com o que foi definido pelos outros níveis de serviço (MAGALHÃES e YAMASHITA, 2008).

Possui basicamente três etapas:

▪ **Implementação:** momento em que os programas, projetos e ações são executados.

▪ **Monitoramento:** consiste na atividade de levantamento e tratamento dos dados, sistematizando as necessidades de informação de cada ator e os dados necessários para as avaliações dos resultados, seja operacional, tático ou estratégico.

▪ **Sistema de avaliação:** desenvolvido para atuar como processo contínuo que compara a situação atual, após a implementação das ações, com os resultados esperados. Abre a oportunidade de avaliação instantânea e, consequentemente, de

rápida correção das ações que não alcançaram os resultados esperados. Também pode atuar no nível tático e operacional.

Conforme a Figura 2.3, as etapas (i) diagnóstico, (ii) objetivo e (iii) avaliação no processo integrado, podem utilizar a metodologia proposta para descobrir padrões de comportamento e associações entre os eventos para extração de informações úteis, para um diagnóstico alternativo como ferramenta de apoio, para a elaboração do planejamento.

Ao analisar cada modelo têm-se algumas considerações acerca das abordagens do planejamento de transporte (MAGALHÃES e YAMASHITA, 2008):

- O modelo de planejamento tradicional contínuo possui uma visão tecnicista e está mais próximo a uma teoria de decisão, segundo a qual o resultado depende das escolhas do planejador;

- O modelo de planejamento estratégico tem um forte foco empresarial e está mais próximo da teoria dos jogos, segundo o qual o resultado depende de um contexto de autores que tomam decisões simultâneas;

- O modelo de planejamento integrado tem um forte foco político-social (discussão política) e principalmente, em seus níveis estratégicos e táticos têm uma forte participação dos tomadores de decisão, com mais respaldado do suporte técnico, de forma que a elaboração do plano deve ser necessariamente um compromisso político-social. Esse modelo tem uma abordagem cooperativa e não competitiva, no sentido de integrar interesses ou ações de diversos atores.

Para entendimento do real estado do objeto do planejamento é de fundamental importância a compreensão do processo de diagnóstico. Então, o próximo tópico apresenta, de forma resumida, a etapa de diagnóstico, realçando não só sua importância, mas também o envolvimento dos planejadores nesse processo.

2.7 DIAGNÓSTICO

Segundo Tedesco (2008), o diagnóstico sempre está presente em todos os modelos de planejamento, mas muito pouco discutido de uma forma sistematizada, conforme a Figura

2.4. O diagnóstico é usualmente a primeira das etapas no processo de planejamento, sem o qual não é possível traçar as metas, objetivos e situação desejada; situação na qual se deseja chegar. Só é possível identificar os problemas e encontrar as soluções mais adequadas por meio de um diagnóstico que reflita o estado do objeto. Assim, a metodologia proposta irá subsidiar a fase de diagnóstico, que é uma etapa fundamental no processo de planejamento, pois precede e define as demais etapas, sendo vital à estruturação do processo de planejamento.

Figura 2.4 : Visão conexão do diagnóstico. Fonte: adaptação de Tedesco (2008)

De acordo com Almeida (2005), o diagnóstico compara o estado encontrado com o estado desejado, avalia-se a eficácia, com base em algum padrão estabelecido, e procuram-se caminhos para diminuir a distância entre a situação existente e a situação desejada. O diagnóstico é parte do desenvolvimento organizacional, é a linha de base para o plano de ação da organização. À luz dos dados, nele levantados, são recomendadas mudanças na organização, que podem referir-se: objetivos e estratégicas; habilidade, conhecimento e atitudes do pessoal; processos interpessoais, dentre outros. Assim, o diagnóstico pode, ainda, recomendar uma série de intervenções que permitam implantar essas mudanças.

A elaboração do diagnóstico de um sistema de transporte, como uma das etapas do planejamento de transportes, é interpretada como a avaliação das condições de atendimento das necessidades de transporte, a partir de parâmetros de referência pré-estabelecidos.

Então, para elaboração, é necessário conhecer os parâmetros a serem utilizados para esta avaliação. A partir do diagnóstico podem ser planejadas ações pontuais ou globais, de forma a melhorar o seu desempenho (TEDESCO, 2008).

Segundo Chiavenato e Sapiro (2004), no planejamento estratégico, o diagnóstico é constituído de:

- **Diagnóstico estratégico interno –** que analisa as condições ambientais importantes, para formulação estratégica que melhor se ajustam ao elemento do ambiente. Relaciona os pontos fortes e fracos, problemas e restrições.

- **Diagnóstico estratégico externo** – que relaciona as oportunidades, potencialidade e ameaças dos eventos externos, contudo está fora de seu controle.

Por fim, no diagnóstico, tem-se uma quantidade enorme de informações, que precisam de técnica de manipulação para conseguir informações mais inteligentes do objeto de estudo. Na próxima seção, será abordado o processo de elaboração do diagnóstico, que está presente em todos os modelos de planejamento.

2.7.1 **Processo de elaboração do diagnóstico**

Para Almeida (2005), o desenvolvimento do diagnóstico exige uma série de atividades, que podem ser agrupadas em três principais fases: preparação; elaboração do projeto do diagnóstico e implementação do diagnóstico.

Fase de preparação

A etapa de preparação do diagnóstico, tem por objetivo estabelecer um cenário organizacional, que encoraje a avaliação e assegure que o pessoal conheça os componentes básicos do processo de avaliação. Neste momento, são realizadas atividades análise de objetivos, metas, identificação dos aspectos das unidades de informação a serem avaliadas, definição da equipe e revisão da literatura.

Fase de elaboração do projeto do diagnóstico

O projeto de diagnóstico é um plano que contém os objetivos do diagnóstico, o problema ou as questões de pesquisa, as hipóteses de trabalho, a metodologia a ser utilizada para a coleta de dados, as medidas de desempenho ou os indicadores de avaliação a serem utilizados e o cronograma do processo.

Fase de implementação do diagnóstico

Esta etapa envolve basicamente duas grandes subetapas: a coleta de dados e a análise e interpretação desses dados.

- **A coleta de dados** é uma etapa eminentemente prática, que consiste na aplicação da metodologia prevista no projeto do diagnóstico.

- **A análise e interpretação dos dados** procuram-se compreender a natureza e as causas dos problemas suscitados, o que pode ser feito aprofundando análise e comparação dos dados levantados, avaliando planos de trabalho e relatórios, analisando o histórico da instituição e da área de informação, ou recorrendo à literatura.

Após discussão acerca da literatura sobre o assunto, pesquisadores do CEFTRU (2008a , 2008b) apresentaram um processo de diagnóstico que é composto de três etapas: (1) Coleta de dados; (2) Comparação dos dados com os parâmetros de referência e (3) Elaboração do diagnóstico.

Etapa 1 – Coleta de dados

Nesta etapa, utilizam-se os instrumentos de coleta de dados com as técnicas pré-determinadas. Segundo as normas técnicas da ABNT - Associação Brasileira de Normas Técnicas (2002c), as formas tradicionais de coleta são: questionário, entrevista, observação e análise de conteúdo. Após a coleta, tratamento, avaliação da qualidade e da viabilidade de sua utilização, os dados se encontram prontos para serem comparados aos parâmetros de referência definidos para os elementos representativos do Sistema.

Etapa 2 – Comparação dados x parâmetros

Nessa etapa de comparação, os dados e suas informações resultantes devem ser adequados à escala em que se encontram os parâmetros, para que possam ser comparados.

Etapa 3 – Elaboração do diagnóstico

O diagnóstico consiste na análise resultante da comparação entre os dados coletados e os parâmetros de referência, ou seja, é o diagnóstico que permite ao planejador definir se o transporte está bom ou ruim, em relação a um determinado elemento do sistema de transporte.

Conforme, todos os modelos de planejamento, têm-se como premissa a elaboração do diagnóstico, que é de fundamental importância para o entendimento do problema e da consequência do direcionamento de todo o planejamento. Então, a metodologia proposta permite rapidez no estudo analítico de grande quantidade de dados coletados, para dar apoio na elaboração do diagnóstico alternativo.

Dessa forma, o presente trabalho vem contribuir, principalmente nessa fase do diagnóstico, para uma análise mais refinada dos elementos do objeto a ser analisado.

2.8 TÓPICOS CONCLUSIVOS EM PLANEJAMENTO DE TRANSPORTE

Este capítulo sintetizou os estudos de transporte e sua relação com o planejamento. Buscou mostrar, que em estudos de transportes, sempre serão contemplados com a realização de diagnósticos e que isso envolve grande quantidade de dados. A manipulação de dados é sempre realizada, conforme a disponibilidade do conhecimento dos técnicos, como também do tempo disponível. Assim, nos estudos de transportes, a necessidade de técnicas que apoiem a análise dos dados, gerando inteligências, é de grande importância para o melhor conhecimento do objeto transportes a ser estudado, que sofrerá as intervenções por meio de políticas públicas.

Buscou-se apresentar os modelos de planejamento e do processo de diagnóstico, que dão suporte à solução de problemas do setor. Foi possível observar que o planejamento é uma

ferramenta poderosa, que permite atingir os objetivos pretendidos no estudo, que se baseia em dados que representam o estado atual.

O processo de planejamento integrado viabiliza uma análise mais rápida nas diversas fontes de dados, que envolvem aspectos cooperativos, e nesse ponto, a aplicação dessa visão no planejamento integrado alinha-se aos objetivos dessa pesquisa.

O processo de planejamento integrado traz uma visão organizacional dos níveis estratégicos, tático e operacional. A interação entre os membros, na troca de informação, acontece em cada um desses níveis, das diversas formas, e o planejamento ajuda a conduzir esses processos.

O planejamento de intervenções do poder público, no setor de transporte, apresenta como obrigatório o conhecimento da situação atual, em que se encontra o sistema de transporte. As decisões e ações, pertinentes a um plano de transporte, têm como resultado o processo de planejamento de transporte. A partir desse conhecimento inicial, do estado atual do transporte, visa-se a um estado novo de transporte. Então, surgem diversas soluções por meio de ações que eliminarão as causas apontadas.

Os modelos de planejamento descritos têm algumas diferenças. A grande vantagem do processo integrado é a capacidade de envolver os diversos níveis de decisões e possuir quatro ciclos de avaliação e revisão. Por outro lado, os processos tradicionais e estratégicos, geralmente, ficam restritos a decisões táticas e operacionais e possuem somente uma etapa de avaliação. Independente do modelo de avaliação, pode-se utilizar da metodologia proposta no capítulo 4, para análise dos dados.

Independente do processo (integrado, estratégico ou tradicional), a etapa de diagnóstico é um elemento fundamental, pois baliza a percepção dos atores, sobre o contexto em que estão atuando, sobre o objeto de estudo.

Considerando os aspectos abordados neste capítulo, conclui-se que o diagnóstico deve ser obtido independente de qual modelo está sendo adotado para planejamento de transporte. E na elaboração do diagnóstico, pode ser incorporada a metodologia proposta, que permite realizar várias analises nos dados coletados, sob um determinado enfoque, que possibilita uma melhor compreensão sobre seu conjunto de dados.

3 SISTEMA INTELIGENTE DE INFORMAÇÃO

3.1 APRESENTAÇÃO

Os sistemas de apoio à tomada de decisão fornecem ferramentas e modelos analíticos para analisar grande quantidade de dados, além de consultas interativas de apoio, para os planejadores enfrentarem situações de tomada de decisão. Dessa forma, para propor soluções para problemas de transportes e apoiar suas atividades, o especialista em transportes, no processo de análise e de tomada de decisão, precisa de dados e informações. Portanto, dados e informações precisam ser coletados e terão valor para os estudos de transportes, dependendo de sua qualidade, consistência e outras características.

Nesse contexto, um dos principais aspectos que se pretende estudar é a questão da mineração de dados, para extrair informação e conhecimento útil em base de dados de transporte. Parte-se do princípio, portanto, de que dados e informações coerentes, e adequados servem para apoiar as decisões e a solução de problemas de transportes.

Diante disso, este capítulo está estruturado em três seções. A primeira seção apresenta a base conceitual do processo de tomada de decisão em sistemas inteligentes de informação. A segunda seção traz alguns conceitos básicos como: dado, informação, conhecimento, inteligência e agentes inteligentes. Finalmente, são apresentados os conceitos e fundamentos teóricos sobre mineração de dados.

3.2 TOMADA DE DECISÃO

Para Laudon e Laudon (2007), umas das principais contribuições dos sistemas de informação têm sido melhorar a tomada de decisão, seja no caso de indivíduos ou de grupos. A tomada de decisão, em uma instituição, costumava limitar-se à diretoria. Hoje, funcionários de níveis mais baixos são responsáveis por algumas dessas decisões, na medida em que os sistemas de informação tornam as informações disponíveis, para níveis mais elementares da empresa (GOMES, 2006).

Assim, nas instituições existem diferentes níveis, cada um desses níveis têm diferentes necessidades de informações, para apoiar suas decisões, e é responsável por diferentes tipos de decisão, que serão descritos no próximo item.

3.2.1 **Tipo de decisão**

Conforme a Figura 3.1, as decisões podem ser classificadas em estruturadas, semi-estruturada e não estruturada (LAUDON e LAUDON, 2007).

Figura 3.1 :Tipos de decisão. Fonte adaptação de Laudon e Laudon (2007)

- **Decisão não estruturada:** são aquelas em que o responsável pela tomada de decisão deve usar seu bom senso, sua capacidade de avaliação e sua perspicácia na definição do problema. Cada uma dessas decisões é inusitada, importante e não rotineira e não há procedimentos bem compreendidos ou predefinidos para tomá-las.

- **Decisão estruturada:** são repetitivas e rotineiras e envolvem procedimentos predefinidos, de modo que não precisam ser tratadas como se fossem novas.

- **Decisão semi-estruturada:** algumas decisões têm características dos dois tipos anteriores. Neste caso, apenas parte do problema tem uma resposta clara e precisa para um procedimento. Em geral, decisões estruturadas são mais corriqueiras nos níveis organizacionais mais baixos, enquanto problemas não estruturados são mais comuns nos níveis altos da instituição.

3.2.2 **Processo de tomada de decisão**

Simon (1969) descreve quatro diferentes estágios no processo de decisão: inteligência, concepção, seleção e implementação, conforme a Figura 3.2

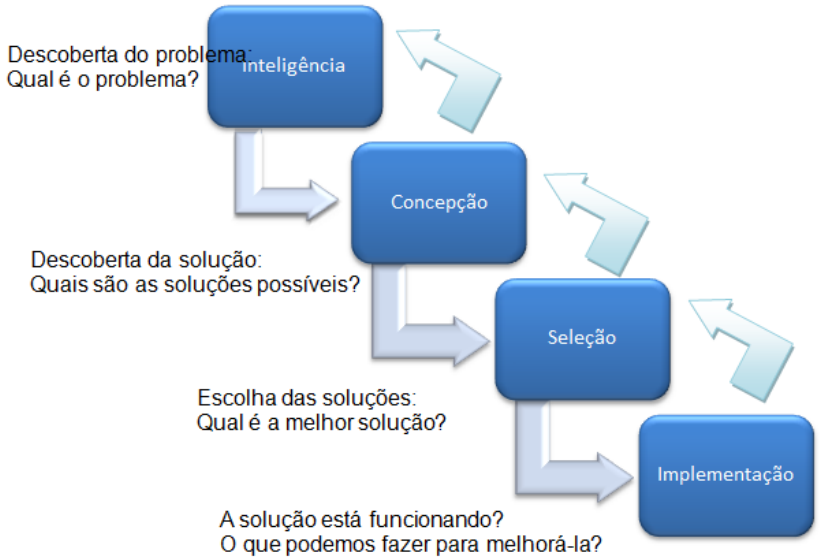


Figura 3.2: Processo de tomada de decisão. Fonte: adaptação de Simon (1969)

Esses estágios correspondem aos quatros passos do processo de resolução de problemas (SIMON, 1969):

- **Inteligência:** consiste em descobrir, identificar e entender os problemas que estão ocorrendo na organização. Logo, porque existe um problema, onde ele está, e qual o seu efeito;
- **Concepção:** envolve a identificação e a investigação das várias soluções possíveis para o problema;
- **Seleção:** consiste em escolher uma das alternativas de solução;
- **Implementação:** envolve fazer a alternativa escolhida funcionar e continuar a monitorar em que medida ela está funcionando.

Logo, o que acontece quando a solução escolhida não funciona? Conforme a Figura 3.2 mostra, é possível voltar ao estágio anterior, no processo de tomada de decisão, e repeti-lo se for necessário.

Assim, na próxima seção são apresentados os conceitos básicos de sistemas inteligentes de informação.

3.3 CONCEITOS BÁSICOS DE SISTEMA INTELIGENTES

Quando se aborda o tema "mineração de dados", alguns termos vêm à tona. São dados, informação, conhecimento e inteligência. Inúmeras definições existem para esses termos, no entanto, é conveniente reduzir seu universo de denotação. Segundo Pinheiro (2008), mineração de dados é um processo de aplicação de técnica estatística, de inteligência artificial e métodos de aprendizagem de máquina para descoberta de conhecimento útil em grande base de dados convencionais ou não.

Os autores Moresi(2000), Nonaka e Takeuchi(1997) afirmam que a informação pode ser compreendida sob variadas formas, sendo que suas classificações podem variar de dados, informação, conhecimento e inteligência, que num contexto decisório, possuem valores diferenciados, conforme a Figura 3.3

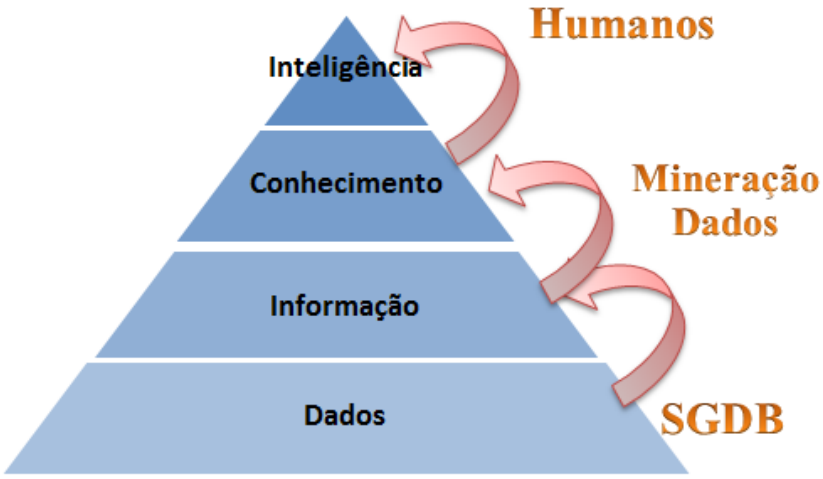


Figura 3.3 : Pirâmide da informação. Fonte: adaptação de Nonaka e Takeuchi(1997) .

- **Dados** é a unidade básica da informação ainda não tratada, sem relevância (NONAKA e TAKEUCHI, 1997; MORESI, 2000; LAUDON K. e LAUDON J., 2007).

- **Informação** é diferente de dados, informação tem significado e está sempre organizada para alguma finalidade; portanto, dado transforma-se em informação quando acrescenta-se significado (NONAKA e TAKEUCHI, 1997; MORESI, 2000; LAUDON K. e LAUDON J., 2007).

- **Conhecimento** é inerente ao ser humano, podendo ser transformado em ação ou ser escrito, explicitado, sob a forma de informação, podendo ser tácito ou explícito (NONAKA e TAKEUCHI, 1997; LAUDON K. e LAUDON J., 2007);

- **Inteligência** é aplicada aos seres humanos a capacidade de combinar o conhecimento adquirido com novas informações e mudar comportamento, a fim de executar determinada tarefa com sucesso ou adaptar-se a uma nova situação (NONAKA e TAKEUCHI, 1997; LAUDON K. e LAUDON J., 2007).

Os **dados** podem ser classificados em **primários e secundários**. Os **dados primári**os são adquiridos por meio de uma coleta de dados em campos e pela utilização de questionários, observações, entre outros. Por outro lato, os **dados secundários** já foram coletados, tabulados e estão disponíveis (MALHOTRA, 1996 *apud* LUCAS, 2001).

De acordo Tait (2000), **a tecnologia de informação** é uma das bases que sustentam os sistemas de informações. Pode se entender como sendo todo software e todo hardware de que uma instituição necessita para atingir seus objetivos organizacionais. O objetivo de um sistema de informação deve ser a integração entre negócios, sistemas e tecnologia da informação.

Para Aflori e Leonn (2004), **os agentes inteligentes** são entidades de software ou hardware que executam algumas tarefas, sem ajuda de usuários, e com algum grau de autonomia. Possuem habilidades para fazer escolhas, planejar, comunicar-se com outros agentes, perceber e adaptar-se a mudanças em um ambiente, além de aprender através da experiência. A implementação destes agentes fazem parte do processo de mineração de

dados, com entendimento dos conceitos básicos sobre sistemas inteligentes de informações.

A próxima seção aborda os conceitos sobre mineração de dados, que é uma etapa de processo maior, chamada descoberta de conhecimento em banco de dados – KDD (*Knowledge Discovery in Databases*) .

3.4 MINERAÇÃO DE DADOS

Atualmente, existem inúmeros conceitos acerca de mineração de dados – MD, com algumas variações em relação ao escopo de suas atividades. Na verdade Fayyad (*et al.*, 1996a, 1996b), afirma que o processo de mineração de dados é uma etapa de um processo maior, chamado descoberta de conhecimento em base de dados – KDD (*Knowledge Discovery in Databases)*. Nesta etapa, a mineração de dados utiliza-se de técnicas e de algoritmos de diferentes áreas do conhecimento, principalmente inteligência artificial, banco de dados e estatística, conforme a Figura 3.4.

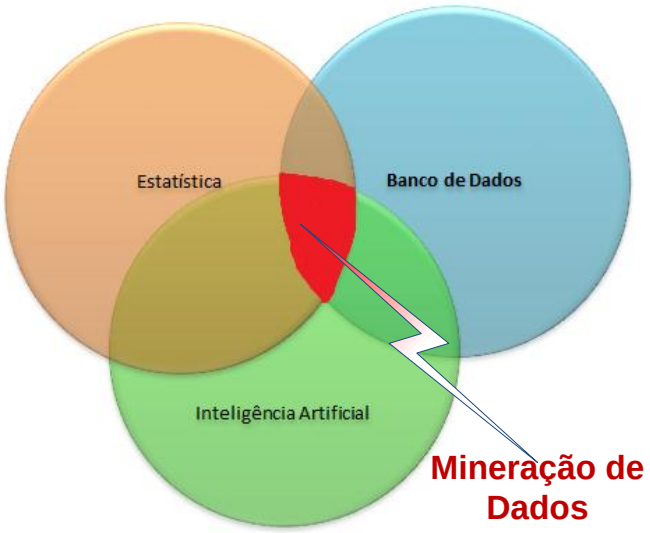


Figura 3.4: Multidisciplinaridade da MD. Fonte: adaptação de Fayyad *et al* (1996a)

Segundo Berry e Linoff (1997), a mineração de dados é a exploração e a análise, por meios automáticos ou semi-automáticos, de grandes quantidades de dados, a fim de descobrir padrões e regras significativas. Nessa visão, a mineração de dados tem, como objetivo

final, a criação de um modelo que possa melhorar a maneira de você ler e interpretar os dados existentes e os seus dados no futuro.

Neste trabalho, define-se ***mineração de dados*** como um processo de aplicação de técnica estatística, inteligência artificial e métodos de aprendizagem de máquina, para descoberta de padrões e regras significativas; com propósito de transformar dados em informações úteis, novas e compreensíveis para apoiar o processo de tomada de decisão ou avaliação de resultados, em grande base de dados convencionais ou não.

Dessa forma, percebe-se claramente que a mineração de dados possui grande relevância, contribuição e abrangência no suporte ao planejamento de transporte, em especial, na etapa de diagnóstico e avaliação das ações implantadas no sistema de transporte. Então, esse processo pode realizar análise analítica dos dados para apoiar o especialista em transporte na avaliação dos resultados.

Nesse contexto, na próxima seção, para uma melhor compreensão dos termos envolvendo mineração de dados, são abordados os seguintes tópicos: (i) conceitos básicos de mineração de dados; (ii) processo KDD (*Knowledge Discovery in Databases*); (iii) atividades; (iv) tarefas; e (v) técnicas.

3.4.1 **Conceitos básicos de mineração de dados**

Neste tópico, são apresentados os conceitos existentes nos métodos de mineração de dados:

A Inteligência Artificial é uma área da ciência da computação, cujo objetivo é habilitar o computador a realizar funções que são efetuadas pelo ser humano, utilizando conhecimento e raciocínio (WITTEN e FRANK, 1999).

Aprendizado de máquina - AM é uma técnica utilizada para obter um novo conhecimento, de forma automática, durante o treinamento que utiliza os algoritmos preventivos e descritivos, para realizarem o aprendizado de forma computacional (WITTEN e FRANK, 1999).

O **treinamento** é massa de dados produzida na etapa de pré-processamento e é subdividida em duas: ***massa de dados de treinamento***, cujos elementos são chamados de

amostra ou exemplos e ***massa de dados para teste***. A primeira delas, massa de dados de treinamento, é utilizada para a aprendizagem do método, por meio de treinamento e a segunda, massa de dados de teste, é utilizada para testar o modelo criado na etapa anterior, para assegurar que o aprendizado realmente seja satisfatório (WITTEN e FRANK, 1999).

A **matriz de confusão** é utilizada para realizar avaliação estatística do modelo idealizado na tarefa de classificação. Logo, o objetivo é quantificar o número de exemplos ou amostras no treinamento, que foram classificados corretamente (representado na diagonal principal) e os exemplos classificados de maneira errada. Na Tabela 3.1 é exemplificada uma matriz de confusão, com os parâmetros utilizados, para calcular a métrica de qualidade do modelo criado na etapa de avaliação (WITTEN e RANK, 1999).

Tabela 3.1 : Estrutura de uma matriz de confusão

	Classificação pelo método	
Classificação atual	**Verdadeiro**	**Falso**
Verdadeiro	A: verdadeiro + verdadeiro (**positivo-verdadeiro**)	B : falso + verdadeiro (**negativo falso**)
Falso	C: verdadeiro + falso (**positivo falso**)	D : falso + falso (**negativo verdadeiro**)

Fonte: Witten e Frank (1999)

onde : cada valor referente à matriz de confusão possui um significado, que se segue:

- A: números de registros para os quais A é classificado com verdadeiro e sua classificação atual é verdadeira (**positivo-verdadeiro**);
- B: números de registros para os quais B é classificado como falso e sua classificação atual é verdadeira, ou seja, estão misturados ***(*negativo-falso**) ;
- C: número de registros para os quais C é classificado com verdadeiro e sua classificação atual é falsa, ou seja, estão misturados (**positivo-falso**);
- D: número de registros para os quais D é classificado com falso e sua classificação atual é falsa (**negativo-verdadeiro**).

A partir destes valores, outras métricas de avaliação de desempenho podem ser geradas: (i) índice de sensibilidade; (ii) índice de especificidade, (iii) taxa de acerto e (iv) taxa de erro.

- O **índice de sensibilidade** mostra a porcentagem dos valores positivos, corretamente classificados, como positivos pelo modelo, conforme a equação 3.1.

$$S = \frac{A}{A + B} \tag{3.1}$$

onde :

S: para índice de sensibilidade;
A: para positivo verdadeiro ;
B: para negativo falso;

- O **índice de especificidade** mostra a porcentagem dos valores negativos, corretamente classificados, como negativos, conforme a equação 3.2.

$$E = \frac{D}{C + D} \tag{3.2}$$

onde :

S: para índice de especificidade;
D: para negativo verdadeiro ;
C: para positivo falso.

- O **taxa de erro** mostra a porcentagem dos valores positivos e negativos, classificados erradamente, conforme a equação 3.3.

$$TE = \frac{B + C}{N} \tag{3.3}$$

onde :

TE: para taxa de erro;
B: para negativo falso;
C: para positivo falso;
N: para representa números de elementos(registros).

- A **taxa de acerto** mostra a porcentagem dos valores positivos e negativos, classificados corretamente, conforme a equação 3.4.

$$TA = \frac{A + D}{N} \tag{3.4}$$

onde :

TA: para acerto;
A: para positivo verdadeiro ;
D: para negativo verdadeiro ;
N: para números de elementos (registros).

Parâmetro é uma medida utilizada para explicar, de forma resumida, uma característica da população (BARBETTA, 2001).

O Conceito de variável representa uma medida que toma um número particular de valores, com a possibilidade de valores diferentes para cada observação. As **variáveis discretas** possuem um conjunto finito de valores distintos. Por outro lado, **variáveis contínuas** podem assumir qualquer valor dentro de um intervalo (BARBETTA, 2001).

O conceito de padrão são unidades de informações que têm uma frequência repetitiva, ou então são sequências de informações que dispõem de uma estrutura repetitiva (BARBETTA, 2001).

Com base no entendimento dos conceitos básicos sobre mineração de dados, a próxima seção descreve o processo de descoberta de conhecimento em banco de dados – KDD (*Knowledge Discovery in Databases)* que tem como objetivo extrair padrões válidos, novos e potencialmente úteis, a partir de grande base de dados.

3.4.2 **Processo de KDD**

Fayyad *et al.* (1996a; 1996b) explica que o processo de descoberta de conhecimento em base de dados – KDD (*Knowledge Discovery in Databases)* é um processo não trivial para identificar padrões válidos, novos, potencialmente úteis e compreensíveis em dados existentes :

- A **validade** dos dados refere-se à representatividade dos padrões, para novos casos, com certo grau de probabilidade;
- **Novos** referem-se a padrões novos, ou seja, estes novos padrões não foram identificados por nenhuma outra ferramenta;
- Os **potencialmente úteis** referem-se à utilização das informações extraídas para dar suporte ao processo de tomada de decisão;
- Os **padrões compreensíveis** referem-se à capacidade do ser humano de absorver as informações extraídas, a fim de executar uma determinada tarefa com sucesso .

O termo KDD é um conjunto de procedimentos composto de: (i) consolidação de dados; (ii) seleção e pré-processamento; (iii) mineração de dados e (iv) interpretação e avaliação, conforme a Figura 3.5.

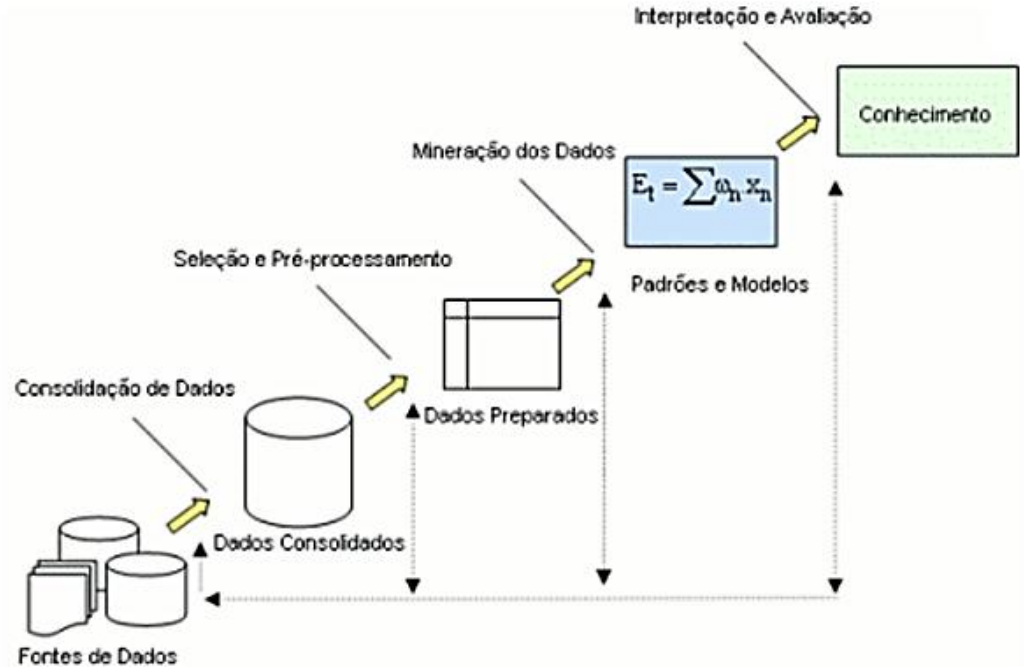


Figura 3.5 : Processo de KDD. Fonte: adaptação de Fayyad *et al*.(1996a)

- **A etapa de consolidação dos dados**: o objetivo é selecionar um conjunto de dados, pertencente a um domínio, contendo todas as possíveis variáveis (atributos) e registros (observações) que farão parte da análise. Essa etapa é bastante complexa, uma vez que os dados podem vir de uma série de fontes diferentes e podem possuir os mais diversos formatos.
- **A etapa de seleção e pré-processamento**: nesta etapa, são executadas as atividades de identificação dos atributos mais relevantes para o processo. A definição da forma de tratamento dos dados incompletos, inconsistente e ruídos. O objetivo é melhorar a qualidade dos dados e o enriquecimento semântico, para evitar possíveis distorções na extração de padrões.
- **A etapa de mineração de dados** (MD) é a responsável pela transformação de dados em informações. Assim, a MD utiliza algoritmo, que busca extrair o conhecimento implícito e potencialmente útil nos dados. A mineração de dados, portanto, é uma descoberta eficiente de informações válidas e não óbvias de uma grande coleção de dados. Dessa forma, o termo MD refere-se ao processo completo

que envolve atividade, tarefa, técnicas, dentre outros. A Figura 3.6, apresenta uma visão hierárquica do processo de MD. Assim, nas próximas seções, serão abordados os conceitos sobre os tópicos ilustrados na Figura 3.6.

Figura 3.6 : Visão hierárquica do processo de mineração de dados

- **A etapa interpretação e avaliação:** é a etapa em que se busca analisar e avaliar os resultados obtidos, a fim de julgar o modelo criado da fase anterior. Caso o resultado não seja satisfatório deve-se retornar a qualquer fase anterior ou ser reiniciado, conforme Figura 3.5.

3.4.3 **Atividades e tarefas de mineração de dados**

Os sistemas de mineração de dados, desenvolvidos para os mais diversos domínios, têm uma variedade de tarefas cada vez mais diversificadas. Essas tarefas possibilitam descobrir diversos tipos de conhecimento, sendo fundamental, decidir no começo do processo, qual o

tipo de conhecimento que deseja que o algoritmo extraia do conjunto de dados (SANTOS e AZEVEDO, 2005).

A Figura 3.7, apresenta a visão das duas atividades desempenhadas no processo de mineração de dados:

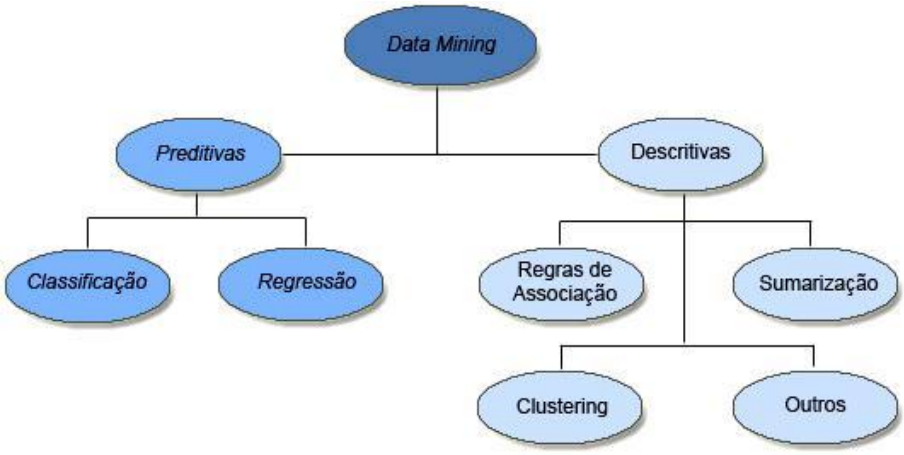


Figura 3.7: Tarefas de MD. Fonte: adaptação de Santos e Azevedo (2005)

- **As atividades preditivas** envolvem o uso de um conjunto de dados para descobrir padrões com o propósito de predizer variáveis de interesse. O objetivo é a redução do esforço na interpretação dos resultados, dessa forma, não é necessário que o público-alvo (usuário final) tenha um profundo conhecimento do domínio de negócio. As tarefas associadas para atividade preditiva são regressão, seleção de atributos e classificação (SANTOS e AZEVEDO, 2005).

- **As atividades descritivas** procuram padrões interpretáveis que são detectados nos dados e fornecem as características gerais dos dados. Nessa abordagem são apenas descritos os resultados que exigem a necessidade de se interpretá-los, e, consequentemente, o domínio do negócio é obrigatório para realizar a análise das regras. As tarefas associadas para a atividade descritiva são o agrupamento e a associação (SANTOS e AZEVEDO, 2005).

A Tabela 3.2 apresenta um resumo das principais tarefas de mineração de dados.

Tabela 3.2 : Tarefas de mineração de dados

Tarefa	Descrição	Exemplo
Classificação	Cria um modelo que possa ser utilizado em dados novos não classificados para categorizá-lo	▪ Aprovação de créditos ▪ Classificação os motoristas de acordo com as infrações cometidas.
Estimativa (regressão)	Estima valor de variável dependente a partir de várias variáveis independentes que quanto em conjunto produzem um resultado.	▪ Estimar custo do transporte escolar. ▪ Estimar o tempo de viagem dos usuários do transporte escolar
Associação	Determina quais fatos ou objetos tendem a ocorrem juntos mesmo eventos.	▪ Determinar quais infrações ocorrem juntas em uma área específica
Agrupamento	Divide um conjunto de dados heterogêneo em vários subgrupos mais homogêneos possíveis entre si.	▪ Identificar o perfil dos motoristas infratores ▪ Construir perfil de carga ▪ Marketing direcionado para motorista

Fonte: Santos e Azevedo (2005)

3.4.3.1 Tarefa de classificação

É o processo realizado em três etapas: (i) **etapa de criação do modelo de classificação** que é constituído de regras extraídas do treinamento, cujos elementos são amostra ou exemplos; (ii) **etapa de verificação** na qual o modelo criado será submetido a um conjunto de dados de teste para verificar sua conformidade de classificação; (iii) **etapa utilização do modelo** com novos dados não classificados. As formas mais comuns de representação dos algoritmos de classificação são regras e árvores de decisão, conforme a Figura 3.8. (WITTEN e FRANK, 1999).

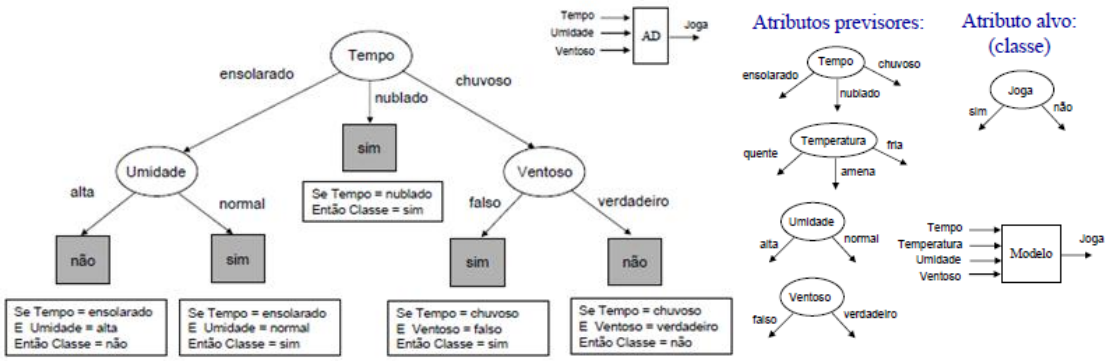


Figura 3.8 : Árvore de classificação. Fonte (WEKA)

3.4.3.2 Tarefa de seleção de atributos

Essa tarefa é de difícil realização, pois depende dos dados de entrada e do algoritmo a ser utilizado. A maioria dos algoritmos de seleção de atributos utilizam princípios estatísticos entre os atributos para cada conjunto de classe. Selecionam-se os atributos que contêm informações relevantes, separando-os dos atributos irrelevantes (WITTEN e FRANK, 1999).

3.4.3.3 Tarefa de regressão (estimativa)

A tarefa de regressão é semelhante a de classificação. A principal diferença é que o atributo classe é contínuo em vez de discreto. O objetivo da tarefa de regressão é prever o resultado do atributo classe (variável dependente), quando informado o conjunto de atributos de entrada (variáveis independentes ou variáveis previsoras).

3.4.3.4 Tarefa de agrupamento

A tarefa de agrupamento procura reconhecer um conjunto finito de categorias, na base de dados, agrupando coleção de objetos em subconjuntos com alta similaridade. Consiste na identificação de grupos homogêneos de objetos onde cada grupo é uma classe. Dentro da mesma classe os objetos são semelhantes e entre as classes são diferentes.

3.4.3.5 Tarefa de associação

A tarefa de associação é uma das mais utilizadas e tem o objetivo de estabelecer regras que interliguem um conceito a outro, ou seja, fatos que tendem a ocorrer juntos em uma transação. Assim, a presença de alguns deles em uma transação, implica na presença de outros na mesma transação, identificando, dessa forma, uma relação ou uma tendência

3.4.4 **Técnica de mineração de dados**

Existem várias técnicas utilizadas para mineração de dados, contudo, não há uma técnica que resolva todos os problemas de mineração de dados. Entre estas, pode-se afirma que não existe uma melhor técnica que a outra para os *diversos* tipos de problema, mas que existe uma técnica melhor que a outra para um *determinado* tipo de problema.

Assim, para cada problema a ser resolvido, deve-se buscar o melhor método, já que não se sabe quais deles atende e modela o problema da melhor forma. A Tabela 3.3, apresenta um resumo das principais técnicas utilizadas neste trabalho (DIAS, 2001, 2002).

Tabela 3.3 : Técnicas de mineração de dados

Técnica	Descrição	Algoritmos
Descoberta de regras de associação	Encontrar associações, correlação estatística entre atributos de dados a partir de um conjunto de dados.	Apriori, AprioriTid,
Árvore de decisão	Criar um modelo na forma de árvore que é utilizada para classificar dados novos sem testar todos os atributos.	C5.0, J.48
Algoritmo genético	Métodos gerais de busca e otimização, inspirados no processo de evolução. Encontrar a solução ideal de um problema específico, após verificar um imenso número de soluções alternativas.	Algoritmo Genético Simples (Goldberg, 1989); Algoritmo de Hillis (Hillis, 1997)
Redes neurais artificias	Modelo tenta reproduzir os padrões de processamento do cérebro humano. São utilizados para solucionarem os problemas complexos para os quais existem uma enorme quantidade de dados coletados.	Perceptron, Rede MLP, Redes de Kohonen, Rede Hopfield, Rede BAM,

Fonte: Dias (2001, 2002)

3.4.4.1 **Regra de associação**

Agrawal e Srikant (1994) definiram que as regras de associação têm a seguinte forma. Sejam $I = \{i_1, i_2, \ldots i_n\}$ um conjunto de i itens distintos e D uma base de dados formada por um conjunto de registros, onde cada registro T é composto por um conjunto de itens, tal que T $\subseteq I$. Uma regra de associação é representada na forma $A => B$, $onde\ A \subset I$, $B \subset I$, $A \neq \theta$, $B \neq \theta\ e\ A \cap B = \theta$. Onde, A e B são respectivamente o antecedente e o consequente da regra. Tanto o antecedente quanto o consequente de uma regra podem ser formados por conjuntos, contendo um ou mais itens. A quantidade de itens pertencentes a um conjunto de itens é chamado de comprimento do conjunto.

A Tabela 3.4 ilustra uma base de dados de registros de compras de um supermercado hipotético, que será utilizado nas próximas seções para ilustrar os exemplos de cálculos de diversas medidas de interesse: confiança, suporte, lift, dentre outros. Cada registro possui uma relação dos produtos adquiridos por um cliente.

Tabela 3.4 : Base de dados hipotética

TID	Lista de itens
1	Arroz, biscoito, chá, feijão
2	Arroz, pão, salaminho
3	Café, pão
4	Chá, pão
5	Arroz, café, feijão, pão
6	café, kiwi, pão

A regra de associação (A => B) possui duas partes: corpo da regra ou antecedente (A) e o resultado (B) ou consequente. O modelo típico para mineração de regras de associação em base de dados consiste em encontrar todas as regras que possuam suporte e confiança maiores ou iguais, respectivamente, a um suporte mínimo e uma confiança mínima, especificados pelo usuário. Por este motivo, o modelo costuma ser referenciado na literatura como *modelo suporte/confiança*:

- **Suporte**: corresponde à frequência com que A e B ocorrem em toda a base de dados. Essa medida é calculada da seguinte forma: quantidade de ocorrência do atributo ou conjunto de atributos dividido pela quantidade de registros. O suporte é interpretado como uma *medida da significância estatística* de uma regra (BERRY e LINOFF, 1997), conforme a equação 3.5.

$$Suporte\ \ (A \Rightarrow B) = \frac{Sup(A \cup B)}{total\ de\ Registros} = \frac{números\ de\ reg.contendo\ A\ e\ B}{Total\ de\ Registros} \qquad (3.5)$$

Exemplo:

$$Suporte\ (Arroz \Rightarrow Feijão) = \frac{2}{6} = 0{,}33 = 33\%$$

- **Confiança**: corresponde a frequência com que B ocorre dentre as instâncias que contêm A. Essa medida é calculada da seguinte forma: quantidade de ocorrência do atributo ou conjunto de atributos dividido pela quantidade de registros que suportam somente o corpo da regra. A confiança é interpretada

como uma *medida da força da regra* (BERRY e LINOFF, 1997), conforme a equação 3.6.

$$Confiança\ (A \Rightarrow B) = \frac{Sup(A \cup B)}{Sup(A)} = \frac{números\ de\ registro\ contendo\ A\ e\ B}{números\ de\ registros\ que\ contêm\ A} \quad (3.6)$$

Exemplo:

$$\text{Confiança (Arroz} \Rightarrow \text{Feijão)} = \frac{2}{3} = 0{,}66 = 66\%$$

3.4.4.2 Árvore de decisão

A estrutura básica de uma árvore de decisão pode ser formada por três tipos de nós: o nó-raiz, que mostra o começo da árvore; os nós-comuns que dividem um determinado item e constrói as ramificações; e os nós-folhas que contemplam as informações de classificação do algoritmo. As ramificações mostram os conjuntos de valores possíveis dos itens indicados no nó, para possibilitar uma melhor compreensão e interpretação (PICHILIANI, 2006). Esse algoritmo identifica quais atributos previsores serão realmente importantes para a classificação e separação dos demais. Sua leitura é feita percorrendo o nó-raiz e demais nós da árvore, de acordo com os valores dos atributos do novo registro, até chegar ao nó-folha, que é a classificação.

3.4.4.3 Rede neurais

De acordo com Souza (2004), as redes neurais podem ser definidas como formalismos computacionais que tentam reproduzir o processo de funcionamento dos neurônios humanos. Para Medeiros (2003), uma rede neural artificial, tal como seu paralelo biológico, é composta de certo número de neurônios que são conectados por ligações. Cada ligação possui uma quantidade associada a um peso. O aprendizado da rede é realizado pela atualização dos pesos.

3.4.5 **Compreender as medidas de interesse das regras associativas**

As medidas de interesse de regras de associação podem ser classificadas como objetivas e subjetivas. As medidas de interesse objetivas são aquelas que dependem apenas da estrutura do conjunto de dados e dos padrões. Não é necessário o conhecimento do domínio ou da aplicação. A maioria dessas medidas é baseada em teorias da probabilidade e estatística. Já, as medidas de interesse subjetivas consideram o conhecimento do domínio, no momento da análise dos padrões. (GENG e HAMILTON, 2006).

3.4.5.1 Medida de interesse objetiva

Gonçalves (2005) afirma que as medidas de interesse objetivas são índices estatísticos utilizados para selecionar regras interessantes, dentre as muitas que possam ser geradas por um algoritmo de mineração de regra de associação. O suporte e a confiança são exemplos de medidas de interesse objetiva.

Neste tópico, serão apresentadas outras medidas de interesse objetivas, desenvolvidas com o objetivo de medir a dependência entre itens de dados. Essas medidas consideram que a regra de associação é interessante, apenas quando o valor do suporte real é maior que o suporte esperado.

Para Geng e Hamilton (2006), as medidas de interesse objetivas podem ser utilizadas de três maneiras diferentes, dentro do processo de mineração de dados, conforme ilustra a Figura 3.9.

- As medidas de interesse podem ser utilizadas durante a mineração, como mecanismo de exclusão dos padrões não interessantes;
- As medidas de interesse podem ser utilizadas também para ordenar os padrões encontrados, de acordo com os valores de interesse;
- Por último, estas medidas podem ser utilizadas durante o pós-processamento, para selecionar apenas os padrões mais interessantes.

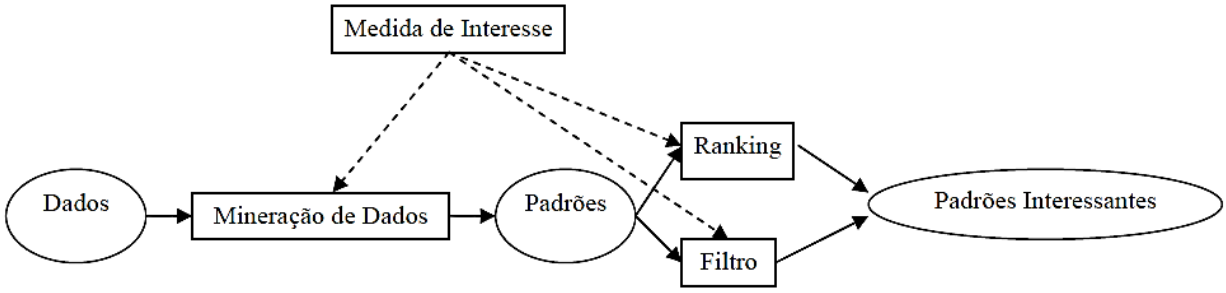


Figura 3.9 : Objetivos das medidas de interesse . Fonte: Geng e Hamilton (2006)

Suporte Esperado: O suporte esperado é computado, baseado no suporte dos itens que compõem a regra. Por definição, seja D uma base de dados de transações definida sobre um conjunto de itens *I*. Sejam $A \subset I \; e \;\; B \subset I$, dois conjuntos não vazios de itens, $A \cap B = \theta$. O suporte esperado (SupEsp) do conjunto $A \cup B$ é representado, conforme a equação 3.7.

$$SupEsp\ (A \Rightarrow B) = Sup(A) * Sup(B) \tag{3.7}$$

Exemplo:

$$SupEsp\ (Arroz \Rightarrow Feijão) = \frac{3}{6} * \frac{2}{6} = 0{,}5 * 0{,}33 \;=\; 0{,}165 = 16{,}5\%$$

Lift: Essa medida é uma das mais utilizada para avaliar dependência entre os itens, ou seja, seu objetivo é mensurar dependência entre os itens. Dada uma regra de associação $A \Rightarrow B$, observa-se, que medida indica o quanto mais frequente torna-se B quando A ocorre (BRIN *et al.*, 1997). Por definição, seja $A \Rightarrow B$ uma regra de associação obtida a partir de D. O valor do lift para $A \Rightarrow B$ é representado, conforme a equação 3.8.

$$Lift\ (A \Rightarrow B) = \frac{Conf(A \cup B)}{Sup(B)} = \frac{Sup(A \cup B)}{Sup(A) * Sup(B)} \tag{3.8}$$

Exemplo:

$$\text{Lift}\ (\text{Arroz} \Rightarrow \text{Feijão}) = \frac{0{,}66}{0{,}33} = \frac{0{,}33}{0{,}165} = 2 \;-\text{dependência positiva}$$

Se $Lift\ (A \Rightarrow B) = 1$, então A e B são Independente. Se $Lift\ (A \Rightarrow B) > 1$, então A e B são positivamente dependente. Se $Lift\ (A \Rightarrow B) < 1$, então A e B são negativamente dependente. Esta medida varia entre $0\ e\ \infty$ e possui intepretação de que quanto maior o valor do lift, maior é o relacionamento entre as variáveis.

Rule Interest: Essa medida pode também ser utilizada para avaliar dependência entre variáveis. Essa medida indica o valor da diferença entre o suporte real e o suporte esperado de uma regra de associação (BRIN *et al.*, 1997). Por definição, seja $A \Rightarrow B$ uma regra de associação obtida a partir de D. O valor do RI para A ⇒ B é representado, conforme a equação 3.9.

$$RI\ (A \Rightarrow B) = \ Sup(A \Rightarrow B) - \ SupEsp(A \Rightarrow B) \tag{3.9}$$

Exemplo:

$$\text{RI}(\text{Arroz} \Rightarrow \text{Feijão}) = \ 0{,}33 - 0{,}165 = 0{,}165 = 16{,}5\% - \text{depedência positiva}$$

Se $RI\ (A \Rightarrow B) = 0$, então A e B são Independente. Se $RI\ (A \Rightarrow B) > 0$, então A e B são positivamente dependentes. Se $RI\ (A \Rightarrow B) < 0$, então A e B são negativamente dependentes. Esta medida varia entre $-0{,}25\ e\ 0{,}25$ e possui interpretação de que quanto maior o valor do RI, maior é o relacionamento entre as variáveis.

Gonçalves (2005) fez uma observação importante quanto ao lift, afirmando que ele consegue destacar com maior facilidade a dependência positiva entre conjuntos de itens, que possuem suporte real baixo. Entretanto, o RI consegue destacar dependência positiva entre conjunto de itens, que possuem suporte real médio ou alto.

Convicção: As medidas Lift e RI são índices que possuem o objetivo de mensurar a dependência entre os itens, ao invés de medir implicação (o sentido da seta "⇒"). Então a medida de convicção é proposta com objetivo de avaliar uma regra de associação, como verdade implicação (BRIN *et al.*, 1997). Por definição, seja $A \Rightarrow B$ uma regra de associação obtida a partir de D. O valor da convicção para $A \Rightarrow B$ é representado, conforme a equação 3.10.

$$Conv\ (A \Rightarrow B) = \frac{Sup(A) * [\ 1 - \ Sup(B)]}{Sup(A) - \ Sup(A \cup B)} \tag{3.10}$$

Exemplo:

$$Conv(Arroz \Rightarrow Feijão) = \ \frac{0{,}5\ *\ 0{,}67}{0{,}17} = 1{,}97\ - regra\ interessante$$

Onde:

$$[\ 1 - \ Sup(B)] = 1 - \ Sup(Feijão) = 1 - 0{,}33 = 0{,}67$$

$$Sup(A) - \ Sup(A \cup B) = Sup(Arroz) - \ Sup(Arroz \Rightarrow Feijão) = 0{,}5 - 0{,}33 = 0{,}17$$

Essa medida varia entre $0\ e\ \propto$. Alguns autores idealizaram o índice e identificaram que as mais interessantes apresentam entre 1,01 e 5. Por outro lado, o valor da convicção maior que 5 representa informações óbvias.

3.4.5.2 Medida de interesse subjetiva

Gonçalves (2005) afirma que as medidas de interesse objetivas identificam, estatisticamente, a força das regras de associação. Entretanto, uma regra pode possuir valores altos para determinadas medidas objetivas e não ser subjetivamente interessante para o analista que examina. Então, em alguns casos, uma regra de associação é interessante para determinado usuário, mas não para outro.

Nestas medidas de interesse subjetivas são identificados dois fatores que podem tornar uma regra de associação subjetivamente interessante para o usuário: utilidade e inesperabilidade (SILBERSCHATZ e TUZHILIN, 2006).

- A **medida de utilidade** estima que um padrão/regra é interessante se o usuário puder fazer algo a partir dela, assim, pode-se tirar vantagem negocial sobre o padrão minerado;

- A **medida de inesperabilidade** considera que um padrão/regra tem grande possibilidade de ser interessante, quando contradiz com as expectativas do usuário no que depende de suas afirmações, portanto, daquilo que ele imagina que esteja em sua base de dados.

Para os autores Silverschats e Tuzhilin (2006), as medidas de utilidade e inesperabilidade são independentes, contudo, na prática os padrões/regras úteis são, na maioria das vezes, inesperados e na maioria das regras inesperadas, também, costumam ser úteis. Por essa razão, a medida de inesperada torna-se uma proveitosa aproximação para a medida de utilidade.

3.4.6 **Entendo característica dos dados**

Uma tarefa específica e os dados disponíveis têm uma relação de dependência com a escolha da técnica. Assim, a identificação das características dos dados em análise tem como finalidade selecionar a técnica de mineração de dados, que minimiza o número de dificuldade de transformação de dados para, a partir destes, obter bons resultados.

A Tabela 3.5 mostra uma lista das modalidades de dados que auxilia na seleção da técnica de mineração (Dias, 2001, 2002).

Tabela 3.5 : Característica de dados

Características	Descrição	Técnica
Variáveis de categorias ou qualitativa (nominal e ordinal)	São campos limitados por um conjunto de valores pré-determinados. Então as resposta podem ser encaixada em categoria.	▪ Descoberta de regras de associação ▪ Árvores de decisão
Variáveis numéricas ou quantitativas (discreta, contínua)	São campos resultantes de contagens, somatórios e ordenados que constituem de um conjunto finito de valores ou conjunto infinitos de valores.	▪ Árvores de Decisão
Muitos campos por registro	É fator crítico na escolha da técnica correta na aplicação específica, pois os métodos diferem na capacidade de processar grandes números de campos de entrada.	▪ Árvores de Decisão
Texto sem formatação	É fator crítico na escolha da técnica correta na aplicação específica, pois a maioria das técnicas é incapaz de manipular texto sem formatação.	▪ Raciocínio baseado em casos (MBR)

Fonte: (Dias, 2001, 2002)

3.4.7 **Validação de resultado**

As técnicas de validação são fundamentais para que os resultados e modelos possam ser avaliados e comparados. Os elementos deste processo são:

- **Os testes de validação** disponibilizam parâmetros de confiabilidade e validade nos modelos gerados.
- **Os indicadores estatísticos** são para auxiliar na análise dos resultados, tais como: matriz de confusão, índice de correção, estatística kappa, sensibilidade, especificidade, taxa de acerto e erros, conforme especificado na seção 3.4.1

3.4.8 **Metodologias**

O processo de mineração de dados pode ser desenvolvido de maneira não sistemática, porém não é aconselhável e geralmente leva a resultados imprevistos. Entretanto, existem metodologias que tem as etapas definidas a serem seguidas na prática de MD. Essas metodologias obrigam que o processo de MD seja desenvolvido de maneira padronizada.

No próximo item, serão apresentadas, de forma resumida, as metodologias CRISP e SEMMA.

3.4.8.1 CRISP-DM

O CRISP-DM (Cross-Industry Standard Process for Data Mining) foi criado 1996, com o objetivo de promover a padronização de conceitos e técnicas na busca de informações específicas para tomada de decisões. O CRISP-DM é definido por seis fases: (i) Compreensão do Negócio; (ii) Compreensão dos Dados; (iii) Preparação dos Dados; (iv) Modelagem; (v) Avaliação e (vi) implantação. Essas fases não precisam ser executadas sequencialmente, porque é um processo interativo e iterativo.

3.4.8.2 SEMMA

Foi idealizado pelo *SAS (Statistical Analysis Software) Instituto*, que define o processo de seleção, exploração e modelagem de grandes quantidades de dados, a fim de descobrir padrões de negócio desconhecidos. A metodologia é composta de cincos etapas: (i) seleção uma amostra representativa do problema a ser estudado; (ii) exploração da informação; (iii) manipulação dos dados para definição do formato adequado aos dados; (iv) criação do modelo, tendo como parâmetros as variáveis explicativas e as variáveis do objeto de estudo com um nível de confiança determinado e (v) análise do modelo.

3.4.9 **Ferramentas**

Atualmente, existe uma variedade de software na área de mineração de dados. Então, para um problema específico existem tarefas, métodos e algoritmos mais apropriados para aquela situação. Seguindo o mesmo entendimento, existem ferramentas mais apropriadas para uma determinada tarefa de mineração de dados. Como ferramentas comerciais têm as seguintes opções: *SAS (Statistical Analysis Software), IBM SPSS (Statistical Package for the Social Sciences), Microsoft SQL Server data mining*. Estas ferramentas podem ser divididas em duas categorias: pacote estatístico e banco de dados. Por outro lado, existem algumas ferramentas de código abertos: *WEKA (Waikato Environment for Knowledge Analysis*) *e rapidMiner(software de mineração de dados).*

Segundo Dias (2001, 2002), para que a escolha de uma ferramenta de mineração de dados tenha sucesso é obrigatória à definição de alguns critérios em consonância com o objetivo da instituição.

No próximo item, serão apresentadas de forma resumida, as ferramentas weka e transporte mining que foram escolhidas para este projeto, considerando a confiabilidade dos algoritmos implementados, custo e o fato das mesmas serem de domínio público, ou seja, software livre. Por outro lado, as ferramentas comerciais foram excluídas, principalmente, no mundo acadêmico, pois o custo de uma licença de um pacote comercial como os mencionados acima pode ser proibido.

3.4.9.1 WEKA

O *software WEKA* (*Waikato Environment for Knowledge Analysis*) é composto de coleção de algoritmo de aprendizado de máquina para diversas tarefas de mineração de dados. Sua interface gráfica é bastante intuitiva e seus algoritmos disponibilizam relatórios com informações sobre análise analíticas e estatísticas da inferência dos dados.

Uma restrição da ferramenta é a escalabilidade da versão atual, que limita a quantidade de dados a ser processada, por causa da dimensão da memória principal (WITTERN e FRANK, 1999). Mesmo assim, é possível realizar a mineração de um conjunto de dados

significativo, tornando o software atrativo para ser utilizado em diversas aplicações acadêmicas e no mercado.

Para realizar a mineração de dados o *software* utiliza arquivos de dados com extensão (.arff), onde devem ser mostradas quais variáveis são permitidas para um projeto (relação) específico, bem como o tipo de dados de cada variável, conforme a Tabela 3.6.

Tabela 3.6: Estrutura do arquivo com extensão *ARRF*

Tipo	Descrição	Exemplo
@relation	Identificar o conjunto de treinamento a ser analisado	**@relation** jogar vôlei
@attribute	Especificar as características de cada variável, ou seja, seu tipo: 1) Nominal – os valores representativos devem estra entre "{}" separados por vírgula. 2) Booleano 3) Numérico ou real	▪ Nominal: **@attribute** previsao {ensolarado, nublado, chuvoso} **@attribute** resposta {1,0} ▪ Booleano: **@attribute** aprovado boolean ▪ Numérico : **@attribute** frequência real
@Data	Consistem as instâncias (registros de dados) do conjunto de treinamento. "O valor de cada atributo para cada registro separado por vírgula e ausencia de atributo deve ser representado pelo símbolo "?". Cada linha representa um único registro.	**@data** ensolarado,quente,alta,?,não? quente,alta,nao_ventando,sim
Simbolo %	As linhas iniciadas com o símbolo de % não serão processadas.	% 1. codigo: real

Fonte: (WITTERN e FRANK, 1999).

3.4.9.2 Transporte mining

Esse software foi desenvolvimento sobre a biblioteca completa do WEKA e foram implementados novos requisitos em alguns algoritmos de mineração de dados. Assim, ele faz uma complementação básica sobre o tema a ser discutido neste trabalho. Sua interface gráfica é muito mais fácil de utilização, porque as informações sobre a opção de seleção do arquivo e dos algoritmos estão em cada tela para uma tarefa específica. No apêndice A, é apresentado o manual na forma simplificada da ferramenta.

Então, esse software foi utilizado neste trabalho, porque foram feitas várias alterações na estrutura lógica do algoritmo *apriori*, para satisfazer os requisitos negociais e as modificações de usabilidade implementada nas suas interfaces gráficas.

3.5 TÓPICOS CONCLUSIVOS DE SISTEMAS INTELIGENTES

Todos os modelos de planejamentos têm uma sub etapa fundamental que é o diagnóstico, que realiza a comparação da situação atual e a desejada em conjuntos de dados disponíveis. Assim, os conceitos sobre mineração de dados têm como objetivo principal dar suporte na elaboração, na análise de banco de dados de transporte, principalmente na realização de diagnóstico para o planejamento. Por isso, essa tecnologia alinha-se aos objetivos dessa pesquisa.

A tecnologia da informação pode ajudar no processo de tomada decisão, pois a aplicação de técnicas de mineração de dados em banco de dados busca identificar regras e padrões válidos, compreensíveis para aquelas pessoas responsáveis pela tomada de decisões e na elaboração do diagnóstico. Assim, esses processos ajudam atingir os objetivos dessa pesquisa.

O usuário de um sistema de descoberta de conhecimento em banco de dados precisa ter um sólido entendimento do negócio, para ser capaz de selecionar corretamente os subconjuntos de dados e as classes de padrões mais interessantes para criação do modelo, para que seja satisfatório.

A mineração de dados transforma um monte de desinformação em informações úteis ao criar modelos e regras. Seu objetivo é utilizar os modelos e regras para prever o comportamento futuro e para melhorar o planejamento de negócio da instituição.

A mineração de dados possui grande relevância, contribuição e abrangência no suporte ao planejamento de transporte, em especial, na etapa de diagnóstico e avaliação das ações implantadas no sistema de transporte.

O processo de mineração de dados utiliza-se de sistema inteligente de informação que faz uso de agentes inteligentes. Esses agentes descobrem padrões e regras, tais como: dividir um conjunto de dados heterogêneos em subgrupos mais similares possíveis; identificar relacionamentos de fatos que tendem a ocorrer juntos e selecionar variáveis mais relevantes entre observações.

A análise prospectiva dos dados prevê futuras tendências, comportamentos ou eventos, com base em dados históricos. Por outro lado, análise retrospectiva dos dados fornece percepções sobre tendências, comportamento ou eventos que já ocorreram.

Essas análises, sob um determinado enfoque, são orientadas a fim de possibilitar uma melhor compreensão sob seu conjunto de dados e gerar informações valiosas e precisas para dar suporte ao planejamento de transportes.

Assim, a escolha da ferramenta depende muito do problema que se deseja solucionar com a técnica de mineração de dados. Para cada problema a ser resolvido deve-se buscar atividade, tarefas, métodos e algoritmos mais adequados. Pois, não se sabe quais deles atendem e modelam o problema da melhor forma. Então, neste trabalho foi escolhida a ferramenta transporte mining, considerando na confiabilidade das novas implementações do algoritmo *apriori*. Além disso, sua interface gráfica é bastante intuitiva, de fácil utilização e possui recursos nativos para visualização e caracterização estatístico dos dados.

4 METODOLOGIA DE APOIO À TOMADA DE DECISÃO EM TRANSPORTE UTILIZANDO O PROCESSO DE MINERAÇÃO DE DADOS

4.1 APRESENTAÇÃO

Os capítulos anteriores revisaram tópicos importantes que serviram de base para o desenvolvimento de uma visão crítica sobre as necessidades de informação, para tomada de decisão em transporte. Como verificado, os estudos levam às necessidades de informação que, por sua vez, levam à coleta de dados em diferentes fontes. Esses dados podem ser analisados pelos técnicos, gerando informações que são utilizadas para as tomadas de decisão em transportes.

Então, os sistemas inteligentes de informação, conforme o capítulo 3, podem extrair regras e padrões significativos de um conjunto de dados até então desconhecidos, para que sejam utilizados no suporte, na tomada de decisão em transporte. Numa tomada de decisão, quanto mais alto na hierarquia estiver o gestor, maior o número de variáveis e mais diversos são os impactos relacionados às suas decisões, tornando, assim, as informações extremamente estratégicas.

A metodologia para elaboração de sistema de apoio à decisão em transporte, utilizando o processo de mineração de dados para descobrir padrões e regras em grande base de dados, é, então, desenvolvida por especialistas da área, de forma a representar a preocupação do gestor na tomada de decisão. Como foi visto, os sistemas inteligentes de informação envolvem todo o processo de decisão, do nível estratégico ao operacional.

O presente capítulo busca sistematizar a metodologia para a elaboração do sistema de apoio à decisão em transporte, que integre o processo de mineração de dados no suporte ao planejamento, controle e operação de transportes.

4.2 METODOLOGIA PROPOSTA DE ELABORAÇÃO DE UM SISTEMA DE APOIO À DECISÃO EM TRANSPORTE UTILIZANDO O PROCESSO DE MINERAÇÃO DE DADOS

No capítulo anterior, foram apresentadas algumas técnicas de mineração de dados em que foi possível perceber que nenhuma delas era completa, e sim, que a sua completude se dá pela integração das diversas técnicas, conforme a metodologia proposta na Figura 4.1, que mostra a utilização da mineração de dados na caracterização de um sistema de apoio à tomada de decisão no planejamento, na gestão e no controle.

A metodologia proposta na Figura 4.1, é composta de 19 atividades divididas em 4 grandes etapas : concepção (requisitos); elaboração (modelagem); construção (implementação) e transição (interpretação).

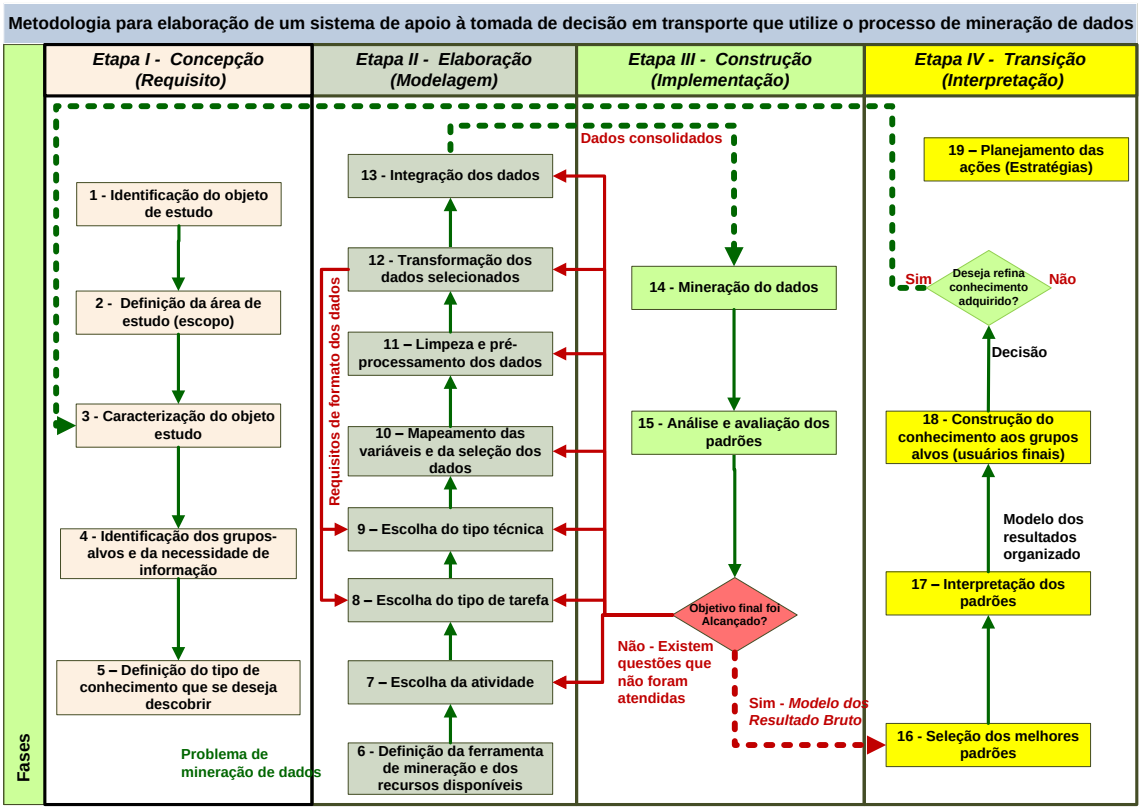


Figura 4.1: Metodologia proposta

A Figura 4.2 representa a visão funcional da metodologia, pois descrevem os processos lógicos ou funções. Cada processo descreve uma sequencia de tarefas e decisões que controlam quando e como elas são realizadas (PENDER, 2004). Essa representação lógica é mais familiarizada como fluxograma, pois ajuda entender a sequencia de atividades que estão embutidas nas quatro etapas, que são concepção, elaboração, construção e transição, as quais são detalhados abaixo:

- **A etapa I de concepção** é composta de cincos atividades - é o momento mais importante, sendo determinante para definição clara do objeto de estudo, o entendimento do problema de mineração de dados, delimitação da área de estudo, identificação dos grupos-alvos e suas necessidades de informações mais adequadas para cada perfil, dentro do processo de tomada de decisão.

- **A etapa II de elaboração** é composta por sete atividades - é a parte mais custosa do processo, a qual consiste na escolha da ferramenta, do tipo de atividade que será seguida no processo de mineração de dados, da limpeza e pré-processamento dos dados selecionados; da transformação e integração dos dados; e da geração do arquivo com os dados selecionados, em formato que a ferramenta compreenda.

- **A etapa III de construção** é composta por duas atividades – consiste na aplicação de método para extração de padrões nos dados, na busca pelo melhor ajuste dos parâmetros do algoritmo, para a tarefa em questão. Em seguida, realiza-se a atividade de pós-processamento com objetivo de identificar os padrões mais interessantes de acordo com o domínio de negócio.

- **A etapa IV transição** é composta por cinco atividades – é a parte que interpreta e avaliam os padrões, sob um determinado contexto de negócio, a fim de gerar informações, conhecimentos úteis e compreensíveis, para que se possa subsidiar o planejador, o controlador e o gestor na elaboração do plano de ação.

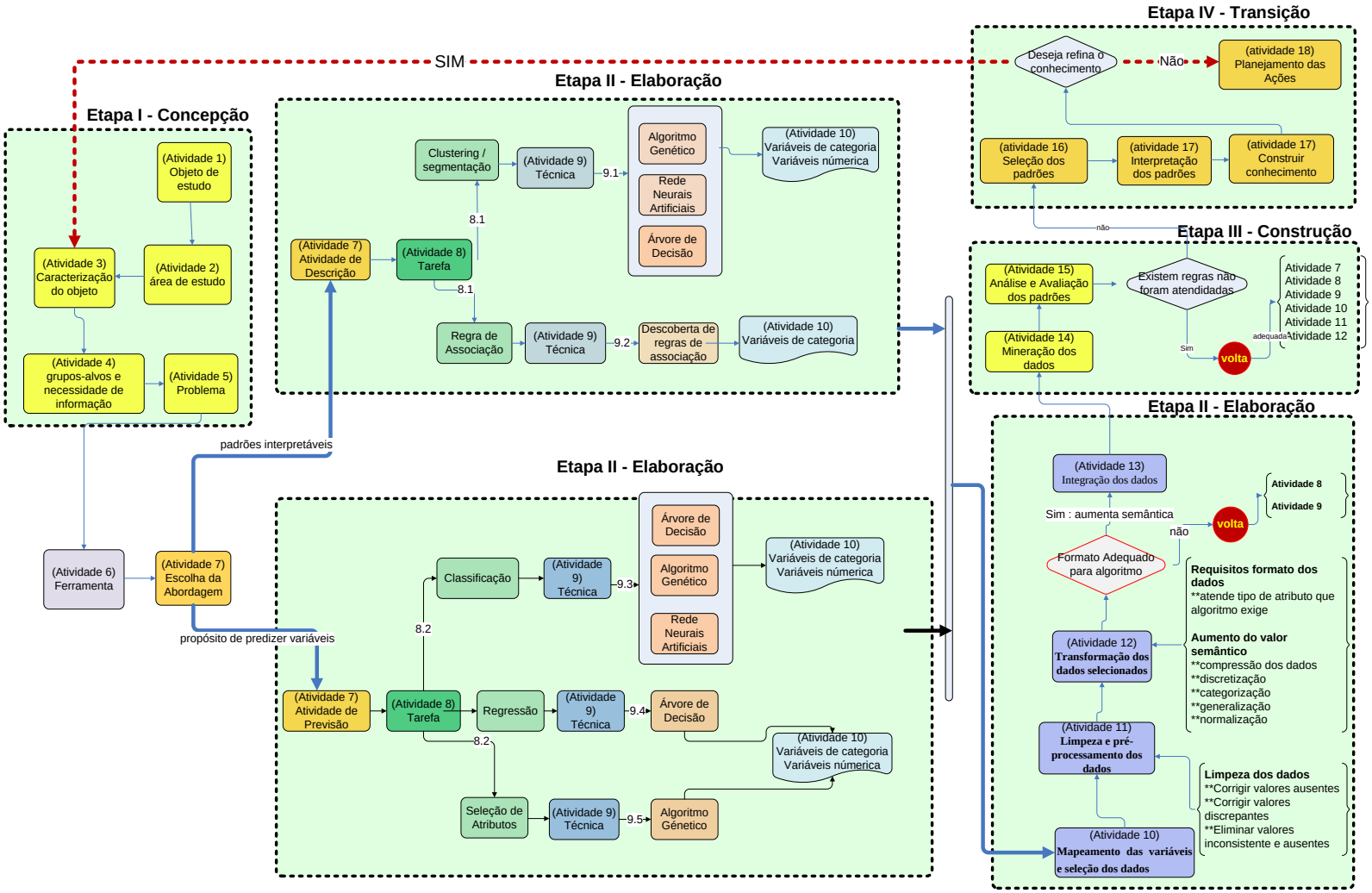


Figura 4.2 : Fluxograma da metodologia proposta

4.3 DESCRIÇÃO DA METODOLOGIA

A seguir, serão apresentadas, detalhadamente, todas as atividades da metodologia.

4.3.1 **Etapa I – Concepção (requisitos)**

Esta etapa é o momento mais importante, sendo determinante para a definição clara do objeto de estudo, entendimento do problema de mineração de dados, delimitação da área de estudo, identificação dos grupos-alvos e das necessidades de informações sistematizadas, mais adequadas a cada perfil dentro do processo de tomada de decisão, conforme a Figura 4.3.

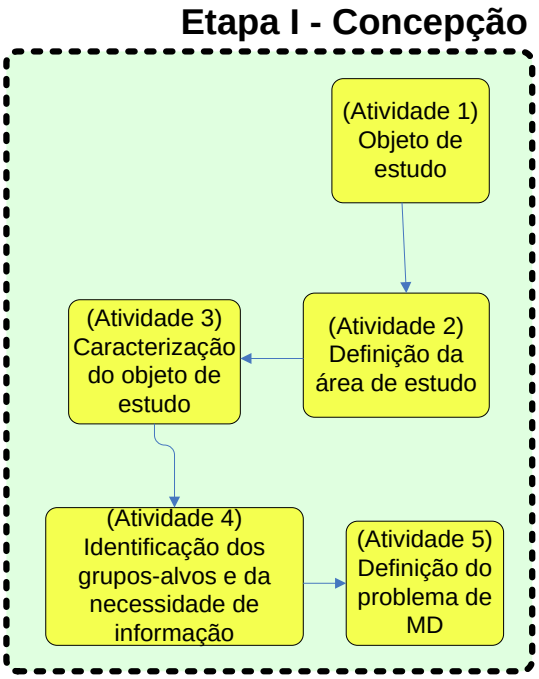


Figura 4.3 : Etapa de concepção da metodologia

Atividade 01: Identificação do objeto de estudo

Para utilizar a técnica de mineração de dados em transportes, a primeira atividade é a definição do objeto *foco de estudo* (sistema), construindo uma estrutura semântica de transporte do modo a ser estudada. Segundo Corradi *et al.* (2001), a rede semântica é uma das formas gráficas de representação do conhecimento, acerca de determinado objeto de

estudo, cuja estrutura é composta por nós e arcos interconectados. No anexo B, é apresentado à rede semântica do sistema de transporte rural na forma simplificada.

Atividade 02: Definição da área de estudo (delimitação)

Caso o objeto de estudo seja um determinado sistema de transportes, é necessário delimitar a área de alcance ou de abrangência deste sistema a ser analisado (limites de sua área de atuação/atendimento), uma vez que os atores, elementos e atividades componentes do sistema afetam o ambiente e são afetados por este, mostrando-se, assim, essencial a definição da área e dos limites físico-geográficos do ambiente.

Atividade 03: Caracterização do objeto de estudo

Esta atividade define os elementos de caracterização do objeto a ser estudado, estabelecido na atividade 1 desta metodologia. Esta atividade especifica as características essenciais para compreensão de todo objeto de estudo, em consonância com a estrutura semântica de transporte construída. No anexo B, é apresentado à rede semântica do sistema de transporte rural na forma simplificada.

Atividade 04: Identificação dos grupos-alvos e da necessidade de informação

Os principais atores (stakeholders) envolvidos com o sistema, ou usuários do sistema, são identificados, como também, as suas necessidades de informação. Assim, determina-se o que cada um precisa saber para desempenhar seu papel dentro do processo de tomada de decisão. Em função desse levantamento, são definidos os elementos a serem representados e os perfis de necessidades de informação. Essa atividade tem como objetivo identificar informações relevantes, sob a perspectiva do domínio de negócio dos grupos-alvos.

No âmbito da Gerência de Projetos, o termo Stakeholder designa as pessoas que influenciam ou são influenciadas pelas ações de um projeto. Para a identificação das partes interessadas(stakeholders), envolvidas com o sistema de transportes estudado, é preciso levantamento de algumas informações que possam auxiliar na confecção de uma lista de

atores envolvidos com o sistema de transportes avaliado. Para isso, existem várias técnicas, dentre as quais: (i) Brainstorming (chuva de idéias); (ii) Grupos focais; (iii) Entrevistas; (iv) Simples observação; (v) Aplicação de questionário eletrônico; (vi) Aplicação de questionário físico (MULCAHY, 2007).

Atividade 05: Definição do tipo de conhecimento que se deseja descobrir

Conforme a necessidade de apoio à tomada de decisão, seja do planejador, do gestor ou do controlador, é definido o tipo de conhecimento necessário que leva à definição dos objetivos. Se a estrutura semântica de representação do conhecimento acerca de um determinado objeto de estudo for sob uma perspectiva de negócio, o tomador de decisão visará ao entendimento dos requisitos desse negócio e dos objetivos do objeto de estudo (ver seção 5.2.1.2- Redes semânticas). Logo depois, determiná-se-a o conhecimento que se deseja descobrir do objeto de estudo, para transformá-lo em um problema, que será decomposto em várias tarefas de mineração de dados.

Esse problema, transformado em uma ou mais tarefas de mineração de dados, será discutido na atividade 07. Por exemplo: o objetivo é reduzir o número de acidentes de trânsito no perímetro urbano, logo um dos problemas de mineração de dados é identificar as características dos motoristas que se envolveram em acidentes de trânsito, descobrir suas razões e dividi-los em grupos, de acordo com os motivos por envolvimento em acidentes.

4.3.2 **Etapa II - Elaboração (modelagem)**

Esta etapa, a parte mais trabalhosa do processo, consiste na escolha da ferramenta, na escolha da tarefa, na escolha da técnica de limpeza e pré-processamento dos dados selecionados; na escolha da técnica de transformação dos dados selecionados; da integração na fonte de dados e na geração do arquivo de dados, conforme a Figura 4.4.

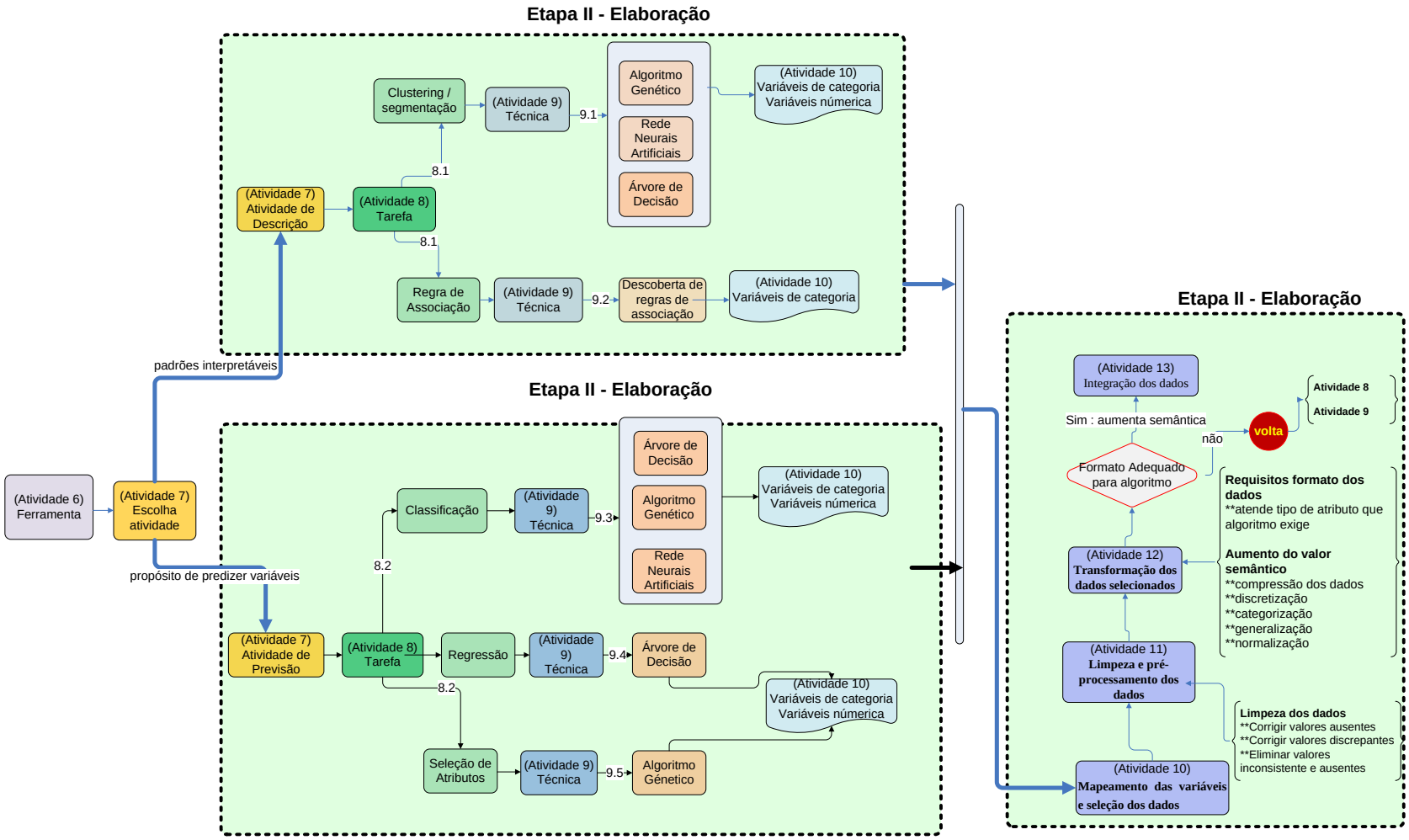


Figura 4.4 : Etapa de elaboração da modelagem

Atividade 06: Definição da ferramenta de mineração e dos recursos disponíveis

Nesta etapa, se avalia e se seleciona ou se desenvolve uma ferramenta que utilize a tecnologia de mineração de dados. Ainda nesta atividade, é feita a avaliação dos recursos disponíveis para o projeto, bem como, recursos computacionais utilizados: hardware, sistema de gerenciamento de banco de dados - SGBD, software de texto, planilha e software de mineração de dados. Segundo Goebel e Gruenwald (1999), as características fundamentais a serem consideradas na escolha de uma ferramenta devem ser:

- Suporte para acesso as diversas fontes de dados;
- A capacidade de reconhecer os diversos tipos de modelagem: modelos de dados orientados a objetos, modelo relacional e modelos não padronizados (tal como multimídia, espacial ou temporal);
- A capacidade de processamento é ilimitada em relação aos números de tabelas, atributos e registros.

Atividade 07: Escolha da atividade

Para o desenvolvimento dessa atividade 07, é fundamental que o problema de mineração de dados esteja bem definido, pois, dependendo do tipo de problema, existem dois fluxos de atividades desempenhados dentro do processo, conforme a Figura 4.4:

- Caso se deseje descobrir *padrões com o propósito de predizer variáveis de interesse*, deve-se selecionar a *atividade preditiva*, cujo objetivo é diminuir o esforço na interpretação dos resultados. A compreensão da atividade de predição, identifica a generalização de observações ou experiência passada, com a resposta conhecida, e cria um modelo (linguagem) capaz de reconhecer a classe de um novo registro (observação). Seu objetivo final é realizar uma análise prospectiva de dados, ou seja, uma análise dos dados que prevê futuras tendências, comportamento ou eventos com base em dados históricos. Por exemplo: perfil dos motoristas que recebem multas gravíssimas com a atividade preditiva, utilizando a tarefa de classificação, somente a título de exemplo, poderia inferir (prever) que motorista do

sexo masculino, com idade entre 18 a 24 anos e peso inferior a 60 quilos recebem multas gravíssimas. Nesse caso, o atributo de '*multas gravíssimas'* é denominado classe, pois é o atributo alvo da classificação;

- Caso se deseje descobrir *padrões interpretáveis e compreensíveis*, a escolha será *descritiva* e obrigatória, na *análise interpretativa dos padrões* e do entendimento do domínio de negócio. O objetivo final da descrição é realizar uma análise retrospectiva dos dados**,** ou seja, uma análise dos dados que forneça uma percepção sobre a tendência, o comportamento ou os eventos que já ocorreram. Por exemplo: Os motoristas que tendem a receber multa gravíssima de avanço de sinal e se envolvem em acidentes graves geralmente são "motoristas do sexo masculino, com idade entre 21 a 24 anos e renda superior a R$ 4.0000". No caso em estudo, essas relações entre os atributos eram, até então, desconhecidas. Logo, percebe-se a importância desse algoritmo na identificação da correlação entre os atributos.

Atividade 08: Escolha do tipo de tarefa

Esta atividade tem uma forte dependência com a atividade 7, pois, a escolha da tarefa está interligada ao tipo de atividade que será utilizado no processo, conforme a Figura 4.4:

Se as atividades forem descritivas as tarefas relacionadas serão:

- **Tarefa de associação**: Determinam-se quais fatos ou objetos tendem a ocorrer juntos, no mesmo evento. Por exemplo: determina quais infrações ocorreram juntas em um determinado horário;

- **Tarefa de agrupamento**: Divide-se um conjunto de dados heterogêneos em vários subgrupos mais homogêneos, possíveis entre si. Por exemplo: identifica o perfil dos motoristas infratores.

Se as atividades forem de predição as tarefas relacionadas serão:

- **Tarefa de regressão**: Estima-se o valor da variável dependente, a partir de várias variáveis independentes, quando em conjunto produzam um resultado. Por exemplo: estimar o tempo de viagem dos usuários do transporte escolar;

- **Tarefa de classificação**: Cria-se um modelo que possa ser utilizado em dados novos, não classificados para categorizá-lo. Por exemplo: a classificação dos motoristas de acordo com as infrações cometidas;

- **Tarefa de seleção de atributos**: Encontram-se os atributos que têm valor significativo de negócio, para uma determinada classe, que serão selecionadas dentre os atributos relevantes para a mineração dos dados, separando-os dos atributos irrelevantes. Por exemplo: no formulário de pesquisa foram confeccionadas 54 perguntas, contudo, somente 15 perguntas foram selecionadas como relevantes.

Atividade 09: Escolha do tipo de Técnica

Segundo Fayyad *et al.* (1996a, 1996b), não existe um método de mineração de dados genérico para todos os problemas, e a escolha de um algoritmo específico para uma determinada aplicação é de certa forma uma arte.

A Figura 4.4 mostra que essa atividade tem uma forte dependência com a atividade 8 (tipo de tarefa), pois a escolha de uma técnica tem uma correlação muito forte com tipo de tarefa escolhida. Como não há uma técnica que resolva todos os problemas de mineração de dados, para cada problema a ser resolvido, deve-se buscar a melhor, já que, não se sabe quais delas atendam e modelem o problema. Então, a escolha da técnica tem uma correlação entre atividade e tarefa, conforme descritos abaixo:

Se a atividade for descritiva com tarefa de agrupamento, então as técnicas serão:

- **Algoritmo k-Means**: realiza a divisão de conjunto de dados heterogêneos em vários subgrupos mais homogêneos possíveis entre si. Essa classificação é baseada em análise e comparação entre os valores numéricos dos dados. Exemplo de algoritmo: *kMeans*;

- **Árvore de decisão**: Cria-se um modelo em forma de árvore, que é utilizado para classificar dados novos sem testar todos os atributos. Exemplo de algoritmo: *C5.0 e J.48;*

- **Algoritmo genético**: Métodos gerais de busca e otimização, inspirados no processo de evolução. Encontra a solução ideal de um problema específico, após verificar um imenso número de soluções alternativas. Exemplo de algoritmo: *Algoritmo Genético Simples e Algoritmo de Hillis*;

- **Redes neurais artificiais**: Modelo que adota os padrões de processamento dos cérebros humanos. São utilizados para solucionar os problemas complexos, para os quais existem uma enorme quantidade de dados coletados. Exemplo de algoritmo: *Perceptron e Rede MLP.*

Se a atividade for descritiva com tarefa de associação, então a técnica será:

- **Descoberta de regras de associação**: Encontra-se associações, correlação estatística entres os atributos de dados, a partir de um conjunto de dados. Exemplo de algoritmo: *Apriori e AprioriTid.*

Se a atividade for preditiva com tarefa de classificação, então as técnicas serão:

- **Árvore de decisão**: Cria-se um modelo em forma de árvore, que é utilizado para classificar os dados novos sem testar todos os atributos. Exemplo de algoritmo: C5.0, J.48;

- **Algoritmo genético**: Métodos gerais de busca e otimização, inspirados no processo de evolução. Encontra-se a solução ideal de um problema específico após verificar um imenso número de soluções alternativas. Exemplo de algoritmo: *algoritmo genético simples e algoritmo de Hillis*;

- **Redes neurais artificias**: Modelo que adota os padrões de processamento do cérebro humano. São utilizados para solucionar os problemas complexos, para os quais existem uma enorme quantidade de dados coletados. Exemplo de algoritmo*: perceptron e rede MLP*.

Se a atividade for preditiva com tarefa de seleção de atributo, então a técnica será:

- **Algoritmo genético**: Métodos gerais de busca e otimização, inspirados no processo de evolução. Encontra-se a solução ideal de um problema específico, após verificar

um imenso número de soluções alternativas. Exemplo de algoritmo: *algoritmo genético simples e algoritmo de Hillis.*

Atividade 10: Mapeamento das variáveis e da seleção dos dados

Essa atividade é decomposta em duas subatividades:

A primeira atividade é o mapeamento das variáveis que precisam do entendimento da rede semântica de transporte (ver seção 5.2.1.2 - Redes semânticas). Conforme afirma Barbetta (2001), para conhecermos certas características dos elementos de uma população, precisamos coletar dados desses elementos (variáveis).

As variáveis caracterizam os elementos de representação da rede. Por exemplo: em um sistema de transportes (objeto de estudo), o automóvel é um elemento da rede semântica. O índice IPK (Índice de Passageiros por Quilômetro) é um de seus elementos de representação. A quantidade de passageiros e a quantidade de quilômetros são as variáveis que qualificam esse elemento de representação.

A segunda atividade é a seleção dos dados, que consiste no levantamento dos dados disponíveis e das possíveis formas de obtenção de novos dados que se mostrem necessários. Então, deve-se aplicar o processo de avaliação da qualidade dos dados, considerando-se os quatros critérios básicos de análise de qualidade: (i) como o dado é coletado; (ii) como o dado é consolidado; (iii) qual a qualidade destes dados; e (iv) se eles atendem às necessidades da etapa de mineração de dados.

Se os dados existentes responderem aos quatro critérios, eles poderão ser aproveitados na atividade de mineração de dados. Caso contrário, será necessário passar pela etapa de execução da pesquisa e tratamento dos dados coletados, ou mesmo pela complementação.

Uma observação muito importante é compreender as características dos dados selecionados, a fim de minimizar as dificuldade de transformação dos dados, para satisfazer os requisitos do algoritmo, conforme mencionado na seção 3.4.6 - Entendo característica dos dados.

Atividade 11: Limpeza e pré-processamento dos dados

Esta atividade é importantíssima e crucial no processo de mineração de dados, pois as qualidades dos dados vão determinar a eficiência dos algoritmos. Nesta etapa, deverão ser realizadas tarefas de eliminação dos campos inconsistentes, redução das discrepâncias (ruidosos) e correção dos campos incompletos. Para que essa atividade tenha sucesso, é preciso uma compreensão melhor sobre os termos utilizados:

- **Inconsistentes** - divergência de valor que está fora da faixa permitida para o campo;
- **Discrepâncias** – dados que possuem valores extremos, atípicos ou com característica bastante distinta dos demais registros;
- **Incompletos** - atributo que não possui valor ou quando o valor do mesmo estiver ausente.

Como estratégias para **correção dos valores incompletos ou ausentes,** podem ser aplicadas as seguintes técnicas (HAN e KAMBER, 2001):

1. Exclusão dos registros;
2. Substituição dos valores ausentes por:
    a. Uma constante global;
    b. A média do atributo;
    c. A média do atributo para todas as instâncias da mesma classe;

Como observação, a técnica 2 deve ser utilizada com cautela, porque podem viciar os dados.

Como estratégias para **correção dos valores discrepantes,** podem ser aplicadas as seguintes técnicas (HAN e KAMBER, 2001):

1. Interpolação;
2. Agrupamento;
3. Inspeção humana e computacional combinadas;

4. regressão.

Como estratégia de **eliminação dos valores inconsistentes,** pode-se corrigir manualmente, através de referências externas. (HAN e KAMBER, 2001).

Atividade 12: Transformação dos dados selecionados

Esta atividade tem uma forte dependência com a atividade 8 e 9, porque a escolha da tarefa e da técnica possuem requisitos ligados ao formato dos dados, conforme mencionado na seção 3.4.6 - Entendo característica dos dados. Frequentemente, é preciso voltar à atividade 8 ou à atividade 9 para selecionar os dados que atendam aos requisitos do algoritmo, que serão utilizados na atividade 14.

Outro ponto de vista importante, nessa atividade é a identificação das características relevantes dos dados. Talvez, seja preciso diminuir a quantidade de variáveis a serem utilizadas no conjunto de dados, tendo como objetivo o aumento do valor semântico das informações utilizadas no processo de mineração de dados.

Nessa atividade, os dados são alterados ou transformados em formato adequado aos requisitos da atividade 8 (selecionar tarefa) e atividade 9 (selecionar técnica). Dentre as transformações de dados existentes, estão listadas algumas técnicas (HAN e KAMBER, 2001):

1. **Compressão de dados (extração de variáveis)**: obtêm novas variáveis através dos atributos iniciais ou reduz o número de variáveis irrelevantes, com a menor perda de informação, ou reduz o tamanho da base de dados, sem prejudicar a qualidade da amostra. Assim, é importante uma análise prévia de quais atributos realmente serão necessários à tarefa de classificação. Por exemplo: telefone e email não influenciam na classificação de conjunto de dados;

2. **Discretização de variáveis**: transforma variáveis contínuas em variáveis categóricas;

3. **Categorização**: atributos que possuem uma enorme variedade de valores, que podem ser reunidos em algumas poucas categorias. Por exemplo, a idade do motorista pode ser agrupada em 4 categorias : <=20, 21..34, 35...60, > 60;

4. **Generalização**: substituem alguns atributos como rua, número e logradouro, por um atributo mais geral como, cidade, aumentando-se o valor semântico;

5. **Normalização**: compreende como escalonar os valores de um atributo de modo que fique dentro de uma faixa curta de valores, por exemplo [-1,1] ou [0,1]. O método de normalização, chamado min-max, é apresentado na equação 4.1.

$$\text{Normalização min} - \text{max} = \frac{\text{V} - \text{Min}}{\text{Max} - \text{Min}} \tag{4.1}$$

onde :

V: valor normalizado;
Min: para o menor valor conjunto de dados ;
Max: para o maior valor conjunto de dados.

Todas as técnicas devem assegurar a representatividade da amostra. Caso a amostra não seja representativa, ou se a quantidade de observações for insuficiente para extrair os padrões escondidos nos dados, os modelos gerados poderão não representar a realidade, tendo como conseqüência a perda de valor da informação, para o processo de tomada de decisão.

Atividade 13: Integração dos dados

Esta atividade é decomposta em duas subatividades:

A primeira atividade é realizada, quando se necessita da integração de diferentes fontes de dados (bancos de dados distintos, arquivos xml, planilhas eletrônica etc.). Essa atividade, quando realizada, deverá ser bastante rigorosa para reduzir e restringir a redundância e a inconsistência no conjunto de dados resultante.

Como **estratégias para a integração de dados** podem ser aplicadas as seguintes técnicas (HAN e KAMBER, 2001):

1. Integração de esquemas dos bancos de dados relacionais (o esquema define as tabelas, os campos em cada tabela e os relacionamentos entre os campos e as tabelas);
2. Redundância de atributos;
3. Identificação e correção de valores de dados conflitantes.

A segunda atividade é a geração do arquivo, com os dados selecionados, em formato específico da ferramenta de mineração de dados.

4.3.3 **Etapa III – Construção (implementação)**

Esta etapa é composta por duas atividades – a primeira consiste na aplicação de método para extração de padrões nos dados, na busca pelo melhor ajuste dos parâmetros do algoritmo, para a tarefa em questão. A segunda realiza a atividade de pós-processamento com objetivo de identificar os padrões mais interessantes, de acordo com o domínio de negócio, conforme a Figura 4.5.

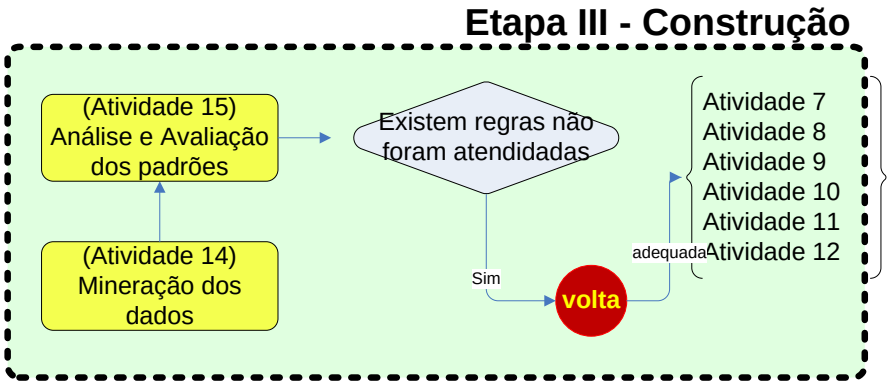


Figura 4.5: Etapa de construção da metodologia

Atividade 14: Mineração de dados

Esta atividade é composta de duas etapas:

1. Realização de teste para identificação dos parâmetros mínimos como também encontrar os melhores ajustes dos parâmetros do algoritmo, para a tarefa em questão;

2. A aplicação do algoritmo no conjunto de dados para geração dos padrões ou modelos.

Atividade 15: Análise e avaliação dos padrões

A atividade de avaliação deve ser executada somente para *atividade preditiva*, porque realiza a verificação do modelo gerado, cujas regras são testadas sobre outro conjunto de dados de teste, diferente do conjunto de dados original. A qualidade do modelo é mensurada pelos registros do banco de teste das regras do modelo, classificadas de forma correta e satisfatória.

Os testes de validação descrevem os parâmetros de validade e a confiabilidade dos modelos gerados, a partir de uma perspectiva de domínio. Por outro lado, os indicadores estatísticos da matriz confusão, do índice de correção entre outros são utilizados para auxiliar na análise dos resultados. Caso estes indicadores mostrem que os modelos são ruins, deverão voltar para algumas das atividades anteriores (atividade 7 até 12).

Cabe ressaltar, que se a atividade de análise dos padrões extraídos não seja interessante, é preciso retornar para alguma etapa da metodologia como, por exemplo: (i) alterar os parâmetros do algoritmo; (ii) selecionar um outro algoritmo; (iii) selecionar um novo conjunto de dado e (iv) outras ações dentro do processo, conforme ilustra na Figura 4.2.

4.3.4 **Etapa IV – Transição (interpretação)**

Nesta etapa é que se interpreta e avalia os padrões, sob um determinado contexto de negócio, a fim de gerar informações e conhecimentos, para que se possa subsidiar o planejador, o controlador e o gestor, na elaboração do plano de ação, conforme a Figura 4.6.

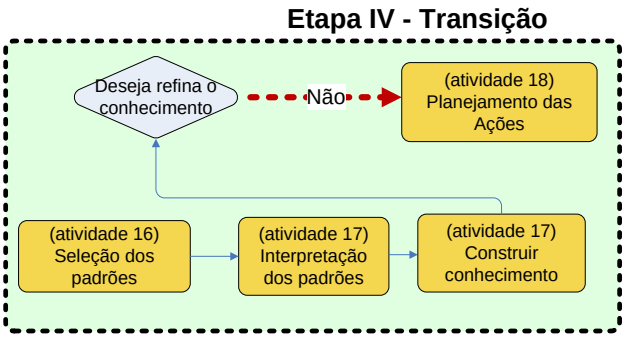


Figura 4.6: Etapa de transição da metodologia

Atividade 16: seleção dos melhores padrões

Os algoritmos geram uma grande quantidade de padrões e muitos deles não são interessantes, faz-se necessário, então, a utilização de algumas técnicas de pós-processamento, com o apoio do especialista de negócio, a fim de prover apenas o conhecimento interessante e útil ao usuário final.

Bruha e Famili (2000) apresentam várias técnicas de pós-processamento utilizadas para tratamento dos padrões extraídos:

- **Filtragem do Conhecimento** – Pode ser realizada por meio da ordenação das medidas de interesse objetivas para a regra de associação;
- **Avaliação** – Pode ser realizada por meio de critérios, como precisão e compreensibilidade;
- **Interpretação e Explanação –** O conhecimento extraído pode ser comparado com o conhecimento prévio do domínio de aplicação e do grau de interesse, entre outros;
- **Integração do Conhecimento** – O conhecimento extraído pode ser integrado a um sistema de apoio à tomada de decisões.

A parte final desta atividade é a organização dos padrões relevantes, sob a perspectiva de negócio.

Atividade 17: Interpretação dos padrões

Nesta etapa, o conhecimento extraído pode ser *simplificado, avaliado, visualizado ou apenas documentado,* para a utilização em processo de tomada de decisão, conforme apresentado na seção 3.2.2 - Processo de tomada de decisão. Dependendo das exigências, esta fase pode ser tão simples, como gerar um relatório, ou complexa, como fazer um banco de dados com os modelos gerados.

Cabe ressaltar que, caso os padrões extraídos não sejam interessantes, então é preciso retornar para alguma das etapas da metodologia como, por exemplo: (i) alterar os parâmetros do algoritmo; (ii) selecionar um outro algoritmo; (iii) selecionar um novo conjunto de dado e (iv) outras ações dentro do processo, conforme ilustra na Figura 4.2.

Atividade 18: Construção do conhecimento aos grupos-alvos (usuário final)

Esta atividade consiste na formação de recursos humanos, capazes de utilizar corretamente o sistema, para apoio à tomada de decisão nas suas atividades. Nesse contexto, são previstos cursos, oficinas e workshops entre outros. Essa metodologia em si, possui duas características relevantes: (i) interativo, pois o usuário pode intervir e controlar o curso das atividades; (ii) iterativo, por ser uma sequência finita de operações cujo resultado de cada uma das operações é dependente dos resultados das que a precedem. Caso o especialista deseje refinar os padrões encontrados, voltará para a atividade 3, conforme ilustra na Figura 4.2.

Atividade 19: Planejamento das ações

Esta atividade é posterior ao processo de mineração de dados. O conhecimento extraído pode ser utilizado pelos grupos alvos (usuários finais) para o apoio ao processo de tomada de decisão, ou análise de resultados, ou incorporado a um sistema de apoio à tomada de decisão, ou reportado a outras partes interessantes, ou ainda utilizado para resolver eventuais conflitos entre o conhecimento pré-existente (especialista de negócio) e o conhecimento obtido, com o processo de mineração de dados.

4.4 TÓPICOS CONCLUSIVOS DA METODOLOGIA PROPOSTA

Este capítulo sintetiza a proposta de uma metodologia, para elaboração de um sistema de apoio à tomada de decisão em transporte, que integre o processo de mineração de dados, para dar suporte ao planejamento, gestão e controle. Logo, os sistemas de transportes geram uma quantidade enorme de informações (dados) em seus diversos segmentos, que precisam ser monitorados e diagnosticados. A manipulação de dados é sempre realizada conforme a disponibilidade do conhecimento dos técnicos, como também do tempo disponível. Assim, nos estudos de transportes, a necessidade de técnicas que apoiem a análise dos dados gerando inteligência é de grande importância para o melhor conhecimento do objeto transportes a ser estudado, o qual sofrerá as intervenções por meio de projetos e planos de ações. Em seguida, serão abordados alguns tópicos conclusivos sobre esse capítulo:

- Essa metodologia demonstrou, de forma clara, que existem outras formas de análise exploratória dos dados. Então, o objetivo dessa pesquisa é exatamente apresentar uma ferramenta em potencial para uma análise analítica, por meio de classificação, agrupamentos e associações dos dados, em complemento as formas tradicionais de análise. Enfim, em um ambiente de análise integrado necessita-se de resultados de ambos os tipos de análises, ou seja, das análises que são *complementares* e *não sobrepostas.*

- A estrutura da metodologia proposta é composta por quatros grandes etapas. Sendo bastante intuitiva e iterativa entre as etapas, como propósito de descobrir padrões válidos, novos e compreensíveis, em base de dados, para apoiar o especialista de domínio na análise dos resultados e no processo de tomada de decisão. Assim, essa metodologia desenvolveu o procedimento de realizar análise de dados, no processo de diagnóstico;

- O desenvolvimento da metodologia é longo, com muitas atividades a serem executadas, porém a agregação de valor da informação serve de base à tomada de decisão em estudo de transporte, durante o processo de planejamento, de gestão e de controle. Nesse sentido, os tomadores de decisão têm uma chance maior de

optar, com segurança, por quais caminhos deverão percorrer para a solução do problema.

- Somente uma técnica de análise exploratória dos dados, utilizando a estatística descritiva, não atende as necessidades de informações de transportes; sendo necessária a integração complementar da análise analítica, com suas técnicas de agrupamento, classificação e associação, para extrair conhecimento escondido em base de dados de transporte, para subsidiar o processo de tomada de decisão. Assim, essa metodologia alinha-se aos objetivos dessa pesquisa.

- O usuário de uma ferramenta de mineração de dados precisa de um bom entendimento sobre a área de negócio; a compreensão e sobre as técnicas implementadas pela ferramenta sobre todo o *processo KDD (Knowledge Discovery in Databases)*.

5 ESTUDO DE CASO: CARACTERIZAÇÃO DO TRANSPORTE ESCOLAR RURAL UTILIZANDO O PROCESSO DE MINERAÇÃO DE DADOS

5.1 APRESENTAÇÃO

Este capítulo tem por objetivo verificar a aplicabilidade da metodologia e dos conceitos dispostos nos capítulos anteriores, os quais revisaram importantes conceitos para o desenvolvimento dessa metodologia. Para isso, será apresentado um estudo de caso que utiliza o sistema de transporte escolar rural. O objeto, transporte escolar rural, foi extensamente estudado pelo grupo de pesquisa do Centro Interdisciplinar de Estudos em Transportes – CEFTRU e Fundo Nacional de Desenvolvimento da Educação - FNDE, com resultados já publicados (CEFTRU/FNDE, 2007a, 2007b, 2007c, 2007d).

Antes de se proceder ao estudo de caso, é preciso se ter claro que a implementação desse sistema de apoio à tomada de decisão se utiliza da modelagem como técnica de mineração, para realização da *análise analítica de dados*. A modelagem é uma proposta de compreensão, cujo objeto de estudo é orientado a fim específico. Em 2006, uma grande pesquisa foi realizada com o objetivo de caracterizar o transporte escolar rural em todo o Brasil (CEFTRU/FNDE, 2007a, 2007b, 2007c, 2007d). Dessa forma, nos diversos municípios do território nacional, esse estudo de caso se utiliza dos dados coletados, na busca de sua caracterização.

Este capítulo foi estruturado em três seções. A seção 5.2 apresentou de forma sucinta a contextualização do objeto de estudo, que é o transporte escola rural (CEFTRU/FNDE, 2007a, 2007b, 2007c, 2007d). A seção 5.3 apresentou de forma geral os dados coletados na pesquisa web realizados em 2007 em todos os municípios brasileiros. A seção 5.4 demonstrou uma aplicação da metodologia, com o objetivo de verificar a viabilidade da metodologia proposta.

5.2 CONTEXTUALIZAÇÃO DO OBJETO DE ESTUDO: TRANSPORTE ESCOLAR RURAL BRASILEIRO

A garantia do acesso à educação dos estudantes, residentes na área rural, é papel fundamental do governo para inclusão social. Nesse sentido, o acesso à educação é uma garantia Constitucional, conforme se pode analisar no art. 206, inciso I e art. 208, inciso VII da Constituição Federal de 1988, que estabelecem que é dever do Estado garantir o acesso à educação e fornecer as condições necessárias para que o aluno chegue à escola; por meio de programas suplementares de material didático-escolar, transporte, alimentação e assistência à saúde (BRASIL, 1988).

Assim, cabe ao Poder Público fornecer os meios de transporte necessários para viabilizar o deslocamento dos alunos aos estabelecimentos de ensino, possibilitando, dessa forma, que uma maior parcela da população tenha acesso à educação.

Nesse sentido, o transporte escolar possui um papel fundamental na viabilização do acesso e na permanência dos estudantes nas escolas, principalmente àqueles que residem em áreas rurais. Assim, ações que visem à melhoria desse tipo de transporte, podem influir no aprendizado dos alunos, que necessitam desse transporte e, com isso, melhorar no desenvolvimento da educação no país, além de possibilitar a permanência desses alunos na área rural (FNDE, 2006; MEC/INEP, 2007).

No entanto, o transporte escolar rural enfrenta problemas em função das características singulares do meio rural. Essas características, tais como o isolamento espacial, a baixa densidade demográfica, a grande parcela da população com baixa renda, o pequeno número de escolas, as vias em condições precárias, a utilização de veículos velhos e com manutenção deficitária, acabam promovendo desconforto e insegurança. (FNDE, 2006; MEC/INEP, 2007).

Segundo Carvalho (2011), o transporte escolar rural nos últimos anos passou a ser observado pelo Poder Público. Contudo, esse serviço recebeu pouco investimento ao longo dos anos, resultado que é visto pelas condições precárias de acessos às unidades de ensino, que interferem tanto na qualidade de sua prestação, como no rendimento escolar dos alunos.

Diante disso, nesta seção, para entendimento do real estado do objeto de estudo, é de fundamental importância a compreensão da representação do conhecimento, do significado de ontologia e da representação da rede semântica (ver seção 5.2.1.2 - Redes semânticas).

5.2.1 **Representação do conhecimento**

A interpretação do conhecimento foi desenvolvida no campo da Inteligência Artificial (IA), com o propósito de tornar os computadores capazes de realizar tarefas e funções similares às desenvolvidas pelos seres humanos. Esses estudos impulsionaram a denominada ciência da cognição, que postula conceitos, para explicar a construção do conhecimento (BREWKA,1996).

Brewka (1996) explica que a representação do conhecimento, como o método utilizado para codificá-lo, é de acordo com a base de conhecimentos de um sistema de informação especialista de área de negócio. Então, a representação do conhecimento, realizada por esses sistemas, proporciona aos homens e aos computadores a compreensão ontológica do conhecimento que se quer representar.

Diante disso, nessa seção são apresentados os conceitos básicos sobre ontologia e rede semântica.

5.2.1.1 Ontologia

De acordo com Sowa (2000), o termo ontologia é originário da filosofia e se constitui em um ramo que lida com a natureza e a organização do ser. Em geral, são conceitos e termos que podem ser usados para descrever alguma área do conhecimento ou construir uma representação do objeto de estudo (RIOS, 2003). Essa representação do conhecimento é um assunto multidisciplinar, que aplica teorias e técnicas de outros três campos (SOWA, 2000):

- **Lógica**: proporciona a estrutura formal e as regras de inferência;
- **Ontologia**: define os tipos de coisas que existem no domínio da aplicação;

- **Computação**: apóia as aplicações, distinguindo a representação do conhecimento, a partir da filosofia pura.

De acordo com Rios (2003), a ontologia pode ser considerada como o ramo da metafísica que trata da natureza do ser e destina-se a criar terminologias únicas, de maneira a contribuir para que o conhecimento possa ser compartilhado e reutilizado. A inteligência artificial adaptou esse termo, utilizando-o para se referir a um conjunto de conceitos ou termos usados, para descrever algumas áreas do conhecimento ou para construir uma representação desse conhecimento em relação a algo.

Contrariamente a esta definição, Guarino (1998), em estudo sobre a relação entre a ontologia formal e os sistemas de informação considera a ontologia como uma disciplina da filosofia, utilizada tanto pela comunidade filosófica, quanto pela comunidade ligada à Inteligência Artificial. O autor define que, no campo da Inteligência artificial, a ontologia se refere a um instrumento de engenharia, constituído por um vocabulário específico, usado para descrever certa realidade, somado a um conjunto de pressupostos explícitos, relacionados ao significado das palavras.

Brewka (1996) expõe que a representação da rede é um conjunto de convenções, sobre a descrição de uma classe de coisas, e é formada por quatro partes essenciais:

- **Parte léxica**: determina a simbologia da representação do vocabulário;
- **Parte estrutural**: determina as restrições na organização dos símbolos;
- **Parte de procedimentos**: estabelece a forma de acesso a fim de criar descrições e responder aos questionamentos por meio dessas descrições;
- **Parte semântica**: estabelece um caminho de associação significativo com as descrições.

Assim, toda forma de representação do conhecimento deve apresentar as partes léxica, estrutural, de procedimentos e semântica, como também deve aplicar teorias e técnicas oriundas dos campos da lógica, da ontologia e da computação. Por isso, os diversos autores consideram a ontologia como um componente de uma coleção de informações, sendo ela o alicerce para a construção do conhecimento (MAGALHÃES, 2010).

No próximo tópico, serão apresentados os conceitos e fundamentos teóricos sobre redes semânticas.

5.2.1.2 Redes semânticas

Quillian (1968) foi precursor do conceito de rede semântica, que forneceu uma representação do conhecimento, baseado no significado das palavras. Então, seu modelo proposto era conhecido como "*memória semântica*", que representava uma estrutura como significado das palavras, que eram representados na memória humana. A estrutura da rede era composta de nós e links, que representavam relações entre as palavras e seus conceitos.

Sowa (1992), explica que o conceito de rede semântica, foi explorado inicialmente pela inteligência artificial. Entretanto, as redes semânticas foram aplicadas também em estudos da filosofia, da psicologia e da linguística, e, recentemente, em várias outras áreas do conhecimento. Essas diversidades de formas de utilização em diversas ciências, se devem às características de uma representação gráfica declarativa. Essa representação pode ser aplicada à inferência sobre esse conhecimento ou representação do conhecimento. A seguir, são apresentados os seis tipos comuns de redes semânticas (SOWA, 1992):

1. **Redes de definição**: enfatizam o subtipo ou a relação "é um", entre um tipo conceitual e seu subtipo recém definido. A rede resultante, também denominada generalização ou hierarquia de submissão, suporta as regras da herança por meio da passagem das propriedades definidas de supertipo para todos os seus subtipos. Uma vez que, as definições são verdadeiras por definição, a informação nessas redes é assumida necessariamente como verdadeira;

2. **Redes de asserção**: são desenhadas para garantir proposições lógicas. Sua informação é considerada contingentemente verdadeira, ao menos que seja marcada com um operador de modo. Algumas redes de asserção têm sido propostas como modelo de estrutura conceitual, ressaltando a linguagem semântica natural;

3. **Redes de implicação**: utilizam a implicação como principal relação entre os nós, geralmente aplicadas para representar padrões de crenças, causalidade e

inferências. A lógica e a probabilística são as principais abordagens aplicadas nesse tipo de rede;

4. **Redes executáveis**: possuem mecanismos para execução de inferências, passagem de mensagens ou busca por padrões e associações, permitindo alteração dinâmica na própria rede;

5. **Redes de aprendizado**: constroem ou estendem a representação, através da aquisição de conhecimento advindo de exemplos. O novo conhecimento altera a rede pela adição ou remoção de nós e ligações e pela alteração de valores numéricos;

6. **Redes híbridas**: que combinam duas ou mais redes em uma única rede ou em redes separadas, mas em interação.

Brewka (1996) afirma que a estrutura da rede era composta de nós que representavam objetos, conceitos ou valores, relacionados a substantivos e adjetivos. Já os arcos, geralmente, representavam verbos e preposições, suas relações mais comuns são: "é um", "parte de" e "é um tipo de". A rede semântica é composta de quatro partes essenciais, que são representadas:

- **Parte léxica**: formada por nós como símbolos dos objetos; arcos simbolizando as relações entre objetos; e rótulos e arcos para relações particulares;

- **Parte estrutural**: geralmente, os nós são representados por retângulos, elipses e círculos; e os arcos são representados por setas;

- **Parte de procedimentos**: constituída por procedimentos construtores, apagadores, e escritores (alteram) de nós e arcos; e procedimentos leitores, para responder questões sobre os nós e os arcos;

- **Parte semântica**: é variável de acordo com a aplicação.

Corradi *et al*. (2001) afirma que a rede semântica é uma forma gráfica de representação do conhecimento, cuja estrutura é composta por nós e arcos interconectados. A noção de rede semântica é proveniente dos estudos realizados pelas ciências da cognição, com foco nas

atividades mentais dos seres humanos para identificar a forma de construção do conhecimento.

Neste trabalho, define-se a ***rede semântica*** como uma forma gráfica de representação do conhecimento, que pode ser utilizada para a organização, disseminação e estruturação do conhecimento produzido. Então, essa representação gráfica facilita a identificação do objeto e seus elementos básicos.

Nos próximos tópicos, são apresentados os conceitos e os fundamentos teóricos sobre o sistema de transporte escolar rural e a rede semântica do transporte escolar rural.

5.2.2 **Conceituando o sistema de transporte escolar rural**

As definições de diversos autores referenciais na área de transporte sobre o conceito de sistema de transporte estão condensadas no seguinte consenso o *sistema de transporte é meio para atingir um fim* (MORLOK, 1978; PAPACOSTAS e PREVEDOUROS, 1993; BANISTER, 1998; ORTÚZAR e WILLUMSEN, 2001; MAGALHÃES, 2010).

O transporte escolar rural, segundo definição da Empresa Brasileira de Planejamento de Transportes - GEIPOT (1995), é "o transporte de passageiros, público ou de interesse social, entre a área rural e a área urbana ou no interior da área rural do município". Esse texto representa uma das primeiras definições sobre o tema, apresentada em um documento de avaliação preliminar do transporte rural (GEIPOT, 1995).

Segundo Lopes *et al.* (2008), define-se ***transporte escolar*** como o "serviço destinado a levar crianças e jovens, que estejam matriculados, de casa para a escola e da escola para casa, permitindo, assim, que todos consigam chegar às unidades de ensino e ter acesso à educação".

De acordo com o CEFTRU/FNDE (2007d), o ***transporte escolar*** é dividido em transporte escolar urbano e rural, e pode ser definido como:

- **Transporte Escolar** - o transporte coletivo de estudantes, para fins educacionais, entre sua residência (ou local específico previamente acordado) e uma instituição educacional de qualquer nível de escolaridade ou dependência administrativa;

- **Transporte Escolar Urbano** - o transporte escolar realizado exclusivamente dentro do perímetro urbano (cidade ou sede de distrito);

- **Transporte Escolar Rural** - o deslocamento que ocorre a partir da intenção dos alunos que residem e/ou que estudem em área rural, e sua finalidade é permitir que o aluno se desloque e possa estudar.

Neste trabalho, define-se o ***Sistema de Transporte Escolar Rural - STER*** como um conjunto de elementos inter-relacionados, que têm como objetivo comum deslocar os alunos, que residam ou que estudem em área rural, entre seu local de residência e um estabelecimento de ensino (CEFTRU/FNDE, 2008a, 2008b). Conforme a Figura 5.1, apresenta-se a relação dos principais elementos existentes nesse sistema:

- **Elementos Físicos do STER**: infra-estrutura, equipamentos, insumos dos equipamentos, recursos financeiros;

- **Elementos Lógicos do STER**: estruturas político-institucional, de planejamento, de gestão e controle, normativas, funcionais, de produção;

- **Atores do STER**: Cliente, Planejador, Gestor, Regulador, Controlador, Prestador de serviço, Provedor de infra-estrutura de transportes, dentre outros

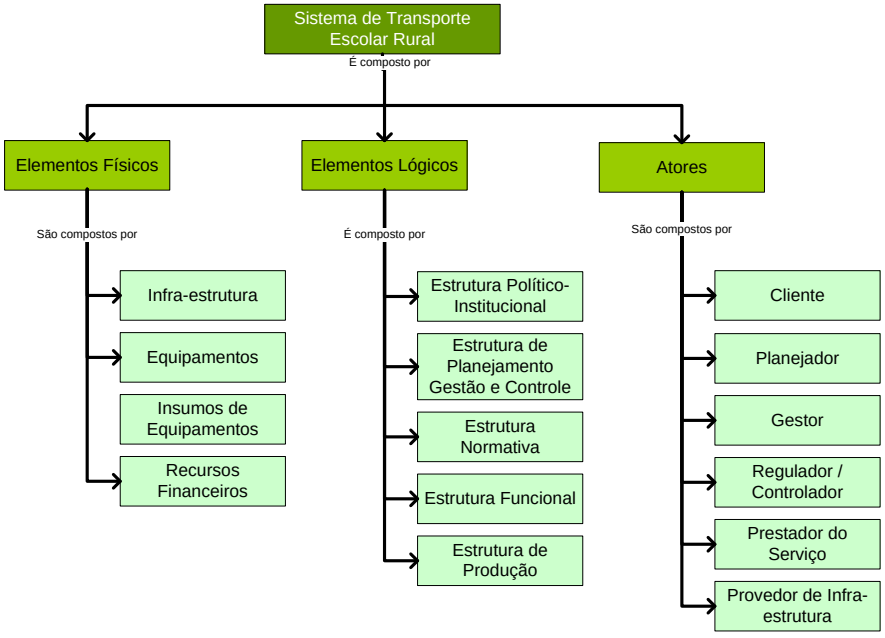


Figura 5.1 : Elementos do STER. Fonte (CEFTRU/FNDE, 2008a, 2008b).

Neste trabalho, define-se o ***Sistema de Transporte*** como um conjunto de elementos, atores, e atividades organizadas e inter-relacionadas que mutuamente se influenciam, e que permitem o deslocamento indispensável. Os atores exercem as atividades a partir dos elementos. Desta forma, a estruturação semântica para um sistema de transporte deve considerar elementos físicos e lógicos (CEFTRU/FNDE, 2008a, 2008b), conforme a Figura 5.1.

5.2.3 **Rede semântica do transporte escolar rural**

Conforme estudos desenvolvidos pelo CEFTRU/FNDE (2008a, 2008b), a respeito das redes semânticas (ver seção 5.2.1.2 - Redes semânticas) pertinentes ao transporte, a estrutura semântica do sistema de transportes organiza os elementos do sistema segundo as categorias, sendo: (i) Elementos lógicos; (ii) Elementos Físicos e (iii) atores, conforme a Figura 5.2. No anexo B, é apresentado à rede semântica do sistema de transporte rural na forma simplificada.

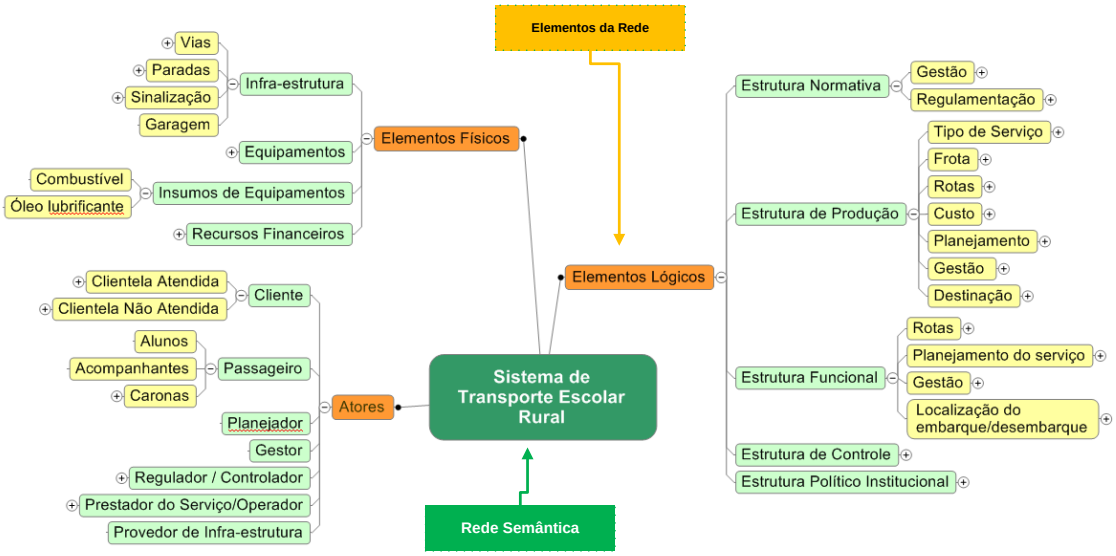


Figura 5.2 : Rede semântica do STER. Fonte (CEFTRU/FNDE, 2008a, 2008b)

Nesta classificação, a Figura 5.3 ilustra uma estrutura semântica do TER que pode ser organizada considerando-se ***elementos físicos e os elementos lógicos.*** A infra estrutura e os equipamentos são os elementos físicos (EF) e as estruturas normativa, funcional, de

produção, de gestão e político institucional os quais são os elementos lógicos (EL). (CEFTRU/FNDE, 2008a, 2008b).

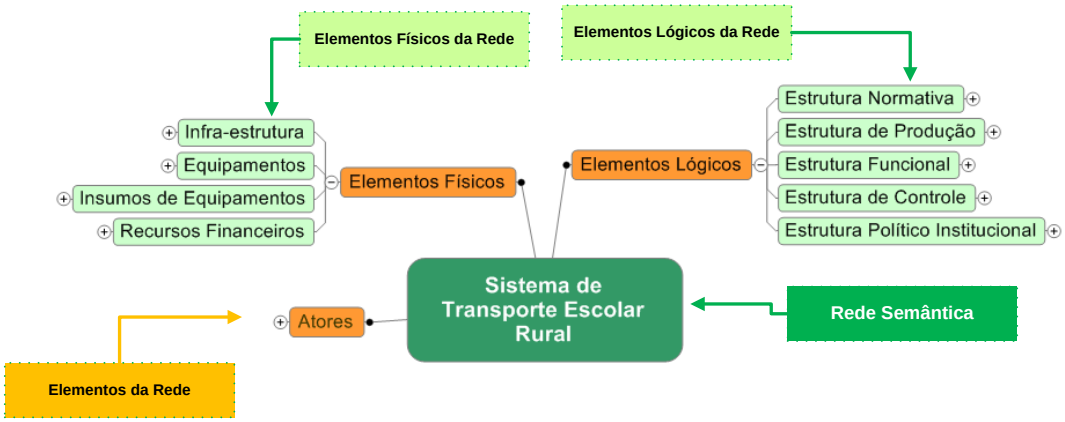


Figura 5.3 : Estrutura semântica do TER. Fonte (CEFTRU/FNDE, 2008a, 2008b)

Os Elementos Físicos e os Elementos Lógicos da Rede são compostos, individualmente, por seus Elementos da Rede Semântica (ERS), como mostra a Figura 5.4.

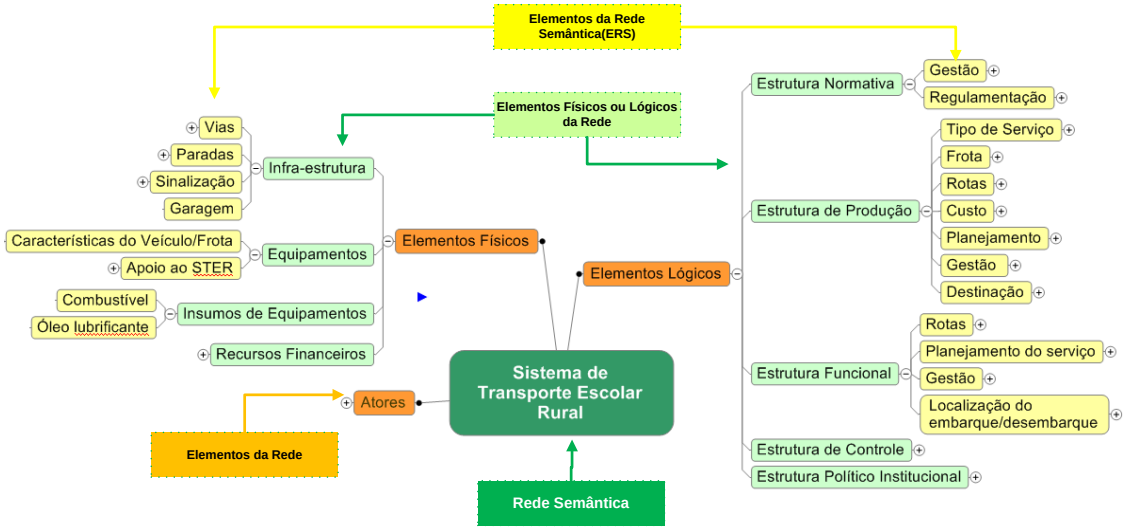


Figura 5.4 : Elementos da rede semântica. Fonte (CEFTRU/FNDE, 2008a, 2008b)

Cada Elemento da Rede Semântica (ERS) é composto por ***Elementos de Representação*** (ERep), que, conjuntamente, caracterizam estes elementos da rede semântica. A Figura 5.5 : Elementos de Representação. Fonte (CEFTRU/FNDE, 2008a, 2008b)Figura 5.5, enumera os ERep, relativos ao ERS - Regulamentação.

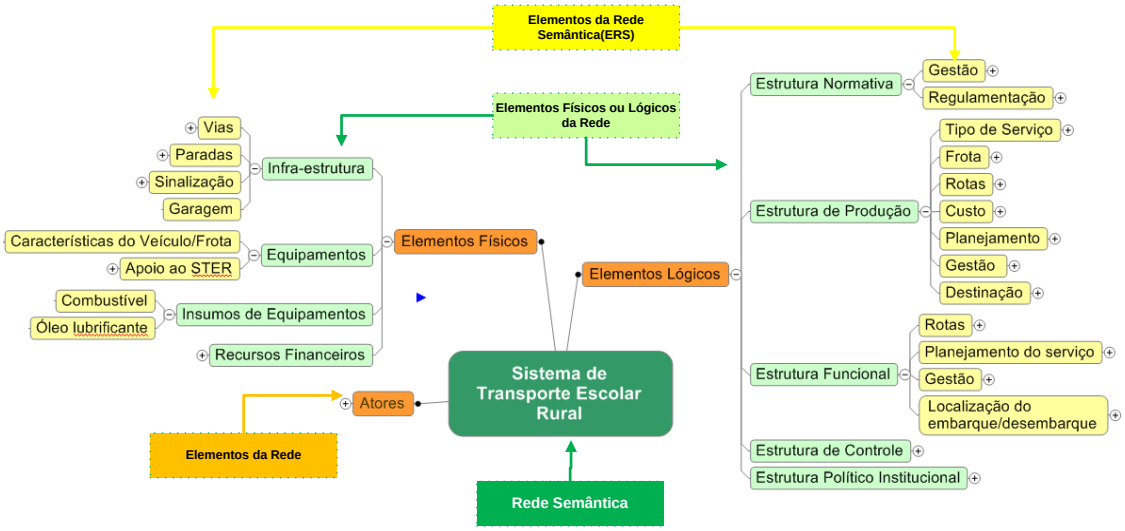


Figura 5.5 : Elementos de Representação. Fonte (CEFTRU/FNDE, 2008a, 2008b)

Assim, essa seção teve como objetivo apresentar a complexidade do mundo do transporte escolar rural. Dessa forma, verificam-se pela rede semântica, os elementos que devem ser conhecidos e dominados, para que o transporte escolar rural possa ser melhorado. Ao mesmo tempo, a rede semântica mostra a grande quantidade de dados/ informações necessárias para se trabalhar no planejamento do transporte escolar rural.

Na próxima seção, mostra-se o estado do transporte escolar rural, tendo como referência a pesquisa realizada pelo CEFTRU/FNDE (2007a, 2007b).

5.3 BASE DE DADOS UTILIZADA: TRANSPORTE ESCOLAR RURAL

Atualmente, poucos estudos sobre o transporte escolar rural foram realizados. O Fundo Nacional de Desenvolvimento da Educação - FNDE, na tentativa de realizar um diagnóstico em 2006, realizou uma pesquisa web com 5.564 municípios brasileiros. A pesquisa foi disponibilizada na internet em dezembro de 2006, permanecendo nesse canal até fevereiro de 2007. Dentre todos os municípios brasileiros, que se encontram aleatoriamente distribuídos pelos estados, somente 2.277 responderam, dentro do prazo, a todas as perguntas do formulário, conforme o apresentado na Tabela 5.1 (CEFTRU/FNDE, 2007a, 2007b). No anexo A, é apresentado o questionário na forma simplificada.

Tabela 5.1: Distribuição dos municípios respondentes por estado

UF	Mun	%	UF	Mun	%	UF	Mun	%
AC (22)	8	36	MG (853)	364	43	RN (167)	58	35
AL (102)	44	43	MS (78)	42	54	RO (52)	22	42
AM (62)	15	24	MT (141)	61	43	RS (496)	252	51
AP (16)	3	19	PA (143)	46	32	SC (293)	130	44
BA (417)	159	38	PB (223)	98	44	SE (75)	36	48
CE (184)	91	49	PE (185)	53	29	SP (645)	271	42
ES (78)	34	43	PI (223)	88	39	TO (139)	37	26
GO (246)	84	34	PR (399)	159	40			
MA (217)	66	30	RJ (92)	56	61			

Nota: os números entre parênteses referem-se ao total de municipios. Fonte: CEFTRU/FNDE (2007a, 2007b)

Essa configuração permite, com determinado grau de confiabilidade, que os resultados obtidos na amostra (2.277 municípios) sejam expandidos aos 5.564 municípios brasileiros. No entanto, a confiabilidade, acima referida, pode ser considerada inválida do ponto de vista estatístico, visto que a pesquisa teve caráter declaratório e voluntário, não tendo sido estabelecidos percentuais de amostragem para seu retorno. Mesmo assim, a amostragem obtida para essa pesquisa, permitiu a obtenção de uma visão ampla da realidade do transporte escolar no Brasil.

Como ressalva, as informações (dados) declaradas no questionário são de inteira responsabilidade dos gestores dos municípios que responderam o questionário. Dessa forma, é preciso considerar a possibilidade de incoerência dos dados fornecidos por parte de determinado município, gerando, assim, uma possível inconsistência nos resultados que, nesse caso, não serão frutos de falhas na análise dos resultados obtidos pela metodologia proposta.

Assim, essa pesquisa teve por objetivo caracterizar o transporte escolar rural, utilizando-se do processo de mineração de dados, para descobrir quais são os procedimentos adotados pelos municípios, para a gestão do transporte escolar, diante da realidade específica de cada município.

A seguir, de forma resumida, detalharam-se os resultados das análises estatísticas descritivas dos dados coletados no questionário web (CEFTRU/FNDE, 2007b). O objetivo dessa apresentação é mostrar como usualmente as análises são desenvolvidas, envolvendo na sua grande maioria das vezes, técnicas da estatística descritiva, organizando os dados em forma de tabelas e gráficos. O objetivo dessa dissertação, é exatamente apresentar uma

ferramenta em potencial para uma análise analítica, por meio de classificação, agrupamentos e associações dos dados.

5.3.1 **Apresentação resumida da caracterização do TER - relatório web**

A caracterização do transporte escolar foi realizada a partir das informações declaradas no questionário. Os dados foram preenchidos pelos responsáveis do setor de transporte escolar de cada município. As informações obtidas foram separadas em três grupos (i) - Serviço, (ii) - Clientela e (iii) Recursos, utilizando técnicas da estatística descritiva e não da análise analítica, por meio de classificação, de agrupamentos e de associações dos dados (CEFTRU/FNDE, 2007b).

5.3.1.1 Serviço

Dentre as informações declaradas por cada município, constituem-se como elementos principais do serviço do transporte escolar: a frota, a gestão e o acompanhamento; e a regulamentação (CEFTRU/FNDE, 2007b).

- A **frota** do transporte escolar, conforme a Figura 5.6-A, era composta principalmente, de quatro **tipos de veículos**: *ônibus, microônibus e vans; kombis e caminhonetes*. Esses quatro tipos de veículos somavam aproximadamente 90% do total da frota. Na Figura 5.6–B, aproximadamente 33% dos municípios utilizavam-se de veículos da própria prefeitura, enquanto 67% dos municípios utilizavam-se da **frota terceirizada** (*caminhonetes e caminhões*). Essas informações demonstraram certa tendência dos municípios a delegarem o serviço a terceiros. Na Figura 5.6–C, mostrou-se a **média da idade** dos veículos por região. Verificou-se que o nordeste concentra os veículos mais velhos e o sudeste os veículos mais novos. Na Figura 5.6–D, apresentou-se os diferentes **tipos de combustíveis** utilizados pelos veículos. Verificou-se que o diesel é o principal deles (69%), seguido da gasolina (29%).

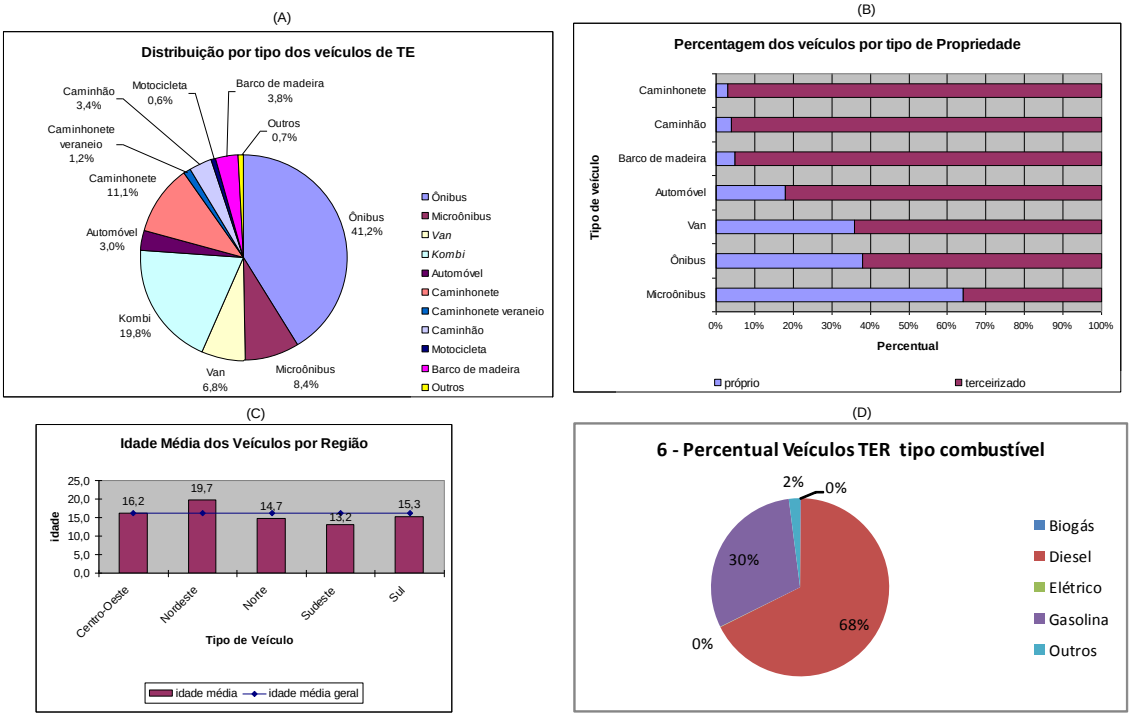


Figura 5.6: Gráficos da frota do TER Fonte: CEFTRU/FNDE (2007b)

- A **gestão e acompanhamento** – São considerados elementos utilizados pelo município para a gerência e o acompanhamento do serviço de transporte escolar, os quais procuram explorar e descrever a rotina do transporte escolar no município. Com base nos dados declarados no questionário, observou-se que (CEFTRU/FNDE, 2007b):

    a. **O critério** mais utilizado para remuneração dos serviços prestados, segundo os dados declarados, foi o de '*valor por quilômetro rodado*' e o menos utilizado foi o de '*valor por aluno*';

    b. **Na periodicidade do serviço de transporte**, cerca de 98% dos municípios oferecem o serviço durante todo período letivo;

    c. **Na utilização dos veículos no transporte escolar**, aproximadamente 27% dos municípios utilizam o veículo para outras finalidades, quando esse não está sendo usado para o transporte de alunos;

d. **No acompanhamento da rotina do transporte escolar**, cerca de 73% dos municípios, declararam ter algum tipo de acompanhamento no desenvolvimento da operação de transporte.

e. **Existem poucos veículos adaptados aos portadores de necessidades especiais**, aproximadamente 6% dos veículos utilizados são adaptados.

- **Regulamentação** – Com base nos dados declarados do questionário, aproximadamente 85% dos municípios, declararam não ter regulamentação própria para o transporte escolar. Essa informação corresponde à verificação da existência de regulamentação municipal específica do transporte escolar, com o objetivo de proporcionar ao FNDE subsídios para a proposição de mecanismos legais que regulamentem o repasse do Programa Nacional de Apoio ao Transporte Escolar – PNATE (CEFTRU/FNDE, 2007b).

5.3.1.2 Clientela

Esse item é formado de *clientela atendida,* composta pelas escolas e pelos usuários que utilizam o serviço de transporte escolar e a *clientela não atendida,* composta pelos usuários que não possuem o serviço, mas que dele necessitam. No entanto, por algum motivo, essa clientela não é atendida pelo serviço e, por isso, no caso de *clientela não atendid*a, não foi possível realizar nenhuma análise (CEFTRU/FNDE, 2007b).

No quesito clientela atendida, conforme a Figura 5.7-A, verificou-se que os municípios das regiões nordeste e norte possuíam a maior quantidade de escolas municipais atendidas pelo serviço de transporte. Na Figura 5.7-B, em que os dados são consolidados por região, foi possível inferir algumas discrepâncias que podem ser observadas na distribuição do número de alunos transportados, de acordo com a dependência administrativa e a sua área de residência (CEFTRU/FNDE, 2007b).

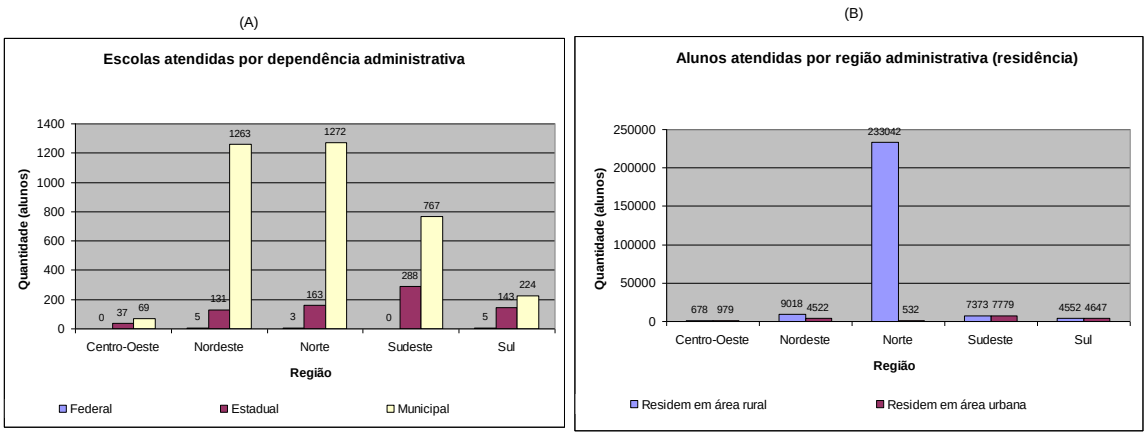


Figura 5.7: Gráficos da clientela atendida no TER. Fonte: CEFTRU/FNDE (2007b)

5.3.1.3 Recursos

Nesse item são apresentadas as informações declaradas pelos gestores dos municípios, com relação aos recursos aplicados na prestação do serviço de transporte escolar (CEFTRU/FNDE, 2007b).

Na **Fonte,** conforme a Figura 5.8-A, o montante gasto pelos municípios, aproximadamente 10%, vêm do PNATE; cerca de 16% dos recursos do estado são repassados ao município; 58% dos recursos são do próprio município e 16% são de outras fontes. Na Figura 5.8-B, verificou-se que os recursos próprios representaram a maior parte dos gastos em transporte escolar. As regiões norte e nordeste, dentre as demais, são as que mais se utilizam dos recursos do Pnate (CEFTRU/FNDE, 2007b).

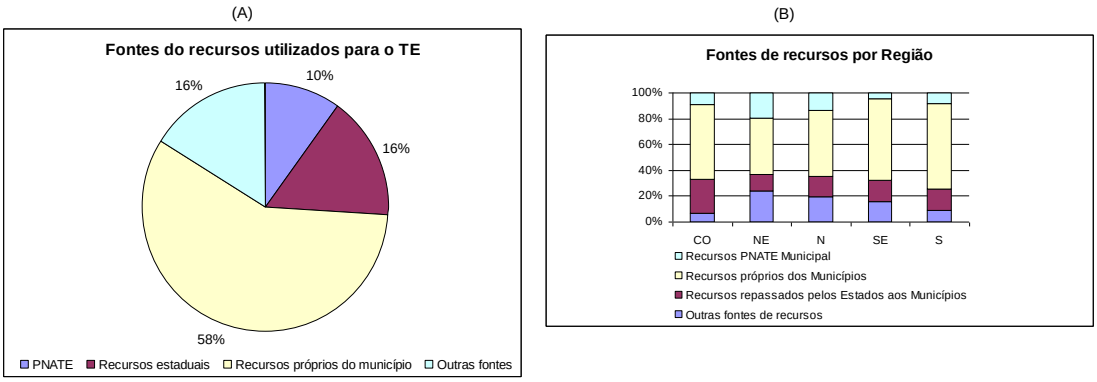


Figura 5.8: Gráficos das fontes de recursos no TER. Fonte: CEFTRU/FNDE (2007b)

Segundo Carvalho (2011), faz-se necessário conhecer melhor o TER oferecido aos alunos, e a partir desse conhecimento, propor ações que melhorem o seu atendimento. Neste

contexto, contribui para o desenvolvimento da educação, a permanência dos alunos na escola, a diminuição da evasão escolar, o auxílio da inclusão social dos alunos, a melhora do rendimento escolar e a perspectiva de futuro melhor, conforme ilustra a Figura 5.9.

Figura 5.9 : Conhecendo o TER. Fonte: adaptação de Carvalho (2011)

Dessa forma, por não ser o foco principal dessa dissertação, *a análise descritiva do objeto do estudo de caso*, contextualizou-se de forma simplificada o objeto. Na próxima seção, detalha-se a aplicação da metodologia que utiliza o *processo de mineração de dados*, em diversas etapas, para caracterizar o transporte escolar rural.

5.4 APLICAÇÃO DA METODOLOGIA PROPOSTA

A fim de simplificar a aplicação do estudo de caso, seu desenvolvimento se restringiu à caracterização do transporte escolar rural, utilizando-se da técnica de mineração de dados, para análise dos dados na pesquisa web, realizada pelo CEFTRU/FNDE (2007b). Cabe ainda destacar que, as informações, ou seja, os dados utilizados foram os declarados pelos gestores do transporte escolar rural nos municípios. A aplicação da metodologia proposta no capítulo anterior, será apresentada, conforme o que se vê na Figura 4.1.

5.4.1 **Etapa I – Concepção (requisitos)**

Atividade 01: Identificação do objeto de estudo

Neste trabalho, considerou-se como objeto de estudo o *transporte escolar rural – TER,* que é fornecido pelo poder público, para o uso exclusivo do aluno, o qual pode ser executado pela prefeitura ou por terceirizado. Sendo então, o sistema de transporte escolar rural como um conjunto de elementos inter-relacionados, que têm como objetivo comum deslocar os alunos que residam ou que estudam em área rural, entre seu local de residência e um estabelecimento de ensino (CEFTRU/FNDE, 2008a, 2008b).

Atividade 02: Definição da área de estudo (delimitação)

A área de estudo refere-se aos municípios brasileiros, onde atua o *transporte escolar rural – TER,* em todo território nacional. Para o estudo de caso, com o objetivo de caracterizar o transporte escolar rural, utilizou-se o processo de mineração de dados nos 2.277 municípios, que responderam integralmente o questionário na internet; tendo como filtro, os dados do ensino fundamental e da área de atuação rural (CEFTRU, 2007a, 2007b).

Atividade 03: Caracterização do Objeto de estudo

A caracterização adotada do objeto de estudo o *transporte escolar rural – TER*, nesta atividade, foi elaborada pelo CEFTRU/FNDE (2007a, 2007b). Consideraram-se os atores e os elementos envolvidos no processo de planejamento e na operação do serviço, para facilitar o entendimento do objeto, a ser analisado nesta pesquisa. Assim, essa estrutura semântica considerou uma caracterização de forma conjunta e complementar, por meio de três objetos (elementos): *serviço, clientela e recurso*s, conforme a Figura 5.10 (CEFTRU/FNDE, 2007a, 2007b).

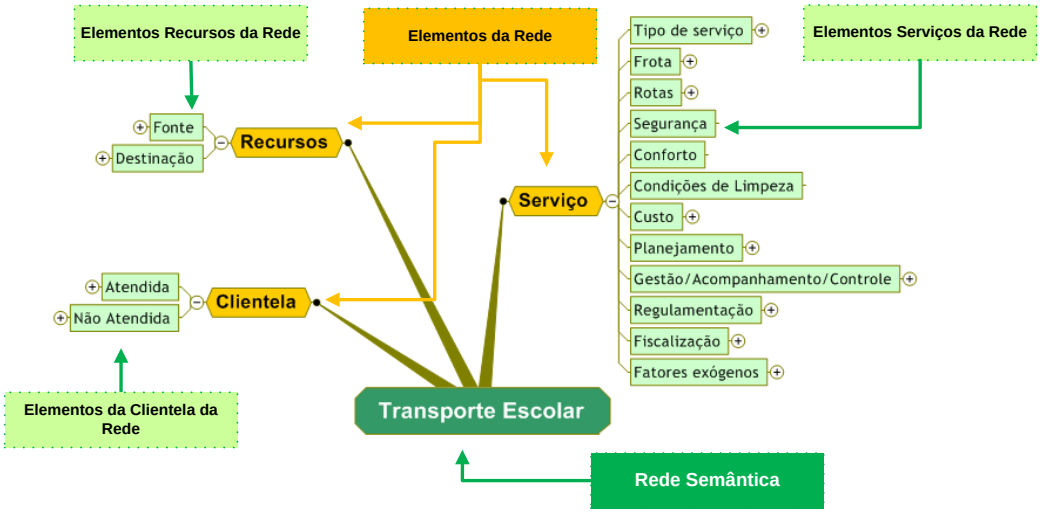


Figura 5.10 : Estrutura semântica do TER. Fonte: CEFTRU (2007a, 2007b)

Esses três objetos (elementos) são compostos, individualmente, pelos elementos da rede semântica (ERS). Cada elemento ERS é composto, por sua vez, pelos elementos de representação (ERep).

I – Serviço

Os principais elementos da rede semântica constituintes da caracterização do serviço resultaram nos seguintes dados do questionário: frota, custo, rotas (roteiros), gestão e acompanhamento; e regulamentação, conforme a Figura 5.11 (CEFTRU/FNDE, 2007a, 2007b).

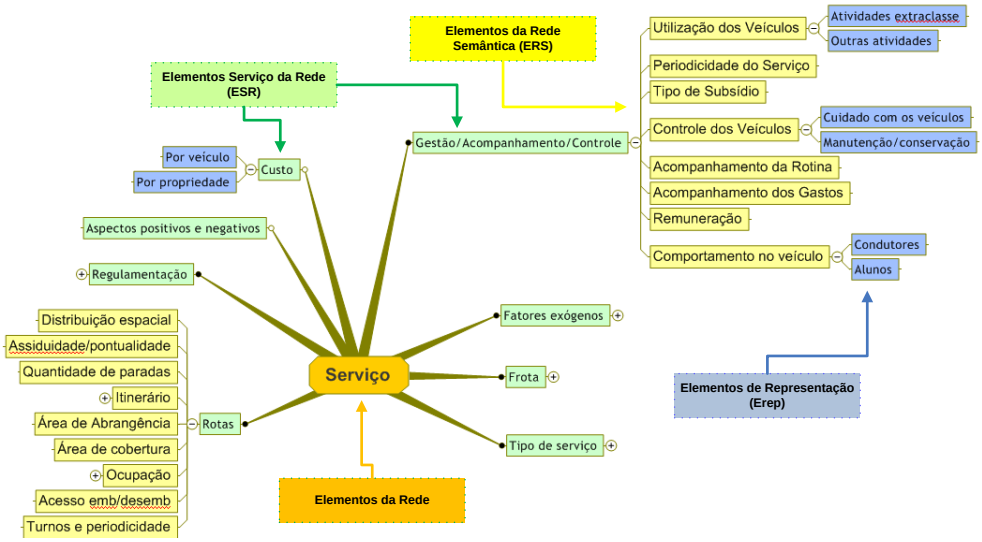


Figura 5.11 : Elementos da rede semântica–serviço. Fonte: CEFTRU (2007a, 2007b)

II – Clientela

A clientela representada pelas escolas e usuários (alunos, acompanhantes, professores, funcionários e outros), refere-se aos beneficiados do serviço de transporte escolar, fornecido pelo município, conforme a Figura 5.12 Figura 5.11 (CEFTRU/FNDE,2007a, 2007b).

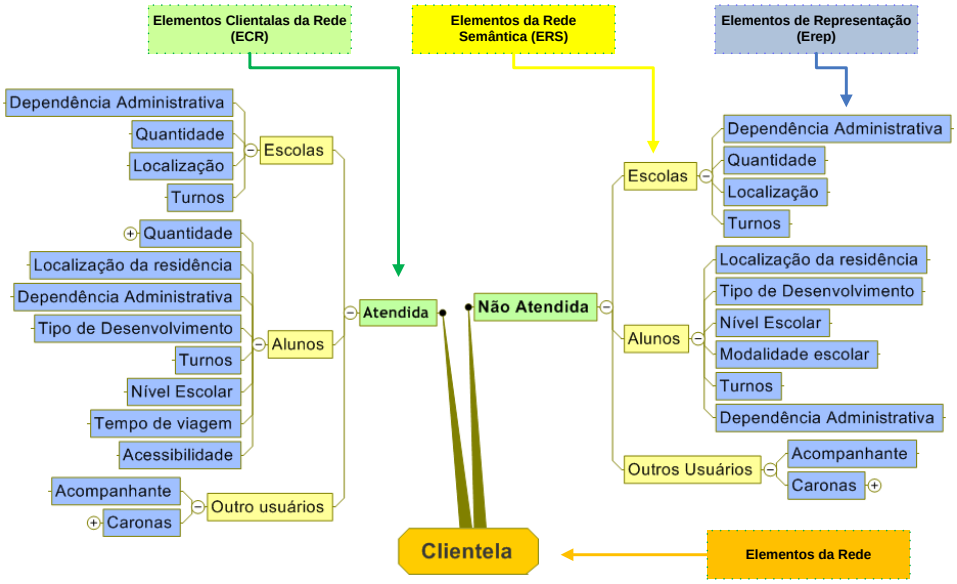


Figura 5.12 : Elementos da rede semântica–clientela. Fonte: CEFTRU (2007a, 2007b)

III – Recursos

Os recursos representam um condicionante, segundo o qual o sistema pode ser oferecido, ou seja, a existência desses recursos é fundamental para a implementação e a operação do transporte escolar rural. Esses recursos foram divididos em fontes e destinação, conforme a Figura 5.13 (CEFTRU/FNDE, 2007a, 2007b).

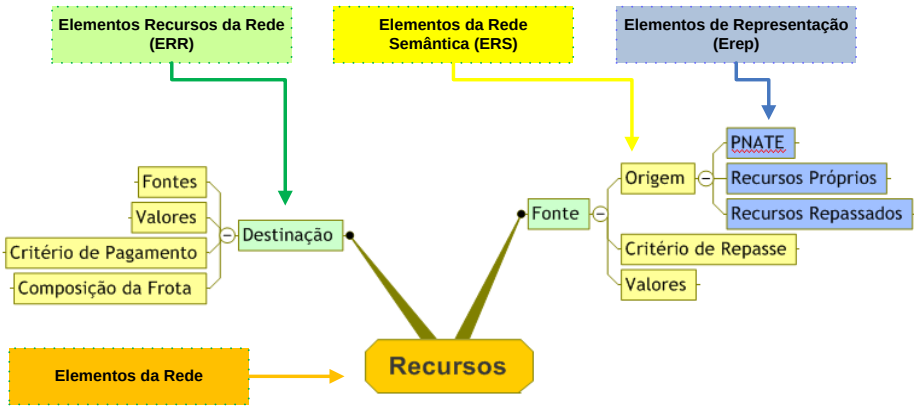


Figura 5.13 : Elementos da rede semântica – recursos . Fonte: CEFTU (2007a, 2007b)

Atividade 04: Identificação dos grupos-alvos e da necessidade de informação

Os atores cliente, planejador, gestor, regulador, prestador de serviço e provedor de infraestrutura, identificados nesta atividade, são de alguma forma interessados no assunto e constam na Figura 5.1.

Atividade 05: Definição do tipo de conhecimento que se deseja descobrir

Para este trabalho, o objetivo foi caracterizar o transporte escolar rural de forma conjunta e complementar, por meio de três objetos: *serviço, clientela e recursos* (CEFTRU/FNDE, 2007a, 2007b). Assim, o problema da mineração de dados considerado foi conhecer as práticas e os procedimentos adotados pelos municípios, para gestão do transporte escolar, diante da realidade específica de cada município.

5.4.2 **Etapa II – Elaboração (modelagem)**

Atividade 06: Definição da ferramenta de mineração e dos recursos disponíveis

Conforme a seção 3.4.9, existem várias ferramentas, e a sua definição se deu em função do seu custo. Assim, as ferramentas utilizadas nesta pesquisa foram:

- O banco de dados do questionário que foi gerenciado pelo *SGDB(sistema de gerenciamento de banco de dados) PostgreSQL 8.1*;
- A decisão de desenvolver a ferramenta *transporte mining* que é uma extensão de outra ferramenta chamada *WEKA*, com algumas implementações específicas das medidas de interesse objetiva. Foram disponibilizados outros fatores relevantes: o custo para a utilização desse *software livre* e a possibilidade de novas implementações. No apêndice A, é apresentado o manual na forma simplificada da ferramenta.

Atividade 07: Escolha da atividade

Nessa atividade, conforme a metodologia apresentada no capítulo 4, a escolha da atividade foi feita de acordo com os objetivos da descoberta de associações entre as práticas e os procedimentos adotados, para gestão do transporte escolar rural. Então, foi escolhida a *atividade descritiva* para identificar a correlação no conjunto de dados.

Atividade 08: Escolha do tipo de Tarefa

Conforme atividade 7, foi escolhida a *tarefa de associação*, para determinar quais seriam os fatos ou objetos que tenderiam a ocorrer juntos, isto é, no mesmo registro.

Atividade 09: Escolha do tipo de Técnica

Conforme a metodologia apresentada no capítulo 4, a técnica a ser utilizada foi a de *descoberta de regras de associação* com o algoritmo *apriori*, dentre as várias opções apresentadas.

Atividade 10: Mapeamento das variáveis e da seleção dos dados

I. Atividade de mapeamento das variáveis

A principal tarefa desempenhada nesta atividade foi o mapeamento das tabelas, cujos dados estavam alocados e familiarizados com a estrutura lógica do registro. Com o auxilio do especialista de domínio, foram identificadas as tabelas que armazenavam as variáveis a serem analisadas.

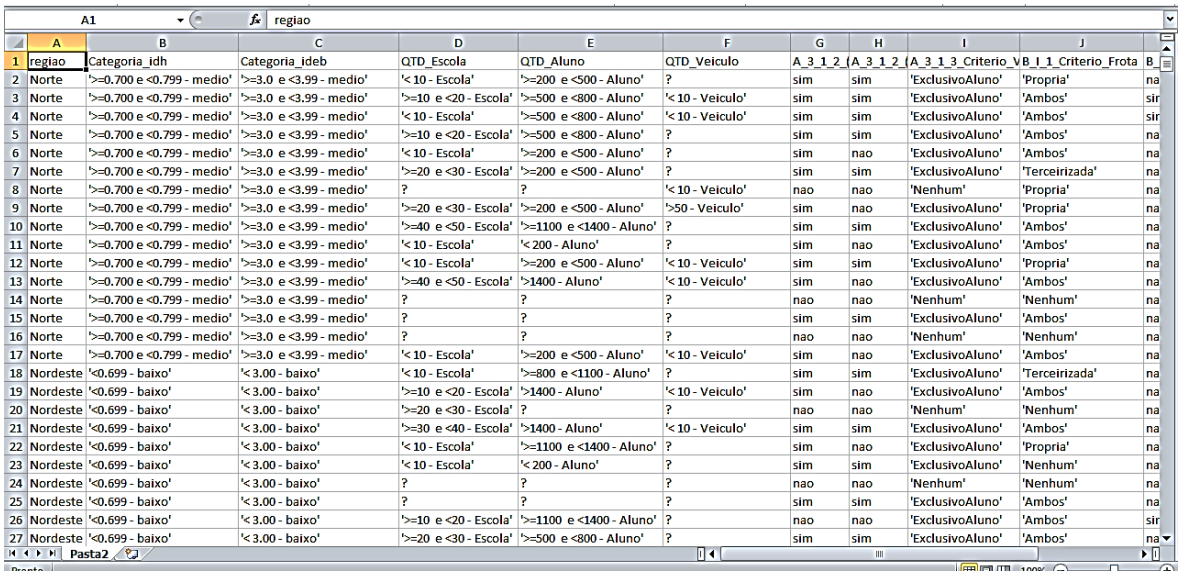
Para esse estudo foram identificadas as 14 variáveis mais relevantes. Conforme a Tabela 5.2, tomou-se como premissa três elementos fundamentais, para a caracterização do transporte escolar rural: *serviço, clientela e recursos,* conforme o objeto definido na atividade 3. Isso permitiu uma visão global das características dos elementos de representação do TER.

Tabela 5.2 : Variáveis identificas no mapeamento

Informação	Elemento	Descrição
1 - Quantidade Escola	clientela - caracterização	Quantidade de escola atendidos pelo serviço transporte escolar
2 - Quantidade alunos	clientela - caracterização	Quantidade de aluno atendidos pelo serviço transporte escolar
3 - Quantidade de veículos	servico - catacterização	Quantidade de veículos utilizado no transporte escolar
4 - Veículos exclusivo	serviço - caracterização	Transporte escolar com veículos exclusivo fornecido pelo município - aluno ou passageiros
5 - Propriedade veículo	serviço - caracterização	A propriedade do veiculo utilizado no transporte escolar - própria ou terceirizada
6 -Fontes de recursos	recursos - fontes	Fontes de recursos utilizados para financiar o transporte escolar fornecido pelo munícipio
7 - Critério de pagamento	recursos - destinação	Critérios utilizados para pagamento do serviço de transporte escolar terceirizado
8 - Alunos atendidos ensino Fundemental	Identificação	Aluno atendidos com transporte escolar com desenvolvimento típico
9 - Alunos atend. TER necess. especiais	Identificação	Existe alunos com necessidades educacionais especiais que utilizada o serviço de transporte
10 - Veículo utilizado outros fins	serviço - caracterização	Veículo exclusivo para transporte escolar está sendo utilizado para outros fins
11 - Servico oferecido todo período letivo	serviço - caracterização	Transporte escolar é oferecido durante todo o período letivo
12 - Veículo adaptado TER	serviço - caracterização	Veículo é adaptado para o transporte escolar
13 - Existe acomp. rotina TER	serviço - caracterização	Existe algum tipo de acompanhamento de rotina do serviço de transporte escolar
14 - Existe regulamento municipal TER	serviço - regulamentação	Existe regulamentação municipal para o transporte escolar no seu municipio

II. Seleção dos Dados

A seleção e a extração dos dados foram realizadas por meio da *Linguagem de Consulta Estruturada – SQL*, no banco de dados, para se obter as informações necessárias para a geração de uma *planilha do excel*, com os dados selecionados. O processo de análise da qualidade dos dados foi realizado nas compilações das *consultas SQL*. A Figura 5.14, mostra uma parte desse arquivo.



	A	B	C	D	E	F	G	H	I	J	
1	regiao	Categoria_idh	Categoria_ideb	QTD_Escola	QTD_Aluno	QTD_Veiculo	A_3_1_2_	A_3_1_2_	A_3_1_3_Criterio_V	B_I_1_Criterio_Frota	B_
2	Norte	'>=0.700 e <0.799 - medio'	'>=3.0 e <3.99 - medio'	'< 10 - Escola'	'>=200 e <500 - Aluno'	?	sim	sim	'ExclusivoAluno'	'Propria'	na
3	Norte	'>=0.700 e <0.799 - medio'	'>=3.0 e <3.99 - medio'	'>=10 e <20 - Escola'	'>=500 e <800 - Aluno'	'< 10 - Veiculo'	sim	sim	'ExclusivoAluno'	'Ambos'	sir
4	Norte	'>=0.700 e <0.799 - medio'	'>=3.0 e <3.99 - medio'	'< 10 - Escola'	'>=500 e <800 - Aluno'	'< 10 - Veiculo'	sim	sim	'ExclusivoAluno'	'Ambos'	sir
5	Norte	'>=0.700 e <0.799 - medio'	'>=3.0 e <3.99 - medio'	'>=10 e <20 - Escola'	'>=500 e <800 - Aluno'	?	sim	sim	'ExclusivoAluno'	'Ambos'	na
6	Norte	'>=0.700 e <0.799 - medio'	'>=3.0 e <3.99 - medio'	'< 10 - Escola'	'>=200 e <500 - Aluno'	?	sim	nao	'ExclusivoAluno'	'Ambos'	na
7	Norte	'>=0.700 e <0.799 - medio'	'>=3.0 e <3.99 - medio'	'>=20 e <30 - Escola'	'>=200 e <500 - Aluno'	?	sim	sim	'ExclusivoAluno'	'Terceirizada'	na
8	Norte	'>=0.700 e <0.799 - medio'	'>=3.0 e <3.99 - medio'	?	?	'< 10 - Veiculo'	nao	nao	'Nenhum'	'Propria'	na
9	Norte	'>=0.700 e <0.799 - medio'	'>=3.0 e <3.99 - medio'	'>=20 e <30 - Escola'	'>=200 e <500 - Aluno'	'>50 - Veiculo'	sim	nao	'ExclusivoAluno'	'Propria'	na
10	Norte	'>=0.700 e <0.799 - medio'	'>=3.0 e <3.99 - medio'	'>=40 e <50 - Escola'	'>=1100 e <1400 - Aluno'	?	sim	sim	'ExclusivoAluno'	'Ambos'	na
11	Norte	'>=0.700 e <0.799 - medio'	'>=3.0 e <3.99 - medio'	'< 10 - Escola'	'< 200 - Aluno'	?	sim	nao	'ExclusivoAluno'	'Ambos'	na
12	Norte	'>=0.700 e <0.799 - medio'	'>=3.0 e <3.99 - medio'	'< 10 - Escola'	'>=200 e <500 - Aluno'	'< 10 - Veiculo'	sim	sim	'ExclusivoAluno'	'Propria'	na
13	Norte	'>=0.700 e <0.799 - medio'	'>=3.0 e <3.99 - medio'	'>=40 e <50 - Escola'	'>1400 - Aluno'	'< 10 - Veiculo'	sim	nao	'ExclusivoAluno'	'Ambos'	na
14	Norte	'>=0.700 e <0.799 - medio'	'>=3.0 e <3.99 - medio'	?	?	?	nao	nao	'Nenhum'	'Nenhum'	na
15	Norte	'>=0.700 e <0.799 - medio'	'>=3.0 e <3.99 - medio'	?	?	?	sim	sim	'ExclusivoAluno'	'Ambos'	na
16	Norte	'>=0.700 e <0.799 - medio'	'>=3.0 e <3.99 - medio'	?	?	?	nao	nao	'Nenhum'	'Nenhum'	na
17	Norte	'>=0.700 e <0.799 - medio'	'>=3.0 e <3.99 - medio'	'< 10 - Escola'	'>=200 e <500 - Aluno'	'< 10 - Veiculo'	sim	nao	'ExclusivoAluno'	'Ambos'	na
18	Nordeste	'<0.699 - baixo'	'< 3.00 - baixo'	'< 10 - Escola'	'>=800 e <1100 - Aluno'	?	sim	sim	'ExclusivoAluno'	'Terceirizada'	na
19	Nordeste	'<0.699 - baixo'	'< 3.00 - baixo'	'>=10 e <20 - Escola'	'>1400 - Aluno'	'< 10 - Veiculo'	sim	nao	'ExclusivoAluno'	'Ambos'	na
20	Nordeste	'<0.699 - baixo'	'< 3.00 - baixo'	'>=20 e <30 - Escola'	?	?	nao	nao	'Nenhum'	'Nenhum'	na
21	Nordeste	'<0.699 - baixo'	'< 3.00 - baixo'	'>=30 e <40 - Escola'	'>1400 - Aluno'	'< 10 - Veiculo'	sim	sim	'ExclusivoAluno'	'Ambos'	na
22	Nordeste	'<0.699 - baixo'	'< 3.00 - baixo'	'< 10 - Escola'	'>=1100 e <1400 - Aluno'	?	sim	nao	'ExclusivoAluno'	'Propria'	na
23	Nordeste	'<0.699 - baixo'	'< 3.00 - baixo'	'< 10 - Escola'	'< 200 - Aluno'	?	sim	sim	'ExclusivoAluno'	'Nenhum'	na
24	Nordeste	'<0.699 - baixo'	'< 3.00 - baixo'	?	?	?	nao	nao	'Nenhum'	'Nenhum'	na
25	Nordeste	'<0.699 - baixo'	'< 3.00 - baixo'	?	?	?	sim	sim	'ExclusivoAluno'	'Ambos'	na
26	Nordeste	'<0.699 - baixo'	'< 3.00 - baixo'	'>=10 e <20 - Escola'	'>=1100 e <1400 - Aluno'	?	nao	nao	'ExclusivoAluno'	'Ambos'	sir
27	Nordeste	'<0.699 - baixo'	'< 3.00 - baixo'	'>=20 e <30 - Escola'	'>=500 e <800 - Aluno'	?	sim	sim	'ExclusivoAluno'	'Ambos'	na

Figura 5.14 : Base de dados na planilha *excel* a partir da consulta *SQL*

Atividade 11: Limpeza e pré-processamento dos dados

Na atividade anterior, durante a extração dos dados, notou-se a necessidade de limpeza dos dados com operações básicas de *eliminação dos valores ausentes*. A principal atividade dessa fase, aplicada na planilha com os dados selecionados, foi o de tratamento dos *campos em branco*, onde os mesmos foram *preenchidos com '?';* para que a ferramenta pudesse interpretar como ausência de informação no campo do registro, conforme ilustra a Figura 5.14.

Outro problema foi o tratamento dos registros com *valores nulos*, no qual foi utilizada a técnica de *eliminação do registro*, por meio do uso da cláusula *is not null* na própria instrução da planilha. Isso se fez necessário por falta de regra de negócio no sistema FNDE, cujos registros não possuíam os atributos preenchidos ou nulos.

Atividade 12: Transformação dos dados selecionados

Como o algoritmo, *apriori*, aceita somente campos nominais, isto é, não trabalha com campos de valores quantitativos, houve a necessidade de utilização da *técnica de discretização* em alguns campos, usando-se apenas dos filtros nas *instruções SQL*.

Como ressalva, foi incluída na base de dados, a variável '*índice do desenvolvimento da educação básica – IDEB'*, referente aos dados do ano 2005, esse indicador utilizou-se da primeira medição de dados, que foram levantados em 2005 (MEC/INEP, 2007). A sua utilização teve como objetivo aumentar o valor semântico das informações, utilizadas no processo de mineração dos dados.

De acordo com MEC/INEP (2007), o *IDEB* foi criado em 2007, para medir a qualidade do ensino no Brasil, numa escala de zero a dez. O indicador é calculado a partir de dois componentes básicos: (i) as taxas de rendimentos escolares (aprovação e evasão) são obtidas a partir do censo escolar realizado anualmente pelo Instituto Nacional de Estudos e Pesquisas Educacionais Anísio Teixeira – INEP (2009); e (ii) no desempenho dos alunos no SAEB (*Sistema Nacional de Avaliação da Educação Básica)* e na avaliação da Prova Brasil.

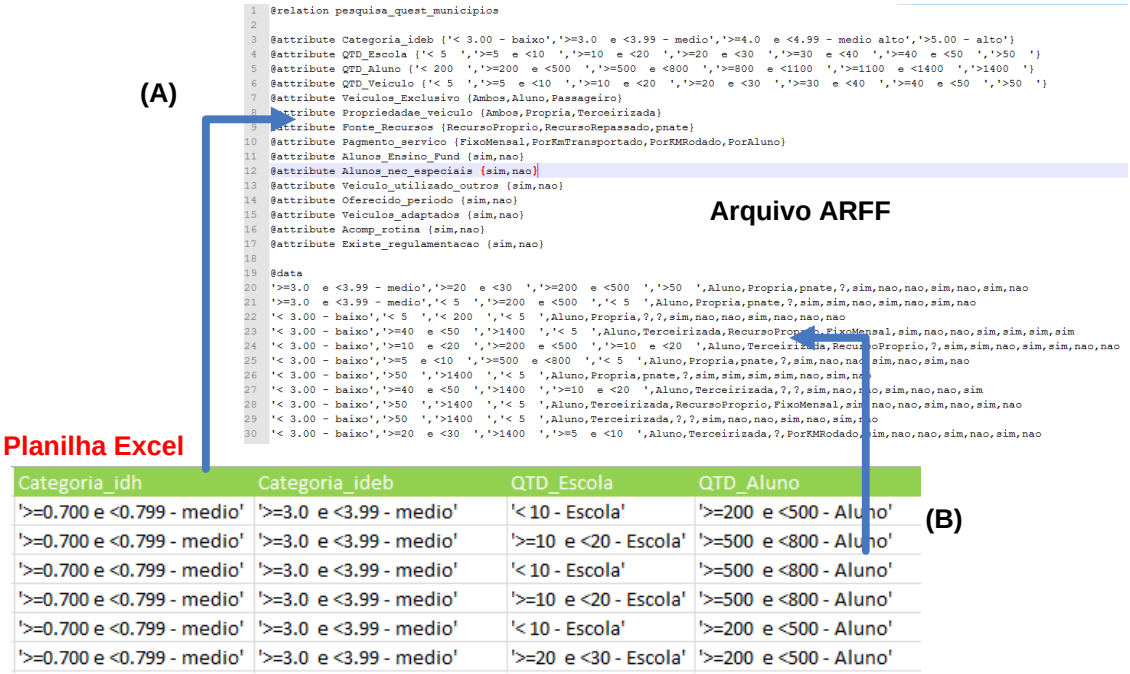
Assim, as transformações foram realizadas nos campos: *'quantidade de escola'; 'quantidade de aluno'; 'quantidade de veículo'; 'índice do IDEB'; 'veículos exclusivos para alunos'; 'fontes de recursos'; 'critério de pagamento';* dentre outros, conforme a Tabela 5.3.

Tabela 5.3: Transformação de diversos campos

	Total de registros		100%
1	cateoria IDEB	< 3.00 - baixo	20.96%
		>=3.0 e <3.99 - medio	70.51%
		>5.00 - alto	0.00%
2	Quantidade Escola	< 5	12.91%
		>=5 e <10	27.91%
		>=10 e <20	35.25%
		>=20 e <30	13.14%
		>=30 e <40	5.22%
		>=40 e <50	2.93%
		>50	2.63%
3	Quantidade Alunos	< 200	8.06%
		>=200 e <500	18.47%
		>=500 e <800	18.66%
		>=800 e <1100	15.08%
		>=1100 e <1400	9.66%
		>1400	30.08%
4	Quantidade veiculos	< 5	62.59%
		>=5 e <10	21.44%
		>=10 e <20	10.36%
		>=20 e <30	3.30%
		>=30 e <40	0.44%
		>=40 e <50	0.67%
		>50	1.20%

	Total de registros		100%
5	Veículos Exclusívo	alunos	90.70%
		Alunos e passageiros	7.13%
6	Propriedade veículo	Frota própria	49.68%
		Frota terceirizada	50.30%
7	Fontes recursos	Recurso Próprio	30.00%
		Recurso Repassado	26.15%
		Pnate	34.18%
8	Critério pagamento servico TER terceirizado	Valor fixo mensal	16.87%
		Por Km transportado	25.48%
		Por KM rodado	32.90%
		Por aluno	5.65%
9	Alunos atendidos TER Fundemental	sim	87.02%
		não	12.98%
10	Alunos atend. TER necess. especiais	sim	73.07%
		não	26.93%
11	Veículo TER é utilizado outros fins	sim	28.38%
		não	71.62%
12	Servico oferecido todo período letivo	sim	99.61%
		não	0.39%
13	Veiculo adaptado TER	sim	9.79%
		não	90.21%
14	Existe acomp. rotina TER	sim	79.86%
		não	20.14%
15	Existe regulamento municipal TER	sim	19.30%
		não	80.70%

Por meio do *excel*, foi possível gerar o arquivo *ARFF (ver seção 3.4.9.1).* Conforme a Figura 5.15- A, as *células-colunas (cor verde)* dos títulos do excel foram transformadas em *colunas* do arquivo *ARFF* . Na Figura 5.15- B, os valores das *células-linhas (cor branca)* do excel foram transformados em *dados* do arquivo *ARFF*. Nesse arquivo, os campos declarados pela *TAG @atribute* devem corresponder fielmente aos campos da *TAG @data*, pois a ferramenta é sensível a este formato.




```
@relation pesquisa_quest_municipios

@attribute Categoria_ideb {'< 3.00 - baixo','>=3.0  e <3.99 - medio','>=4.0  e <4.99 - medio alto','>5.00 - alto'}
@attribute QTD_Escola {'< 5  ','>=5  e <10  ','>=10  e <20  ','>=20  e <30  ','>=30  e <40  ','>=40  e <50  ','>50  '}
@attribute QTD_Aluno {'< 200  ','>=200  e <500  ','>=500  e <800  ','>=800  e <1100  ','>=1100  e <1400  ','>1400  '}
@attribute QTD_Veiculo {'< 5  ','>=5  e <10  ','>=10  e <20  ','>=20  e <30  ','>=30  e <40  ','>=40  e <50  ','>50  '}
@attribute Veiculos_Exclusivo {Ambos,Aluno,Passageiro}
@attribute Propriedadae_veiculo {Ambos,Propria,Terceirizada}
@attribute Fonte_Recursos {RecursoProprio,RecursoRepassado,pnate}
@attribute Pagmento_servico {FixoMensal,PorKmTransportado,PorKMRodado,PorAluno}
@attribute Alunos_Ensino_Fund {sim,nao}
@attribute Alunos_nec_especiais {sim,nao}
@attribute Veiculo_utilizado_outros {sim,nao}
@attribute Oferecido_periodo {sim,nao}
@attribute Veiculos_adaptados {sim,nao}
@attribute Acomp_rotina {sim,nao}
@attribute Existe_regulamentacao {sim,nao}

@data
'>=3.0  e <3.99 - medio','>=20  e <30  ','>=200  e <500  ','>50  ',Aluno,Propria,pnate,?,sim,nao,nao,sim,nao,sim,nao
'>=3.0  e <3.99 - medio','< 5  ','>=200  e <500  ','< 5  ',Aluno,Propria,pnate,?,sim,sim,nao,sim,nao,sim,nao
'< 3.00 - baixo','< 5  ','< 200  ','< 5  ',Aluno,Propria,?,?,sim,nao,nao,sim,nao,nao,nao
'< 3.00 - baixo','>=40  e <50  ','>1400  ','< 5  ',Aluno,Terceirizada,RecursoProprio,FixoMensal,sim,nao,nao,sim,sim,sim,sim
'< 3.00 - baixo','>=10  e <20  ','>=200  e <500  ','>=10  e <20  ',Aluno,Terceirizada,RecursoProprio,?,sim,sim,nao,sim,sim,nao,nao
'< 3.00 - baixo','>=5  e <10  ','>=500  e <800  ','< 5  ',Aluno,Propria,pnate,?,sim,nao,nao,sim,nao,sim,nao
'< 3.00 - baixo','>50  ','>1400  ','< 5  ',Aluno,Propria,pnate,?,sim,sim,sim,sim,nao,sim,nao
'< 3.00 - baixo','>=40  e <50  ','>1400  ','>=10  e <20  ',Aluno,Terceirizada,?,?,sim,nao,nao,sim,nao,nao,sim
'< 3.00 - baixo','>50  ','>1400  ','< 5  ',Aluno,Terceirizada,RecursoProprio,FixoMensal,sim,nao,nao,sim,nao,sim,nao
'< 3.00 - baixo','>50  ','>1400  ','< 5  ',Aluno,Terceirizada,?,?,sim,nao,nao,sim,nao,sim,nao
'< 3.00 - baixo','>=20  e <30  ','>1400  ','>=5  e <10  ',Aluno,Terceirizada,?,PorKMRodado,sim,nao,nao,sim,nao,sim,nao
```

Categoria_idh	Categoria_ideb	QTD_Escola	QTD_Aluno
'>=0.700 e <0.799 - medio'	'>=3.0 e <3.99 - medio'	'< 10 - Escola'	'>=200 e <500 - Aluno'
'>=0.700 e <0.799 - medio'	'>=3.0 e <3.99 - medio'	'>=10 e <20 - Escola'	'>=500 e <800 - Aluno'
'>=0.700 e <0.799 - medio'	'>=3.0 e <3.99 - medio'	'< 10 - Escola'	'>=500 e <800 - Aluno'
'>=0.700 e <0.799 - medio'	'>=3.0 e <3.99 - medio'	'>=10 e <20 - Escola'	'>=500 e <800 - Aluno'
'>=0.700 e <0.799 - medio'	'>=3.0 e <3.99 - medio'	'< 10 - Escola'	'>=200 e <500 - Aluno'
'>=0.700 e <0.799 - medio'	'>=3.0 e <3.99 - medio'	'>=20 e <30 - Escola'	'>=200 e <500 - Aluno'

Figura 5.15 : Arquivo gerado no formato *ARFF*

NOTA- A ferramenta escolhida na atividade 6, trabalha com um formato próprio, o arquivo com extensão *ARFF*, o qual tem de descrever o domínio do atributo. Este é um arquivo *ASCII* usado para definir atributos e seus valores: '*atributo=valor*'.

Atividade 13: Integração dos dados

Nesta atividade não foi executada nenhuma integração de dados, porque existe uma única fonte de dados.

5.4.3 **Etapa III – Construção (implementação)**

Atividade 14: Mineração de dados

I. Teste e Ajuste

Conforme o mencionado na atividade 9, foi utilizado o algoritmo *apriori*, com algumas modificações na estrutura lógica, para a geração da regra de associação. Essas alterações foram para que o algoritmo implementasse a especificação das *medidas de interesse*

objetiva, *suporte*, *confiança* e *lift*. Essa implementação teve por objetivo reduzir significativamente a chance de que fosse minerado um número excessivo de regras de associações óbvias.

A Figura 5.16-A, apresentam-se os resultados de vários testes com alternâncias do intervalo de *confiança*, *suporte e lift*, com o propósito de identificar os valores ideais dos parâmetros mínimos dessas medidas, para gerar somente as regras interessantes. Para cada análise diferente, esses parâmetros de entrada do algoritmo receberam valores diferentes, para que um resultado satisfatório fosse alcançado. Na Figura 5.16-B, mostra os parâmetros mínimos que foram identificados com ajuda do especialista de negócio, esses parâmetros foram repassados para o algoritmo como parâmetros de entrada. Assim, as regras de descobertas nesse procedimento possuem os valores de *lift*, *confiança* e *suporte* maior ou igual aos parâmetros mínimos especificados.

(A)

Modelo confiança e suporte		
Medida Objetivas	**Suporte**	**N. Regras**
confiança [20;40[	*suporte [10;20[*	133353
	suporte [20;40[	15192
	suporte [40;60[	0
	suporte [60;80[	0
	suporte [80;100[	0
confiança [40;60[	*suporte [10;20[*	84143
	suporte [20;40[	20198
	suporte [40;60[	1260
	suporte [60;80[	0
	suporte [80;100[	0
confiança [60;80[	*suporte [10;20[*	66064
	suporte [20;40[	13733
	suporte [40;60[	1959
	suporte [60;80[	131
	suporte [80;100[	0
confiança [80;100[	*suporte [10;20[*	43457
	suporte [20;40[	9153
	suporte [40;60[	1519
	suporte [60;80[	278
	suporte [80;100[	19

Modelo confiança, suporte e lift			
Medidas Objetivas		**Suporte**	**N. Regras**
Lift > 1.1	*confiança [20;40[*	*suporte[10;20[*	365
		suporte [20;40[	0
		suporte [40;60[	0
		suporte [60;80[	0
		suporte [80;100[	0
	confiança [40;60[	*suporte [10;20[*	7977
		suporte [20;40[	404
		suporte [40;60[	0
		suporte [60;80[	0
		suporte [80;100[	0
	confiança [60;80[	*suporte [10;20[*	6834
		suporte [20;40[	**361**
		suporte [40;60[	**6**
		suporte [60;80[	0
		suporte [80;100[	0
	confiança [80;100[	*suporte [10;20[*	1544
		suporte [20;40[	**9**
		suporte [40;60[	0
		suporte [60;80[	0
		suporte [80;100[	0

(B)

Intervalo Escolhido			
Medidas de Interesse Objetiva			N. Regras
lift > 1.1	***conf [60;100[***	***sup [30;60[***	45

Figura 5.16 : Resultados dos testes na base de dados da pesquisa web

A Figura 5.17, mostra a tela de parâmetro do software: ‘*suporte mínimo’* de 30% e ‘*suporte máximo’ de 60%; ‘confiança mínima’ de 60% e ‘confiança máxima’ de 100%; e ‘lift mínimo’ de 1,10*, com parâmetros de corte das regras não interessantes ou óbvias.

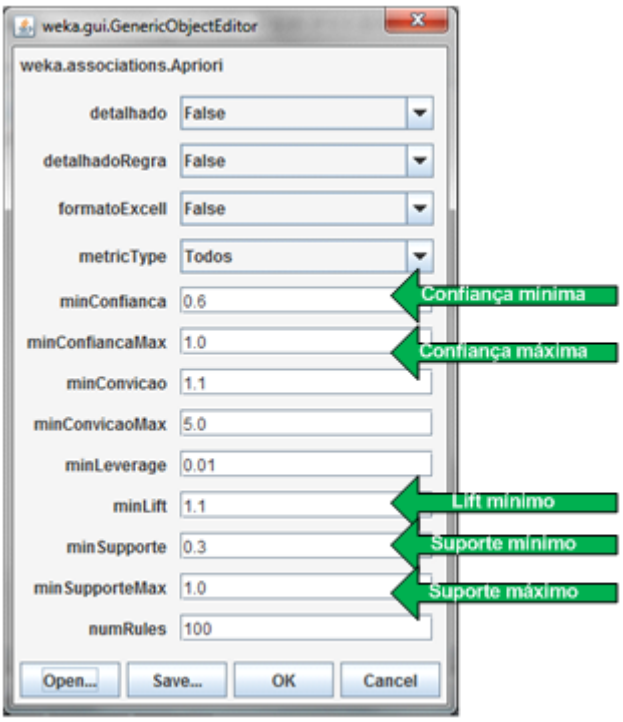


Figura 5.17 : Tela de parâmetro de entrada do algoritmo *apriori*

I. Aplicabilidade

Como resultado, o programa apresentou um *relatório de saída* com 45 regras de associação, geradas a partir da base de dados, conforme a Figura 5.18. Esse número pequeno de regras é explicado pela utilização da ‘*lift’; ‘confiança’* e *‘suporte’;* como parâmetros de entrada, para se evitar a geração de regras envolvendo *itens independentes* e *itens com dependência negativa*, conforme explicado na seção 0.

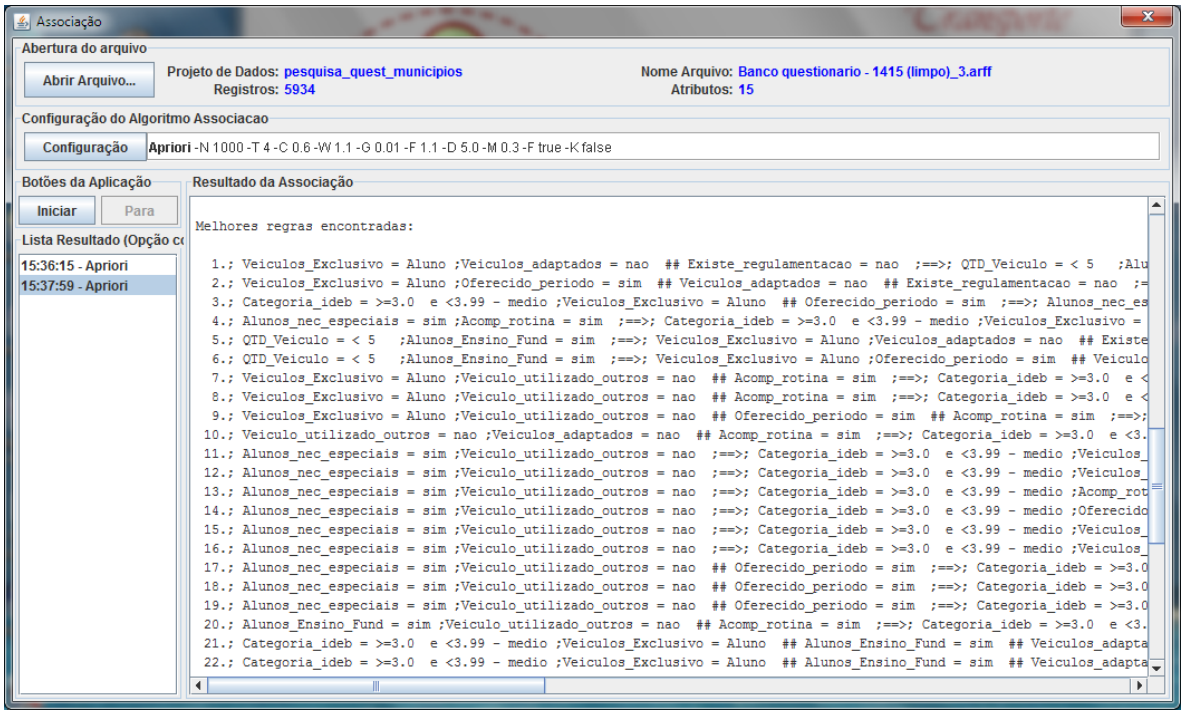


Figura 5.18 : Tela de saída do algoritmo *apriori*

Como ressalva, a quantidade de regras geradas depende da *quantidade de registros*, da *quantidade de atributos* e dos *parâmetros mínimos especificados*. Na grande maioria das vezes, torna-se inviável a interpretação de todas as regras geradas, para obtenção de um conhecimento útil e compreensível, que possa auxiliar no processo de tomada de decisão.

Atividade 15: Analise e avaliação dos padrões

Esta atividade não foi executada neste trabalho, uma vez que a atividade selecionada na atividade 7 foi a *descritiva.*

5.4.4 **Etapa IV – Transição (interpretação)**

Atividade 16: Seleção dos melhores padrões

As regras de associação geradas pela ferramenta de mineração de dados, foram transferidas para um *arquivo do excel,* a fim de facilitar a análise e descobrir as regras mais significativas, de acordo com cada medida de interesse, conforme Tabela 5.4.

Tabela 5.4 : Resultado após execução do algoritmo *apriori*

	Conjunto A			Conjunto B		Medidas de Interesse		
N	Item A	Item B		Item C	Item D	sup(Y):	conf:(Y)	Lift:(Y)
1	Veiculos_Exclusivo = Aluno	Veiculos_adaptados = nao , Existe_regulamentacao = nao	==>	QTD_Veiculo = < 5	Alunos_Ensino_Fund = sim	41%	61%	1.1
2	Veiculos_Exclusivo = Aluno	Oferecido_periodo = sim , Veiculos_adaptados = nao , Existe_regulamentacao = nao	==>	QTD_Veiculo = < 5	Alunos_Ensino_Fund = sim	41%	61%	1.1
3	índice ideb - medio	Veiculos_Exclusivo = Aluno , Oferecido_periodo = sim	==>	Alunos_nec_especiais = sim	Acomp_rotina = sim	42%	66%	1.1
4	Alunos_nec_especiais = sim	Acomp_rotina = sim	==>	índice ideb - medio	Veiculos_Exclusivo = Aluno , Oferecido_periodo = sim	42%	70%	1.1
5	QTD_Veiculo = < 5	Alunos_Ensino_Fund = sim	==>	Veiculos_Exclusivo = Aluno	Veiculos_adaptados = nao , Existe_regulamentacao = nao	41%	74%	1.1
6	QTD_Veiculo = < 5	Alunos_Ensino_Fund = sim	==>	Veiculos_Exclusivo = Aluno	Oferecido_periodo = sim , Veiculos_adaptados = nao , Existe_regulamentacao = n	41%	74%	1.1
7	Veiculos_Exclusivo = Aluno	Veiculo_utilizado_outros = nao , Acomp_rotina = sim	==>	índice ideb - medio	Alunos_Ensino_Fund = sim , Veiculos_adaptados = nao	32%	61%	1.11
8	Veiculos_Exclusivo = Aluno	Veiculo_utilizado_outros = nao , Acomp_rotina = sim	==>	índice ideb - medio	Alunos_Ensino_Fund = sim , Oferecido_periodo = sim , Veiculos_adaptados = nao	32%	61%	1.11
9	Veiculos_Exclusivo = Aluno	Veiculo_utilizado_outros = nao , Oferecido_periodo = sim , Acomp_rotina = sim	==>	índice ideb - medio	Alunos_Ensino_Fund = sim , Veiculos_adaptados = nao	32%	61%	1.11
10	Veiculo_utilizado_outros = nao	Veiculos_adaptados = nao , Acomp_rotina = sim	==>	índice ideb - medio	Veiculos_Exclusivo = Aluno , Alunos_Ensino_Fund = sim , Oferecido_periodo =	32%	62%	1.1
11	Alunos_nec_especiais = sim	Veiculo_utilizado_outros = nao	==>	índice ideb - medio	Veiculos_Exclusivo = Aluno	37%	71%	1.1
12	Alunos_nec_especiais = sim	Veiculo_utilizado_outros = nao	==>	índice ideb - medio	Veiculos_Exclusivo = Aluno , Oferecido_periodo = sim	36%	70%	1.1
13	Alunos_nec_especiais = sim	Veiculo_utilizado_outros = nao	==>	índice ideb - medio	Acomp_rotina = sim	33%	64%	1.11
14	Alunos_nec_especiais = sim	Veiculo_utilizado_outros = nao	==>	índice ideb - medio	Oferecido_periodo = sim , Acomp_rotina = sim	33%	64%	1.11
15	Alunos_nec_especiais = sim	Veiculo_utilizado_outros = nao	==>	índice ideb - medio	Veiculos_Exclusivo = Aluno , Alunos_Ensino_Fund = sim	32%	62%	1.11
16	Alunos_nec_especiais = sim	Veiculo_utilizado_outros = nao	==>	índice ideb - medio	Veiculos_Exclusivo = Aluno , Alunos_Ensino_Fund = sim , Oferecido_periodo =	32%	62%	1.1
17	Alunos_nec_especiais = sim	Veiculo_utilizado_outros = nao , Oferecido_periodo = sim	==>	índice ideb - medio	Veiculos_Exclusivo = Aluno	36%	70%	1.1
18	Alunos_nec_especiais = sim	Veiculo_utilizado_outros = nao , Oferecido_periodo = sim	==>	índice ideb - medio	Acomp_rotina = sim	33%	65%	1.11
19	Alunos_nec_especiais = sim	Veiculo_utilizado_outros = nao , Oferecido_periodo = sim	==>	índice ideb - medio	Veiculos_Exclusivo = Aluno , Alunos_Ensino_Fund = sim	32%	62%	1.1
20	Alunos_Ensino_Fund = sim	Veiculo_utilizado_outros = nao , Acomp_rotina = sim	==>	índice ideb - medio	Veiculos_Exclusivo = Aluno , Oferecido_periodo = sim , Veiculos_adaptados = na	32%	63%	1.1
21	índice ideb - medio	Veiculos_Exclusivo = Aluno , Alunos_Ensino_Fund = sim , Veiculos_adaptados = nao	==>	Veiculo_utilizado_outros = na	Acomp_rotina = sim	32%	64%	1.1
22	índice ideb - medio	Veiculos_Exclusivo = Aluno , Alunos_Ensino_Fund = sim , Veiculos_adaptados = nao	==>	Veiculo_utilizado_outros = na	Oferecido_periodo = sim , Acomp_rotina = sim	32%	64%	1.1
23	índice ideb - medio	Veiculos_Exclusivo = Aluno , Alunos_Ensino_Fund = sim , Oferecido_periodo = sim , Veiculos_adaptados = n	==>	Veiculo_utilizado_outros = na	Acomp_rotina = sim	32%	64%	1.11
24	Veiculos_Exclusivo = Aluno	Alunos_nec_especiais = sim , Veiculo_utilizado_outros = nao	==>	índice ideb - medio	Acomp_rotina = sim	31%	65%	1.11
25	Veiculos_Exclusivo = Aluno	Alunos_nec_especiais = sim , Veiculo_utilizado_outros = nao	==>	índice ideb - medio	Oferecido_periodo = sim , Acomp_rotina = sim	31%	64%	1.11
26	Veiculos_Exclusivo = Aluno	Alunos_nec_especiais = sim , Veiculo_utilizado_outros = nao , Oferecido_periodo = sim	==>	índice ideb - medio	Acomp_rotina = sim	31%	65%	1.11
27	Alunos_nec_especiais = sim	Veiculo_utilizado_outros = nao , Veiculos_adaptados = nao	==>	índice ideb - medio	Veiculos_Exclusivo = Aluno	33%	71%	1.11
28	Alunos_nec_especiais = sim	Veiculo_utilizado_outros = nao , Veiculos_adaptados = nao	==>	índice ideb - medio	Veiculos_Exclusivo = Aluno , Oferecido_periodo = sim	33%	71%	1.11
29	Alunos_nec_especiais = sim	Veiculo_utilizado_outros = nao , Veiculos_adaptados = nao	==>	índice ideb - medio	Alunos_Ensino_Fund = sim	31%	67%	1.1
30	Alunos_nec_especiais = sim	Veiculo_utilizado_outros = nao , Oferecido_periodo = sim , Veiculos_adaptados = nao	==>	índice ideb - medio	Veiculos_Exclusivo = Aluno	33%	71%	1.11
31	Alunos_Ensino_Fund = sim	Alunos_nec_especiais = sim , Veiculo_utilizado_outros = nao	==>	índice ideb - medio	Veiculos_Exclusivo = Aluno	32%	71%	1.1
32	Alunos_Ensino_Fund = sim	Alunos_nec_especiais = sim , Veiculo_utilizado_outros = nao	==>	índice ideb - medio	Veiculos_Exclusivo = Aluno , Oferecido_periodo = sim	32%	70%	1.1
33	Alunos_Ensino_Fund = sim	Alunos_nec_especiais = sim , Veiculo_utilizado_outros = nao , Oferecido_periodo = sim	==>	índice ideb - medio	Veiculos_Exclusivo = Aluno	32%	71%	1.1
34	índice ideb - medio	Alunos_Ensino_Fund = sim , Veiculos_adaptados = nao , Acomp_rotina = sim	==>	Veiculos_Exclusivo = Aluno	Veiculo_utilizado_outros = nao	32%	73%	1.11
35	índice ideb - medio	Alunos_Ensino_Fund = sim , Veiculos_adaptados = nao , Acomp_rotina = sim	==>	Veiculos_Exclusivo = Aluno	Veiculo_utilizado_outros = nao , Oferecido_periodo = sim	32%	73%	1.11
36	índice ideb - medio	Alunos_Ensino_Fund = sim , Oferecido_periodo = sim , Veiculos_adaptados = nao , Acomp_rotina = sim	==>	Veiculos_Exclusivo = Aluno	Veiculo_utilizado_outros = nao	32%	73%	1.11
37	Alunos_nec_especiais = sim	Veiculo_utilizado_outros = nao , Acomp_rotina = sim	==>	índice ideb - medio		33%	78%	1.11
38	Alunos_nec_especiais = sim	Veiculo_utilizado_outros = nao , Acomp_rotina = sim	==>	índice ideb - medio	Oferecido_periodo = sim	33%	78%	1.11
39	Alunos_nec_especiais = sim	Veiculo_utilizado_outros = nao , Acomp_rotina = sim	==>	índice ideb - medio	Veiculos_Exclusivo = Aluno	31%	72%	1.13
40	Alunos_nec_especiais = sim	Veiculo_utilizado_outros = nao , Acomp_rotina = sim	==>	índice ideb - medio	Veiculos_Exclusivo = Aluno , Oferecido_periodo = sim	31%	72%	1.13
41	Alunos_nec_especiais = sim	Veiculo_utilizado_outros = nao , Oferecido_periodo = sim , Acomp_rotina = sim	==>	índice ideb - medio		33%	78%	1.1
42	Alunos_nec_especiais = sim	Veiculo_utilizado_outros = nao , Oferecido_periodo = sim , Acomp_rotina = sim	==>	índice ideb - medio	Veiculos_Exclusivo = Aluno	31%	72%	1.13
43	Veiculos_Exclusivo = Aluno	Alunos_nec_especiais = sim , Veiculo_utilizado_outros = nao , Acomp_rotina = sim	==>	índice ideb - medio		31%	78%	1.1
44	Veiculos_Exclusivo = Aluno	Alunos_nec_especiais = sim , Veiculo_utilizado_outros = nao , Acomp_rotina = sim	==>	índice ideb - medio	Oferecido_periodo = sim	31%	78%	1.11
45	Veiculos_Exclusivo = Aluno	Alunos_nec_especiais = sim , Veiculo_utilizado_outros = nao , Oferecido_periodo = sim , Acomp_rotina = sim	==>	índice ideb - medio		31%	78%	1.1

Atividade 17: Interpretação dos padrões

Foram utilizadas as técnicas de filtragem e de interpretação do conhecimento, para que fossem identificados os resultados mais interessantes, sob a ótica de negócio da gestão do transporte escolar rural. Então, na análise dos resultados, procuraram-se as regras de associação de maior interesse para o estudo de caso.

Cabe ressaltar, que as regras geradas são resultantes de um conjunto de atributos relevantes, bem como das variações desses atributos, juntamente com a alternância do intervalo das medidas de interesse da *lift, confiança e suporte* do algoritmo.

Para melhor entendimento das regras de associação, será necessária uma compreensão da leitura do *relatório de saída* do algoritmo dessas regras**.**

COMPREENDENDO A ESTRUTURA DA INTERPRETAÇÃO DA REGRA DE ASSOCIAÇÃO

O algoritmo *apriori* de mineração de dados de regras de associação gerou um relatório com resultados das regras, no formato específico a seguir, conforme a Tabela 5.5 .

Tabela 5.5: Leitura da regra de associação

	Conjunto A			Conjunto B	Medida Objetiva		
N.	**atributo=valor**	**atributo=valor**	**⇒**	**atributo=valor**	**sup**	**conf.**	**lift**
(A)	**(B)**	**(C)**		**(D)**	**(E)**	**(F)**	**(G)**

A leitura da regra neste formato apresenta dois conjuntos. Sendo o conjunto A, o corpo da regra (antecedente) e separado por uma seta(⇒) e o segundo, conjunto B, que é o resultado da regra (consequente).

Assim prosseguindo, o número da regra (N) está representado no campo A. Os campos B e C representam os atributos que compõem o corpo da regra, que neste caso, são dois atributos ou pode ser formado por um ou mais atributos. Separada por uma seta (⇒), tem-se o campo D, representante do atributo que compõe o resultado da regra, que neste caso, é um atributo ou pode ser formado por um ou mais atributos. Nos campos -E, F, G- têm-se as

medidas de interesses objetivas, as quais representam os índices estatísticos da regra. Essas medidas são '*suporte (E)*', '*confiança (F)*' e '*lift(G)*'.

Para melhor entendimento, são apresentadas duas formas de interpretação dos dados: (i) Análise estatística descritiva para descrever o objeto de estudo, tendo os resultados esquematizados em forma de gráficos e tabelas; (ii) Análise analítica consiste em um processo de mineração de dados para explorar grande quantidade de dados, para transformar dados em informação e conhecimento úteis, tendo como resultados regras e padrões significativos.

I. Análise descritiva – estatística descritiva: variáveis da regra n.99

Na Figura 5.19-A, cerca 24% dos municípios, oferecem o serviço de transporte escolar rural para alunos com necessidades educacionais especiais. Na Figura 5.19-B, aproximadamente 93% dos municípios, utilizam de veículos não adaptados para portadores de necessidades especiais. Na Figura 5.19-C, aproximadamente 27% dos municípios, utilizam os veículos para outras finalidades, quando estes não estão sendo usado para o transporte escolar rural (CEFTRU/FNDE, 2007b).

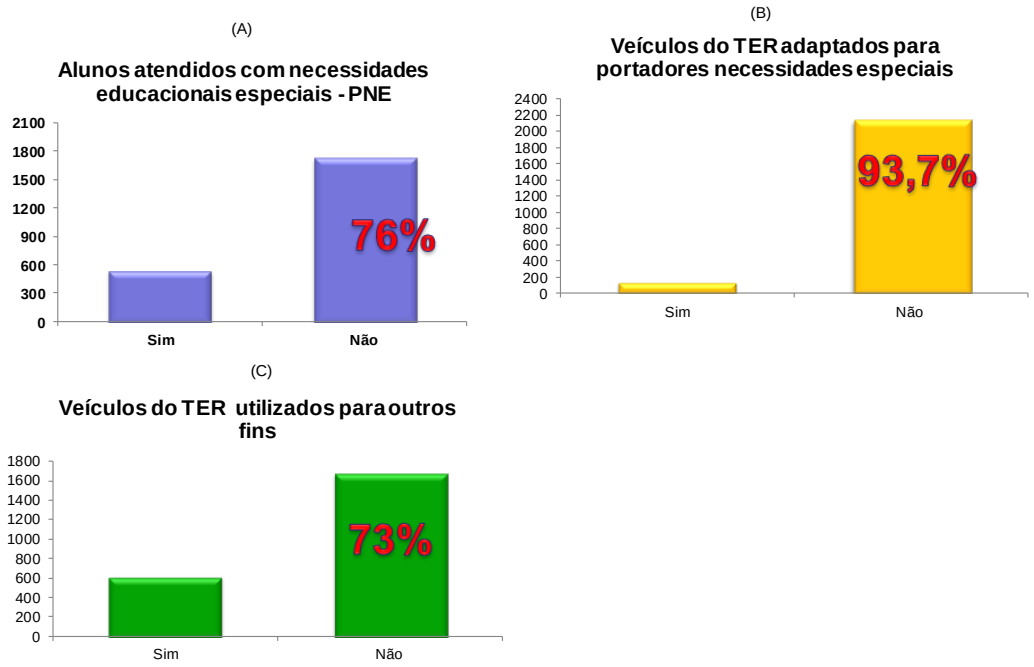


Figura 5.19: Gráficos das variáveis da regra 99 . Fonte: CEFTRU/FNDE (2007b)

II. Análise analítica – Mineração de dados: regra de associação n. 99

Tabela 5.6 Exemplo de regra de associação hipotética

R:99	*'existe aluno com necessidade especial = sim', 'veiculo é adaptado = não'* **Medidas de interesse** *'suporte (R)'* = 56% *'confiança(R)'* = 84% *'lift(R)'* = 3,1	⇒	*'veiculo é exclusivo para aluno = sim'*

Analisando a regra de número 99, com os atributos do corpo da regra *'existe aluno com necessidade especial'* e '*veiculo é adaptado';* obtem-se como resultado *'veiculo é exclusivo para aluno*'; têm-se as possíveis interpretações:

- O valor, *medida de suporte*, indica que 56% de todos os gestores oferecem o transporte escolar rural exclusivo para alunos, mas transportam alunos portadores de necessidades especiais em veículos não adaptados;
- O valor, *medida de confiança*, indica que a probabilidade de um gestor usar veículos não adaptados, para o transporte escolar rural de portadores de necessidades especiais e serem de uso exclusivo para alunos é de 84%;
- O valor, *medida de lift*, indica que o uso de veículo exclusivo para alunos é 3,1 vezes maior que o uso de veículo não adaptado para alunos portadores de necessidades especiais.

Essas regras mostram uma correlação entre essas variáveis, ou seja, se gestor oferecer um serviço para uma clientela, que necessita de cuidados especiais e disponibiliza um veículo inadequado, então essa prática adotada pelo município mostra-se inadequada.

A análise que acaba de ser apresentada demonstra que cada uma das *medidas de interesse objetivas* é empregada para fornecer uma informação específica, a respeito de uma regra de associação. Esse fato evidencia que todas essas medidas são importantes para o usuário de uma ferramenta de mineração de dados.

OBSERVAÇÕES IMPORTANTES PARA COMPREENSÃO DAS INTERPRETAÇÕES

Cabe destacar que todas as vezes que na análise forem utilizados *gestores*, entende-se que foram aqueles que responderam o questionário.

Cabe esclarecer, que uma tarefa de associação identifica os fatos que tendem a ocorrer juntos em uma transação. Assim, a presença de alguns deles em uma transação implica na presença de outros na mesma transação, identificando dessa forma, uma relação ou uma tendência. Assim, duas variáveis estão associadas se conhecimentos de uma delas alterar a probabilidade de ocorrência de algum resultado da outra. Contudo, não implica necessariamente, uma relação de causa-efeito entre as variáveis.

Outra ressalva é quanto ao '*índice do desenvolvimento da educação básica – IDEB'*, referente aos dados do ano 2005; esse indicador utilizou-se na primeira medição, os dados que foram levantados em 2005 (MEC/INEP, 2007). De acordo com MEC/INEP (2007), o *IDEB* foi criado em 2007, para medir a qualidade do ensino no Brasil, numa escala de zero a dez.

Como ressalva, as regras de associação geradas possuem os valores dos *'suporte calculado'* maior ou igual ao '*suporte mínimo'* de 30% e menor ou igual ao '*suporte máximo*' de 60%; e da '*confiança calculada'* maior ou igual a '*confiança mínima'* de 30% e menor ou igual a '*confiança máximo'* de 100%; e do '*lift calculo*' maior ou igual ao '*lift mínimo'* de 1,10.

A seguir serão realizadas as duas formas de análises, conforme o item anterior.

I. Análise descritiva – estatística descritiva: variáveis da regra n.30

Na Figura 5.20-A, cerca de 24% dos municípios oferecem o serviço de transporte escolar rural para alunos com necessidades educacionais especiais. Na Figura 5.20-B, aproximadamente 27% dos municípios, utilizam os veículos para outras finalidades, quando esse não está sendo usado para o transporte escolar rural. Na Figura 5.20-C, mostra que na periodicidade do serviço, cerca de 98% dos municípios que oferecem o serviço durante todo o ano letivo. Na Figura 5.20-D, aproximadamente 93% dos municípios, utilizam de veículos não adaptados para portadores de necessidades especiais. Na Figura 5.20-D, aproximadamente 93% dos municípios, utilizam de veículos não adaptados para portadores de necessidades especiais (CEFTRU/FNDE, 2007b). Na Figura 5.20-E, segundo MEC/INEP (2007), cerca de 65% dos municípios, apresentam o índice IDEB (>=3.0 e <4) médio de desempenho nas avaliações do INEP, referente o ano de 2007.

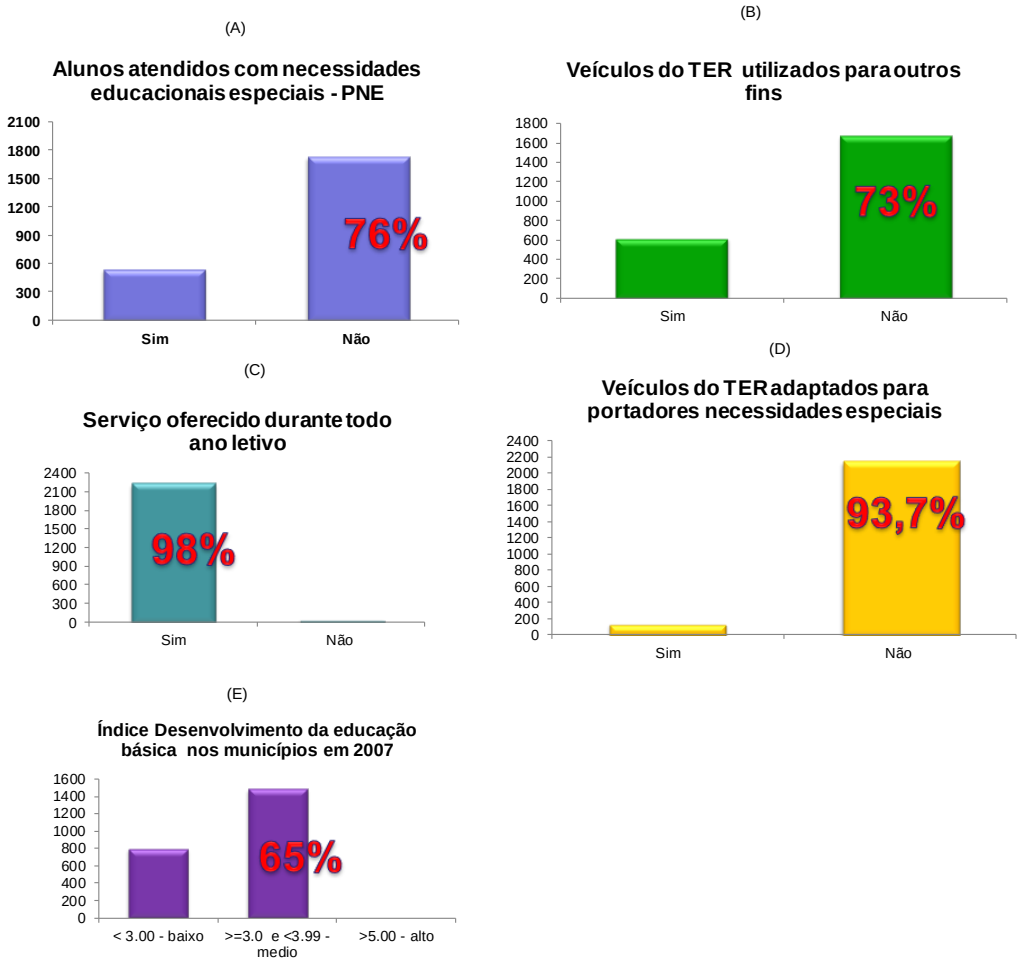


Figura 5.20: Gráficos das variáveis da regra 30 . Fonte: CEFTRU/FNDE (2007b)

II. Análise analítica – Mineração de dados: regra de associação n. 30

A Tabela 5.7 mostra a regra de número 30, com as seguintes interpretações:

Tabela 5.7 : estrutura da regra 30

R: 30	*'existe aluno com necessidade especial = sim'*, *'veiculo é utilizado para outros fins = não'*, *'serviço é oferecido no período letivo = sim'*, *'veiculo é adaptado = não'*	⇒	*'indice IDEB é médio'*, *'veiculo é exclusivo para aluno = sim'*

Medidas de interesse
'suporte (R)' = 33%
'confiança(R)' = 71%
'lift(R)' = 1,1

Analisando a regra de número 30, com os atributos do corpo da regra '*existe aluno com necessidade especial'; 'veículo é utilizado para outros fins'; 'serviço é oferecido em todo período letivo'* e *'veiculo é adaptado';* obtem-se como *resultado 'índice IDEB(>=3.0 e <4) – médio'* e *'veiculo é exclusivo para aluno';* têm-se as possíveis interpretações:

- O valor, *medida de suporte*, indica que 33% de todos os municípios apresentam o '*índice IDEB (>=3.0 e <4) – médio';* oferecem o transporte escolar rural exclusivo para aluno em veículos não adaptados aos portadores de necessidades especiais, não utilizados para outros fins e transportando em todo período letivo;

- O valor, *medida de confiança*, indica que a probabilidade de um município, que apresenta o *'índice IDEB (>=3.0 e <4) – médio';* usar veículos não adaptados aos portadores de necessidades especiais, não utilizados para outros fins, para o transporte escolar rural exclusivo em todo período letivo é de 71%;

- O valor, *medida de lift*, indica que os municípios com o '*índice IDEB (>=3.0 e <4) – médio'* e com uso de veículos exclusivos para alunos é 1,1 vezes maior que os municípios que fazem uso de veículos não adaptados aos alunos portadores de necessidades especiais, não utilizados para outros fins, de uso exclusivo e transportando em todo período letivo.

I. Análise descritiva – estatística descritiva: variáveis da regra n.6

Na Figura 5.21-A, aproximadamente 27% dos municípios, utilizam os veículos para outras finalidades, quando esse não está sendo usado para o transporte escolar rural. Na Figura 5.21-B, mostra que na periodicidade do serviço, cerca de 98% dos municípios que oferecem o serviço durante todo o ano letivo. Na Figura 5.21-C, aproximadamente 93% dos municípios, utilizam de veículos não adaptados para portadores de necessidades especiais. Na Figura 5.21-D, cerca de 85% dos municípios declararam não ter regulamentação própria para o transporte escolar. Na Figura 5.21-E, cerca de 63% dos municípios utilizam menos de cincos veículos para transporte escolar rural (CEFTRU/FNDE, 2007b).

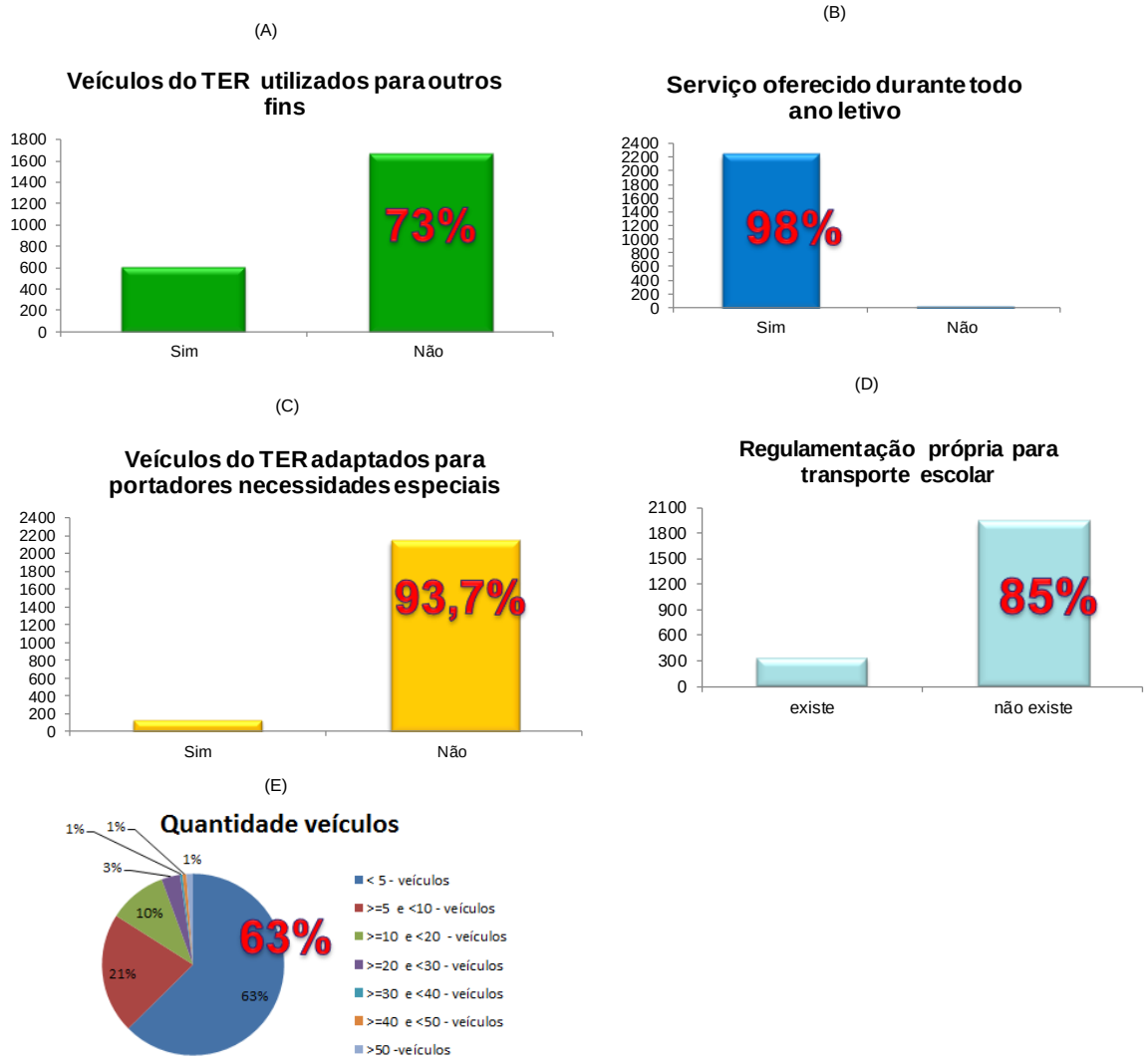


Figura 5.21: Gráficos das variáveis da regra 6 . Fonte: CEFTRU/FNDE (2007b)

II. Análise analítica – Mineração de dados: regra de associação n. 6

A Tabela 5.8 mostra a regra de número 6, com algumas interpretações:

Tabela 5.8 : estrutura da regra 6

R: 6	*'quantidade de veículos utilizados < 5'* **Medidas de interesse** *'suporte (R)'* = 41% *'confiança(R)'* = 74% *'lift(R)'* = 1,1	⇒	*'veiculo é exclusivo para aluno = sim',* *'serviço é oferecido no período letivo = sim',* *'veiculo é adaptado = não',* *'existe regulamentação = não'*

Analisando a regra de número 6, com o atributo do corpo da regra *'quantidade de veículos utilizados';* obtem-se como resultado '*veiculo é exclusivo para aluno'; 'serviço é oferecido em todo período letivo'; 'veiculo é adaptado' e 'existe regulamentação'*; têm-se as possíveis interpretações:

- O valor, *medida de suporte*, indica que 41% de todos os municípios, onde não existem regulamentação própria, utilizam menos de cinco veículos não adaptados ao transporte escolar rural exclusivo para aluno e transportando em todo período letivo;

- O valor, *medida de confiança,* indica que a probabilidade de um município, onde não exista regulamentação própria, utilizar menos de cinco veículos não adaptados para transporte escolar rural exclusivo para aluno e transportando em todo período letivo é de 74%;

- O valor, *medida de lift,* indica que os municípios, onde não existem regulamentação própria; e com uso de veículos não adaptados para transporte escolar rural exclusivo para aluno e transportando em todo período letivo é 1,1 vezes maior que os municípios com uso de menos de cinco veículos para o transporte escolar rural.

I. Análise descritiva – estatística descritiva: variáveis da regra n.23

Na Figura 5.22-A, aproximadamente 27% dos municípios utilizam os veículos para outras finalidades, quando esse não está sendo usado para o transporte escolar rural. Na Figura 5.22-B, mostra que na periodicidade do serviço, cerca de 98% dos municípios oferecem o serviço durante todo o ano letivo. Na Figura 5.22-C, aproximadamente 93% dos municípios se utilizam de veículos não adaptados, para portadores de necessidades especiais. Na Figura 5.22-D, cerca de 73% dos municípios declararam ter algum tipo de acompanhamento das rotinas do transporte escolar rural (CEFTRU/FNDE, 2007b). Na Figura 5.22-E, segundo MEC/INEP (2007), cerca de 65% dos municípios apresentam o índice IDEB (>=3.0 e <4) médio de desempenho nas avaliações do INEP, referente o ano de 2007

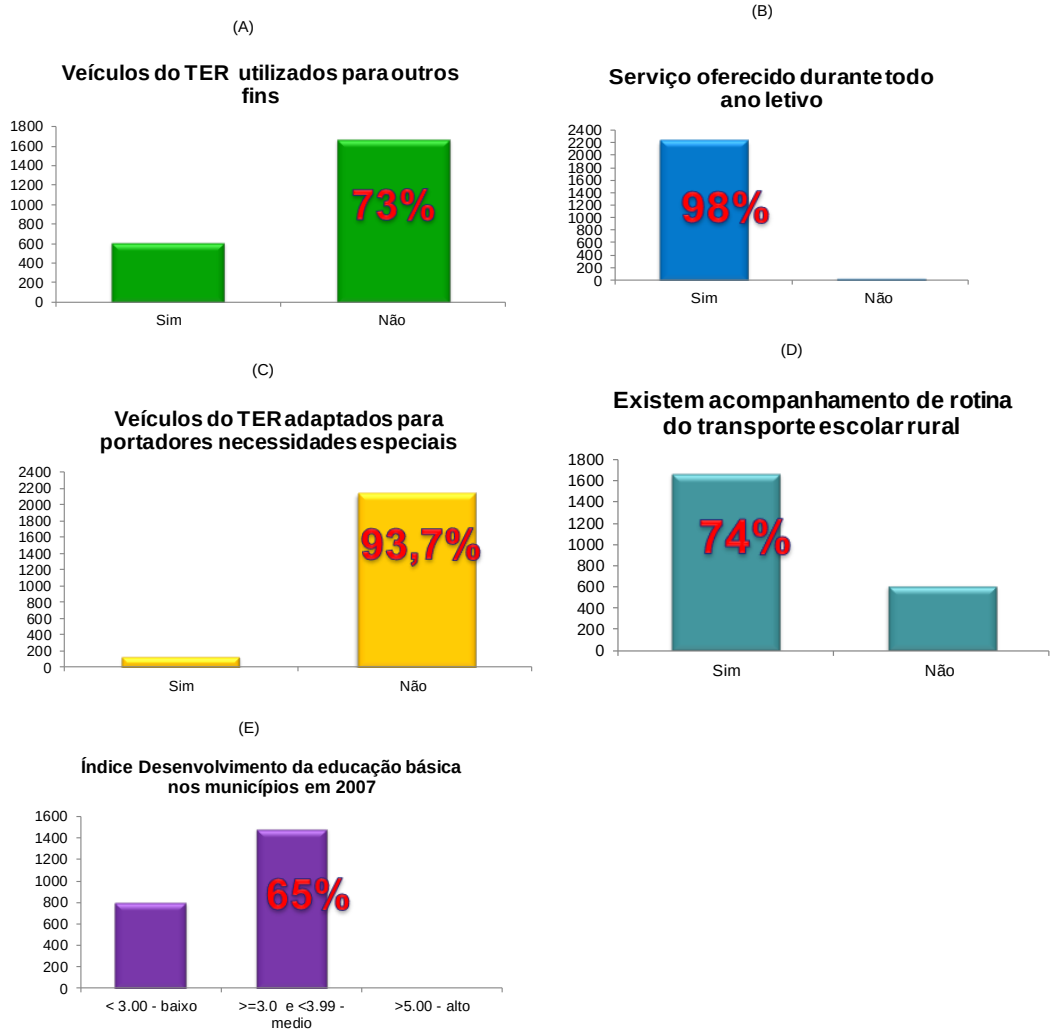


Figura 5.22: Gráficos das variáveis da regra 23. Fonte: CEFTRU/FNDE (2007b)

II. Análise analítica – Mineração de dados: regra de associação n. 23

A Tabela 5.9 mostra a regra de número 23, com as seguintes interpretações:

Tabela 5.9 : estrutura da regra 23

R: 23	*'índice IDEB é médio',* *'veiculo é exclusivo para aluno = sim',* *'serviço é oferecido no período letivo = sim',* *'veiculo é adaptado = não'* **Medidas de interesse** *'suporte (R)'* = 32% *'confiança(R)'* = 64% *'lift(R)'* = 1,1	⇒	*'veiculo é utilizado para outros fins = não',* *'existe acompanhamento de rotina= sim'*

Analisando a regra de número 23, com os atributos do corpo da regra '*índice IDEB(>=3.0 e <4) – médio' ; 'veiculo é exclusivo para aluno' ; 'serviço é oferecido em todo período letivo'* e *'veiculo é adaptado';* obtem-se como resultado '*veículo é utilizado para outros fins'* e '*existe acompanhamento de rotina';* têm-se as possíveis interpretações:

- O valor, *medida de suporte*, indica que 32% dos municípios, que apresentam o '*índice IDEB(>=3.0 e <4) – médio',* existem acompanhamento de rotina nos veículos do transporte escolar rural não adaptados para portadores de necessidades especiais; não utilizados para outros fins, para o transporte exclusivo de aluno e transportando em todo período letivo;

- O valor, *medida de confiança*, indica que a probabilidade de um município, que apresenta '*índice IDEB(>=3.0 e <4) – médio';* apresentar acompanhamento de rotina nos veículos do transporte escolar rural não adaptados aos portadores de necessidades especiais; não utilizados para outros fins, para o transporte exclusivo de aluno e transportando em todo período letivo é de 64%;

- O valor, *medida de lift*, indica que nos municípios, onde existem acompanhamento de rotina, e uso de veículos não utilizados para outros fins é 1,17 vezes maior que nos municípios que apresentam o '*índice IDEB(>=3.0 e <4) – médio';* com uso de veículos não adaptados aos portadores de necessidades especiais, para o transporte exclusivo de aluno e transportando em todo período letivo.

I. Análise descritiva – estatística descritiva: variáveis da regra n.1

Na Figura 5.23-A, aproximadamente 27% dos municípios, utilizam os veículos para outras finalidades, quando estes não estão sendo usados para o transporte escolar rural. Na Figura 5.23-C, aproximadamente 93% dos municípios, utilizam de veículos não adaptados para portadores de necessidades especiais. Na Figura 5.23-C, cerca de 85% dos municípios, declararam não ter regulamentação própria para o transporte escolar. Na Figura 5.23-D, cerca de 63% dos municípios, utilizam menos de cincos veículos para transporte escolar rural (CEFTRU/FNDE, 2007b).

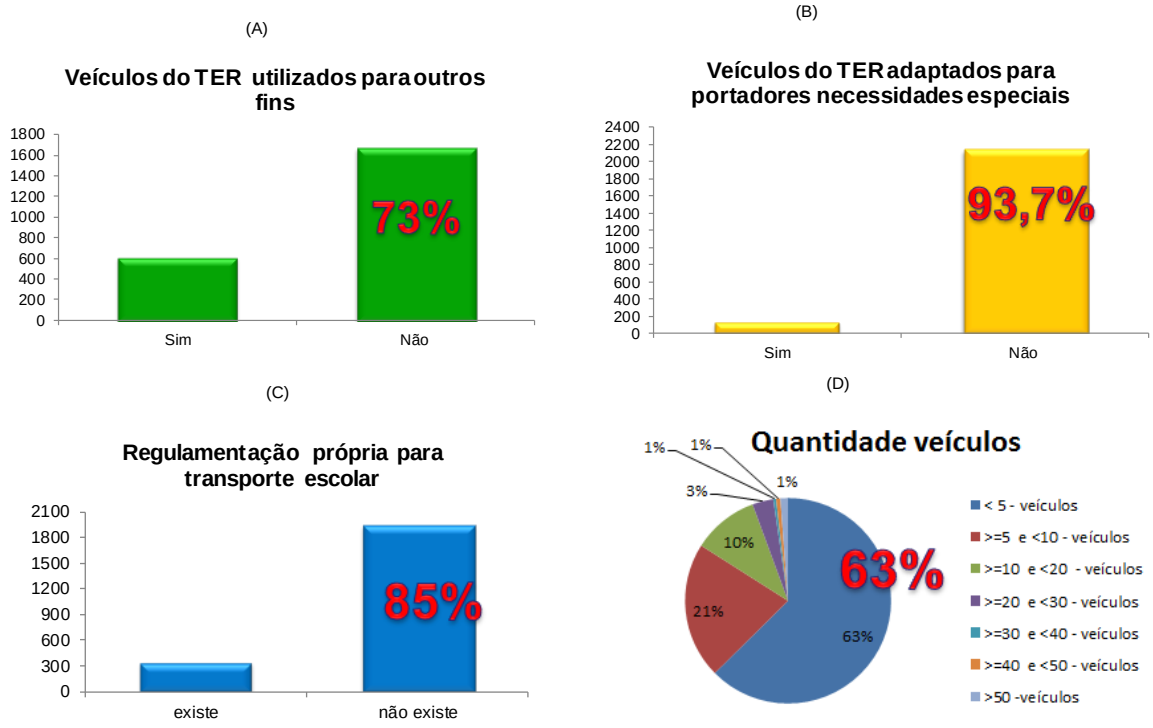


Figura 5.23: Gráficos das variáveis da regra 1. Fonte: CEFTRU/FNDE (2007b)

II. Análise analítica – Mineração de dados: regra de associação n. 1

A Tabela 5.10 mostra a regra de número 1, com algumas interpretações:

Tabela 5.10 : estrutura da regra 1

R: 1	*'veiculo é exclusivo para aluno = sim'*, *'veiculo é adaptado = não'*, *'existe regulamentação = não'*	⇒	*'quantidade de veículos utilizados < 5'* *'aluno é do ensino fundamental=sim'*

Medidas de interesse
'suporte (R)' = 41%
'confiança(R)' = 61%
'lift(R)' = 1,1

Analisando a regra de número 1, com o atributo do corpo da regra '*veiculo é exclusivo para aluno'; 'veiculo é adaptado' e 'existe regulamentação'*; obtem-se como resultado *'quantidade de veículos utilizados*' e '*alunos do ensino fundamental*'; têm-se as possíveis interpretações:

- O valor, *medida de suporte*, indica que 41% de todos os municípios, onde não existem regulamentação própria, utilizam menos de cinco veículos não adaptados aos portadores de necessidades especiais, para o transporte escolar rural exclusivo para aluno do ensino fundamental.

- O valor, *medida de confiança*, indica que a probabilidade de um município, onde não exista regulamentação própria, utilizar menos de cinco veículos não adaptados aos portadores de necessidades especiais, para transporte escolar rural exclusivo ao aluno do ensino fundamental é de 61%.

- O valor, *medida de lift,* indica que os municípios, com uso de menos de cinco veículos, para o transporte escolar rural ao aluno do ensino fundamental é 1,1 vezes maior que os municípios que fazem uso de veículo não adaptado aos alunos portadores de necessidades especiais; de uso exclusivo e que não possuem regulamentação própria.

I. Análise descritiva – estatística descritiva: variáveis da regra n.9

Na Figura 5.24-A, aproximadamente 27% dos municípios, utilizam os veículos para outras finalidades, quando estes não estão sendo usados para o transporte escolar rural. Na Figura 5.24-B, mostra que na periodicidade do serviço, cerca de 98% dos municípios, oferecem o serviço durante todo o ano letivo. Na Figura 5.24-C, cerca de 73% dos municípios, declararam ter algum tipo de acompanhamento das rotinas do transporte escolar rural. Na Figura 5.24-D, cerca de 65% dos municípios, apresentam o índice IDEB (>=3.0 e <4) médio de desempenho nas avaliações do INEP, referente o ano de 2007 (MEC/INEP, 2007). Na Figura 5.24-C, aproximadamente 93% dos municípios, utilizam de veículos não adaptados para portadores de necessidades especiais (CEFTRU/FNDE, 2007b).

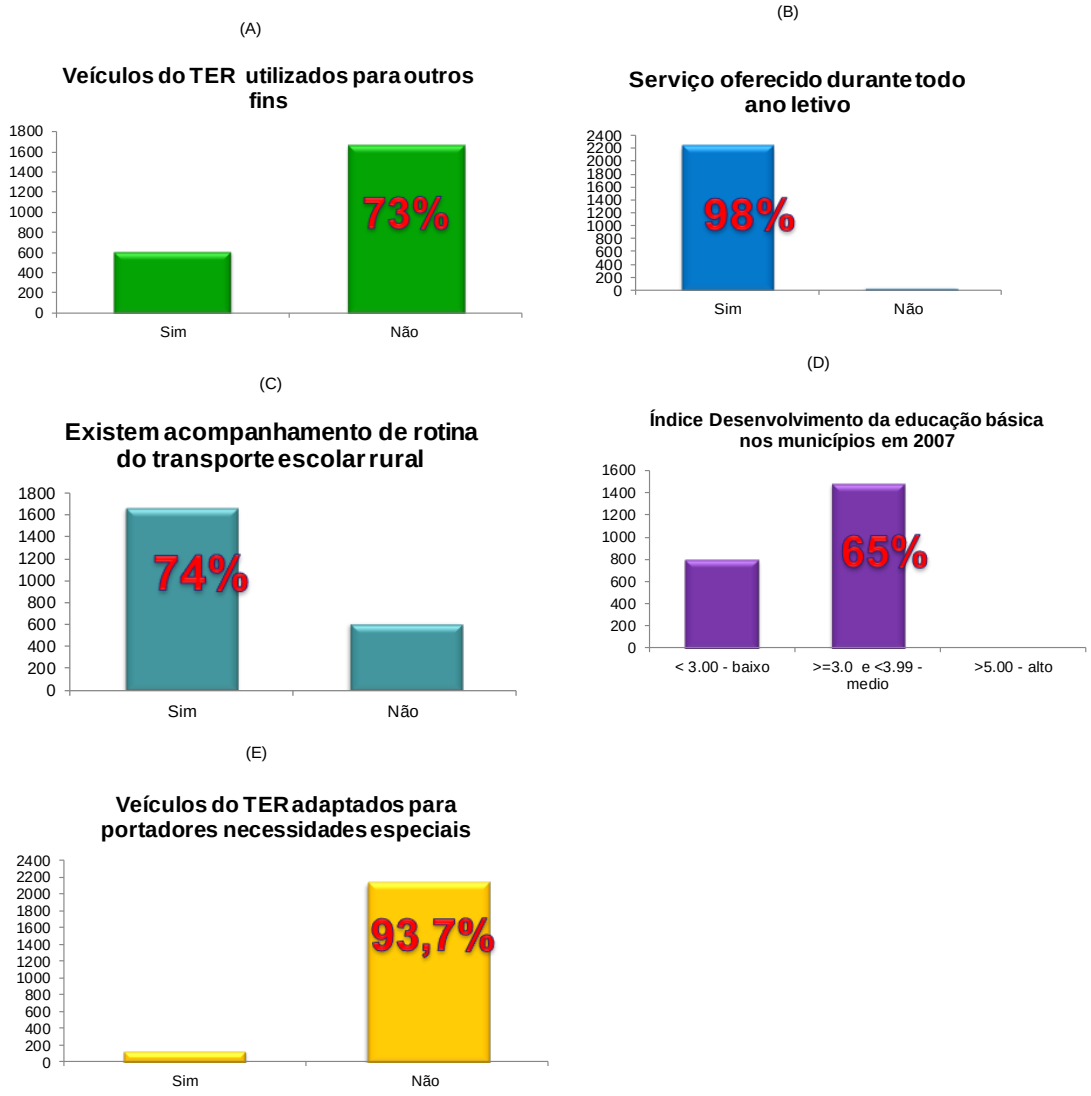


Figura 5.24: Gráficos das variáveis da regra n.9. Fonte: CEFTRU/FNDE (2007b)

II. Análise analítica – Mineração de dados: regra de associação n. 9

A Tabela 5.11 mostra a regra de número 9, com as seguintes interpretações:

Tabela 5.11 : estrutura da regra 09

R: 09	*'veiculo é exclusivo para aluno = sim'* *'veiculo é utilizado para outros fins = não'* *'serviço é oferecido no período letivo = sim',* *'existe acompanhamento de rotina= sim',* **Medidas de interesse** *'suporte (R)'* = 32% *'confiança(R)'* = 61% *'lift(R)'* = 1,1	⇒	*'índice IDEB é médio',* *'aluno é do ensino fundamental=sim'* *'veiculo é adaptado = não'*

Analisando a regra de número 9, com os atributos do corpo da regra *'veiculo é exclusivo para aluno' ;* '*veículo é utilizado para outros fins'; 'serviço é oferecido em todo período letivo'* e '*existe acompanhamento de rotina';* obtem-se como resultado '*índice IDEB(>=3.0 e <4) – médio';* '*alunos do ensino fundamental*' *e 'veiculo é adaptado';* têm-se as possíveis interpretações:

- O valor, *medida de suporte*, indica que nos 32% dos municípios onde o '*índice IDEB(>=3.0 e <4) – médio'*, apresentam o acompanhamento de rotina no desenvolvimento da operação de transporte (gastos de insumo, quilometragem percorrida, dentre outros) escolar rural exclusivo ao aluno do ensino fundamental, com veículos não adaptados aos portadores de necessidades especiais; não utilizados para outros fins e transportando em todo período letivo.

- O valor, *medida de confiança*, indica que a probabilidade de um município que apresenta '*índice IDEB(>=3.0 e <4) – médio';* apresentar o acompanhamento de rotina no desenvolvimento da operação de transporte (gastos de insumo, quilometragem percorrida, dentre outros) escolar rural exclusivo ao o aluno do ensino fundamental; com veículos não adaptados aos portadores de necessidades especiais, não utilizados para outros fins e transportando em todo período letivo é de 61%.

- O valor, *medida de lift*, indica que os municípios com o '*índice IDEB (>=3.0 e <4) – médio;' e* com uso de veículos não adaptados aos alunos portadores de necessidades especiais e transportando alunos do ensino fundamental é 1,1 vezes maior que os municípios que apresentam acompanhamento de rotina no

desenvolvimento da operação de transporte (gastos de insumo, quilometragem percorrida, dentre outros) escolar rural e fazem uso de veículos exclusivos para alunos; não utilizados para outros fins e transportando em todo período letivo.

Atividade 18: Construção do conhecimento aos grupos-alvos

Esta atividade consiste na formação de recursos humanos, capazes de utilizar corretamente a metodologia, para dar suporte as suas atividades e ao processo de tomada de decisão. Assim, todos esses conhecimentos extraídos podem ser utilizados pelo *gestor principal,* para apoio ao processo de tomada de decisão, ou suporte na formulação de plano de ação; com o propósito de aumentar o nível da qualidade dos serviços oferecidos e que esse nível seja compatível com as características físicas dos alunos de modo geral. Portanto, é essencial que o transporte escolar não apenas exista, mas que também seja realizado com qualidade e segurança, pois a prestação de serviços precários e de baixa qualidade tem impacto direto no aproveitamento das aulas pelos estudantes.

Atividade 19: Planejamento das ações

Essas regras de associação extraídas na base de dados, tais como: *veículos não adaptados e alunos com necessidades especiais,* podem ser utilizadas pelos *gestores principais,* para servir de base à tomada de decisão em estudo de transporte durante o processo de planejamento. Portanto, foi alcançado o objetivo da metodologia que é o de caracterização do serviço de transporte escolar, utilizando-se da mineração de dados com o entendimento dos três elementos: *serviço, clientela e recursos*; para que esses elementos possam ser utilizados pelos *gestores principais* no planejamento de suas ações.

5.5 TÓPICOS CONCLUSIVOS DO ESTUDO DE CASO

Esta seção apresenta algumas análises conclusivas sobre a aplicabilidade da metodologia proposta, com propósito de realizar as interpretações dos padrões extraídos da base de dados do questionário web (CEFTRU, 2007b); tendo como objetivo realizar a caracterização do transporte escolar rural nos municípios brasileiros, utilizando-se da mineração de dados, com a aplicação da tarefa de regras de associação, para descobrir quais são as práticas e procedimentos adotados na gestão do transporte escolar rural.

- O sistema de transporte escolar rural, por suas características, envolve grande quantidade e variedades de informações necessárias para seu entendimento. Dessa forma, verificam-se pela rede semântica, os elementos que devem ser conhecidos e dominados, para que o transporte escolar rural possa ser melhorado. Ao mesmo tempo, a rede semântica mostra a grande quantidade de dados/informações necessárias, para se trabalhar no planejamento do transporte escolar rural.

- Ficou claro que existe pouco banco de dados, sobre o transporte escolar rural, oferecido nos municípios brasileiros. Atualmente, o que se tem de mais completo, foi realizado entre CEFTRU e FNDE, no estudo sobre o Transporte Escolar Rural; com o objetivo de traçar um retrato da situação desse serviço no Brasil. (CEFTRU, 2007a, 2007b, 2007c, 2008a, 2008b).

- Ficou evidente, que o transporte escolar rural é forte aliado da integração social e espacial no meio rural; caracterizando-se como elemento importante para garantir os seus direitos enquanto cidadãos. Nesse sentido, a educação é um direito garantido na Constituição Federal (BRASIL, 1988) pela obrigatoriedade do fornecimento do transporte escolar gratuito pelo poder público.

- Dado o tempo disponível para o desenvolvimento da pesquisa e com as limitações enfrentadas, foi possível utilizar a base de dados da pesquisa web (CEFTRU, 2007b) com algumas adequações e adaptações no banco de dados, para atender aos requisitos da *fase de elaboração da metodologia proposta*. Cabe destacar, que os dados utilizados foram os declarados pelos municípios nos questionários e que são de inteira responsabilidade dos seus gestores e não foram coletados *in loco.*

- Acerca das dificuldades enfrentadas, para a utilização dos dados descritos na *etapa de elaboração da metodologia proposta*, demandou-se um tempo considerável de ajuste manual; pré-processamento dos dados selecionados; transformação dos dados e geração do arquivo de dados, de acordo com a especificação da ferramenta de mineração de dados. Isso mostra a importância da necessidade de uma banco de dados modelado adequadamente;

- Por meio da aplicação do estudo de caso e dos resultados alcançados, foi possível concluir que a mineração de dados contribui para tornar tangível o conhecimento implícito em base de dados de sistema de informação em transporte;

- A metodologia, utilizada no estudo de caso, permitiu realizar de forma resumida o diagnóstico do sistema de transporte escolar rural, utilizando-se da técnica de mineração de dados. Dado o tempo disponível para o desenvolvimento da pesquisa, limitou-se o escopo a entender e conhecer as práticas adotadas pelos municípios, para a gestão do transporte escolar rural diante da realidade e, principalmente, identificar os procedimentos adotados pelos gestores, para resolverem as dificuldades e os problemas de rotina do transporte escolar rural exclusivo para alunos, incluindo os portadores de necessidades especiais. Mostrando, assim, o potencial da ferramenta, enquanto a sua não utilização permite somente análise descritiva, como apresentado pelo CEFTRU (2007b).

6 CONCLUSÕES

6.1 APRESENTAÇÃO

Este trabalho teve como objeto de estudo a proposição de uma metodologia do sistema de apoio à tomada de decisão, que se utiliza da mineração de dados para exploração e análise de grandes bases de dados, com o objetivo de descobrir padrões e regras significativas, adequadas às necessidades de informação, que sirvam de base à tomada de decisão em estudo de transportes, durante todo o processo de planejamento, gestão e controle.

Os principais objetivos da mineração de dados são descobrir relacionamentos entre os dados e fornecer subsídios, para que se possa fazer uma previsão de tendências futuras baseadas no passado. Esse processo pode ser incorporado a alguma etapa do planejamento e os seus resultados serão essenciais para a estruturação das etapas subsequentes, dentro do processo de tomada de decisão.

Este capítulo mostra os principais resultados observados ao longo do desenvolvimento dessa pesquisa e foi estruturado em três seções. Na seção 8.2, apresentam-se as principais considerações em relação à aplicação da metodologia proposta. Logo depois, na seção 8.3, apresentam-se as principais avaliações e conclusões relacionadas à própria metodologia proposta e suas etapas, já se considerando sua aplicação no estudo de caso, como forma de validação da hipótese adotada no capítulo 1, bem como o cumprimento dos objetivos para seu desenvolvimento. E, finalmente na seção 8.4, apresentam-se algumas recomendações e as sugestões para trabalhos futuros.

6.2 CONSIDERAÇÕES SOBRE A APLICABILIDADE DA METODOLOGIA

A metodologia proposta foi aplicada na base de dados da pesquisa web (CEFTRU, 2007b), com objetivo de realizar a caracterização do transporte escolar rural nos municípios brasileiros, utilizando-se da mineração de dados com aplicação da tarefa de regras de associação, para descobrir quais são as práticas e os procedimentos adotados na gestão do transporte escolar rural.

Alguns gestores, indicados pelos prefeitos, tiveram grandes dificuldades no preenchimento de algumas questões do questionário (veículos utilizados; qualidade dos serviços; quantidade de alunos atendidos pelo transporte escolar etc.). Assim, alguns dados não puderam ser incorporados ao banco de dados. Como ressalva, declare-se que os dados apresentados no questionário são de inteira responsabilidade dos municípios. Assim, é preciso considerar a possibilidade de incoerência nos dados fornecidos, gerando inconsistência nos resultados que, neste caso, não serão frutos de falhas na metodologia proposta.

O principal mérito dessa metodologia foi estabelecer uma estrutura de *análise analítica* que buscasse representar, de forma clara, os elementos envolvidos na caracterização do transporte escolar rural, segundo as práticas e procedimentos adotados pelos municípios, para gestão do transporte escolar rural, no que se refere ao planejamento e a operação do serviço, tendo como referencial o ano de 2006.

Essa caracterização de procedimentos poderá servir de base de apoio à tomada de decisão na formulação de políticas públicas ao setor, com o objetivo de melhorar o serviço ofertado, além de permitir um maior conhecimento, por parte dos gestores, da situação atual do transporte escolar rural brasileiro, tendo com o referencial o ano de 2006.

Após várias interações na base de dados do transporte escolar rural (CEFTRU, 2007b), utilizando-se de diferentes valores de *'confiança', 'suporte' e 'lift'*; chegou-se às seguintes considerações: a aplicação da tarefa de regra de associação mostrou uma correlação entre algumas variáveis, visto que na maioria dos gestores municipais, observou-se um serviço de transporte escolar rural exclusivo, com veículos não adaptados aos alunos com necessidades especiais. Então, após essa análise, fica evidente que a prática adotada por esses gestores mostra-se, de modo geral, inadequada e incompatível às características físicas dos alunos.

Essa metodologia demonstrou, de forma clara, que a análise exploratória dos dados é independente. Na seção 5.3.1 - Apresentação resumida da caracterização do TER - relatório web, utilizou-se a *análise estatística descritiva*, para descrever o objeto de estudo, que é o transporte escolar rural, tendo resultados esquematizados em forma de gráficos e tabelas. Por outro lado, na seção 5.4.4 - Aplicação da Metodologia Proposta, foi utilizada outra forma de análise, a *analítica*, que buscou transformar dados em *informação* e

conhecimento útil e compreensível; para dar suporte à tomada de decisão no planejamento, na gestão e no controle. Enfim, em um ambiente de análise completo, necessita-se de resultados de ambos os tipos de análises. Logo, as análises são *complementares* e *não sobrepostas.*

A metodologia para descoberta de padrões e regras significativas, abordada no capítulo 4, poderá estar associada a outras etapas do processo de planejamento, para um sistema de transporte. Nos instrumentos de pesquisa, especialmente os questionários, antes da sua confecção, puderam utilizar-se da técnica de seleção de variáveis, para aperfeiçoar a relação de informação entre as *entradas* e as *saída*s de algum modelo, reconhecendo informações relevantes, a respeito da situação desejada, de forma conjunta com as informações da situação real.

6.3 AVALIAÇÃO DA METODOLOGIA PROPOSTA

Esta seção teve por objetivo a avaliação da viabilidade da metodologia proposta, no sentido de verificar a comprovação da hipótese e a constatação do cumprimento dos objetivos traçados para a pesquisa.

A hipótese adotada nesse trabalho baseou-se em sistema de apoio à tomada de decisão que utilizasse o processo de mineração de dados, para analisar grande quantidade de dados, a fim de descobrir padrões e regras significativas, que pudessem gerar conhecimentos úteis e compreensíveis, para subsidiar o processo de tomada de decisão ao diversos níveis de planejamento. Desse modo, permitiu-se a caracterização do estado do objeto de estudo, que contemplasse as diferentes percepções dos grupos alvos envolvidos no sistema, a fim de que se auxiliasse na elaboração de planos de ações para melhorar o serviço oferecido.

A aplicação da metodologia demonstrou a comprovação da hipótese formulada no capítulo 1, ou seja, pôde-se considerar que a inclusão do processo de mineração de dados, na etapa de planejamento, pôde contribuir para identificar padrões de procedimentos entre os municípios, na gestão do transporte escolar rural. Assim, a pesquisa alcançou seu objetivo de estruturar uma metodologia, para identificar padrões e regras significativas em banco de dados dos sistemas de informação de transporte, inseridos dentro de um contexto mais amplo de planejamento de transportes.

A metodologia proposta, no capítulo 4, possibilita confirmar a integração dos conceitos de planejamento de transporte e de diagnóstico, com a aplicação dos conceitos de sistema de apoio à tomada de decisão, que utiliza o processo de mineração de dados, para analisar grande quantidade de dados nos sistemas de informação de transporte.

A metodologia demonstrou, também, a importância e a necessidade de elaboração da estrutura semântica pertinente ao objeto de estudo, a partir do qual, foram definidos os seus elementos constituintes e os seus respectivos elementos de representação, o que permitiu estruturar a caracterização e o reconhecimento dos padrões de procedimentos relacionais; conforme o que foi explicitado na seção 5.2.1 - Representação do conhecimento, a qual expôs a existência de uma correlação entre algumas variáveis (ex. alunos com necessidade especiais, veículo exclusivo para aluno, veículos sem adaptação adequada aos portadores de necessidades especiais).

Quanto ao objetivo principal, apresentou-se o desenvolvimento de uma metodologia para elaboração de um sistema de apoio à tomada de decisão, que utilizasse o processo de mineração de dados, a fim de gerar conhecimento útil e compreensível, para subsidiar as decisões mais fundamentadas e inteligentes, nos diversos níveis de planejamento de transporte. No que se refere a essa técnica, foi possível o seu cumprimento, haja vista que, como apresentado no capítulo 1 e comprovado no capítulo 5, a metodologia proposta foi estruturada a partir de bases consistentes, levantadas no referencial teórico e sistematizadas em forma de uma estrutura metodológica viável.

Quanto aos objetivos específicos, foram atendidos de acordo com a definição e implementação da etapa 4 da metodologia proposta. As atividades envolvidas, nessa etapa, basearam-se na interpretação dos padrões extraídos na base de dados da pesquisa web (CEFTRU, 2007b).

6.4 SUGESTÕES PARA FUTURAS PESQUISAS

Como sugestões para futuras pesquisas, apontam-se:

- O desenvolvimento de novas ferramentas de mineração de dados poderá utilizar-se de outras medidas de interesse objetiva em regras de associação e de aplicação, destas medidas em problema específico de transporte, com o objetivo de atenuar a dependência à técnica de mineração de dados em relação ao conhecimento do domínio de negócio;

- O desenvolvimento de novas metodologias de representação gráfica apresenta-se como fator de grande motivação, para pesquisa de processo de mineração de dados, com suporte em banco de dados espaciais (ex. aplicação em elementos geográficos, imagens de sensoriamento remoto para análise de rede de transporte) e dados complexos e semiestruturados (ex. imagens, textos, grafos, web, multimídia);

- Na ausência de recursos computacionais, recomenda-se o desenvolvimento de linguagem que especifique consultas e processos (iteratividade e interatividade das tarefas) em *ambiente de descoberta de conhecimento de base de dados – KDD*, uma vez que, o uso de diversas ferramentas e do controle do fluxo do processo necessita de um grande esforço por parte do analista de negócio.

- Devido ao enorme número de algoritmos de regras de associação, recomenda-se ainda, nas bases de dados grande e complexa, um estudo comparativo entre aqueles existentes para aplicações em transporte;

- Estudos sobre novos perfis de profissionais da área de transportes com conhecimentos em informática, principalmente, *padronização, compatibilização de dados e tecnologia inteligentes*, tornam-se importantes e necessários à viabilização do uso e o do compartilhamento de dados, para estudos de transportes ao longo do tempo. Esses profissionais poderão contribuir com a integração, a comunicação entre as diversas fontes de dados e principalmente com a realização de análise exploratória de dados, *análise prospectiva e retrospectiva de dados*, para subsidiar a tomada decisão.

REFERÊNCIAS BIBLIOGRÁFICAS

ABNT (2002a) NBR 6023 – *Referências bibliográficas*. Associação Brasileira de Normas Técnicas, Rio de Janeiro, 2002. 24p.

ABNT (2002b) NBR 10520 – *Apresentação de citações em Documentos*. Associação Brasileira de Normas Técnicas, Rio de Janeiro, 2002. 7p.

ABNT (2002c) NBR 10520 – *Técnicas de coleta de dados*. Associação Brasileira de Normas Técnicas, Rio de Janeiro, 2002. 7p.

AFLORI, Cristian e LEON, Florin.(2004) *Efficient Distributed Data Mining using Intelligent Agents*. In: Proceedings of the 8th International Symposium on Automatic Control and Computer Science.

AGRAWAL, R. e SRIKANT, R. (1994) *Fast algorithms for mining association rules*. Proc. of the 20th Int'l Conference on Very Large Databases. Santiago, Chile.

ALMEIDA, Maria Christina Barbosa de. (2005) *Planejamento de bibliotecas e serviços de informação*. 2ª. ed. Brasília: Briquet de Lemos/livros.

BANISTER, D. (1998) *Transport Policy and the Enviroment*, E & FN Spon, London, England and New York, USA

BARBETTA, P. A. (2001) *Estatística aplicada a ciências sociais*. Florianópolis: Editora da UFSC.

BERRY, M. e LINOFF, G. (1997) *Mastering Data Mining*: The Art and Science of Customer Relationship Management; John Wiley e Sons, Inc.; USA, 1997.

BRASIL (1988) *Constituição da República Federativa do Brasil*. Brasília, DF: Senado. Diário Oficial da República Federativa do Brasil, Brasília, 5 de outubro de 1988.

BREWKA, G. (1996) *Well-founded semantics for extended logic programs with dynamic preferences.* Journal of Artificial Intelligence Research, 4, 1996. Disponível em: <http://www.jair.org/papers/paper284.html>. Acesso em: 10/03/2012.

BRIN, S. et. al.(1997). *Dynamic Itemset Counting and Implication Rules for Market Basket Data.* Estados Unidos: ACM SIGMOD. Disponível em: http://www.liaad.up.pt/~amjorge/docs/AssociationRules/Dynamic_Itemset_Counting.pdf > Acesso em: 06 jan. 2012.

BRUHA, I. e FAMILI, A. (2000) *Postprocessing in machine learning and data mining*. ACM SIGKDD Explorations Newsletter, v. 2, p. 110-114.

BRUTON, M. (1979) *Introdução ao planejamento dos transportes*. Tradução: João Bosco Furtado Arruda, Carlos Braune, César Cals de Oliveira Neto. São Paulo : Editora da Universidade de São Paulo.

CARVALHO, Willer Luciano. (2011) *Metodologia de Análise para a localização de escola em áreas rurais*, Tese de Doutorado. Faculdade de Tecnologia, Departamento de Engenharia Civil e Ambiental, Universidade de Brasília, DF, 215p.

CEFTRU/FNDE (2007a) Centro Interdisciplinar de Estudos em Transporte e Fundo Nacional de desenvolvimento da Educação. *Projeto: Transporte Escolar Rural - volume I - Metodologia de caracterização do transporte escolar rural*. Brasília, DF. 2007.

CEFTRU/FNDE (2007b) Centro Interdisciplinar de Estudos em Transporte e Fundo Nacional de desenvolvimento da Educação. *Projeto: Transporte Escolar Rural - volume II - Levantamento de dados para a caracterização do transporte escolar – questionário web*. Brasília, DF. 2007.

CEFTRU/FNDE (2007c) Centro Interdisciplinar de Estudos em Transporte e Fundo Nacional de desenvolvimento da Educação. *Projeto: Transporte Escolar Rural - volume III – Tombo I - Caracterização do transporte escolar nos municípios visitados*. Brasília, DF. 2007.

CEFTRU/FNDE (2007d) Centro Interdisciplinar de Estudos em Transporte e Fundo Nacional de desenvolvimento da Educação. *Projeto: Transporte Escolar Rural - Relatório da Base de Fundamentos e Critérios para a Avaliação, Aperfeiçoamento e Desenvolvimento de Indicadores*. Brasília, DF. 2007.

CEFTRU/FNDE (2008a) Centro Interdisciplinar de Estudos em Transporte e Fundo Nacional de desenvolvimento da Educação. *Projeto: Transporte Escolar Rural – volume I - Diagnóstico do Transporte Escolar Rural*. Brasília, DF. 2008.

CEFTRU/FNDE (2008b) Centro Interdisciplinar de Estudos em Transporte e Fundo Nacional de desenvolvimento da Educação. *Projeto: Transporte Escolar Rural – volume II - Diagnóstico do Transporte Escolar Rural*. Brasília, DF. 2008.

CHIAVENATO, I. e SAPIRO, A. (2004) *Planejamento Estratégico: Fundamentos e*

Aplicações. Rio de Janeiro: Elseviver.

CHIAVENATO, I.(2003) *Teoria geral da administração: introdução*. 7. ed. Rio de Janeiro: Campos.

CORRADI, F. M. *et al*. (2001) *Nós, links, e redes*. Revista de Biologia e Ciências da Terra, v.1, n.1

CRISP-DM - Cross Industry Standard Process for Data Mining. Disponível em < www.crisp-dm.org>. Acesso em: 12 mar. 2012.

DIAS, Maria Madalena. (2001) *Um modelo de formalização do processo de desenvolvimento de sistemas de descoberta de conhecimento em banco de dados.* Tese Doutorado - Curso de Pós-Graduação em Engenharia de produção, Universidade Federal de Santa Catarina.

DIAS, Maria Madalena. (2002) *Parâmetros na escolha de técnica e ferramenta de mineração de dados*. Acta Scientiarum, Vol. 24, n.6, p.1715-1725.

FAYYAD, U. e Piatetsky-Shapiro, G. e Smyth, P. (1996a) *The KDD Process for Extracting Useful Knowledge from volumes of data*. Communications of the ACM. November 1996. Vol. 39. No. 11.

FAYYAD, U. M.; Piatesky-Shapiro, G. e Smyth, P. (1996b) *From Data Mining to Knowledge Discovery: An Overview*. In: Advances in Knowledge Discovery and Data Mining, AAAI Press.

FERRARI, C. (1979) *Curso de Planejamento Municipal Integrado*. 2ª. ed. Pioneira, São Paulo.

FNDE (2006) Fundo Nacional de Desenvolvimento da Educação. C*artilha de Planejamento para o Transporte Escolar*. Disponível em: < http://www.fnde.gov.br/index.php/programas-transporte-escolar> Acesso em: 12 mar. 2011.

GEIPOT (1995). Empresa Brasileira de Planejamento de Transporte. *Avaliação preliminar do Transporte Rural* – destaque para o segmento rural. Brasília, 278p.

GENG, L. e HAMILTON, H. J.(2006) *Interestingness Measures for Data Mining*: A Survey. ACM Computing Surveys, v. 38, n. 3. Disponível em: http://mmlab.ceid.upatras.gr/courses/ais_site/files/3%5CInterestingness%20measures%20for%20data%20mining-%20A%20survey.pdf > Acesso em: 02 jan. 2012.

GIACAGLIA, M. E. (1998) *Modelagem de dados para planejamento e gestão operacional de transportes*. São Paulo. 2008. Tese (Doutorado) — D. Engenharia de Transportes. Escola Politécnica da Universidade de São Paulo.

GOEBEL, M. e GRUENWALD, L. (1999). *A survey of data mining and knowledge discovery software tools*. ACM SIGKDD, San Diego, v. 1, n. 1, p. 20-33, 1999.

GOMES, Luiz Flavio *et al.* (2006). *Tomada de Decisão Gerencial: um enfoque multicritério*. 2 ed. São Paulo: Atlas, 2006.

GONÇALVES, Eduardo Corrêa. (2005) *Regras de associação e suas medidas de interesse objetiva e subjetiva*. UFF – Universidade Federal Fluminense. Disponível em : <http://www.dca.fee.unicamp.br/~gudwin/courses/IA009> Acesso em: 28 mar. 2011.

GUARINO, N. (1998) *Formal Ontology and Information Systems*. Amended version of a paper appeared in N. Guarino (ed.), Formal Ontology in Information Systems. Proceedings of FOIS'98, Trento, Italy, 6-8 June 1998. Amsterdam, IOS Press, pp. 3-15.

GUELL, J. M. F (1997) *Planificación Estratégica de Ciudades*. Barcelona: Editorial Gustavo Gili.

HAN, J. e KAMBER, M. (2001) *Data mining: concepts and techniques*. Morgan Kaufmann.

INEP (2009) Instituto Nacional de Estudos e Pesquisas Educacionais Anísio Teixeira. *Censo Escolar*. Disponível em: < http://portal.inep.gov.br/basica-censo> Acesso em 05/02/2012

LAUDON, K. C. e LAUDON, J. P. (2007) *Sistema de informação gerencial*. São Paulo: Pearson Prentice Hall.

LOPES, E. P. *et al.* (2008) *Transporte escolar como instrument de viabilização do acesso à educação: o que estabelecem as leis?* Texto para discussão, nº 1 in : CEFTRU - Centro Interdisciplinar de Estudos em Transporte, UNB, Elemento mínimo para a regulação do transporte escolar rural. Brasília. 2008.

LUCAS, M. E. C. (2001) *Contribuição para o Desenho de um Sistema de Informação de Inteligência Estratégica para Empresas Operadoras do Transporte Urbano: Elementos do Projeto Lógico*. Brasília. 2001. Dissertação (Mestrado em Engenharia de Transporte) — Faculdade de Tecnologia, Universidade de Brasília.

MAGALHÃES, M. T. Q. (2010) *Fundamentos para a pesquisa em transporte: reflexões*

filosóficas e contribuições da ontologia de Mário Bunge. Tese (Doutorado em Transportes) - Faculdade de Tecnologia, Universidade de Brasília, Brasília.

MAGALHÃES, M. T. Q. e YAMASHITA, Y. (2008) *Repensando o Planejamento.* Texto para discussão n.4. CEFTRU/ UnB: Brasília, 2008.

MARCONI, Marina de Andrade e LAKATOS, Eva Maria. (2009) *Metodologia do Trabalho Científico: procedimentos básicos pesquisa*. 7.ed. São Paulo : Atlas.

MEC/INEP(2007) Ministério da Educação. Disponível em: < http://portal.mec.gov.br/index.php?option=com_content&view=article&id=273&Itemid=345 > Acesso em: 05/03/2012.

MEDEIROS, L. F. (2003) *Redes Neurais em Delphi*. Florianópolis: Visual Books.

MELLO, J. C. (1981) *Planejamento dos Transportes Urbanos*. Rio de Janeiro: Campus.

MORESI, E.A.D. (2000) *Delineando o valor do sistema de informação de uma organização*. Ciência da Informação, Jan./Apr. 2000, Vol. 29.

MORLOK, E. K. (1978) *Introduction to Transportation Engineering and Planning*. New York: McGraw-Hill.

MPOG (2006) – Ministério de Planejamento Orçamento e Gestão. *Manual de Elaboração de Programas - Plano Plurianual 2004-2007*. Brasília, Brasil.

MULCAHY, Rita. (2007) *Preparatório para exame de PMP. 3.ed. Tradução : Roberto Pon.*. EUA: RMC Publications Ins.

NONAKA, I e TAKEUCHI, H. (1997) *Criação de conhecimento na empresa*. Rio de Janeiro: Campus.

ORTÚZAR, J. D.; WILLUMSEN, L. G. (2001) *Modelling Transport.* (3a. ed) Wiley-Blackwell.

PAPACOSTAS C. S. e PROVEDOUROS, P. D. (1993) *Transportation Engineering and Planning*. 2ª. ed. New Jersey: Prentice Hall.

PENDER, Tom. (2004) *UML, a Bíblia*. Tradução por Daniel Vieira. Rio de Janeiro: Elsevier, 2004.

PICHILIANI, M.(2006) *DataMining na Prática: Árvores de Decisão.* Disponível em: < http://imasters.com.br/artigo/5130/sql_server/data_mining_na_pratica_arvores_de_decisao/ > Acesso em: 14 jul. 2011.

PINHEIRO, Carlos André Reis (2008). *Inteligência Analítica: Mineração de dados e descoberta de conhecimento*. Rio de Janeiro: Editora Ciência Moderna.

QUILLIAN, M. R. (1968) *Semantic memory*. In: Minsky, M. (ed.). Semantic information processing. MIT Press, Cambridge, MA, USA.

RIOS, J. A. (2003) *Ontologias: alternativa para a representação do conhecimento explícito organizacional.* VI CINFORM - Encontro Nacional de Ciência da Informação, Salvador.

SANTOS, M. Filipe e AZEVEDO S. Carla. (2005) *Data Mining - Descoberta de Conhecimento em Bases de Dados*. Portugal : Editora FCA.

SEMMA. SAS Enterprise Miner .Disponível em <http://www.sas.com/offices/europe/uk/technologies/analytics/datamining/miner/semma.html>. acesso em: 18 jul. 2011.

SILBERSCHATZ, A e TUZHILIN (2006) *On subjective measures of Interestingness in knowledge discovery. ACM* Computing, *Montreal*, p. 275-281. Disponível em : < http://aaaipress.org/Papers/KDD/1995/KDD95-032.pdf> Acesso em: 28 dez. 2011.

SIMON, H. A. (1969) *The new science of management decision*. New Yord: Harper e Row.

SOUZA, Welder Maurício de. (2004) *Aplicação de mineração de dados para o levantamento de critérios do Programa Nacional do Transporte Escolar. Brasília*. 168f. Dissertação (Mestrado em Engenharia de Transporte) - Faculdade de Tecnologia, Universidade de Brasília.

SOWA, J. F. (1992) *Semantic Networks*. Versão revisada e extendida de um artigo redigido para a Encyclopediaof Artificial Intelligence, editada por Stuart C. Shapiro, Wiley, 1987. Disponível em: <http://www.jfsowa.com/pubs/semnet.htm>. Acesso em: 10/03/2012.

SOWA, J. F. (2000) *Knowledge Representation: logical, philosophical, and computational foundations*. Pacific Grove: Brooks, Cole.

TAIT, T. F. C. (2000) *Um Modelo de Arquitetura de Sistemas de Informação para Setor Público*: *Estudo em Empresas Estatais Prestadoras de Serviços de Informática*. Florianópolis. 2000. Tese (Doutorado em Engenharia de Produção) - Engenharia de Produção e Sistemas, Universidade Federal de Santa Cantarian - UFSC.

TEDESCO, G. M. I. (2008) *Metodologia para Elaboração do Diagnóstico de um Sistema*

de Transporte. Brasília. 2008. 215p. Dissertação (Mestrado em Engenharia de Transporte) — Faculdade de Tecnologia, Universidade de Brasília.

TEIXEIRA, G. L. (2003) *Uso de Dados Censitários para Identificação de Zonas homogêneas para Planejamento de Transportes utilizando Estatística Espacial*. Brasília Tese(Doutorado) — Universidade de Brasília.

VASCONCELLOS, E. A. (2000) *Transporte Urbano nos Países em Desenvolvimento:* reflexões e propostas. 3ª. ed. São Paulo: FAPESP.

WEKA - University of Waikato. *Weka 3.6 – Machine Learning Software in Java.* Disponível em < http://www.cs.waikato.ac.nz/ml/weka> Acesso em: 10/02/2011.

WITTEN, I. H., AND FRANK E. (1999) *Data Mining: Practical Machine Learning Toolsand Techniques with Java Implementations*. San Francisco, 1999.

APÊNDICES

A - MANUAL SIMPLIFICADO DO SOFTWARE MINERAÇÃO DE DADOS

Software : Transporte Mining

Esse software foi desenvolvimento sobre a biblioteca completa do *weka* e possui uma interface gráfica amigável e seus algoritmos fornecem *relatórios com dados analíticos* e *estatísticos* do domínio minerado. Sua interface gráfica é muito mais fácil de utilização porque as informações sobre a opção de seleção do arquivo e os algoritmos estão em cada tela para uma tarefa específica de mineração.

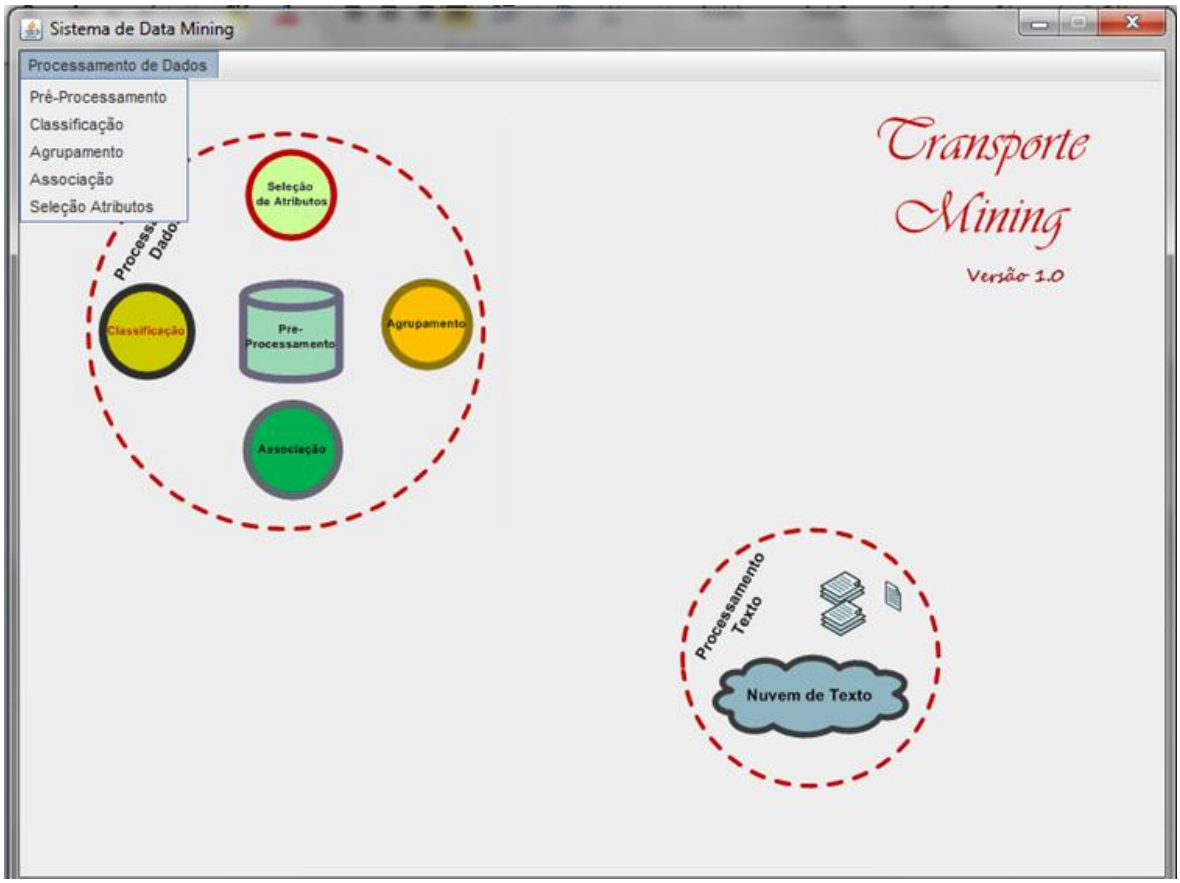


Figura Apêndice A. 1: Tela principal do software

Telas principais

Embora seja intuitiva, para uma abordagem inicial faz-se necessário identificar os elementos principais das interfaces.

Tela de pré-processamento

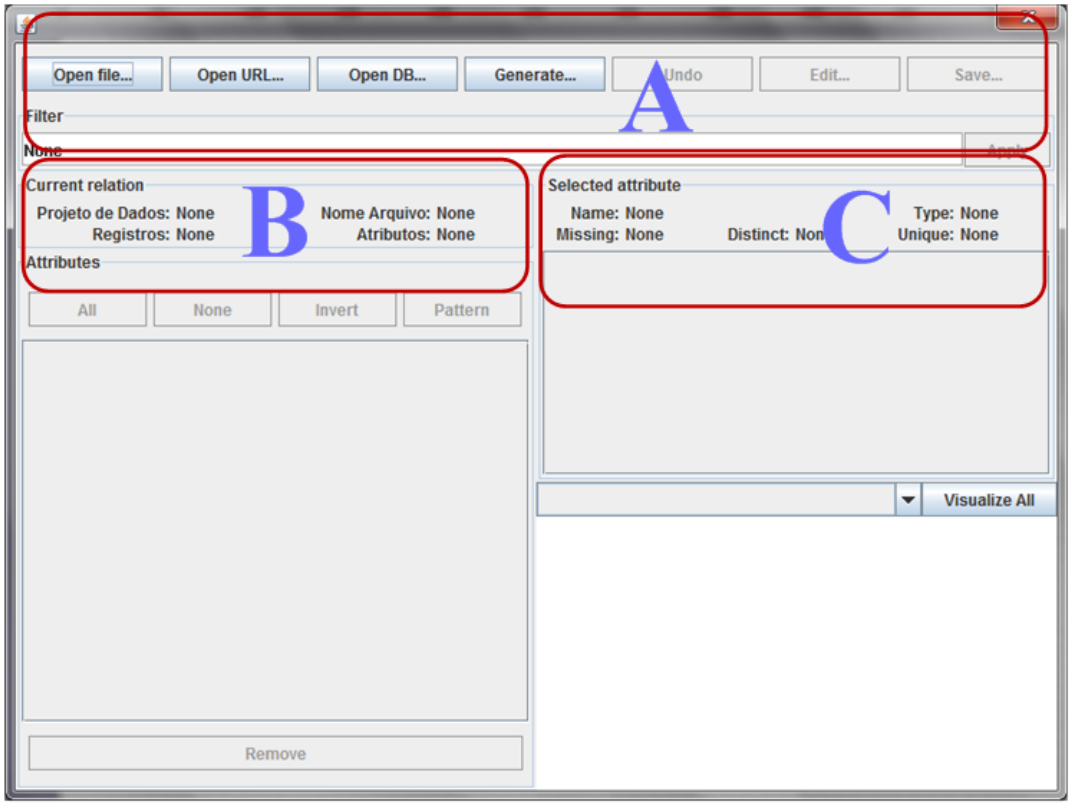


Figura Apêndice A. 2: Tela da tarefa pré-processamento

(A) Open *File*, *Open URL*, *Open DB*: através destes botões é possível selecionar, respectivamente, bases de dados a partir de arquivos locais (formato .arff), bases remotas (Web), e diferentes bancos de dados (via JDBC).

(B) No botão filter é possível efetuar sucessivas filtragens de atributos e instâncias na base de dados previamente carregada (seleção, discretização,normalização, amostragem, dentre outros);

(C) Navegando interativamente pelos atributos (quadro Attributes) é possível obter informações quantitativas e estatísticas sobre os mesmos (quadro Selected attribute);

Tela de Classificação:

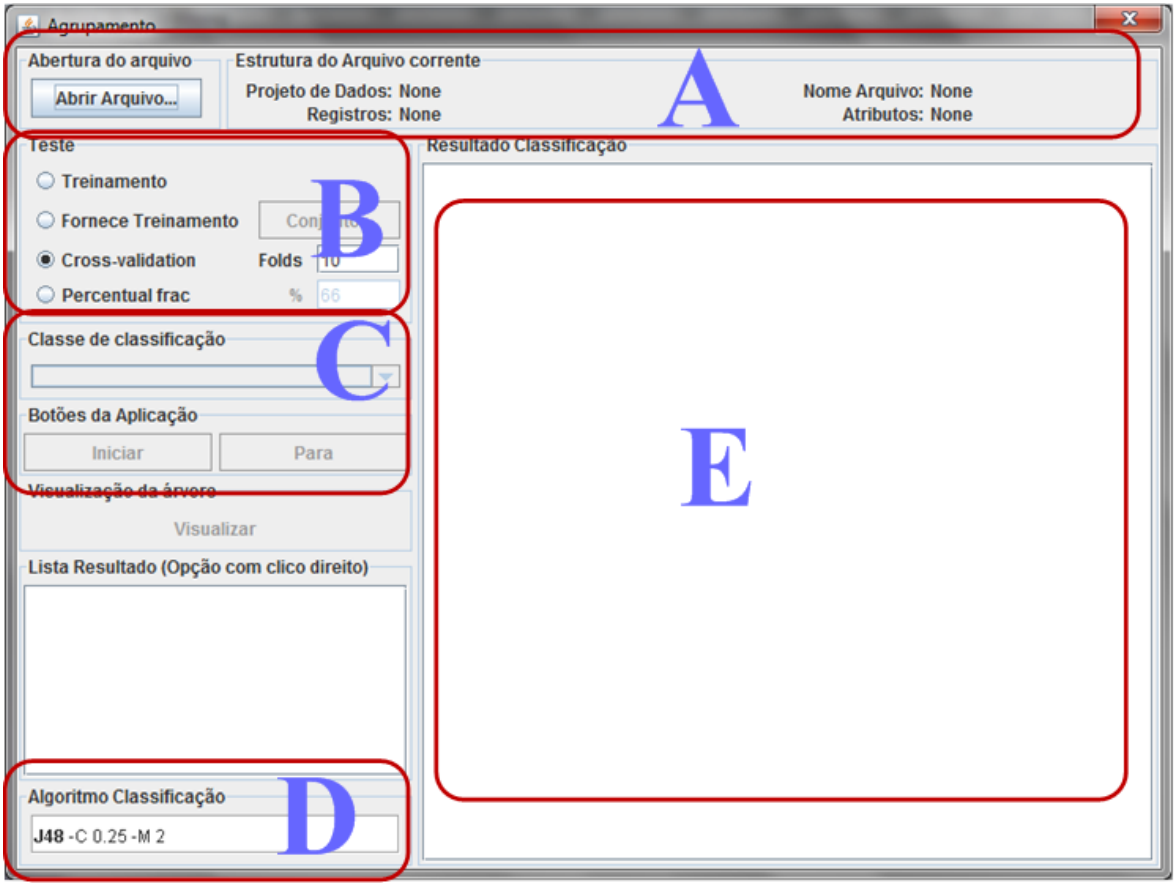


Figura Apêndice A. 3: Tela da tarefa de classificação

(D) O botão abrir arquivo: carrega bases de dados a partir de arquivos locais (formato.arff).

(E) Permite selecionar a opção de teste e validação do modelo gerado (o próprio conjunto de dados do treinamento, fornece de conjunto teste, crossvalidation, percentual conjunto de treinamento para teste);

(F) Seleção do atributo classe para a tarefa de classificação e botões da aplicação;

(G) Mostra o algoritmo a ser utilizado (J48);

(H) Resumo da tarefa efetuada, com dados estatísticos, modelo, matriz de confusão etc.

Tela de Agrupamento:

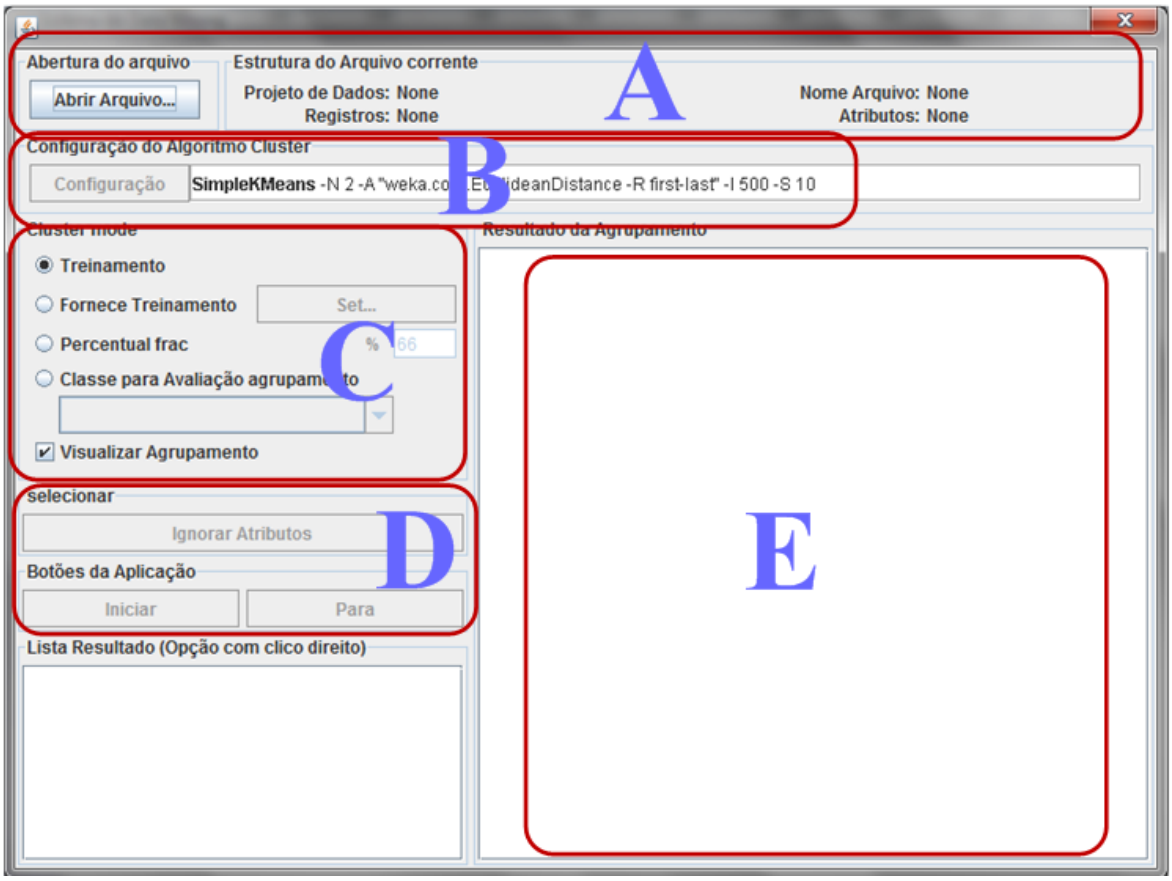


Figura Apêndice A. 4: Tela da tarefa de agrupamento

(A) O botão abrir arquivo: carrega bases de dados a partir de arquivos locais (formato.arff).

(B) Parametrização do algoritmo K Means;

(C) Permite selecionar a opção de teste e validação do modelo gerado (o próprio conjunto de dados do treinamento, outro conjunto só para testes, cross-validation, separar parte do conjunto de treinamento para teste);

(D) Seleção do atributo classe para a tarefa de agrupamento e botões da aplicação;

(E) Resumo da tarefa efetuada, com dados estatísticos, modelo etc.

Tela de associação:

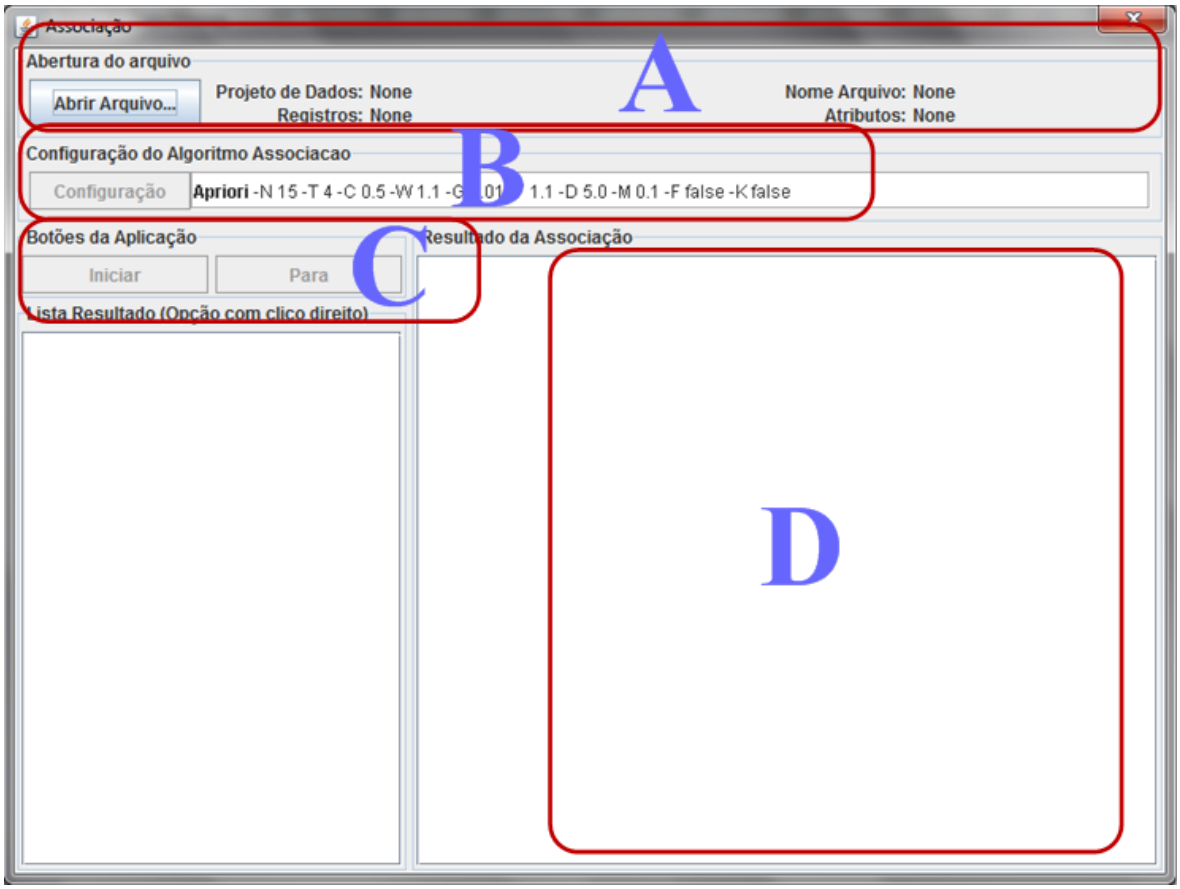


Figura Apêndice A. 5: Tela da tarefa de associação

(A) O botão abrir *arquivo*: carrega bases de dados a partir de arquivos locais *(formato.arff).*

(B) Parametrização do *algoritmo* a*priori*

(C) botões da *aplicação*;

(D) Resumo da tarefa efetuada, com *dados estatísticos, modelo* etc.

ANEXOS

A – QUESTIONÁRIO WEB SIMPLIFICADO

PARTE A: IDENTIFICAÇÃO

Nesta primeira parte, você deverá fornecer dados sobre a identificação de seu município e do responsável pelo preenchimento do formulário e a caracterização do transporte escolar fornecido.

1. Identifique o seu município preenchendo os campos abaixo com a sigla do estado e o nome de seu município.

Estado:	Município:

2. Identifique o responsável pelas informações (setor ou órgão onde trabalha e o nome) e o responsável pelo preenchimento do formulário (nome e cargo, e-mail e telefone para contato).

2.1. Responsável pelas informações Setor/Órgão: ☐ Órgão Municipal de Transportes ☐ Órgão Municipal de Educação ☐ Prefeitura Municipal (Gabinete do Prefeito) Nome:	
Cargo:	
2.2. Responsável pelo preenchimento do formulário Nome:	
Cargo:	
Telefone para contato: ( )	e-mail:

3. Informe se há transporte escolar rural e/ou urbano fornecido por seu município. Informe o mês e ano de início do transporte escolar (rural e/ou urbano), o tipo de aluno atendido e seu nível escolar, assim como o tipo de serviço oferecido.

3.1. Transporte Escolar Rural

3.1.1. Existe transporte escolar rural em seu município? ☐ Não ☐ Sim Desde: / (Mês/Ano)
3.1.2. Quais alunos são atendidos pelo transporte escolar rural no seu município? ☐ Alunos com desenvolvimento típico ☐ Educação Infantil ☐ Ensino Fundamental ☐ Ensino Médio ☐ Ensino Superior ☐ Educação de Jovens e Adultos ☐ Alunos com necessidades educacionais especiais ☐ Educação Infantil ☐ Ensino Fundamental ☐ Ensino Médio ☐ Ensino Superior ☐ Educação de Jovens e Adultos
3.1.3. Qual o tipo de serviço de transporte escolar rural oferecido no seu município? ☐ Transporte escolar com veículo exclusivo, fornecido pelo município: ☐ Exclusivo para alunos ☐ Alunos e Passageiros Comuns ☐ Transporte escolar com veículo exclusivo, fornecido por particulares (contrato entre o aluno e o prestador do serviço) ☐ Transporte coletivo regular ☐ Outros:

Figura Anexo A. 1: Questionário web – Parte A. Fonte (CEFTRU, 2007a, 2007b)

PARTE B: SERVIÇO DE TRANSPORTE ESCOLAR

Nesta parte, serão levantados dados sobre os serviços de transporte escolar de seu município referentes ao ano de **2005**. Você deverá preencher todos os campos, mesmo aqueles cuja informação não esteja facilmente disponível. Neste caso, consulte os documentos e/ou a(s) pessoa(s) mais indicada(s) para fornecer a informação correta e, somente após isso, digite os dados solicitados.

I – Caracterização do serviço fornecido pelo município

1. Informe qual a propriedade do veículo utilizado para o serviço de transporte escolar fornecido pelo município.

☐ Frota própria (veículos do município) ☐ Frota terceirizada (contratada pela prefeitura)

2. Informe se no município os veículos exclusivos para o transporte são utilizados para outros fins quando não estão transportando escolares.

2.1. Quando o veículo exclusivo para transporte escolar não está sendo usado para transportar os alunos, ou pq eles estão dentro da escola (em aula), ou pq é feriado ou final de semana, ele é utilizado para outros fins? ☐ Sim ☐ Não
2.2. Em caso afirmativo, especifique para quais fins os veículos do transporte escolar estão sendo utilizados. 1. _____ 2. _____ 3. _____

3. Informe se o município oferece o transporte escolar durante todos os dias do ano em que os alunos têm aula.

3.1. O transporte escolar é oferecido durante todo o período letivo? ☐ Sim ☐ Não
3.2. Em caso negativo, diga os motivos pelo qual o serviço de transporte escolar não é oferecido durante todo o período letivo. 1. _____ 2. _____ 3. _____

Figura Anexo A. 2: Questionário web – Parte B. Fonte (CEFTRU, 2007a, 2007b)

PARTE C: CLIENTELA

Nesta parte, você deverá preencher dados relativos às escolas, aos alunos e aos professores que foram ou não atendidos pelo transporte escolar do seu município, em **2005**. Para isso, você poderá consultar dados do censo escolar 2005 e usá-los quando necessário.

I – Tipo de Clientela

1. Informe o número de escolas que têm alunos atendidos pelo transporte escolar em seu município, de acordo com a localização.

	Quantidade de escolas de acordo com sua localização			
☐ Transporte escolar com veículo exclusivo, fornecido pelo município	☐ Rural	_____	☐ Urbana	_____
☐ Transporte escolar com veículo exclusivo, fornecido por particulares (contrato entre o aluno e o prestador do serviço)	☐ Rural	_____	☐ Urbana	_____
☐ Transporte coletivo regular	☐ Rural	_____	☐ Urbana	_____

2. Informe a quantidade de alunos atendida pelo transporte escolar com veículo exclusivo fornecido pelo município, de acordo com a área de sua residência (rural ou urbana). Especifique por turno em que os alunos são atendidos e informe, também, quantos professores, servidores e outras pessoas utilizam o transporte escolar.

2.1. Turno da manhã	Quantidade de alunos (e outros) por local de residência			
☐ Alunos	☐ Rural	_____	☐ Urbano	_____
☐ Professores (quando houver)	☐ Rural	_____	☐ Urbano	_____
☐ Servidores da escola (quando houver)	☐ Rural	_____	☐ Urbano	_____
☐ Outros: _____	☐ Rural	_____	☐ Urbano	_____

2.2. Turno da tarde	Quantidade de alunos (e outros) por local de residência			
☐ Alunos	☐ Rural	_____	☐ Urbano	_____
☐ Professores (quando houver)	☐ Rural	_____	☐ Urbano	_____
☐ Servidores da escola (quando houver)	☐ Rural	_____	☐ Urbano	_____
☐ Outros: _____	☐ Rural	_____	☐ Urbano	_____

2.3. Turno da noite	Quantidade de alunos (e outros) por local de residência			
☐ Alunos	☐ Rural	_____	☐ Urbano	_____
☐ Professores (quando houver)	☐ Rural	_____	☐ Urbano	_____
☐ Servidores da escola (quando houver)	☐ Rural	_____	☐ Urbano	_____
☐ Outros:	☐ Rural	_____	☐ Urbano	_____

Figura Anexo A. 3: Questionário web – Parte C. Fonte (CEFTRU, 2007a, 2007b)

PARTE D: RECURSOS UTILIZADOS

Os dados a serem preenchidos nesta parte dizem respeito aos recursos utilizados para o pagamento do transporte escolar fornecido pelo município, no ano de **2005**.

I – Fontes de recursos utilizados

1. Quais os valores e as fontes (origem) de recursos utilizados para financiar o transporte escolar fornecido pelo seu município, no ano de **2005**?

☐ Valor do PNATE repassado ao município pela União:	R$
☐ Recurso do Estado repassado ao município:	R$
☐ Recurso próprio do município:	R$
☐ Outras fontes. Quais?	R$

2. Informe se existe convênio entre município e o estado para o transporte de alunos da dependência administrativa estadual.

2.1. Existe convênio entre o município e o Estado?		
☐ Sim	☐ Não	
2.2. Se existe, qual é o critério de repasse do recurso do estado e qual é o valor?		
☐ Valor por aluno	_____	(R$/aluno)
☐ Valor por quilômetro rodado	_____	(R$/km)
☐ Valor por quilômetro rodado transportando aluno	_____	(R$/km)
☐ Valor fixo mensal	_____	(R$)
☐ Outros:	_____	( )

II – Destinação dos recursos utilizados

1. Informe o valor e qual é a fonte (origem) dos recursos destinados ao transporte escolar fornecido pelo seu município. Especifique o valor, por área de abrangência, referente ao ano letivo de 2005 e também 2006 (até o dia 30 de junho).

No ano de 2005:	
☐ Transporte Escolar Rural	R$
Origem:	
☐ Município	R$
☐ Estado	R$
☐ União	R$
☐ Transporte Escolar Urbano	R$
Origem:	
☐ Município	R$
☐ Estado	R$

Figura Anexo A. 4: Questionário web – Parte D. Fonte (CEFTRU, 2007a, 2007b)

B – REDE SEMÂNTICA DO SISTEMA DE TRANSPORTE ESCOLAR RURAL SIMPLIFICADA

A Rede Semântica do Sistema de Transporte Escolar foi baseada no método produzido e aplicado na construção da Rede Semântica do Planejamento de Transportes, desenvolvida em projeto para o Ministério dos Transportes (CEFTRU, 2007a). A construção da Rede Semântica do Sistema de Transporte Escolar Rural foi feita uma compatibilização entre a Rede Semântica do Sistema de Transporte (Ceftru, 2007d) e a Árvore Conceitual do Transporte Escolar Rural (Ceftru, 2007c).

Segundo CEFTRU (2008a, 2008b), a Rede Semântica proposta para o Sistema de Transporte Escolar Rural – STER possui três elementos: (i) Elementos Físicos; (ii) Elementos Lógicos e (iii) Atores, conforme a Figura Anexo B. 1.

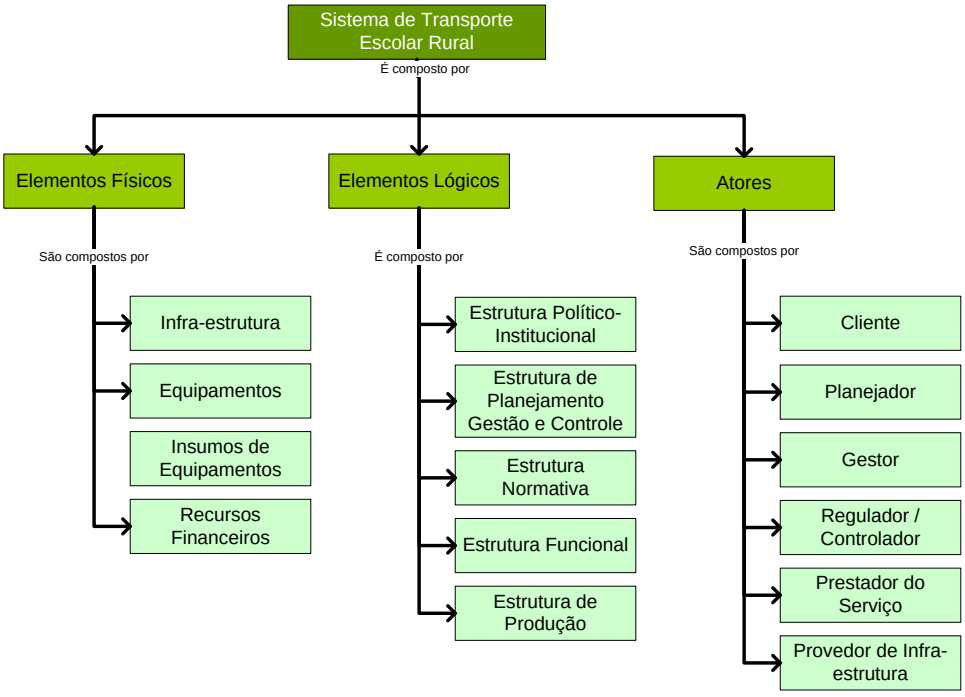


Figura Anexo B. 1: Rede semântica do STER. Fonte: (CEFTRU, 2008a, 2008b)

A seguir, são detalhados os elementos que compõem a Rede Semântica do STER.

1 - Elementos Físicos

Segundo CEFTRU (2008a, 2008b), os Elementos Físicos do Sistema de Transporte Escolar Rural compreendem: (i) Infra-estrutura; (ii) Equipamentos; (iii) Insumos dos equipamentos; e (iv) Recursos financeiros, conforme a Figura Anexo B. 2.

2 - Elementos Lógicos

Segundo CEFTRU (2008a, 2008b), os Elementos Lógicos do STER são as estruturas lógicas obrigatórias para execução dos processos internos do sistema. Estas estruturas se referem às relações institucionais, ao planejamento, à gestão, ao controle, à forma de organização, ao funcionamento e à operação do sistema, conforme a Figura Anexo B. 3.

3 - Elementos Lógicos

Segundo CEFTRU (2008a, 2008b), os Elementos Lógicos do STER são as estruturas lógicas obrigatórias para execução dos processos internos do sistema. Estas estruturas se referem às relações institucionais, ao planejamento, à gestão, ao controle, à forma de organização, ao funcionamento e à operação do sistema, conforme a Figura Anexo B. 3.

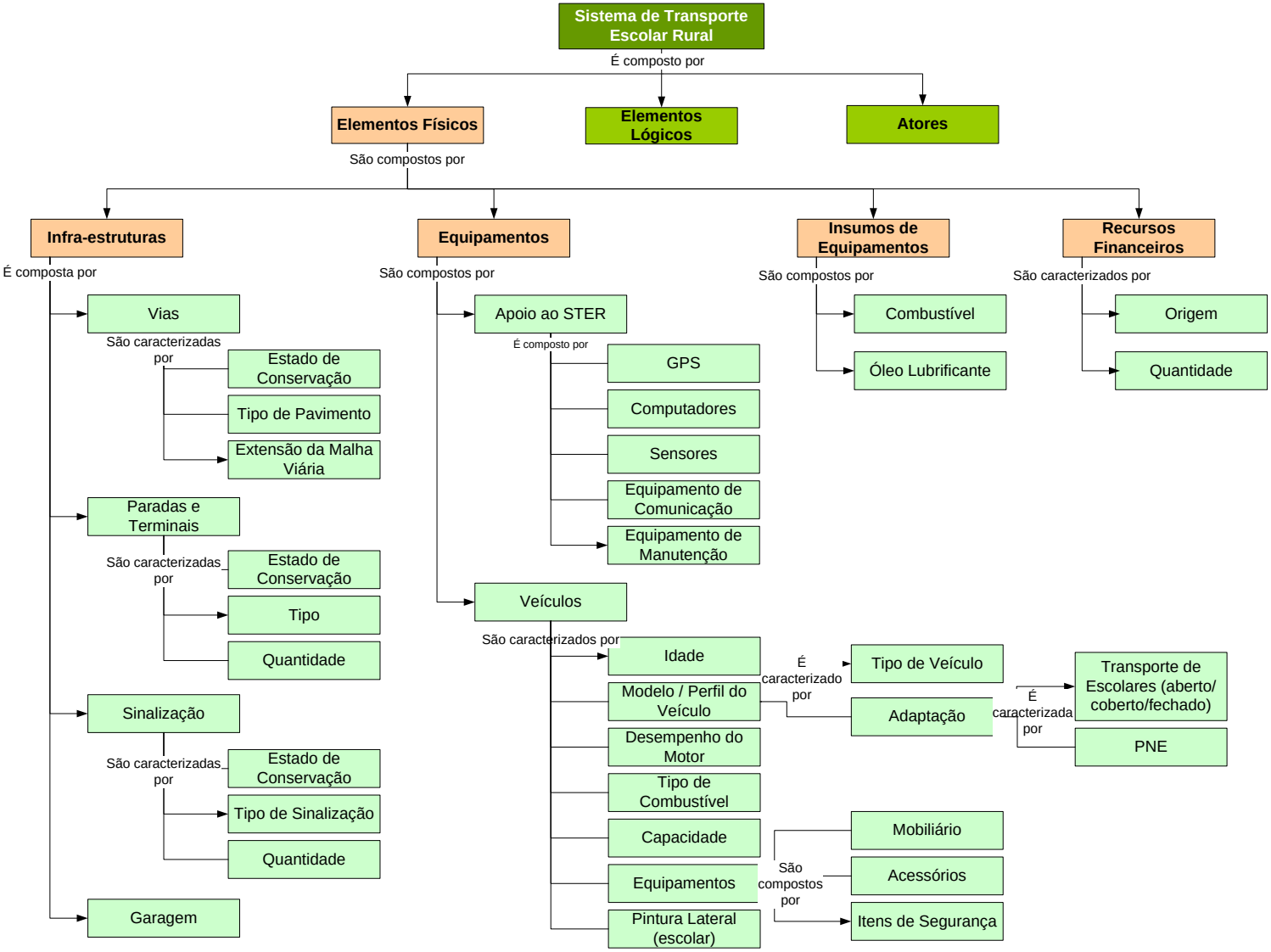


Figura Anexo B. 2: Elementos físicos. Fonte (CEFTRU, 2008a, 2008b)

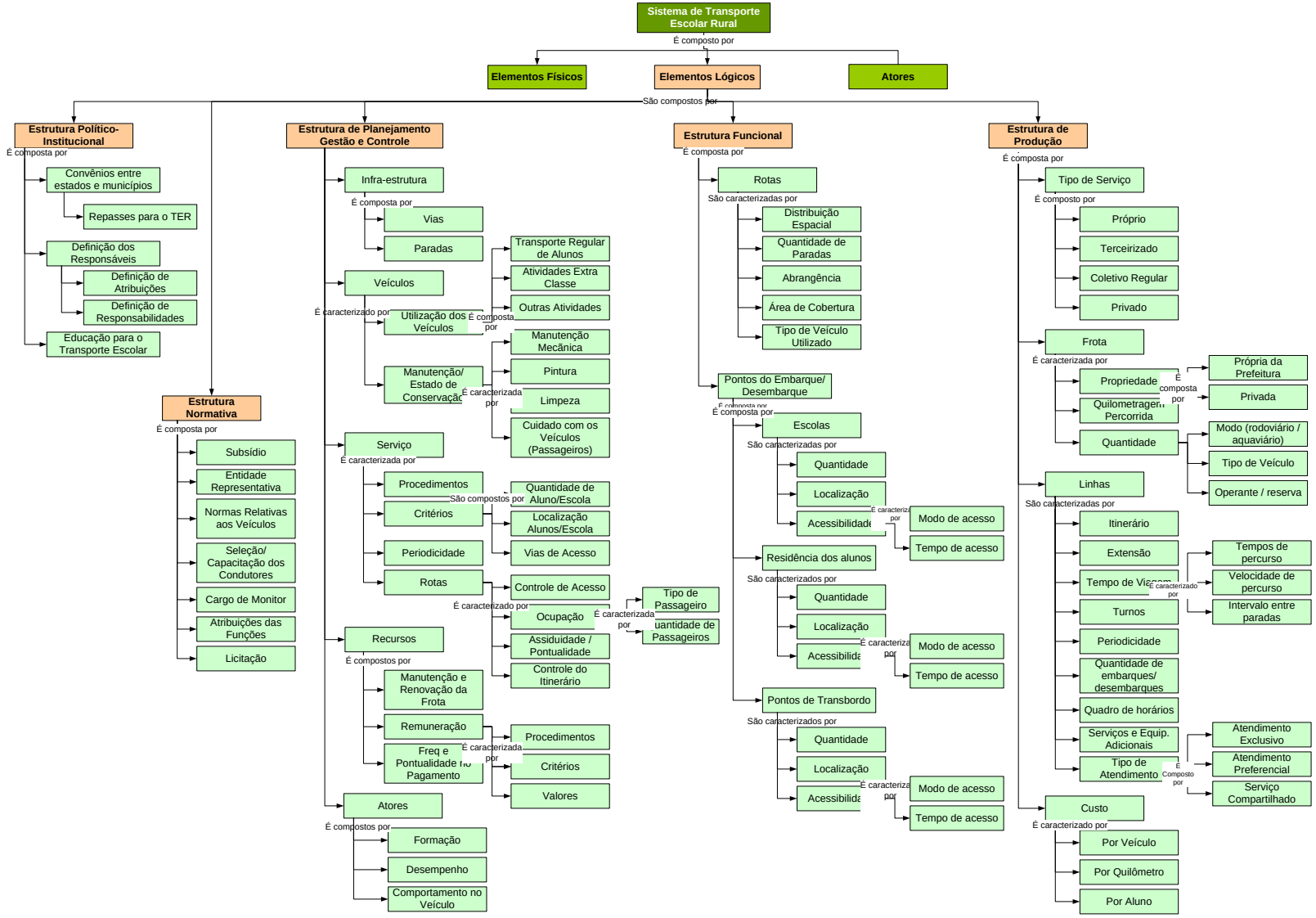


Figura Anexo B. 3: Elementos lógicos. Fonte (CEFTRU, 2008a, 2008b)

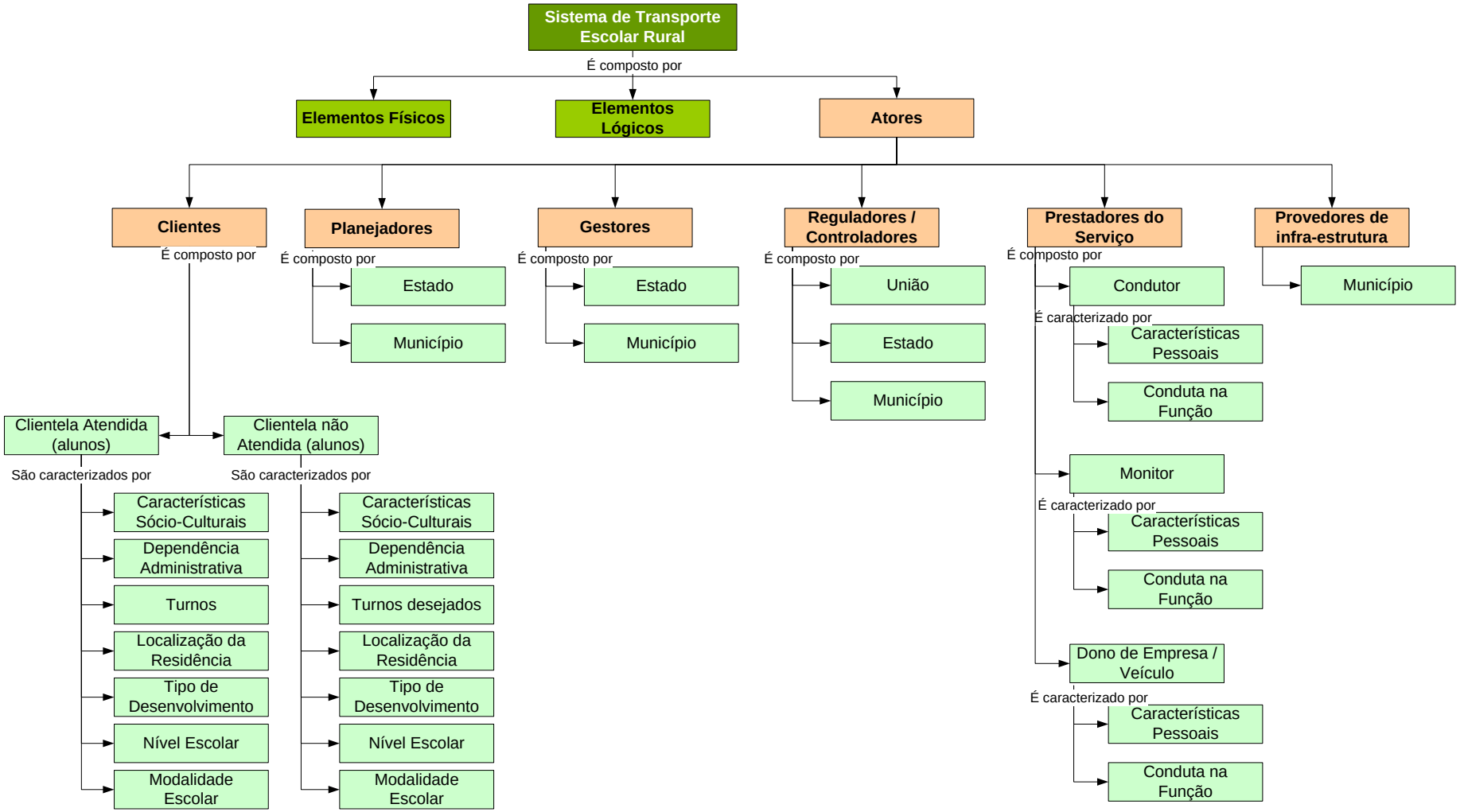


Figura Anexo B. 4: Atores. Fonte (CEFTRU, 2008a, 2008b)