PONTIFÍCIA UNIVERSIDADE CATÓLICA DE GOIÁS
PROGRAMA DE MESTRADO EM ENGENHARIA DE PRODUÇÃO E SISTEMAS

MINERAÇÃO DE DADOS APLICADA A CLASSIFICAÇÃO DOS CONTRIBUINTES DO ISS.

TIAGO LEVERGGER PICCIRILLI

2013

PONTIFÍCIA UNIVERSIDADE CATÓLICA DE GOIÁS
PROGRAMA DE MESTRADO EM ENGENHARIA DE PRODUÇÃO E SISTEMAS

MINERAÇÃO DE DADOS APLICADA A CLASSIFICAÇÃO DOS CONTRIBUINTES DO ISS.

TIAGO LEVERGGER PICCIRILLI

Dissertação de Mestrado apresentada ao Programa de Pós-Graduação em Engenharia de Produção e Sistemas da Pontifícia Universidade Católica de Goiás, como parte dos requisitos para obtenção do título de Mestre em Engenharia de Produção e Sistemas.

Orientador: Prof. Sibelius Lellis Vieira, Dr.

Goiânia
Abril, 2013.

MINERAÇÃO DE DADOS APLICADA A CLASSIFICAÇÃO DOS CONTRIBUINTES DO ISS.

Tiago Levergger Piccirilli

Esta dissertação julgada adequada para obtenção do título de Mestre em Engenharia de Produção e Sistemas, e aprovada em sua forma final pelo Programa de Pós-Graduação em Engenharia da Produção e Sistemas da Pontifícia Universidade Católica de Goiás em Abril de 2013.

_____

Prof. Ricardo Luiz Machado, Dr.
Coordenador do Programa de Pós-Graduação em Engenharia de Produção e Sistemas

Banca examinadora:

_____

Prof. Sibelius Lellis Vieira, Dr.
Orientador

_____

Prof. Alencar de Melo Junior, Dr.

_____

Profa. Maria Jose Pereira Dantas, Dra.

_____

Profa. Solange da Silva, Dra.

GOIÂNIA - GOIÁS

ABRIL DE 2013

Piccirilli, Tiago Levergger.
P589m Mineração de dados aplicada a classificação dos contribuintes do ISS [manuscrito] / Tiago Levergger Piccirilli. – 2013.
116 f. ; il. ; 30 cm.

Dissertação (mestrado) – Pontifícia Universidade Católica de Goiás, Programa de Mestrado em Engenharia de Produção e Sistemas, 2013.
"Orientador: Prof. Dr. Sibelius Lellis Vieira".

Referências bibliográficas: p. 98-103.

1. Mineração de dados (Computação). 2. Imposto sobre serviços. I. Título.

CDU: 004.45:336.22(043)

Dedico este trabalho a Deus por ter-me agraciado com saúde, fé e persistência, subsídios essenciais para realizar a condução deste trabalho.
A minha família em especial a minha mãe, maior exemplo de vida e base da minha sustentação.

AGRADECIMENTOS

A todos os amigos e familiares que, direta ou indiretamente, contribuíram para a construção deste trabalho.

Ao meu orientador professor Dr. Sibelius Lelis Vieira pela enorme presteza e disposição apresentada no decorrer de todo o processo do desenvolvimento deste trabalho de mineração de dados.

Aos professores componentes da banca examinadora dessa dissertação pelas contribuições que porventura aperfeiçoarão a qualidade deste trabalho.

Aos meus colegas do Departamento de Desenvolvimento de Sistemas, do Departamento de Arrecadação e diretores da AMTEC pela enorme disposição na colaboração deste trabalho e pelo companheirismo: Osmar, Miralho, Cortes, Cesar e Gabriel.

Em especial ao chefe do Departamento de Arrecadação, Cortes, por prover os recursos humanos necessários no entendimento do problema e extração dos dados resultando na oportunidade de desenvolver um conhecimento voltado ao Departamento Fiscal do Município.

“Nenhum obstáculo é grande demais quando confiamos em Deus”

(Aristóteles)

Resumo da Dissertação apresentada ao MEPROS/PUC Goiás como parte dos requisitos necessários para a obtenção do grau de Mestre em Engenharia da Produção e Sistemas (M.Sc.)

MINERAÇÃO DE DADOS APLICADA A CLASSIFICAÇÃO DOS CONTRIBUINTES DO ISS.

Tiago Levergger Piccirilli

Abril 2013.

Orientador: Prof. Sibelius Lellis Vieira, Dr.

A administração pública é responsável pela instituição, recebimento e controle de tributos pagos pelos contribuintes. Este recurso é imprescindível para manutenção de sua estrutura administrativa e estabelecimento de políticas públicas. Para aperfeiçoar o controle realizado pela administração é necessário investimento em novas tecnologias, visto que o departamento de fiscalização recebe constantemente inúmeros dados da movimentação econômica dos contribuintes e de regularização cadastral. Os recursos computacionais atuais armazenam informações com capacidade superior à condição humana de manipulação e extração de conhecimento. Nesse contexto, surge na ciência uma área denominada Mineração de Dados, especifica para extrair conhecimento e padrões desconhecidos por meio de bases de dados. Este trabalho apresenta um modelo para classificar os contribuintes do Imposto Sobre Serviços de Qualquer Natureza (ISS) que apresentaram alguma irregularidade, de posse dos recursos e técnicas da mineração. O trabalho foi realizado no Município de Goiânia na Secretaria de Finanças especificamente no departamento de Arrecadação, abrangendo o cenário apresentado no ano de 2011. Entre os modelos construídos com algoritmo de árvore de decisão, apresentou como resultado, a classificação dos contribuintes irregulares com um índice de acertos de 92,03%.

Palavras-chave: Mineração de dados, detecção de irregularidade, *ISS*.

Summary of Thesis submitted to MEPROS/PUC Goiás as part of the requirements for the degree of Master in Production and Systems Engineering (M. Sc.)

DATA MINING APPLIED TO THE CLASSIFICATION OF THE TAXPAYERS OF THE TOWN OF GOIÂNIA.

Tiago Levergger Piccirilli

April, 2013

Advisor: Prof. Sibelius Lellis Vieira*, Doctor.*

The Public Administration is responsible for the institution, receiving and control of taxes paid by taxpayers. This feature is indispensable to maintenance of its administrative structure and establishment public policies. To improve the control performed by the administration, it's necessary to invest in new technologies since the inspection department constantly receives large data movement economic and regularization of taxpayers. The current computational resources store information with a larger human perception of manipulation and knowledge extraction. In this context, appears in science an area called data mining, specific to extract unknown patterns and knowledge through databases. This study aimed to develop a model to classify taxpayers Tax Services (ISS) which showed some irregularity, with resources and techniques of data mining. The study was performed in the city of Goiania in finance secretary specifically of the Department of Revenue, covering the scenario presented in the year 2011. Among the models built with decision tree algorithm, presented as a result, the classification of irregular contributors with a hit rate of 92,03%.

Keywords*: Data Mining, irregularity detection, ISS.*

SUMÁRIO

LISTA DE FIGURAS

LISTA DE TABELAS

LISTA DE SIGLAS

AMTEC - Agência Municipal de Tecnologia e Inovação

ARRF - A*tributte-relation File Format*

BI - *Business Intelligence*

CAE - Cadastro de Atividades Econômicas

CC - Código Civil

CF - Constituição Federal

CNAE - Cadastro Nacional de Atividades Economicas

CTM - Código Tributário do Município

CRISP-DM - *Cross-Industry Standard Process for Data Mining*

DCBD - Descoberta de Conhecimento em Bases de Dados

DMS - Declaração Mensal de Serviços

DW - *Data WareHouse*

EPP - Empresa de Pequeno Porte

FBI - *Federal Boreau of Investigation*

GNU - *General Public Licence*

KDD - *Knowledge Discovery in Databases*

IBGE - Instituto Brasileiro de Geografia e Estatística

IBM - *International Business Machines*

IBPT - Instituto Brasileiro de Planejamento Tributário

ICMS - Imposto Sobre Circulação de Mercadoria e Serviços

IPTU - Imposto Predial e Territorial Urbano

ITBI - Imposto sobre a Transmissão de Bens Imóveis

ISS - Imposto sobre Serviços de Qualquer Natureza

IR - Imposto de Renda

MD - Mineração de Dados

ME - Micro Empresa

MIT - *Massachusettts Institute of Tecnology*

MLP – *Multilayer Perceptron*

NFS-e - Nota Fiscal Eletrônica de Serviços

OLAP - *On-line Analytical Processing*

RCTM - Regulamentação do Código Tributário do Município

SGDB - Sistema de Gerenciamento de Banco de Dados

STF - Supremo Tribunal Federal

TIC - Tecnologia da Informação e Comunicação

I INTRODUÇÃO

O município na figura de tutor da sociedade representa, dentro da estrutura definida pela Constituição Federal, a camada mais próxima do cidadão para atender aos interesses da coletividade e bem comum. Para seu funcionamento, necessita também dos recursos oriundos dos contribuintes, na instituição de impostos, taxas e contribuições visando o desempenho das funções administrativas e no estabelecimento de políticas públicas que garanta aos cidadãos os direitos básicos declarados na Constituição.

Figura como direitos básicos dos cidadãos e responsabilidade do município a prestação de serviços pertinente à manutenção de vias públicas, o controle do trânsito, a manutenção da educação infantil, a coleta de lixo, a saúde básica, o transporte coletivo, a regularização e fiscalização das atividades comerciais no perímetro do município, o adensamento e crescimento da cidade e controle das condições de vida, dentre elas os registros de nascimentos e óbitos ocorridos.

A prestação de serviços públicos tem custos e a arrecadação de impostos implica na qualidade dos serviços prestados pela administração, uma vez que o estabelecimento do planejamento anual dos cronogramas de projetos apresentados e aprovados na câmara legislativa depende deste recurso. Assim, a arrecadação tributária delegada aos municípios tem caráter econômico e social, uma vez que o seu aprimoramento influencia na eficiência e continuidade dos serviços prestados a população.

Assim, os recursos financeiros necessários para manter as atividades que sustentam a administração pública em funcionamento, juntamente com a manutenção dos serviços prestados com qualidade a população, dependem diretamente da arrecadação de impostos.

1.1 O PROBLEMA DA SONEGAÇÃO

O relatório estatístico apresentado pelo Instituto Brasileiro de Geografia e Estatística (IBGE) mostra um forte crescimento populacional apresentado nas cidades brasileiras nos últimos 10 anos (IBGE, 2007). Esse crescimento implica em maior carência dos serviços públicos por parte da população oriunda de novos adensamentos populacionais e consequentemente, no provimento de suporte ao acréscimo de novos contribuintes.

Ocorre que a estrutura da máquina pública não acompanha o crescimento e demanda da população, apresentando uma limitação, como exemplo, dos setores de fiscalização. Tal defasagem acaba contribuindo indiretamente para um maior índice de inadimplência e sonegação. Visto que cabe à administração pública a verificação dos cumprimentos legais relativos à tributação, torna-se imprescindível para o setor de arrecadação prover mecanismos de controle e recuperação dos impostos sonegados.

O problema da sonegação está presente nos departamentos fiscais brasileiros, como mostra estudos realizados no setor tributário. O primeiro estudo apresenta dados resultantes da pesquisa realizada pelo Instituto Brasileiro de Planejamento Tributário[1] (IBPT) em 2004 com 7.437 empresas. Dentre os vários setores econômicos estudados, um percentual de 29,45% das empresas pesquisadas apresenta fortes indícios de sonegação Fiscal (FUTEMA, 2005). Já o trabalho realizado por SIQUEIRA e RAMOS (2005) foi mais abrangente ao apresentar estudos que visaram quantificar os percentuais relativos à sonegação observados em vários países. Na sua conclusão, a aplicação de diversos métodos de mensuração sugere que nos países industrializados ocidentais a sonegação de impostos atinge 5% a 25%

[1] Organização não Governamental especialista em governança tributária que acompanha e analisa a arrecadação tributária no país com divulgações periódicas de cálculos globais.

da arrecadação tributaria potencial, dependendo da técnica adotada no país, com percentuais mais elevados (até 30% ou 40%) para países menos desenvolvidos.

A sonegação envolve todo ato que, realizado conscientemente ou inconscientemente, de forma ilegal, culmina no não pagamento ou pagamento a menor de imposto. Quando, observado suas lacunas ou fazendo planejamento fiscal, resultando no não recolhimento de imposto, ocorre à elisão. Já a evasão fiscal caracteriza-se pelo claro atento a lei, utilizando-se de meios ilícitos para evitar o pagamento de impostos (YAMASHITA, 2005).

Independente da forma de sonegação, voluntária ou involuntária, quando comprovada, tem como consequência à aplicação de penalidades pela administração tributária. Tais penalidades envolvem desde a aplicação de multas pecuniárias a perda de vantagens e benefícios (ANDRADE FILHO, 2005).

Diminuir ao máximo a inadimplência e sonegação de impostos torna-se de vital importância para a administração pública e anseio contínuo das administrações fazendárias no sentido de prover ao município inúmeros benefícios, tais como:

- Prover recursos equivalentes ou maiores que o acréscimo advindo do crescimento econômico para investimento na qualidade do serviço prestado pela administração pública;

- Favorecer a competitividade do mercado no que tange a equivalência de obrigações por parte dos contribuintes perante a administração fazendária, atuando de forma regular e recolhendo seus impostos;

- Conter subsídios claros e suficientes para prover reformas tributárias favorecendo diversos setores econômicos, pulverizando o crescimento econômico e social.

1.2 MOTIVAÇÃO E IMPORTÂNCIA

A busca pela diminuição da sonegação pode ser auxiliada pela implantação e uso de tecnologia, que provê agilidade, eficiência e principalmente controle pela administração pública das atividades realizadas pelos contribuintes sob sua competência. Para melhor eficiência é necessário analisar as implicações de diferentes processos de controle (GIRIOLI, 2010).

No controle da arrecadação do imposto, o município controla diversos dados relativos a informações socioeconômicas e fiscais dos contribuintes, provendo subsídios para estabelecer as alíquotas e valores dos impostos devidos dos mesmos. Também, realiza em outra vertente, o controle das atividades realizadas pelo Departamento de Fiscalização no sentido de recuperação do imposto sonegado e inadimplente.

O respectivo controle para a prestação de serviços é respaldado pela lei federal complementar 116, de 31 de Julho de 2003, que dispões de regras relativas ao exercício da tributação municipal. Desta forma, o município tem competência para deliberar sobre a prestação dos serviços, instituição dos impostos e deliberação das atividades econômicas, através de lei complementar e regras que envolvem os procedimentos comuns, tais como Nota Fiscal, Livro de Registro Fiscal, Cadastro Fiscal, Guia de Recolhimento, Autorização para Documento Fiscal e Certidões (NASCIMENTO, 2010).

Para realizar o controle, o departamento fiscal recebe mensalmente um grande volume de informações socioeconômicas e fiscais relativos aos movimentos dos contribuintes, além dos dados de procedimentos efetuados para regularização cadastral. Todo esse controle culmina no crescimento contínuo das bases de dados,

tornando-se um desafio no que tange a utilização do potencial conhecimento armazenado, decorrente tanto para organizações públicas quanto privadas.

John Naisbitt, ex executivo da *Internacional Business Machines* (IBM), atualmente pesquisador e especialista em tendências globais, afirma que após o controle da tecnologia, “estamos afogados em informações, mas famintos por conhecimento” (FARIA e QUONIAN, 2002). Na era digital a capacidade de armazenar conteúdo é infinitamente superior a nossa capacidade de extrair informação e vislumbrar conhecimentos através dos métodos convencionais de bases de dados, visto que em determinados áreas de negócios, como exemplo na medicina, temos bases de dados em que seus atributos já atingiram a ordem de 10^3 e na astrologia 10^9.

Dentro desta realidade, a área de Mineração de Dados (MD) também conhecida com o acrônimo *Data Mining* tem atraído atenção dos profissionais de Tecnologias da Informação e Comunicação (TIC) e da comunidade acadêmica em geral, resultado das experiências e colaboração de publicações voltadas ao descobrimento de informações em grandes massas de dados, na descoberta de padrões inclusive na classificação de contribuintes que apresentem irregularidades.

Na aplicação da MD na área fiscal vale destacar o protótipo realizado no Estado de São Paulo com 70 empresas, onde foram detectadas 3.700 inconsistências que representariam um montante de R$ 15.000.000,00 de imposto a recuperar concernente ao Imposto Sobre Circulação de Mercadoria e Serviços (ICMS) (CORVALÃO, 2009). Outro importante trabalho realizado no Texas (EUA) relata a comparação entre a forma tradicional de seleção de contribuintes para auditoria e o modelo baseado em MD proposto pelos autores Micci-Barreca e Ramachandran (2006) – o trabalho apresentou uma melhoria média em torno de 16% com a implantação do novo modelo.

A MD pode ser decisiva no auxilio da detecção de irregularidades por possuir caráter multidisciplinar contendo técnicas de banco de dados, estatística e aprendizado de máquina, apresentando grandes utilidades na descoberta de regras ou padrões, na previsão de futuras tendências e comportamento de grupos similares. A detecção de irregularidade tornou-se uma das principais aplicabilidades da MD e a principal tarefa reside na construção de modelos ou perfis que indiquem comportamento possivelmente fraudulento, possibilitando sua averiguação (BONCHI *et al*., 1999).

A motivação desta pesquisa visa a necessidade de investimento em mecanismos tecnológicos pelo setor público que auxilie o departamento de fiscalização na tomada de decisões pertinentes ao cenário dos contribuintes visualizado. A essência desta pesquisa visa utilizar o contexto da MD para construir um modelo, a partir da técnica de classificação, para identificar o perfil dos contribuintes com irregularidades, com o intuito de planejar estratégias de auditorias na recuperação de impostos sonegados e estabelecimento de políticas públicas e sociais.

O conhecimento do perfil dos contribuintes permite analisar os mecanismos de recuperação de impostos e também propiciar o estabelecimento de políticas publica e sociais. Através do cenário dos contribuintes e seu comportamento é possível visualizar quais atividades econômicas tem maior incidência de atuação informal, revisar e estabelecer novas alíquotas e incentivos fiscais.

Neste trabalho é utilizado para construir o modelo de classificação dados dos contribuintes de prestação de serviços do município de Goiânia mediante a ocorrência de irregularidades. Esse processo de elaboração de um modelo classificador para visualizar o problema em questão representa um fato inovador ao abordar o contexto dos contribuintes de prestação de serviços mediante seu domínio socioeconômico e fiscal.

Os desafios correntes justificam a importância do trabalho ao abordar o processo de seleção de atributos no contexto da fiscalização de serviços e a classificação do perfil dos contribuintes possivelmente irregulares no intuito de prover subsídios para auxiliar na indicação do escopo de contribuintes a ser fiscalizado visando melhor eficiência mediante o número reduzido de fiscais. A qualidade do resultado do modelo está condicionada a disposição das informações e técnicas empregadas para obter o modelo de classificação.

1.3 OBJETIVO DO TRABALHO

A pesquisa de MD é um processo de descobrimento de informação, disposto em etapas, necessitando de uma metodologia para sua organização. A classificação é uma tarefa da MD utilizada para problemas de detecção de irregularidades. Visando o problema em questão o objetivo da pesquisa foi estabelecido na seguinte disposição:

1.3.1 Objetivo Geral

O objetivo geral é elaborar, através de uma metodologia especifica com um processo de MD, um modelo de classificação capaz de auxiliar o Departamento de Fiscalização a direcionar seus esforços de auditoria com base nos registros dos contribuintes com possíveis irregularidades.

1.3.2 Objetivos Específicos

Para alcançar a proposta deste trabalho, o estudo dos seguintes objetivos específicos deve ser realizado:

- Definir o conjunto de MD que melhor identifique o perfil dos contribuintes através da classificação mediante a natureza das informações socioeconômica e fiscais apresentadas;

- Preparar e ajustar os dados do setor de serviços extraídos junto a Secretaria de Finanças para que seja passível de mineração através de uma metodologia;

- Delimitar dentre as ocorrências realizadas na auditoria as modalidades de relevância pelo setor de fiscalização;

- Aplicar o modelo proposto num cenário real pertencente ao município de Goiânia, no intuito de classificar o perfil dos contribuintes com irregularidades.

1.4 ESTRUTURA DO TRABALHO

O trabalho é constituído por seis capítulos, assim estruturados:

O capitulo 2 abrange a fundamentação teórica, apresentando o contexto de MD, metodologias de MD, tarefas e técnicas mais utilizadas com ênfase na classificação, tarefa utilizada neste trabalho. Finaliza com a apresentação de trabalhos correlatos a MD aplicada na detecção de irregularidades e fraudes.

O capitulo 3 abrange a fundamentação teórica relativa ao contexto do domínio da informação utilizada neste projeto relativa a imposto e tributo, apresentando características e fundamentação legal.

O capitulo 4 descreve o modelo proposto para realizar a classificação do perfil dos contribuintes, os materiais, métodos e técnicas utilizadas.

O capitulo 5 ilustra o modelo estabelecido mediante o conjunto de dados da Secretaria de Finanças do Município de Goiânia registrado no ano de 2011. Utilizando-se esse conjunto de dados foi criado um processo de MD para elaborar um modelo classificador através do algoritmo árvore de decisão para avaliar o perfil dos

contribuintes com irregularidades. O capitulo descreve o processo realizado juntamente com os métodos de avaliação.

Por fim, o capitulo 6 apresenta a conclusão obtida na pesquisa, restrições e trabalhos futuros.

II. REFERENCIAL TEÓRICO

Este capítulo visa apresentar os conceitos referentes ao contexto da MD, metodologia, suas principais tarefas e técnicas de mineração. O capítulo aborda a importância e relevância da adesão de um processo para direcionar a extração de conhecimento mediante o denso volume dos conjuntos de dados atuais. O esclarecimento do contexto de MD é imprescindível para os fins metodológicos deste trabalho.

2.1 DESCOBERTA DE CONHECIMENTO EM BASE DE DADOS (DCBD)

A fase inicial do processo de informatização preocupou-se em automatizar os processos rotineiros e repetitivos, objetivando dar agilidade aos procedimentos das diversas modalidades de negócios. O avanço e disseminação dos recursos computacionais culminaram no constante crescimento das bases de dados atuais. Estima-se que a quantidade de informação no mundo dobra a cada 2 anos e que o tamanho e a quantidade das bases de dados crescem com velocidade ainda maior (GANTZ e REINSEL, 2011). Este denso volume de dados desperta o interesse das organizações em investir em projetos de MD que auxiliem a descoberta de conhecimento oriundo de bases de dados.

A demanda crescente na busca por conhecimento aliada a fatores científicos e econômicos propulsionam o uso da tecnologia através do desenvolvimento de técnicas e ferramentas para auxiliar o trabalho humano na análise de densos volumes de dados no intuito de descobrir padrões e correlações. LAROSE (2006) exemplificou a relevância da utilização da MD na descoberta de conhecimento ao relatar a declaração realizada em novembro de 2002, pelo então presidente dos Estados Unidos da América, Bill Clinton, que logo após os incidentes de 11 de setembro de 2001, os

agentes do *Federal Bureau of Investigation* (FBI) examinaram uma enorme quantidade de dados de consumidores e encontraram informações de cinco terroristas em tais bases, inclusive um deles utilizava em torno de trinta cartões de crédito com volume de $250.000 em movimentação. A análise também identificou que o referido suspeito residia há dois anos na cidade. Para realizar tal análise e alcançar os resultados mencionados é necessário empregar técnicas que modelem a quantidade densa de dados ao nível de compreensão humana.

A necessidade recorrente de explorar densos conjuntos de dados culminou no surgimento de metodologias com a finalidade de Descoberta de Conhecimento em Base de Dados (DBCD) que é acrônimo do termo inglês *Knowledge Discovery in Databases* (KDD). Segundo FAYYAD *et al.* (1996), o processo KDD foi relatado pela primeira vez na literatura 1989 para enfatizar que o conhecimento é objetivo final de uma descoberta em bases de dados. Posteriormente, desponta como um campo de pesquisa voltado especificamente para a descoberta de conhecimento a partir de conjuntos de dados, com metodologia própria e baseada em um conjunto de etapas, no qual MD é a etapa mais relevante. É definido por FAYYAD *et al*. (1996) como um processo, não trivial, de descoberta de padrões válidos, novos, úteis e acessíveis.

O processo DCDB possui uma abrangência maior relacionada à atividade MD. Para que seja possível a aplicação de mineração, o DCDB dispõe de etapas que envolvem todo o processo de extração e exploração dos dados, enquanto a MD enfatiza a descoberta de padrões desconhecidos ou ocultos vasculhando os dados explorados, representando apenas uma etapa do processo DCDB como ilustra a figura 1. A principal vantagem do processo DCDB é a possibilidade de avaliar os resultados obtidos após a aplicação de técnicas de MD. No entendimento de FAYYAD *et al*. (1996), o DCDB se refere a todo o processo de descoberta de conhecimento

mediante o processamento de dados, e envolve também a limpeza e preparação, incorporação de conhecimento e apresentação de resultados.

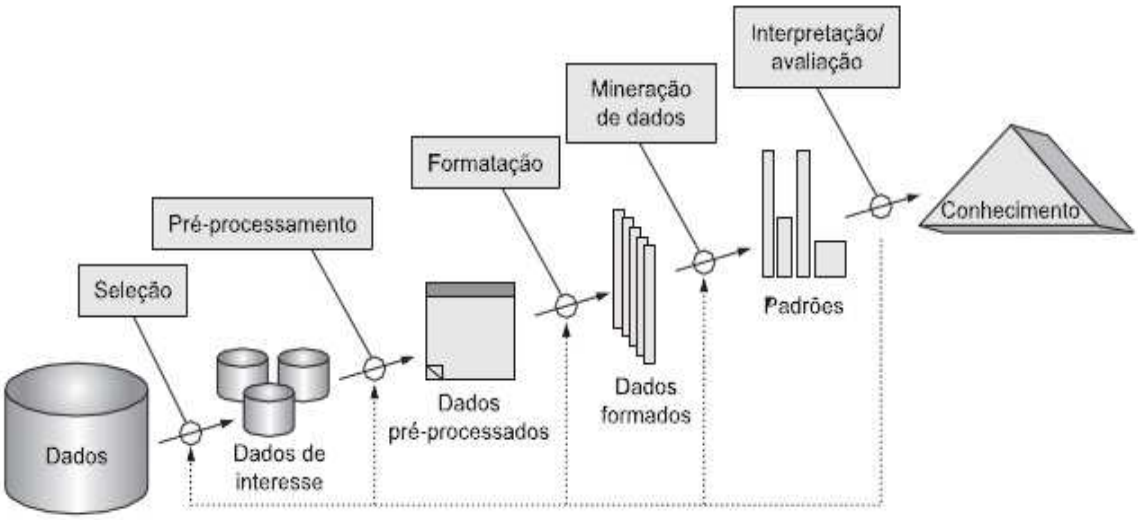


Figura 1 - Processo DCBD. Fonte FAYYAD *et al*. (1996).

O DCDB como um processo é identificado pelas seguintes fases: Seleção dos dados; Pré-processamento e Limpeza; Formatação; Mineração e Interpretação conforme ilustrado na figura 1. As três primeiras fases refletem o esforço necessário para obter qualidade dos dados objeto da MD. O processo DCBD pode ser visto como a aplicação do método científico para a descoberta de conhecimento e padrões não explícitos nos dados por possibilitar a verificação dos resultados e repetição dos experimentos (ROGER e GEATZ, 2003).

Inicia-se o processo DCDB com um pleno entendimento do domínio do problema e dos objetivos almejados. Posteriormente uma atividade de seleção é necessária para determinar quais dados são de relevância para o objeto de pesquisa. A fase de pré-processamento visa correção de inconsistências, ruídos, dados faltantes e normalizações com o intuito de obter padronização no conjunto de dados.

Para realizar análise em grandes massas de dados é importante indicar uma área correlata à atividade de MD voltada ao pré-processamento de dados, *Data*

Warehousing. A criação de um *Data Warehouse* (DW) é considerada como um dos primeiros passos para viabilizar a análise em grandes massas de dados (REZENDE, 2005). O *DW* é uma coleção de dados integrados, orientados por assunto, variáveis com o tempo e não voláteis, usados para suporte ao processo gerencial de tomada de decisões (INMON e HACKATHORN, 1997). Sua utilização traz benefícios importantes ao processo DCDB ao viabilizar extração dos dados de um ambiente transacional e analítico a um ambiente que permita a análise de diversas maneiras e de forma organizada. A etapa de formatação concentra-se na eliminação de redundâncias, categorização e granularidades de dimensões visando a aplicação das técnicas de MD. As etapas de pré-processamento e formatação consomem o maior esforço. Podem levar até 80% do tempo necessário para todo o processo (SILVER, 1996).

Posteriormente ao pré-processamento, o processo desencadeia a fase de mineração, principal atividade do processo DCBD. Nessa fase, diversas técnicas podem ser utilizadas através das tarefas selecionadas para realizar extração de conhecimento, os quais são passíveis de interpretação e avaliação mediante dados novos e métodos escolhidos para averiguar a eficácia do conjunto utilizado. As técnicas mais utilizadas são apresentadas no decorrer deste capitulo, dentre elas técnicas voltadas à atividade de classificação, foco deste trabalho.

A fase de avaliação não finda a utilização do processo DCDB. Se porventura os resultados não forem satisfatórios, o processo pode ser realimentado modificando o conjunto de dados de entrada e reprocessando algumas fases. Assim a visualização dos dados resultantes e o que fazer com eles se torna desafio aos gestores do processo. Ainda segundo FAYYAD *et al*. (1996), DCDB é um processo complicado de identificação de modelos válidos, potencialmente útil e atuais em processamento de dados. Os autores ressaltam que o processo compreende muitas etapas entre a exploração e avaliação dos dados, e todas elas repetidas em múltiplas interações.

Frequentemente, alguns autores tratam a MD como o processo de DBCD. Para outros, um passo particular do processo realizado através de algoritmos específicos para a extração de conhecimentos nos dados. As fases adicionais do processo, tais como preparação, seleção e limpeza nos dados, incorporação de conhecimento prévio adequado e interpretação dos resultados da MD, asseguram que o conhecimento útil seja derivado dos dados (MITRA e ACHARYA, 2003).

É importante ressaltar que apesar dos avanços tecnológicos e da capacidade de realizar o processo de forma automatizada, a intervenção humana acionada ativamente em cada fase do processo de *DCBD* continua essencial.

O processo de DCDB é utilizado em projetos de MD para auxiliar na descoberta de conhecimento em diversas áreas. Dentre elas, as mais destacadas são: detecção de fraudes e irregularidades, finanças (especialmente investimentos), seguros, marketing, saúde, indústria, segurança e telecomunicações (LAROSE, 2006).

2.2 MINERAÇÃO DE DADOS

A MD é uma fase do processo DCDB com objetivo de analisar informações de grandes conjuntos de dados no intuito de descobrir correlações e padrões que sejam úteis para os patrocinadores do projeto. Pode ser realizado de forma automática ou mais frequentemente, de forma semi-automática. Aborda a resolução de problemas através de analises de dados já presentes em banco de dados e os padrões descobertos devem ser significativos, na medida em que leva a alguma vantagem, normalmente econômica (WITTEN e FRANK, 2005).

Outra definição apresentada por GIUDICI (2003) define *MD* como o processo de seleção, exploração, e modelagem de grandes quantidades de dados para descobrir padrões ou relações que são em primeira análise desconhecidos, com o objetivo de obter resultados claros e úteis para o dono do banco de dados. O processo de MD

tornou-se alvo de interesse de empresas privadas e públicas de diversos segmentos devido à constante evolução do volume de dados e da dificuldade de extrair conhecimentos escondidos ou ocultos, sem o auxílio de um modelo que norteie o processo. Com as técnicas de MD esse conhecimento pode ser descoberto, extraído e acessado, transformando as tarefas de base de dados voltadas a armazenamento e recuperação para aprender e extrair conhecimento (AL-RADAIDEH, NAGI, 2012).

Segundo a revista eletrônica de tecnologia *Technology Review* (MIT, 2001*)*, a MD foi escolhida entre as 10 tecnologias emergentes que mudarão o mundo, relevância destacada pelo fato da sua aplicabilidade generalizada, estendendo-se a áreas surpreendentes e pela característica multidisciplinar incorporada ao seu conceito, como ilustra a figura 2.

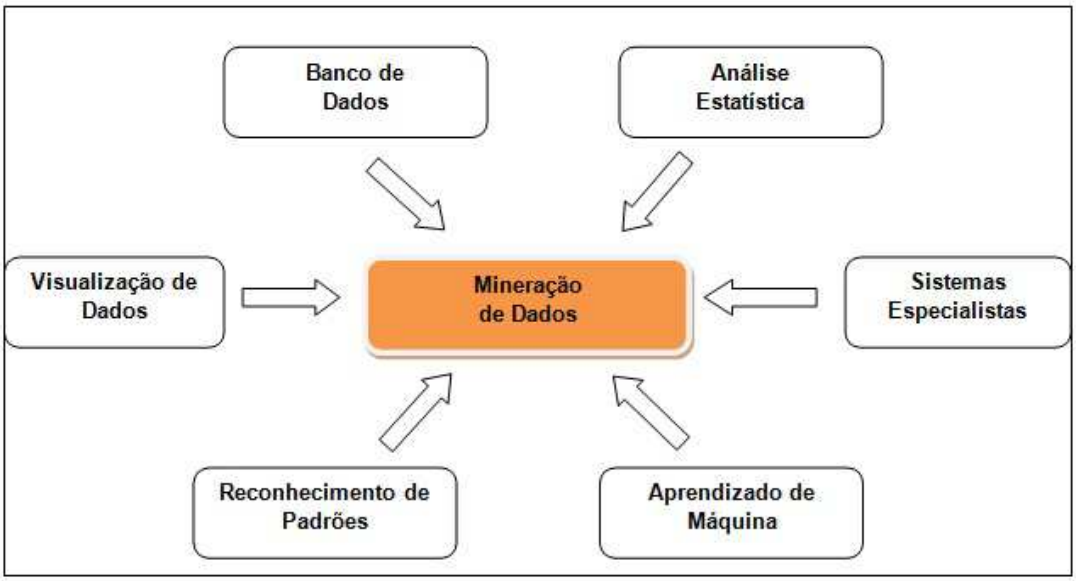


Figura 2 - Característica multidisciplinar da MD. Adaptado (NETO, 2010).

A MD é um campo interdisciplinar contendo técnicas de aprendizado de máquina, reconhecimento de padrões, estatística e matemática, aquisição de conhecimento para sistemas especialistas, base de dados e visualização para abordar a questão da extração de conhecimento de grandes bases de dados (NETO *et al.*, 2010). Segundo

CABENA *et al*. (1998), a atividade de minerar é constituída pela intersecção de diversos campos de pesquisas.

A atividade de minerar pode ser realizada através de várias tarefas que visam extrair informações e conhecimento do conjunto de dados tais como classificação, predição, análise de agrupamentos e associação. A definição da tarefa utilizada está condicionada aos objetivos do projeto de MD e da disposição do domínio dos dados. A MD também envolve a utilização de diversas técnicas, materializada por algoritmos computacionais, necessárias para realizar as tarefas de mineração.

A classificação é uma das mais utilizadas para predição de dados futuros. Uma determinada tarefa pode ser realizada por diversas técnicas. A classificação, por exemplo, pode ser efetivada pelas técnicas árvore de decisão, algoritmo bayesiano ou redes neurais. As técnicas utilizadas na classificação são rotuladas de aprendizado supervisionado por ser direcionado por um atributo alvo de referência para um conjunto de classes pré-definidas, e é uma das mais utilizadas no mundo para a predição futura de um conjunto de dados (AL-RADAIDEH, NAGI, 2012).

2.3 TAREFAS DE MINERAÇÃO DE DADOS

No decorrer do processo de mineração a definição de uma tarefa de MD representa um marco que restringe a continuidade do projeto. A tarefa auxilia a dar transparência ao modelo de MD*,* à medida que os resultados dos modelos descrevem padrões claros, passíveis de interpretação intuitiva e explicação. As tarefas de MD são classificadas em duas categorias conforme ilustra a figura 3, dependendo dos objetivos que se queira atingir com a utilização do processo.

Figura 3 - Classificação das tarefas de MD. Adaptado (DOMINGUES, 2004)

Tarefas preditivas: As preditivas têm intuito de prever o valor de um determinado atributo baseado nos valores de outros. O atributo a ser previsto é comumente conhecido como atributo alvo ou dependente e o direcionamento é realizado a classes previamente conhecidas.

Tarefas descritivas: As descritivas visam identificar padrões intrínsecos com o intuito de derivar padrões (correlações, tendências, grupos, trajetórias e anomalias) que resumam os relacionamentos subjacentes em um conjunto de dados, sendo que esses padrões não são direcionados pela figura de um atributo alvo como o disposto nas tarefas de predição (TAN *et al*., 2009).

As principais tarefas de MD, detalhamentos e respectivas aplicabilidades estão destacadas na sequência, nas próximas subsecções.

2.3.1 Associação

Concentra-se em descobrir regras para determinar a relação entre dois ou mais atributos de um conjunto de dados indicando aqueles que ocorrem de forma frequente. Objetiva descobrir padrões que subsidiam a tomada de decisões baseado em fatos que ocorrem de forma conjunta em determinada operação. Intuitivamente, essa tarefa consiste em encontrar conjuntos de itens que ocorram simultaneamente e de modo frequente em um conjunto de dados (VIGLIONI, 2007). A tabela 1 ilustra a utilização desta tarefa com o objetivo de investigar correlações de produtos sob a ótica do consumidor, realizada a partir de observações de operações de um supermercado.

Tabela 1 - Exemplo de transações de um supermercado. Adaptado (LAROSE, 2006).

TID	Itens
1	{Pão, Leite}
2	{Pão, Fraldas, Cerveja, Ovos}
3	{Leite, Fraldas, Cerveja, Cola}
4	{Pão, Leite, Fraldas, Cerveja}
5	{Pão, Leite, Fraldas, Cola}

{Fraldas –> Cerveja}

Com base na tabela 1, nota-se um relacionamento forte entre a venda de fraldas e a venda de cerveja porque muitos consumidores que compraram fraldas também compraram cervejas. Esse conhecimento admitido pelas regras de associação auxiliaria o supermercado na disposição de produtos, facilitando o acesso dos clientes aqueles produtos que apresentem um mesmo grau de similaridade de compra (TAN *et al*., 2009).

As regras de associação normalmente são do tipo “se antecedente, então conseqüente”. No comércio em geral, investiga a similaridade de produtos adquiridos pelo consumidor (LAROSE, 2006).

2.3.2 Classificação

Concentra-se na classificação ou associação de um determinado registro a uma classe previamente definida. Suponha-se que uma variável alvo categórica[2] seja utilizada para delimitar a faixa de renda de consumidores. Nesse caso pode-se dividi-los em três classes ou categorias: baixa, media e alta renda.

A classificação estabelece uma nova observação a um conjunto de classes previamente rotuladas. Objetiva descobrir algum relacionamento entre os atributos da classe de entrada e as classes de saída, de forma que esse conhecimento possa ser utilizado para prever em qual classe um novo registro não conhecido possa ser classificado (SUMATHI e SIVANANDAM, 2006). Considere o resumo de dados da tabela 2.

Tabela 2 - Resumo de dados para classificação de renda. Adaptado (LAROSE, 2006).

Identificador	Idade	Gênero	Ocupação	Faixa de renda
001	47	F	Engenheiro de software	Alta
002	28	M	Consultor de Marketing	Media
003	35	M	Desempregado	Baixa
⋮				

Conforme o exemplo mencionado, suponha que o objetivo da pesquisa seja classificar a faixa de renda com base nas características associadas à pessoa, tais como idade e sexo.

Assim, o algoritmo de classificação examina o conjunto de dados com as variáveis de predição selecionadas e realiza um aprendizado de quais combinações estariam associadas às faixas de renda. Esse conjunto de dados é denominado conjunto de

[2] Variáveis qualitativas que representam características não quantificáveis assumindo valores nominais ou ordinais.

treinamento. Após esse processo de aprendizado, o algoritmo, ao analisar um novo conjunto de dados, é capaz de realizar a classificação de um novo registro.

Diversos trabalhos de classificação com o objetivo de predição de comportamentos são apresentados na literatura envolvendo várias áreas da ciência, tecnologia da informação, biologia, recursos humanos e medicina (LAROSE, 2006). A classificação pode ser utilizada para resolver problemas de diversos gêneros, tais como:

- Determinar se uma transação de cartão de credito é fraudulenta;
- Posicionar um novo estudante em uma faixa particular mediante suas necessidades especiais;
- Diagnosticar a presença de uma doença específica;
- Identificar se certo comportamento financeiro ou pessoal de um indivíduo indica uma ameaça de um possível terrorista.

Suponha que o interesse seja classificar a prescrição do tipo de medicamento receitar a um determinado paciente baseado em suas características tais como idade e relação de sódio versus potássio (NA/K). A figura 4 ilustra um gráfico de dispersão de 36 pacientes classificado pela relação de idade apresentada na parte inferior horizontal versus sódio/potássio (NA/K), apresentado no eixo esquerdo vertical. As três linhas destacadas representam quatro distintas regiões e referenciam qual medicamento receitar: MEDIC A, MEDIC B, MEDIC e MEDIC D. A primeira região MEDIC A, localizada na parte inferior do gráfico especifica qual medicamento é mais indicado para os pacientes com idade entre 15 e 80 anos e com uma relação de sódio/potássio (NA/K) próximo a 10. Os gráficos auxiliam o entendimento de relacionamento entre duas ou três variáveis, mas normalmente, os algoritmos de classificação estão preparados para trabalhar com o relacionamento constituído de diversas variáveis.

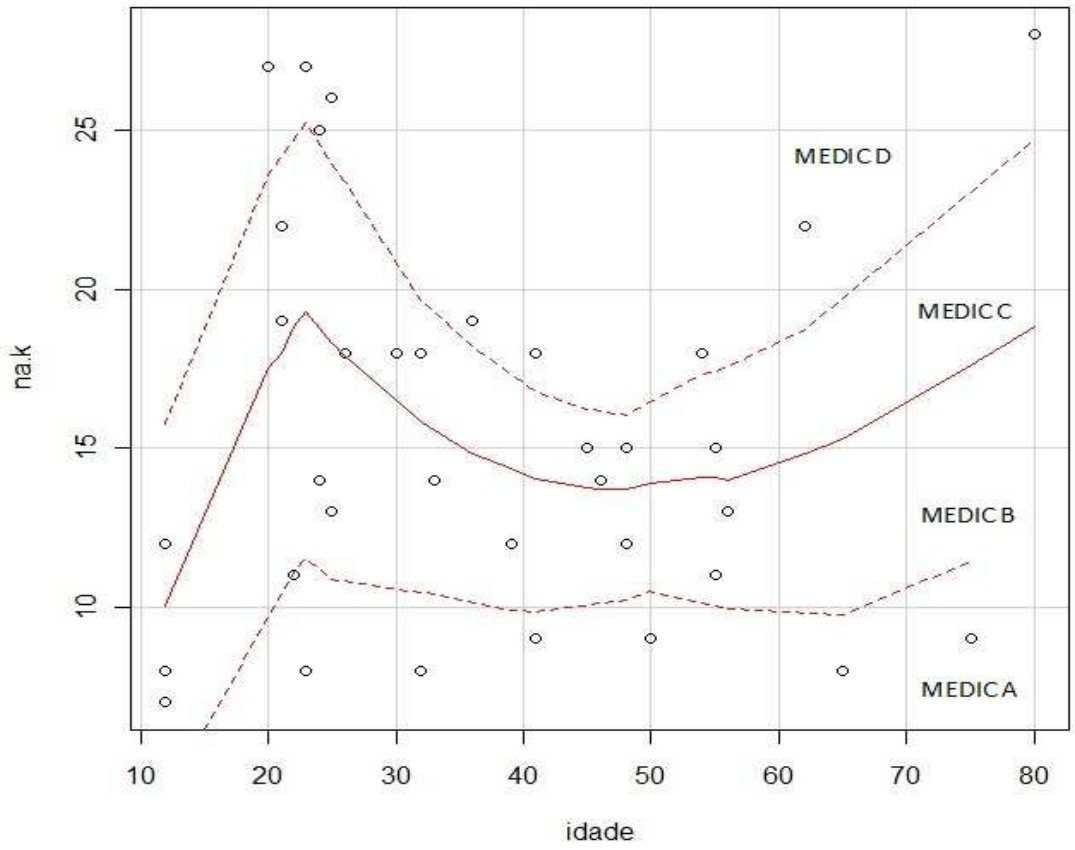


Figura 4 - Gráfico de dispersão idade versus sódio/potássio. Adaptado (LAROSE, 2006).

2.3.3 Agrupamento (*Cluster*)

Análise de grupo ou *cluster* concentra-se no agrupamento de registros que apresentem similaridades. O algoritmo aproxima os objetos semelhantes (*cluster*) e distancia os que apresentem poucas similaridades. Procura encontrar observações relacionadas de modo que os objetos em um mesmo grupo sejam mais similares entre si do que aos que pertencem a outros grupos, como ilustra a figura 5.

Difere da classificação porque o agrupamento é realizado baseado na similaridade dos dados dos objetos observados, sem a utilização de um atributo alvo para nortear o processo de agrupamento à uma classe pré-determinada (LAROSE, 2006).

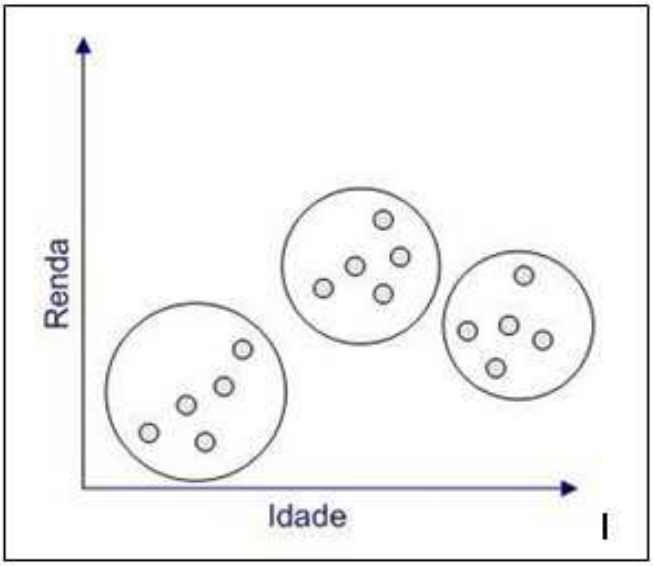


Figura 5 - Exemplo de análise de agrupamento. Extraído (CORVALÃO, 2009).

A análise de agrupamentos averigua a correlação dos atributos. Esta tarefa pode ser utilizada para projetar uma campanha de marketing, direcionar as finalidades de uma auditoria financeira ou reduzir a dimensão de um determinado conjunto de dados quando o mesmo possuir muitos atributos.

2.3.4 Estimação

A tarefa de estimação procura obter um valor para um atributo em relação aos demais observados ou a distribuição do valor de determinado atributo em um conjunto de dados. A estimação é similar à classificação, diferindo-se porque na estimação o atributo que direciona a tarefa é um atributo numérico e não categórico. No campo da analise estatística, os métodos de estimativa são largamente utilizados para determinar pontos estatísticos, intervalos de confiança, regressão linear simples, correlações e regressão múltipla. A estimação também pode utilizar a técnica de rede neural (LAROSE, 2006).

2.3.5 Predição

A tarefa de predição é similar à estimação, diferindo-se porque na predição os resultados visam analisar estados futuros. Em algumas circunstâncias, os métodos e técnicas utilizadas na estimação podem também ser utilizados na predição, auxiliando em pesquisas que visam:

- Prever o preço de um produto estocado meses posteriores;
- Prever o número de acidentes em um período posterior caso a velocidade de uma via seja aumentada;
- Prever o campeão do campeonato brasileiro de futebol baseado na comparação estatística de cada equipe.

2.3.6 Sumarização

A tarefa de sumarização visa aplicar métodos para prover uma descrição compacta de um subconjunto de dados. Essa tarefa é frequentemente utilizada no pré-processamento de dados para uma exploração intuitiva, quando valores inválidos são determinados através de métodos estatísticos como exemplo, a tabulação da média e desvio padrão, ou em casos mais sofisticados, através da distribuição de frequência de valores (FAYYAD, PIATESK-SHAPIRO e SMUTY, 1996). Técnicas de sumarização mais sofisticadas são imprescindíveis para se obter um entendimento, muitas vezes intuitivo, do conjunto de dados (SFERRA E CORRÊA, 2003).

2.3.7 Detecção de anomalias ou *outliers*

A detecção de anomalias visa identificar objetos que possuam características significativamente diferentes dos demais registros do conjunto de dados. As técnicas

de detecção de anomalias utilizam uma função média para representar o comportamento normal do sistema e assim, avaliar possíveis desvios (FAYYAD *et al*., 1996). Um bom algoritmo de detecção implica no descobrimento verdadeiro de anomalias evitando afirmações errôneas sobre desvios encontrados. Incluem a pesquisa por desvios temporais (mudanças significativas nos dados de séries temporais) e desvios em grupos (diferenças não esperadas entre dois subconjuntos de dados) (SUMATHI E SIVANDAM, 2006).

As aplicações de detecções de anomalias são voltadas para detecções de fraudes, intromissões na rede, padrões incomuns de doenças e perturbações do meio ambiente.

2.4 TÉCNICAS MAIS UTILIZADAS

As técnicas ou algoritmos de MD representam o processo de tratamento dos dados, evidenciando possíveis tipos de relacionamento. Para que o processo de MD seja eficiente é necessário conhecimento do tipo de resultado esperado para definição da técnica almejada.

A definição de uma técnica não é uma tarefa trivial. Segundo Harrison (1998), a escolha é definida com base na tarefa selecionada e na disposição dos dados utilizados para o processo de MD. Sugere ainda que critérios sejam utilizados para estabelecer a definição de uma técnica como a tradução do problema de negócio em séries de tarefas de mineração e a compreensão da natureza dos dados disponíveis, observados campos, tipos de dados e estrutura de relações dos registros.

Existem tanto algoritmos simples como complexos e determinada técnica pode ser aplicada em diversas tarefas. Dentre as mais utilizadas despontam as desenvolvidas inicialmente para outras áreas de conhecimento, tais como modelos estatísticos e probabilísticos, redes neurais e algoritmos genéticos.

2.4.1 Árvores de decisão

Árvore de decisão é uma das mais atrativas técnicas usada para classificação. Baseia-se na estrutura em formato de árvore que representa conjuntos de decisões que geram regras para a classificação de um conjunto de dados (SUMATHI E SIVANDAM, 2006).

O funcionamento da árvore baseia-se em uma série de questões cuidadosamente organizadas sobre os atributos de um determinado registro. Essa série de questões e suas possíveis respostas são organizadas em uma estrutura hierárquica constituída de nodos e arestas direcionadas verticalmente. Iniciando no nodo raiz, normalmente apresentado por convenção no topo do diagrama de decisão, os atributos são testados pelos nodos internos, cada qual com seu conjunto de alternativas possíveis, criando novo nível de ramificação até alcançar os nodos folha ou terminais os quais recebem o rótulo de uma classe. A figura 6 ilustra uma classificação de risco de crédito.

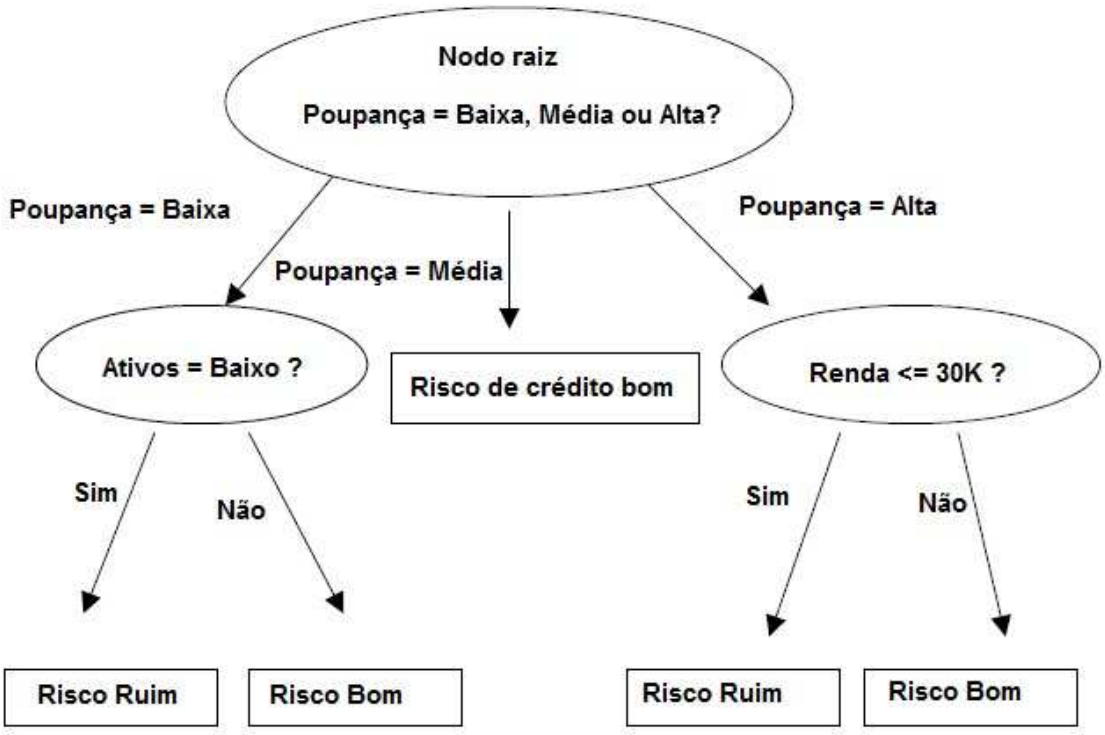


Figura 6 - Exemplo de uma árvore simples. Adaptado (LAROSE, 2006).

O exemplo ilustrado na figura 6 demonstra o algoritmo utilizado para classificar o atributo alvo risco de crédito, de um determinado conjunto de clientes. A classificação inicia no nodo raiz com o atributo poupança, juntamente com os nodos ativos e renda. No exemplo ilustrado na figura 6, o cliente que possui uma poupança média é considerado bom, sem a necessidade de novos nodos para tal afirmação. O nodo raiz que representa o início da árvore é definido pelo algoritmo de classificação. Já os nodos subseqüentes, são calibrados pelos resultados de acertos do algoritmo de classificação (LAROSE, 2006). Assim, determinados requisitos devem ser acordados antes da aplicação do algoritmo da árvore de decisão:

- Os algoritmos de árvore de decisão representam o aprendizado supervisionado, necessitando de atributos alvo pré-classificados. Neste caso, um conjunto de dados de treinamento deve ser aplicado para estabelecer valores aos atributos alvos.

- O conjunto de treinamento deve ser rico e variado, abrangendo todas as possibilidades de registros que possam ocorrer no futuro. Algoritmos de classificação aprendem com exemplos, e se os exemplos estiverem incompletos ou ausentes, o resultado da classificação fica comprometido.

- O atributo alvo deve conter valor discreto e não contínuo para possibilitar a definição de pertencer a uma classe ou não.

Um conjunto de atributos pode ser utilizado para construir árvores de diversas maneiras e apesar do recurso computacional, encontrar a árvore ótima torna-se inviável pelo tamanho exponencial do campo de pesquisa. Os algoritmos utilizam uma estratégia de crescimento tomando uma serie de decisões locais sobre qual atributo particionar. O algoritmo *Hunt* é um dos pioneiros utilizado para indução de árvore de

decisão é foi utilizado como referência para criação do ID3, *CART* e o algoritmo C4.5 (TAN *et al*., 2009).

A técnica da mineração através de árvore de decisão consiste na definição dos atributos preditivos que serão utilizados para compor a árvore responsável pela classificação. Os primeiros algoritmos propostos para criação de árvores tinham como requisitos básicos a discretização[3] dos atributos selecionados para predição como exemplo, o algoritmo ID3. Os algoritmos C4.5 e C5.0 têm capacidade de manipular propriedades nominais, ordinais e numéricas além da capacidade de trabalhar com atributos ausentes (QUINLAN, 1993).

Segundo QUINLAN (1993) a classificação é a principal tarefa de sistemas baseado em aprendizado de máquina e pontua também o reconhecimento da contribuição do algoritmo de árvore de decisão C4.5 para essa descoberta de conhecimento em base de dados. O algoritmo após a análise de propriedades nominais e numéricas formula padrões em forma de árvore de decisão com capacidade para classificar novos itens, enfatizando a criação de modelos compreensíveis bem como sua acurácia.

2.4.2 Algoritmo de Associação

As regras de associação, também conhecida como análise de associação têm o objetivo de descobrir atributos ou características que ocorrem com determinada frequência e sempre juntos. Os relacionamentos descobertos são identificados na forma de regra associativa ou conjunto de itens frequentes. Os algoritmos de associação quantificam as regras por meio da ocorrência do relacionamento entre dois ou mais atributos, através de uma medida de suporte e confiança (LAROSE, 2006).

[3] Redução do domínio de informação de um atributo a um conjunto discreto de valores.

Uma regra de associação é uma expressão de implicação no formato $X \rightarrow Y$, em que *X* e *Y* são conjuntos disjuntos de itens. O suporte determina a frequência da regra aplicável a um conjunto de dados e a confiança, a frequência de *Y* em observações que contenham *X*. O suporte é utilizado porque uma regra com baixo suporte pode indicar uma casualidade e também grande probabilidade de não haver interesse para a perspectiva do negócio. Já a confiança mede a confiabilidade da inferência feita por determinada regra. Suponha-se que em um supermercado em particular verificou-se que dentre 1.000 consumidores do quinto dia da semana no período noturno, 200 compraram fraldas, e dos 200 que compraram fraldas, 50 compraram cerveja. Assim a regra de associação seria “se comprar fraldas, compra cerveja” com um suporte de 200/1.000 = 20% e confiança 50/200 = 25% (LAROSE, 2006).

A definição da utilização de um algoritmo de associação está intimamente relacionada ao volume do conjunto de dados e da quantidade possíveis de regras de associações aplicáveis. A MD através de regras de associação foi proposta por *Agrawal* em 1993. Os algoritmos mais populares para as técnicas de regras de associação são o *Apriori*, *Eclat* e o algoritmo *FP-Growth* (ANGELINE, 2012).

2.4.3 Agrupamentos (*clusters*)

Análise de agrupamento consiste na divisão de registros de um conjunto de dados mediante as semelhanças de seus atributos. As técnicas de agrupamentos diferem da classificação porque não utilizam um atributo alvo para direcionar a criação dos grupos. Os algoritmos de agrupamento pesquisam todos os atributos de um registro procurando agrupá-los mediante sua homogeneidade, assim, maximizando a similaridade dentro de um grupo e minimizando a similaridade dos demais. Os agrupamentos são baseados em medidas de similaridades ou modelos probabilísticos

(SFERRA e CORRÊA, 2003). A definição de proximidade pode não estar claramente definida pelos atributos. No entanto, diversos algoritmos são utilizados para auxiliar na definição de agrupamentos através de medidas de proximidade tais como o algoritmo *K-means*, redes *de Kohonen e DBSCAN*.

Devido ao grande volume dos conjuntos de dados das bases atuais, em alguns casos, análise de agrupamentos é utilizada como uma etapa preliminar no processo de MD. Essa etapa preliminar visa criar um resumo de dados ou prepará-los para execução de outra técnica de mineração, como exemplo redes neurais. Análise de agrupamentos tem contribuído para exploração de conhecimento em diversas áreas tais como: psicologia e outras ciências sociais, estatística, biologia, recuperação de informações, reconhecimento de padrões, aprendizado de máquina e MD (TAN *et al*., 2009).

2.4.4 Redes neurais

A rede neural é uma tentativa de reproduzir o aprendizado não linear que ocorre na rede de neurônios do sistema neural biológico. É utilizada para tarefas que buscam estimação ou predição. Apesar da estrutura de um neurônio ser relativamente simples, uma rede pode apresentar uma conjuntura densa conectada por inúmeros neurônios com o intuito de executar tarefas complexas como classificação e reconhecimento de padrões (LAROSE, 2006). Podem ser definidas como sistemas computacionais compostos por inúmeros elementos de processamento, interconectados de acordo com uma topologia específica (arquitetura) e com capacidade de modificar seus pesos de conexão e parâmetros dos elementos de processamento (aprendizado) (ZORNETZER *et al*, 1994).

Um neurônio é um elemento fundamental no processamento de uma rede neural (HAYKIN, 1999). O processo de aprendizado é o conceito que norteia a rede neural, e

é realizado através dos neurônios. A figura 7 ilustra o funcionamento de uma rede neural que contém a seguinte estrutura básica:

- Um conjunto de sinapses ou conexões de entrada ponderadas com um peso sináptico. Desse modo, um sinal x_j na entrada da sinapse *j* conectada ao neurônio *k* é multiplicado pelo peso w_{kj};
- Uma junção de soma responsável pela adição dos sinais de entrada ponderados pelos respectivos pesos do neurônio;
- Uma função de ativação geralmente não linear representando a ativação de saída do y_k neurônio.

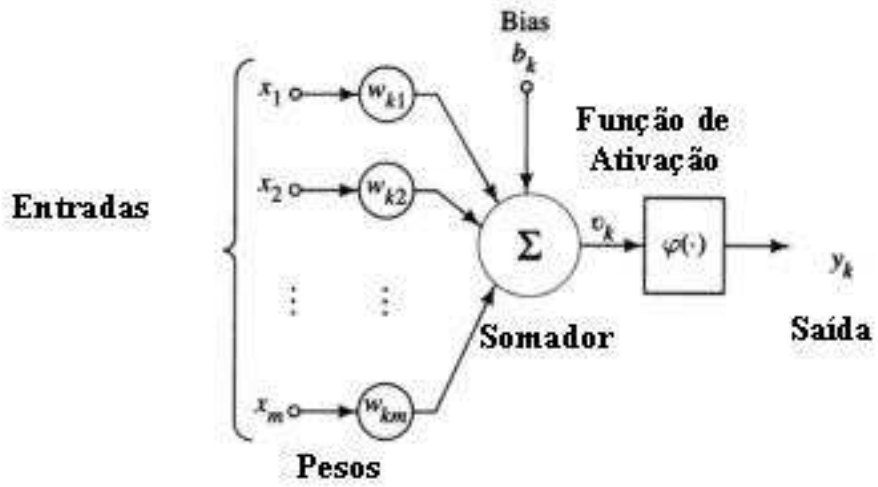


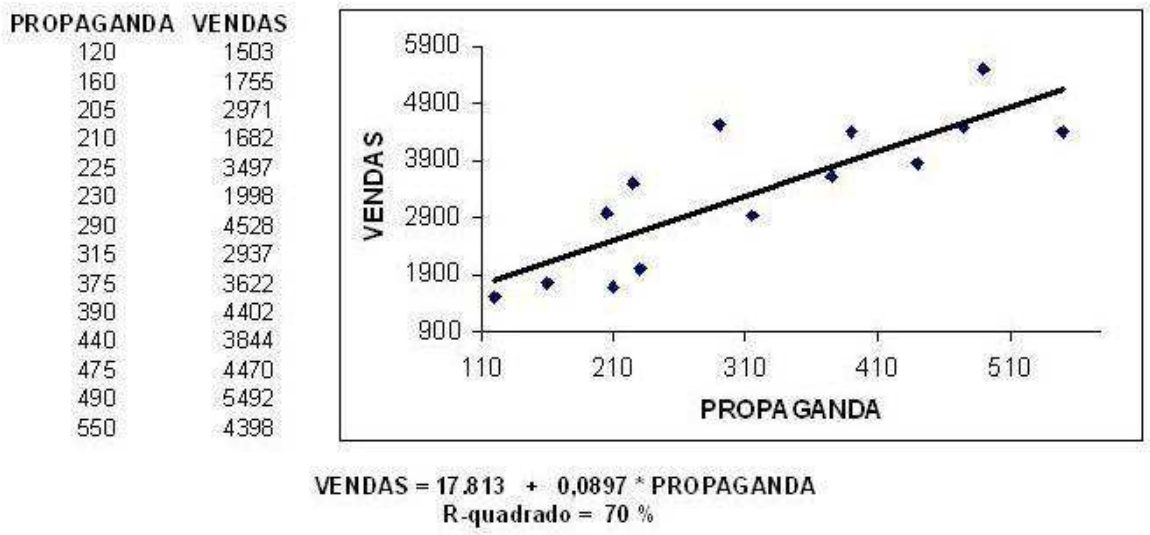
Figura 7 - Exemplo de uma rede neural. Adaptado (YODA, 2000).

O processo que busca a melhor calibração dos pesos w_{kj} é conhecido como processo de aprendizado ou treinamento da rede. A rede está em treinamento ou aprendizado para aprimorar e aproximar das informações desejadas através de iterações e ajustes nos pesos após as entradas informadas. O algoritmo de retropropagação é o mais utilizado para melhorar a predição dos dados.

2.4.5 Regressão

A regressão é uma técnica de modelagem preditiva classificada como método bivariado. Em alguns casos o objetivo de interesse é utilizar o valor de uma variável para determinar o valor de outras variáveis em determinada linha do tempo (LAROSE, 2006). A análise de regressão é a técnica de quantificar a dependência entre variáveis dependentes e independentes (CHIU e TAVELLA, 2008). Tais dependências são observadas através de funções continuas que modelam as regressões lineares simples, lineares múltiplas e regressão logística.

A técnica de regressão linear destaca-se ao utilizar variável alvo contínua para examinar o relacionamento entre uma variável dependente e demais variáveis independentes através de uma equação de regressão, como exemplo, prever o índice de indicadores econômicos. A figura 8 ilustra a aplicação da técnica de regressão linear para avaliar o comportamento da relação entre vendas e propagandas. O objetivo é prever qual será o quantitativo de vendas baseado no gasto realizado com propaganda.



PROPAGANDA	VENDAS
120	1503
160	1755
205	2971
210	1682
225	3497
230	1998
290	4528
315	2937
375	3622
390	4402
440	3844
475	4470
490	5492
550	4398

Figura 8 - Exemplo de regressão linear. Extraído de (RUD, 2001).

O exemplo da figura 8 ilustrou que 70% da variação apresentada nas vendas pode ser explicada pela relação com a variável propaganda inclusa no modelo. O indicador apontado pela variável R-quadrado com valor de 70% é conhecido como método dos mínimos quadrados ou soma dos quadrados residuais que tem objetivo de minimizar a soma do erro quadrado.

A técnica de regressão logística é uma variação da regressão linear destacando-se pelo uso de uma variável binária indicando a possibilidade probabilística de o evento ocorrer. A figura 9 ilustra a aplicação da regressão logística para observar a probabilidade de um indivíduo desenvolver uma determinada doença.

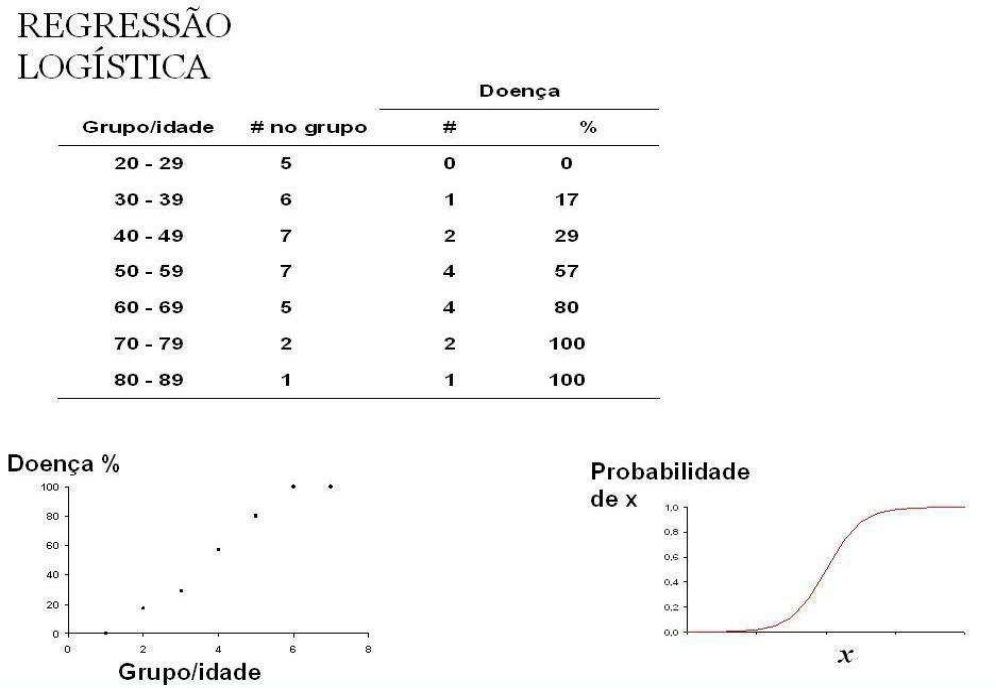


Figura 9 - Exemplo de regressão logística. Extraído (CORVALÃO, 2009).

Em alguns casos, a variável dependente é qualitativa e tem duas possibilidades, assumindo assim um valor binário zero e um. As variáveis independentes podem possuir valores discretos ou contínuos. Neste caso, o método de mínimos quadrados

não oferece estimadores satisfatórios e uma boa aproximação é obtida pela regressão logística para calcular ou prever a probabilidade ocorrência de um evento especifico (FIGUEIRA, 2006).

2.4.6 Análise estatística

Dentre as várias técnicas tradicionais utilizadas na análise de dados, a análise estatística é a que mais se aproxima de MD.

Assim como estimação e predição são tarefas de mineração, a análise estatística tem realizado a atividade de MD no ultimo século através de métodos estimativos e intervalo de confiança (LAROSE, 2006). No entanto, a análise estatística é orientada para validar hipóteses. Nesse sentido, a maioria das técnicas de estimativas populares requer o desenvolvimento prévio de uma hipótese, juntamente com o desenvolvimento manual de uma equação que atenda a hipótese (CABENA *et al*., 1998).

Contudo, a estatística tem um papel fundamental na maioria dos projetos de mineração, e a melhor estratégia é utilizá-la em conjunto com abordagens complementares de mineração.

2.4.7 Algoritmos genéticos

A computação natural é uma recente aposta para criação de modelos visando resolver os problemas atuais de otimização através de processos evolutivos e uma ferramenta de otimização, inspirado em fenômenos naturais. Os algoritmos genéticos podem ser definidos como uma técnica de otimização baseada nos conceitos de combinação genética, mutação e seleção natural (SUMATHI e SIVANANDAM, 2006).

Nos últimos tempos inúmeras técnicas inspiradas na natureza foram desenvolvidas e inseridas no contexto de mineração de dados, dentre elas às redes neurais, lógica

nebulosa, e computação evolutiva, no intuito de observar e comparar dispositivos naturais e artificiais.

Os algoritmos genéticos são técnicas inspiradas na teoria da evolução natural de Darwin, utilizada para resolução ou modelagem de processos evolutivos, mediante aplicação de uma heurística de seleção natural. O processo é adaptativo e após um estado inicial, inúmeras iterações são realizadas com o objetivo de melhorar os resultados iniciais, promovendo uma competição e seleção de indivíduos para uma nova população (PIZZIRANI, 2003). O desenvolvimento da melhor solução simula a seleção natural do processo de evolução mediante operadores de seleção, mutação e cruzamento para prover sucessíveis geradores de solução.

2.5 MODELOS DE DCBD

A eficiência da aplicação do processo de MD está condicionada ao uso de um método sistemático, que contenha regras e padrões formalizados para auxiliar e direcionar sua realização. A utilização de um modelo possibilita a abstração e representação das atividades do processo, como também possibilitar o gerenciamento e acompanhamento da sua execução.

A metodologia *Cross-Industry Standard Process for Data Mining* (CRISP-DM) foi desenvolvida em 1996 por um consórcio de especialistas representando *Daimler-Chrysler*, SPSS e NCR (SHEARER, 2000). Os idealizadores utilizaram sua experiência profissional para prover um processo padrão não proprietário e gratuito que incorporasse objetivos da organização e conhecimento. De acordo com o CRISP-DM, o ciclo de vida do processo de MD é composto de seis fases, como ilustra a figura 10.

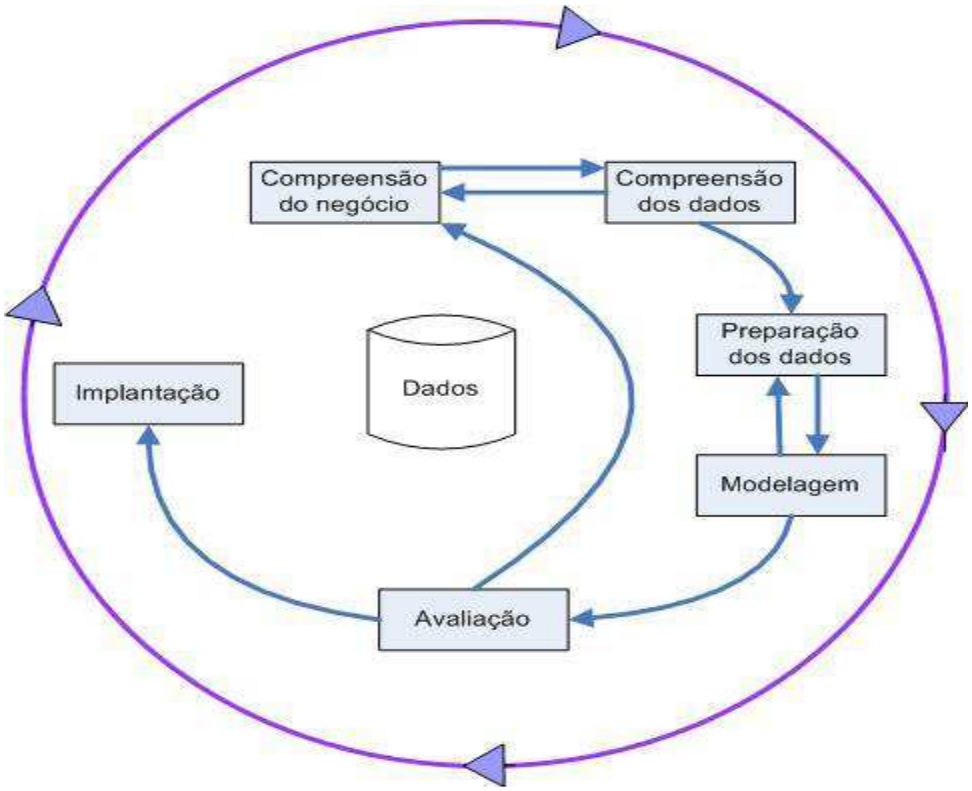


Figura 10 - CRISP-DM, como um processo iterativo e adaptativo. Adaptado (LAROSE, 2006).

Observe que o processo é adaptativo, uma vez que a sequência da próxima fase pode ser determinada pelos resultados da fase anterior. Suponha que a fase de modelagem não contenha modelos satisfatórios. Desse modo, a fase de preparação pode ser acionada novamente, antes do processo prosseguir para a fase de avaliação. A linha externa da figura 10 demonstra a natureza iterativa do CRISP-DM.

2.5.1 Compreensão do negócio

Representa o inicio do modelo, tornando-se elemento norteador para o processo de MD. Compreende o pleno entendimento do negócio entre todos participantes para determinar quais os objetivos almejados do negócio e possíveis restrições, preparando uma estratégia preliminar em busca dos objetivos através de um plano de projeto de MD.

Em um setor de fiscalização, esta fase envolve a compreensão da composição do sistema de tributação, legalidade e competências, a execução da auditoria e abrangências, informações registradas e melhorias que possam ser apresentadas visando combater as dificuldades enfrentadas pelo departamento de fiscalização na identificação de empresas com indícios de irregularidades. Esta informação deve ser adicionada na definição do problema e no projeto de MD (CORVALÃO, 2009).

2.5.2 Compreensão dos dados

A fase de compreensão envolve a coleta, exploração e familiarização com os dados pelos integrantes do projeto objetivando averiguar se são passíveis de mineração.

Utiliza a análise exploratória para familiarizar com os dados e avaliar fontes, qualidade e características que auxiliem a descobrir percepções iniciais que ajudem a moldar o projeto de MD. É uma fase crucial, pois fornece subsídios para montagem dos modelos de MD (LAROSE, 2006). Uma seleção cuidadosa de atributos independentes pode facilitar e viabilizar a construção de modelos para obtenção de conhecimento (OLSON e DELEN, 2008).

Esta fase necessita de um trabalho mais apurado, pois neste ponto pode-se declarar a viabilidade ou não do projeto de MD. Devido ao fator criticidade, esta fase pode consumir muito tempo, mas é crucialmente importante para o sucesso do projeto (CORVALÃO, 2009).

2.5.3 Preparação dos dados

A fase de preparação objetiva uniformizar os dados, tornando-os mais claros e compreensíveis. Normalmente, a coleta de dados pode advir de várias fontes distintas com formato e tecnologias diferentes, o que influi no padrão de armazenagem.

A fase de preparação dos dados envolve um pré-processamento visando modelar dados extraídos de fontes diversas para um conjunto definido que é utilizado nas fases subsequentes. Essa padronização é importante para viabilizar a construção do modelo e envolve a seleção das variáveis que serão apropriadas para análise, limpeza, transformação, integração e formatação (LAROSE, 2006). Através dos objetivos traçados na fase de compreensão do negócio, o analista determina quais tipos de dados são relevantes e quais técnicas de mineração utilizar. Esta fase também trata possíveis problemas com os dados, como dados faltantes (MICCI-BARRECA e RAMACHANDRAN, 2006).

2.5.4 Modelagem

A fase de modelagem concentra-se na concepção de algoritmos de MD que visem extrair o conhecimento por meio dos dados do conjunto preparado na fase anterior. Foca na seleção e aplicação de algoritmos, técnicas de modelagem apropriadas ao problema selecionado e na calibração do modelo para melhorar os resultados. Diferentes técnicas podem ser utilizadas para um mesmo problema de MD e cada técnica, requer tipos específicos de dados. Se necessário, o processo pode retornar a fase de preparação para trazer dados compatíveis a um determinado tipo de técnica particular (LAROSE, 2006).

A saída da fase de modelagem resulta na criação de um modelo ou um conjunto de modelos contendo o conhecimento descoberto em um formato apropriado (MICCI-BARRECA e RAMACHANDRAN, 2006). É nesta fase que se estabelece o modelo de solução do problema.

2.5.5 Avaliação

Esta fase objetiva interpretar, avaliar e comparar os resultados obtidos dos modelos idealizados na fase anterior no que concerne a sua qualidade e eficiência, verificando se o modelo atinge os objetivos de negócio definidos na primeira fase (LAROSE, 2006).

Os algoritmos de mineração podem gerar um numero ilimitado de padrões e em alguns casos, irrelevantes aos objetivos traçados pelo projeto de MD. A saída desta fase alicerça a decisão sobre quais modelos utilizar baseado nos resultados obtidos e averiguar sua possibilidade de implantação.

2.5.6 Implantação

Na fase de implantação o objetivo é fazer uso do modelo proposto e avaliado e a ocorrência de sua disponibilização em ambiente organizacional não significa que o projeto esteja finalizado. A fase de implantação também envolve processos repetitivos para o aperfeiçoamento do modelo ou sua recalibração (MICCI-BARRECA e RAMACHANDRAN, 2006). E importante que os gestores do processo tenham em mente a dinâmica apresentada pelos negócios e produzam modelos que sejam passíveis de adequação e possibilite obtenção de novos resultados. Como exemplo, as frequentes modificações que a legislação tributária sofre ao longo do tempo (CORVALÃO, 2009).

A atividade de disponibilizar o resultado do modelo ao ambiente organizacional pode não ser trivial. Pode ser simples tal como a criação de um relatório, ou complexa quanto à implantação de um processo de MD em toda empresa (MICCI-BARRECA e RAMACHANDRAN, 2006). Esse processo de reengenharia e customização de uma ferramenta que faça interpretação dos dados minerados como suporte de apoio a

decisão é imprescindível para que os gestores do negócio utilizem a informação de forma correta.

2.6 LIMPEZA DE DADOS E PRÉ-PROCESSAMENTO

Os tópicos anteriores dedicaram à introdução dos conceitos de MD juntamente com o modelo de processo padrão CRISP-DM. O modelo ilustra a primeira fase e demonstra uma ideia de como aplicá-la na obtenção da compreensão do negócio. O núcleo de um projeto de MD é a fase de modelagem, e para construí-la os dados necessitam de uma limpeza e pré-processamento para garantir uma padronização.

Os dados oriundos de conjuntos de dados diversos podem estar sujeitos a ruídos. Podem conter dados faltantes, redundantes, obsoletos, fora de limites estabelecidos (*outliers*) ou em um padrão inadequado para avaliação dos modelos de MD. As atividades de limpeza e pré-processamento viabilizam as finalidades do projeto de MD ao sintetizar o conjunto ideal para atividade de modelagem e pode consumir ate 60% do esforço total empreendido no projeto. (LAROSE, 2006).

A limpeza dos dados concentra-se na uniformização ao estabelecer um domínio a determinado atributo, evitando que dados sejam equivocadamente descartados. A uniformização é estabelecida através de um domínio que descreve quais tipos de dados são válidos para um determinado atributo. Assim, uma convenção singular é utilizada para representar os dados do sistema, detectando possíveis erros.

2.7 TRABALHOS CORRELATOS

A MD tornou-se campo de pesquisa e objeto de estudo pela necessidade de extrair conhecimento de conjuntos de dados através de um processo. Na literatura foram encontrados diversos trabalhos relacionados ao tema proposto, realçando a

importância da MD na detecção de irregularidades, em especial na declaração de impostos e fraudes.

A tarefa mais utilizada nessa linha de estudo incide sobre a classificação. Alguns trabalhos utilizaram técnicas diversas de classificação tais como árvore de decisão, redes neurais e redes bayesianas com objetivo de classificar contribuintes com indícios de irregularidades e outros, através da regressão e modelos de séries temporais, mensurar possível desfalque de devedores ou previsão de arrecadação. No entanto não foi encontrado um trabalho que envolvesse especificamente a classificação de contribuintes com irregularidades do Imposto Sobre Serviços de Qualquer Natureza (ISS). O trabalho aqui proposto visa classificar os contribuintes mediante suas particularidades socioeconômicas tais como modalidade jurídica, tipo de atividade, tempo de atividade e número de empregados diferente dos trabalhos com ICMS que utilizaram o contexto econômico e fiscal através de atributos como inventário de entradas e saídas, mapeamento fiscal e registros de postos fazendários como atributos preditivos para a classificação.

A seguir são discutidos trabalhos baseados em classificação no contexto de contribuintes com irregularidades.

- CORVALÃO (2009) descreve um trabalho cujo objetivo é classificar os contribuintes do Imposto Sobre Circulação de Mercadorias e Serviços (ICMS) junto a Secretaria da Fazenda de Santa Catariana com indícios de irregularidades. A similaridade é a utilização da metodologia CRISP-DM para condução do processo de MD e escolha da tarefa de classificação. O trabalho proposto pelo autor foi evidenciado em duas fases. Uma para realizar o agrupamento dos contribuintes, motivado pelo volume de dados selecionado e pelas características do experimento realizado, e posteriormente, outra para a classificação. Para realizar a classificação dos

contribuintes irregulares o autor realizou a priori a tarefa de agrupamentos por meio da técnica de análise de cluster *two-step* para criar grupos de empresas baseado nas características regionais e econômicas no intuito de reduzir a dimensão dos dados. A escolha desta técnica de agrupamento foi escolhida pelo autor pelo fato da mesma manipular grandes volumes de dados e da capacidade de processar atributos contínuos quanto categóricos. Apresentou como resultado 21 clusters, contendo a quantidade de empresas e média de faturamento de cada cluster. Esta atividade inicial no processo de MD possibilitou a análise multivalorada das empresas pelos atributos microrregião, faturamento e atividade econômica. Posteriormente, o autor utilizou a técnica de regressão logística através de modelos probabilísticos para prever a classificação mediante a variável categoria de notificação, a partir de demais variáveis geradas pelo cluster.

- Já BRAGA (2010) apresenta um trabalho cuja proposta é realizar um comparativo do resultado das técnicas rede neural e regressão linear para prever o valor a ser declarado pelos contribuintes do Imposto de Renda (IR) válido para Pessoas Jurídicas, com intuito de identificar irregularidades na Receita Federal do Brasil. A similaridade é a utilização da metodologia CRISP-DM para condução do processo de MD e objetivar a classificação dos contribuintes. O autor visou comparar o resultado de duas técnicas que permitem classificação de contribuintes mediante a comparação da receita bruta prevista e a declarada pelos contribuintes. O autor utilizou a rede neural *perceptron* com múltiplas camadas (MLP) e regressão linear através do atributo alvo receita. Os resultados obtidos pelas duas técnicas demonstram indícios de irregularidades mediante as variáveis utilizadas, visualizado uma vantagem de abrangência adquirida pela técnica de rede neural mediante a natureza dos dados selecionados.

- ANDRADE (2009) apresenta em seu trabalho proposta para indicar possíveis sonegadores do ICMS junto a Secretaria da Fazenda do Ceará. A similaridade é a utilização da metodologia CRISP-DM para condução do processo de MD e escolha da tarefa de classificação. O autor propôs dividir a fase de modelagem do CRISP-DM nas subfases: análise de agrupamentos, seleção de atributos e classificação. O autor utilizou a técnica de rede neural do tipo mapa auto-organizável para realizar a montagem dos *clusters* agrupando os registros em conjuntos baseados em suas próprias características e para realizar a classificação dos contribuintes irregulares, posteriormente utilizou novamente a técnica de rede neural do tipo *perceptron* MLP. Entre essas duas fases, o autor utilizou árvore de decisão para descartar dados não significativos para o processo e otimizar o tempo gasto com treinamento das redes. A proposta do autor visa classificar os contribuintes através de uma rede neural pelo fato do atributo alvo utilizado ser uma referencia de coordenadas x, y dos agrupamentos criados pela rede neural.

A seguir são discutidos trabalhos baseados na detecção de contribuintes com irregularidades.

- VENKATESWAR RAO *et al*. (2005) apresenta um trabalho para detectar a evasão fiscal através do processo de MD com dados dos contribuintes da Índia. No processo de MD, os autores apresentaram um novo algoritmo de busca inovador que detecta variações no nome dos contribuintes, convertendo os caracteres para uma combinação numérica e utilizando operações matemáticas para determinar a combinação do resultado da busca. O algoritmo apresentado visa superar as limitações de precisão, escalabilidade e velocidade das buscas convencionais em grandes bases

de dados. Os autores demonstraram a aplicabilidade da busca dos nomes dos contribuintes na geração do perfil do imposto de imposto de renda e correlacionar com os dados recebidos pelo departamento fiscal para detectar a evasão.

- CLEARY (2011) apresenta um trabalho realizado com MD para auxiliar o departamento de auditoria na melhor seleção dos contribuintes do imposto de renda na Irlanda. O modelo proposto pelo autor foi construído pela ferramenta SAS *Enterprise Miner* (SAS, 2013) e utilizado no ano de 2011, para predizer a probabilidade de um contribuinte sofrer uma intervenção de auditoria. O autor utilizou a metodologia SEMMA (SEMMA, 2013) juntamente com modelo preditivo de mineração. Após a utilização do modelo em campo pelo departamento de auditoria, o departamento verificou uma média de 75% de acerto dos contribuintes informados pelo modelo.

A seguir são discutidos trabalhos baseados em árvore de decisão no contexto de contribuintes com irregularidades

- BONCHI *et al*., (1999) apresenta um trabalho que enfatiza a importância da MD na detecção de fraude ressaltando uma das principais aplicações de MD. O autor ressalta que o principal objetivo do trabalho é a construção de modelos ou perfis que indiquem o comportamento fraudulento para possibilitar otimizar o departamento de fiscalização. O resultado da proposta de MD apresentado visa um trabalho preventivo na detecção de fraudes e na obtenção de redução de custos no planejamento de estratégias de auditorias. Para realizar o estabelecimento das estratégias que maximize os benefícios da auditoria e minimize os custos, o autor

utilizou o DCDB para elucidar o processo de MD juntamente com a classificação através da técnica de árvore de decisão.

Ainda relacionado à aplicação da técnica de árvore de decisão na detecção de fraude e irregularidades YU, QIN, e JIA (2006) apresenta um trabalho intitulado *Data Mining Application Issues in Fraudulent Tax Declaratin Detection* que descreve processo de MD no intuito de criar modelos que propõe a descoberta de fraude fiscal através bases de dados.

III IMPOSTO E TRIBUTAÇÃO

A arrecadação de tributos compõe o conjunto de receitas públicas correspondente ao recolhimento efetuado ao erário em número ou espécie que represente valor. Tal receita tem caráter essencial e compreende o montante de recursos que possibilitará ao ente público realizar despesas e planejamentos sob sua responsabilidade.

Este capítulo visa apresentar o ramo do direito positivo brasileiro com normas jurídicas denominadas normas jurídico-tributárias que dispõem sobre instituição, arrecadação e fiscalização de tributos.

3.1 O CONCEITO DE TRIBUTO

Conforme a Lei n° 4.320, de 17 de Março de 1964, qu e institui normas gerais do Direito Financeiro, dispõe que:

> *"Art 9- Tributo é a receita derivada, instituída pelas entidades do direito público, compreendendo os impostos, taxas e contribuições nos termos da Constituição e das leis vigentes em matéria financeira, destinando-se o seu produto ao custeio de atividades gerais ou específicas exercidas por essas entidades."*

Já o Código Tributário Nacional (CTN), dispõe que:

> *"Art 3- Tributo é toda prestação pecuniária compulsória, em moeda ou cujo valor nela se possa exprimir, que não constitua sanção de ato ilícito, instituída em lei e cobrada mediante atividade administrativa plenamente vinculada"*.

Tendo em vista a importância do termo tributo e a variação de conceitos dispostos na literatura diante da noção geral oferecida pela Constituição Federal (CF) de 1988, para o contexto deste trabalho o termo tributo é utilizado na acepção de norma tributária em sentido estrito, restrito, objetivo, ou seja, observado o princípio da legalidade em que todos os tributos são criados por leis. O termo tributo é norma

jurídica que orienta o comportamento de o particular entregar determinada quantia em dinheiro ao erário, quando se realizar o fato lícito descrito em sua hipótese normativa (TOMÉ, 2005).

3.2 IMPOSTO

O CTN em seu artigo 16 institui imposto com a seguinte definição:

> *"Art. 16- o tributo cuja obrigação tem por fato gerador uma situação independente de qualquer atividade estatal específica, relativa ao contribuinte"*.

A CF de 1988 distingue aos entes da federação a instituição de impostos de sua competência. Especificamente aos municípios, o artigo 156 com redação dada pela emenda constitucional nº3 de 1993, dispõe os impostos que o mesmo tem competência de arrecadar e fiscalizar. Importante ressaltar que a receita pública do município é constituída pelo Imposto Predial e Territorial Urbano (IPTU), Imposto sobre a Transmissão de Bens Imóveis (ITBI) e Imposto sobre Serviço de Qualquer Natureza (ISS). Essa composição é validada na CF de 1998 que dispõe:

> *"Art. 156- Compete aos Municípios instituir impostos sobre:*
>
> *I - propriedade predial e territorial urbana;*
>
> *II- transmissão* inter vivos*, a qualquer título, por ato oneroso, de bens imóveis, por natureza ou acessão física, e de direitos reais sobre imóveis, exceto os de garantia, bem como cessão de direitos a sua aquisição;*
>
> *III- serviços de qualquer natureza, não compreendido no art. 155, II, definidos em lei complementar."*

3.3 IMPOSTO SOBRE SERVIÇOS DE QUALQUER NATUREZA – ISS

O ISS é instituído pelos municípios. O Sistema Tributário Nacional através da CF de 1988 em seu art. 156, III, definiu que compete aos municípios instituir o imposto sobre serviços de qualquer natureza salvo sobre os serviços de transporte municipal e

interestadual e de comunicação cuja competência é dos Estados e do Distrito Federal. Em matéria tributária a constituição não foi genérica e sintética. O legislador delineou minuciosamente a atividade tributária que se inicia com a instituição do tributo, fiscalização, e inserção da entrada de receita aos cofres públicos. Delimitou também em inúmeros preceitos jurídicos o exercício de tributar vinculado aos entes políticos (CUNHA, 2007).

3.3.1 Conceito de serviço

Assim, quando a regra matriz de competência delega aos municípios a possibilidade de instituir o imposto sobre serviços, está delimitando que o objeto tributação é o serviço. No entendimento do Supremo Tribunal Federal (STF) “a constituição, quando atribui competência impositiva ao Município para tributar serviços de qualquer natureza, não compreendidos na competência das outras pessoas políticas, exige que só se alcancem, mediante incidência do ISS, os atos e fatos que possam lhe qualificar, juridicamente como serviços” (CUNHA, 2007).

A CF utilizou o ordenamento civil para qualificar juridicamente o vocábulo “serviço”. Porém, no Código Civil (CC) não há uma definição precisa do que venha a ser serviço, mas apenas relata a regulamentação de sua contratação. O artigo 594 do CC define da seguinte forma: “toda espécie de serviço ou trabalho lícito, material ou imaterial, pode ser contratado mediante retribuição”. O Código do Consumidor, Lei n° 8.078, de 11 de setembro de 1990, define serviço como qualquer atividade fornecida no mercado de consumo, mediante remuneração, inclusive de natureza bancária, financeira, de crédito e securitária, salvo as decorrentes das relações de caráter trabalhista.

Destas fontes, pode-se destacar que o ato serviço envolve uma atividade a terceiro, a prestação de um esforço humano físico ou intelectual na execução de uma

atividade qualquer, que leva em consideração a qualidade e a técnica de ser humano, e não o resultado final de sua execução (CUNHA, 2007). Assim, a prestação de serviços não se destaca pelo produto final, mas pela destreza de pessoas envolvidas na sua execução. A prestação de serviço pode resultar em uma atividade meramente intelectual, com a produção de um bem imaterial como também pode produzir um bem material.

3.3.2 Fator econômico

Outro fator relevante é o fator econômico. Os serviços desprovidos de conteúdo econômico não podem ser objeto de tributação por não representarem riqueza econômica. A CF arrolou fatos que pode presumir-se em riqueza. Consequentemente, serviços sem conotação econômica não podem receber tributação tais como gratuitos ou de cortesia, em regime familiar, filantrópicos, religiosos e os altruísticos (CUNHA, 2007).

O Código do Consumidor reforça tal entendimento ao definir serviço como “qualquer atividade fornecida no mercado de consumo, mediante remuneração”. Neste sentido, sem o conteúdo econômico, é inexistente o fator essencial do critério quantitativo da hipótese de incidência, que é a base de cálculo, sem a qual a incidência não completa.

De acordo com a CF, extrai-se que o serviço objeto de tributação do ISS é aquele regido pelo direito privado, não incluindo o serviço público que é prestado sem qualquer vinculo de subordinação, salvo os serviços públicos específicos, por disposição da CF, os quais são objetos de tributação por meio de taxa (art. 145, III). O serviço público é todo aquele prestado pela Administração Pública, ou por seus delegados, sob normas e controles estatais, para satisfazer necessidades essenciais

ou secundárias da coletividade, ou por simples conveniência do Estado (MEIRELLES, 2004).

3.3.3 Expressão “lei complementar”

A CF ao instituir o art. 156, III, definiu que compete aos municípios instituir o imposto sobre serviços de qualquer natureza, não compreendidos no art. 155, II, “definidos em lei complementar”.

No entendimento doutrinário, a CF outorgou aos municípios competência para legislar sobre impostos sobre serviços de qualquer natureza, definidos em lei complementar, e não para onerar todo e qualquer serviço. Caso haja inexistência da lista de serviço baixada por lei complementar não pode o município escolher tais serviços por lei ordinária municipal (CUNHA, 2007). A lei federal complementar n° 116, de 31 de julho de 2003, estabelece a referência nacional para catálogo de serviços.

3.4 SISTEMA TRIBUTÁRIO DE GOIÂNIA

O sistema tributário do município de Goiânia é regido pelo Código Tributário do Município (CTM), mantido pela Secretaria de Finanças. O CTM é composto de duas partes:

- Parte I – Código Tributário Municipal através da Lei n° 5040, de 20 de Novembro de 1975, que dispõe sobre o Código Tributário do Município de Goiânia e demais providencias.
- Parte II – Regulamento do Código Tributário através do Decreto n° 2.2173 de 13 de Agosto de 1996, que aprova o regulamento do Código Tributário Municipal de Goiânia.

Dentre outras providencias, o CTM especifica o catálogo de serviços constantes da lei complementar n° 116/2003, fato gerador, incidên cia de calculo e isenção, alíquotas, contribuintes incidentes na legislação tributária, contribuinte prestador, contribuinte substituto, apuração e recolhimento do ISS.

O CTM especifica também as obrigações acessórias, infrações e penalidades sujeitas ao regime da legislação. As obrigações acessórias incidem as taxas que atestam a regularidade das empresas atuantes na execução de serviços, tal como taxa de licença ou exercício de atividade especial. Esclarece também a competência e responsabilidade das autoridades fiscais e de fiscalização como também das orientações do processo administrativo tributário.

O sistema de arrecadação de Goiânia é composto de sistemas informatizados para auxiliar no gerenciamento dos contribuintes regulares prestadores de serviços no município. Antes do inicio de suas atividades, qualquer pessoa jurídica ou física que exerça atividade econômica no município de Goiânia, sejam elas atividades comerciais, prestacionais ou industriais, devem se inscrever no sistema de Cadastro de Atividades Econômicas (CAE), independentemente se são sujeitas ou não a incidência de recolhimento do ISS, taxas, imunes e tributáveis. O CTM dispõe de informações sobre documentação necessária para a empresa jurídica e quais processos necessários às pessoas físicas, normalmente reservadas a autônomos, para se regularizar. O sistema CAE gerencia todas as informações cadastrais dos contribuintes regulares.

Na parte fiscal, o município dispõe do sistema de Declaração Mensal de Serviços (DMS) para realizar o monitoramento contínuo, formalizado e informatizado a partir de 2005, substituindo o Livro de Serviços Prestados na sua forma originária de preenchimento. Esse mecanismo eficiente facilita a apresentação das informações referentes aos serviços prestados pelos contribuintes na forma eletrônica, com

apuração automática do ISS devido. O sistema DMS registra todas as notas fiscais emitidas pelas empresas enquadradas em determinado período, inclusive as canceladas. A partir de 2012 as empresas estão se adequando a Nota Fiscal de Serviços Eletrônica (NFS-e).

O setor de fiscalização dispõe de sistema informatizado para averiguar o andamento regular das empresas cadastradas no CAE. A fiscalização tem duas formas de averiguação:

- Descumprimento de obrigações acessórias – visa averiguar regularização da empresa mediante notas fiscais, diários, taxas de funcionamento ou expediente, registros de empregados etc;

- Irregularidades de recolhimento – visa averiguar se o informado pelo contribuinte condiz com o realizado normalmente sob divergências na retenção ou omissão de impostos e taxas.

O setor de fiscalização desempenha papel importante no que tange ao aspecto social; além do objetivo de diminuir a sonegação e consequentemente aumentar a arrecadação, através da inteligência fiscal mediante os sistemas informatizados, busca a formalidade, a legalidade e angariar recursos para o município. Através de tal análise e possível reaver alíquotas e benefícios a segmentos que estão atuando na informalidade ou então conceder benefícios a segmentos que realizam investimentos em projetos sociais como o Projeto Estação Digital (DIGITAL, 2013), ao conceder benefícios às empresas do segmento de Tecnologia da Informação.

IV METODOLOGIA

A referência para conduzir as atividades deste projeto de pesquisa inclui métodos e técnicas disponíveis na literatura voltados a pesquisa científica em engenharia de produção e uma metodologia direcionada a organizar as atividades do modelo proposto de MD.

O projeto é classificado de acordo com seus objetivos de natureza exploratória e explicativa com abordagem predominantemente qualitativa, recorrendo a análises quantitativas utilizando-se do método de estudo de caso.

O trabalho utilizou dados dos contribuintes do ISS devidamente regularizados, registrado pela Secretaria de Finanças do Município de Goiânia através dos departamentos de Auditoria e Arrecadação, contemplados entre Janeiro e Dezembro de 2011, com o intuito de avaliar o perfil dos mesmos.

4.1 APLICAÇÃO DO MÉTODO ESTUDO DE CASO

Segundo YIN (2003), o método estudo de caso é definido como uma investigação empírica que investiga um fenômeno contemporâneo dentro de seu contexto da vida real, especialmente quando os limites entre o fenômeno e o contexto não estão claramente definidos. Nessa visão, a análise da perspectiva dos contribuintes do Município em relação ao domínio de irregularidades torna-se uma relação complexa. O método de estudo de caso pode apresentar variações de estratégias de pesquisa dividindo-se em estudo de caso exploratório, descritivo ou explanatório podendo basear-se em uma mescla de provas quantitativas e qualitativas (TURRIONE e MELLO, 2011).

YIN (2003) destaca ainda o desafio de natureza técnica intrínseco do Método de estudo de caso mediante a abrangência do contexto que fatalmente nesse método haverá mais variáveis de interesse do que ponto de dados.

Do objetivo para o qual é utilizado, um estudo de caso pode ser exploratório: “uma espécie de estudo piloto feito para testar as perguntas norteadoras do projeto, hipóteses e principalmente os instrumentos e procedimentos utilizados”; descritivo: “objetiva mostrar ao leitor uma realidade que ele não conhece”; explanatório: “tem por objetivo explicar não apenas uma realidade, mas também explicá-la em termos de causa-efeito” (TURRIONE e MELLO, 2011).

O uso do método estudo de caso é preferido ao se examinar acontecimentos contemporâneos, mas quando não se podem manipular comportamentos relevantes. O diferencial do método está na sua capacidade de lidar com uma ampla variedade de evidencias, podendo ser quantitativas, qualitativas ou ambas (TURRIONE e MELLO, 2011). Visão acompanhada por CHIZZOTTI (2008) ao defender que o estudo de caso faz uso de diversos meios de coletar informação e também GIL (2009) ao se referir ao teor pluralista do método estudo de caso.

Visto a natureza dos dados, a necessidade de contar com diversas fontes de evidências, as condições multivariadas culminaram na utilização do método estudo de caso neste projeto de pesquisa.

4.2 ESTRATÉGIA DE MD

Na condução do presente trabalho de MD é utilizada metodologia guiada por estratégias, métodos e técnicas disponíveis na literatura, como sugere (TURRIONE e MELLO, 2011).

Existem diversas metodologias especificas para MD, em destaque, o CRISP-DM que se tornou de fato referencia mundial para auxiliar o processo de desenvolvimento de um projeto de MD (JACKSON, 2002).

A pesquisa objetiva desenvolver através do modelo DCBD: uma estrutura em árvore que alcance resultados satisfatórios na aplicação da tarefa de classificação através de um algoritmo de árvore de decisão, não sendo objeto deste estudo, desenvolver novas técnicas ou ferramentas computacionais que auxiliem na classificação de contribuintes. O algoritmo de árvore de decisão é indicado para realizar a classificação quando o atributo que distingue as classes previamente conhecidas de um conjunto de dados é idealizado apenas por um atributo alvo, nesse caso, o atributo que distingue os contribuintes regulares dos que apresentaram alguma irregularidade.

Para realizar a estratégia de MD, os dados precisam de um processo de preparação disposto nas seis fases do modelo DCBD: seleção dos dados, pré-processamento e limpeza, formatação, mineração e interpretação. No auxílio da execução das fases do modelo DCBD ferramentas computacionais são utilizadas como *PENTAHO* e *WEKA*. Como os dados originais estão dispostos em sistemas que operam com linguagem de *mainframe NATURAL/ADABAS*, uma ferramenta computacional é idealizada e construída para permitir a extração das informações das bases de origem a uma base de dados que possibilite a execução das fases do modelo DCBD.

4.3 RECURSOS UTILIZADOS

A metodologia orienta os passos necessários para almejar os objetivos do trabalho.

4.3.1 Linguagem de Mainframe NATURAL/ADABAS

O ADABAS é um Software Gerenciador de Banco de Dados (SGBD) idealizado e mantido pela empresa Software AG, disponível inicialmente para *mainframes*, uma espécie de computador de grande porte dedicado ao processamento de um volume grande de informações (ADABAS, 2012).

Despontou no mercado na década de 70, sendo apontado na literatura como um dos pioneiros sistemas de gerenciamento de dados produzido comercialmente, inicialmente lançado pela gigante da computação IBM. Como é uma base que opera com listas invertidas, não é considerado um SGBD relacional, o que inviabiliza a utilização de diversas ferramentas *On-line Analytical Processing* (OLAP) existentes no mercado para extração de dados. Mediante tal situação a linguagem NATURAL é uma linguagem de *mainframe* utilizada para ler os arquivos e registros no intuito de extrair e gerar a massa de dados para um ambiente de baixa plataforma em formato texto, contendo na primeira linha a descrição dos campos extraídos, padronização necessária para condução das fases do modelo DCBD.

4.3.2 Software de *Business Inteligence* (BI)

Após o processo de extração dos dados, a ferramenta computacional de BI *PENTAHO* é utilizada para auxiliar a fase de preparação dos dados do modelo DCBD. É uma solução de código aberto, mantido pela comunidade de software livre mundial pela modalidade *General Public Licence* (GPL). E um software baseado na linguagem Java, composto de diversos módulos formalizando uma suíte de solução BI. A ferramenta manipula massas de dados possuindo recursos que contemple diversas fases do modelo DCBD dentre eles: seleção da fonte de dados, integração, DW,

extração de dados, MD e análise mediante diversas opções de relatórios e painéis indicadores (PENTAHO, 2009).

Apesar da amplitude da ferramenta em atender a demanda das fases do modelo DCBD, viabilizando a realização das fases desde o pré-processamento até a avaliação dos dados, a mesma não possui algoritmos de classificação baseado em arvores de decisão; desse modo, a utilização da ferramenta WEKA será necessária para realizar a tarefa de MD.

4.3.3 Ferramenta *WEKA*

A ferramenta *WEKA* começou a ser idealizada em 1993 por um grupo de pesquisadores da universidade de *Waikato,* localizada na Nova Zelândia. Tal ferramenta foi escrita na linguagem de programação Java e é uma solução específica para realizar pesquisas voltadas para MD. Ao longo dos anos se consolidou como a ferramenta de mineração mais utilizada no meio acadêmico. Dentre seus recursos, contém uma coleção de algoritmos de aprendizado de máquina específico para as atividades de MD que podem ser aplicados diretamente a um conjunto de dados ou mediante customização para ser integrada a outra solução construída sobre a linguagem de código aberto Java (WEKA, 2012).

O *WEKA* é um software livre dentro das especificações *General Public License* (GPL). O recurso potencial da ferramenta incide sobre a tarefa de classificação, mas disponibiliza recursos para também realizar as tarefas de pré-processamento dos dados, regressão, clusterização, regras de associação e visualização.

A figura 11 ilustra a relação entre as fases do modelo DCDB e os recursos utilizados para realizá-la.

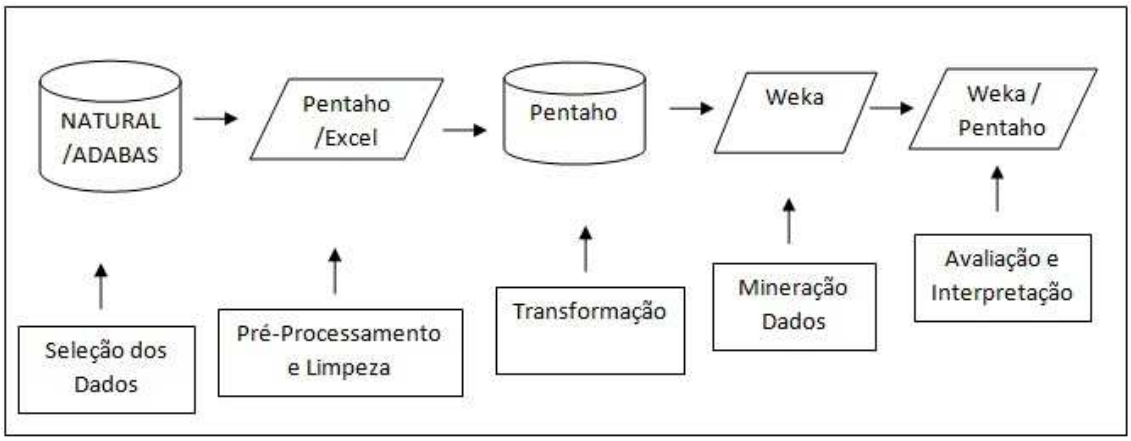


Figura 11 - Relação dos recursos utilizados mediante as fases do modelo DCDB.

Neste trabalho, a ferramenta *WEKA* é utilizada para realizar uma análise computacional em um conjunto de dados previamente estabelecido, recorrendo aos algoritmos de MD, de modo indutivo e mediante os padrões analisados, para gerar hipóteses de resolução do objeto em questão. Tal análise será realizada por meio da tarefa de classificação, seguindo um algoritmo de árvore de decisão visto que a ferramenta disponibiliza outras possibilidades de classificação, tal como as redes neurais.

4.4 COMPOSIÇÃO DO MODELO

A MD é a principal ciência para estabelecer os objetivos do estudo deste trabalho. Para viabilizar a atividade de mineração um minucioso trabalho deve ser realizado para a disposição dos dados. O modelo explora a obtenção dos dados, a preparação, a composição da árvore através do algoritmo de mineração, e por fim, a classificação dos contribuintes.

4.4.1 Origem dos dados

As empresas aptas a prestarem serviços no município de Goiânia necessitam se regularizar no departamento municipal responsável, neste caso a Secretaria de Finanças, visando realizar os devidos procedimentos de habilitação de funcionamento. Neste processo de regularização é requisitada a devida documentação, juntamente com respectivas declarações do comportamento funcional das empresas na abertura, no acompanhamento e na fiscalização.

A primeira etapa do modelo visa definir o grupo de dados que serão necessários para realizar a classificação dos contribuintes. Dentre as informações apresentadas pelos contribuintes, é possível dividi-las em três grupos que formarão a base:

- Dados estruturais: contendo informações do tempo de funcionamento, atividades econômicas, natureza e tipo de empresa, tipo de isenção, percentual de imposto, dentre outras;
- Dados econômicos: contendo informações do volume de movimentação e outras informações de contexto econômico-fiscal;
- Dados de auditorias: contendo informações do registro de auditorias realizadas mencionando irregularidades caso ocorram.

Cabe ressaltar que, após a extração, formatação e definição inicial do conjunto de dados, uma analise estatística auxiliará a definição da relevância dos atributos através da incidência de preenchimento de alguns campos catalogados, visto que o preenchimento em determinadas circunstâncias não é realizado de forma obrigatória.

A identificação das fontes de dados juntamente com a definição do conjunto de atributos é primordial para garantir assimilação realizada por modelos quantitativos (BERRY e LYNOFF, 2004). Esta fase é composta de atividades relacionadas ao pré-

processamento dos dados, justificada pela irregularidade com os quais possivelmente são armazenados, contendo ruídos, alem da necessidade dos algoritmos em exigir uma formatação especial das informações (LAROSE, 2006).

4.4.2 Extração dos dados

A extração é uma atividade de extrema importância ao relacionar quais informações estão no domínio dos sistemas envolvidos e com que regularidade é informada. Esta atividade tem como saída a identificação das fontes de dados que fará parte da MD, quais atributos são necessários e de que forma será extraída para um conjunto especial de dados a ser utilizado no processo.

4.4.3 Limpeza

As fontes de dados podem possuir informações oriundas de migrações ou inseridas de maneira inválida. Após a extração, o conjunto de dados é submetido a atividade que atesta a qualidade dos dados ao averiguar se pertencem a um mesmo domínio, por exemplo, se são numéricos ou alfanuméricos, se são íntegros e se são verídicos. Esta atividade procura eliminar ruídos através de atributos obsoletos, redundantes, incompletos ou inconsistentes. O algoritmo, ao instituir a árvore, realiza a leitura do domínio dos atributos, e caso não estejam padronizados pode interferir na qualidade do modelo construído.

4.4.4 Identificação de relevância

O algoritmo de classificação avalia variáveis categóricas e os atributos precisam de uma composição mínima válida e consistente. Esta atividade busca identificar atributos que possam causar erros no algoritmo de classificação. Visa eliminar os atributos que não são preenchidos de forma obrigatória e em determinados casos, os atributos que

não são válidos para todos os tipos de registro do conjunto de dados. A manipulação de dados ausentes é um dos principais motivos que inviabiliza uma análise de dados. Determinados atributos quantitativos são requisitados dependendo da natureza dos contribuintes. Recursos gráficos como histogramas e diagramas de *ploter* auxiliarão na avaliação da relevância dos atributos selecionados na constituição do algoritmo de classificação. Esta atividade é marcada por definir o conjunto de dados que será utilizado na MD através do algoritmo de classificação.

4.4.5 Transformação

A etapa de transformação visa identificar atributos com disparidade grande de valores, os chamados *outliers*. Essa transformação visa melhorar a capacidade preditiva dos modelos visto que em alguns algoritmos de MD essa disparidade pode influenciar os resultados, causando uma tendência nas variáveis que apresentam valores dispersos fora de uma regularidade. Na ocorrência de disparidade, uma normalização deve ser realizada para padronizar o conjunto de valores.

Os algoritmos de MD, para serem aplicados, necessitam de um conjunto em forma de matriz. Esta etapa visa realizar a transformação de atributos multivalorados ou compostos em um modelo em que seja passível a execução das técnicas de MD.

A tabela 3 demonstra as saídas efetivadas ao finalizar as atividades da etapa responsável por identificar os dados.

Tabela 3 - Relação dos artefatos gerados na etapa de origem dos dados.

Atividades	Saídas
Extração dos dados	Relação das fontes e dados
Limpeza	Datamart
Identificação de relevância	Relação dos atributos definitivos
Transformação	Conjunto de dados normalizado

4.5 Método de classificação

O domínio dos dados utilizado neste trabalho é composto das informações socioeconômicas e auditoria dos contribuintes, tais como: tipo de empresa, tempo de atividade, atividades econômicas e registro de empregados. Pela natureza dos dados apresentados composta de atributos ordinais e um atributo alvo categórico, a classificação mediante o algoritmo árvore de decisão foi selecionada para realizar a classificação do perfil dos contribuintes. A classificação através do algoritmo de árvore de decisão é indicada quando as variáveis preditivas são compostas de valores nominais ou ordinais e o atributo alvo e composto por apenas uma variável, de preferência, categórica. O algoritmo de árvore de decisão necessita de um atributo alvo para direcionar o aprendizado as classes previamente estabelecidas. Resultante da padronização, a variável IRREGULARIDADE será utilizada para direcionar através dos demais atributos preditivos.

A ferramenta WEKA disponibiliza diversos algoritmos de classificação dentre eles algoritmos bayesianos, regras de associação, regressão logística e árvore de classificação. Dentre os algoritmos disponíveis o mais indicado para esse tipo de projeto de pesquisa é o algoritmo J48, que foi utilizado neste trabalho por ser a implementação da ferramenta para o algoritmo de árvore de decisão baseado no C4.5 (WEKA.2013).

V ESTUDO DE CASO

Este capítulo descreve o processo de preparação dos dados utilizando o modelo DCDB para formalizar um conjunto pré-processado e formatado, necessário para a aplicação do modelo proposto de MD para a tarefa de classificação. Pela natureza dos dados disponíveis composta de atributos ordinais, um atributo alvo categórico conforme apresentando do item 4.5 e em função dos objetivos estabelecidos nesta pesquisa estabeleceu-se a classificação através de um algoritmo de árvore de decisão.

5.1 DOMÍNIO DOS DADOS

Os contribuintes considerados aptos a prestarem serviços no município de Goiânia necessitam de regularização mediante um processo de cadastro no órgão responsável designado Secretaria de Finanças. No ano de 2011, referência para este trabalho, a Secretaria continha um quantitativo de 38.700 contribuintes ativos com habilitação para realizar a prestação de serviços na capital. Desse quantitativo, excluem-se dos trabalhos do departamento de auditoria as empresas classificadas como Microempresas (ME) ou as Empresas de Pequeno Porte (EPP), que são optantes do Simples Nacional, modalidade que entrou em vigor a partir de primeiro de julho de 2007 por meio da lei complementar n° 123 de 14 de D ezembro de 2006. Neste caso, a respectiva arrecadação é realizada por um comitê gestor composta de órgãos da União. Desta forma, os contribuintes que estão regulares e aptos a receberem fiscalização pelo município de Goiânia e, consequentemente, dentro do domínio de objeto de estudo deste trabalho, totalizam um quantitativo de 28.700 contribuintes e estão classificados nas modalidades de atuação, como mostra o gráfico da figura 12.

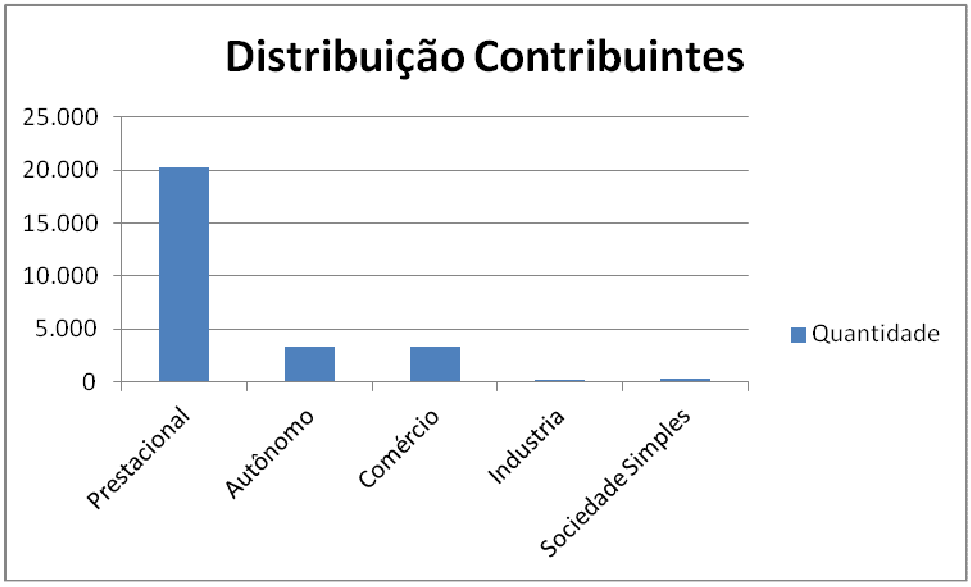


Figura 12 - Distribuição das modalidades dos contribuintes não optantes do SIMPLES Nacional.

5.2 COLETA DOS DADOS

Os dados necessários à realização deste trabalho foram coletados dos Sistemas Cadastro de Atividades Econômicas (CAE) e do Sistema de Registro de Auditoria. O sistema CAE registra todas as informações socioeconômicas do contribuinte e serve também de referência para a efetivação do sistema de Auditoria na medida em que subsidia o fiscal na averiguação da real situação dos contribuintes fiscalizados. O sistema de Auditoria registra informações das atividades realizadas pelo fiscal nos contribuintes e na ocorrência de alguma irregularidade catalogada, tais como a ausência de obrigações acessórias, falta de recolhimento de taxas ou falta de recolhimento de impostos. Nestes casos, o fiscal registra a fundamentação legal da infração ocorrida juntamente com o respectivo valor devido. As informações do registro da fiscalização são necessárias e imprescindíveis para moldar o conjunto de dados ao prever a possibilidade de distinção de contribuintes regulares e irregulares pelo algoritmo de classificação.

Visto que as ferramentas de MD necessitam de um conjunto de dados formatado e padronizado em forma de matriz, um programa foi escrito na Linguagem NATURAL/ADABAS, pelo Departamento de Desenvolvimento da Agência Goiana de Tecnologia e Inovação (AMTEC), especificamente para realizar a extração das informações utilizadas nesta pesquisa, uma vez que as fontes de obtenção contem tecnologia diferente de registro de dados baseado em arquivos através de registros e índices. O programa criado por analistas do Departamento de Desenvolvimento da AMTEC teve como finalidade filtrar os contribuintes não optantes do SIMPLES Nacional e excluir as fiscalizações que não estavam finalizadas, coletando somente os atributos que continham relevância para o objeto desta pesquisa, ficando de fora aqueles dados que poderiam personificar ou identificar qualquer contribuinte.

5.3 PRÉ-PROCESSAMENTO E FORMATAÇÃO

A etapa de pré-processamento e formatação é imprescindível ao sucesso do processo de MD ao estabelecer um conjunto apto à tarefa de mineração. Após a coleta, é necessário padronizar e formatar os dados visando eficiência do algoritmo de MD escolhido, tratando os dados na forma mais conveniente possível através de suas restrições.

5.3.1 Redução e limpeza dos dados.

Para reduzir o universo do domínio das informações disponíveis, algumas restrições foram utilizadas para reduzir e definir o real conjunto de dados utilizado neste trabalho para desenvolver o processo de MD. Tal atividade possui elevada importância ao retirar do contexto dos dados registros anormais que poderiam ocasionar distorções nas análises do resultado apresentado pela tarefa de MD. A primeira restrição imposta por meio legal incide sobre a exclusão dos contribuintes

optantes do SIMPLES Nacional do contexto dessa pesquisa, visto que tais contribuintes se adequaram a uma legislação própria, submetendo-se a averiguação da União mesmo sendo, primariamente, contribuintes de prestação de serviços, com responsabilidades de atuação dos Municípios e do Distrito Federal. Desta forma o universo dos contribuintes é formado pelos demais contribuintes não optantes do SIMPLES Nacional distribuídos nas modalidades apresentadas na figura 12.

Posteriormente, foi identificado um período para reduzir o universo do domínio do conjunto de dados. Essa definição foi efetivada em consonância com o registro de fiscalizações efetuadas, visto que tais informações são imprescindíveis para avaliar o comportamento dos contribuintes (se o mesmo incorreu em alguma irregularidade ou não). Nesse sentido, foram selecionados os registros realizados pelo departamento de auditoria no ano de 2011, totalizando o quantitativo de 1.074 contribuintes distribuídos nas modalidades de atuação apresentadas na figura 13.

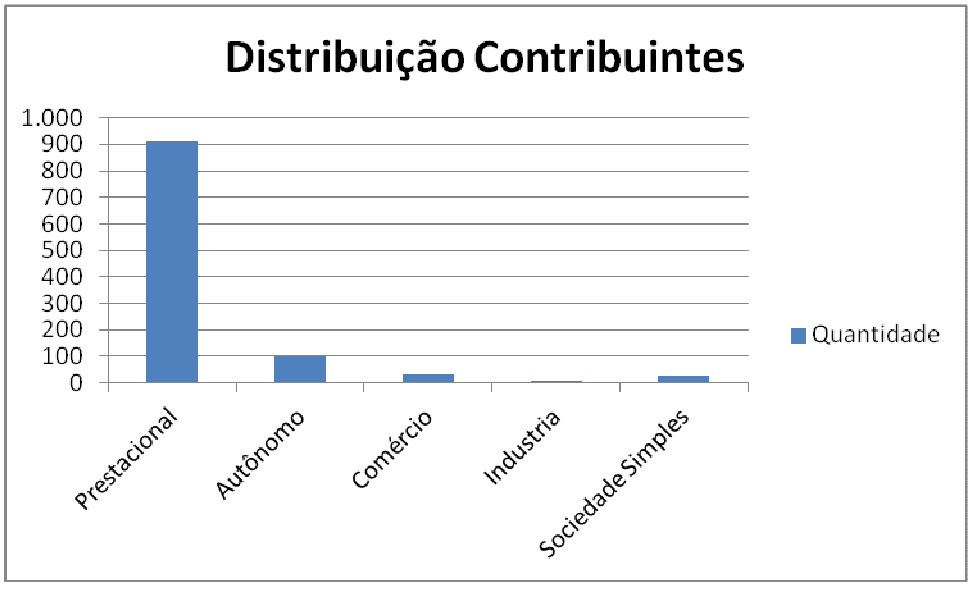


Figura 13 - Distribuição das modalidades de atuação dos contribuintes do ano de 2011.

Analisando-se o conjunto de dados, apresentado na figura 14, observa-se que os contribuintes da modalidade prestacional e autônomo compõe um percentual de 93,9% dos contribuintes fiscalizados. Assim, o percentual restante das demais

modalidades foi retirado do conjunto de dados selecionado, priorizando os contribuintes com finalidade primária de serviços e visto que tal percentual poderia influenciar na eficiência do processo de MD. O conjunto final ficou estabelecido com 1.040 registros, distribuídos de acordo com a figura 14.

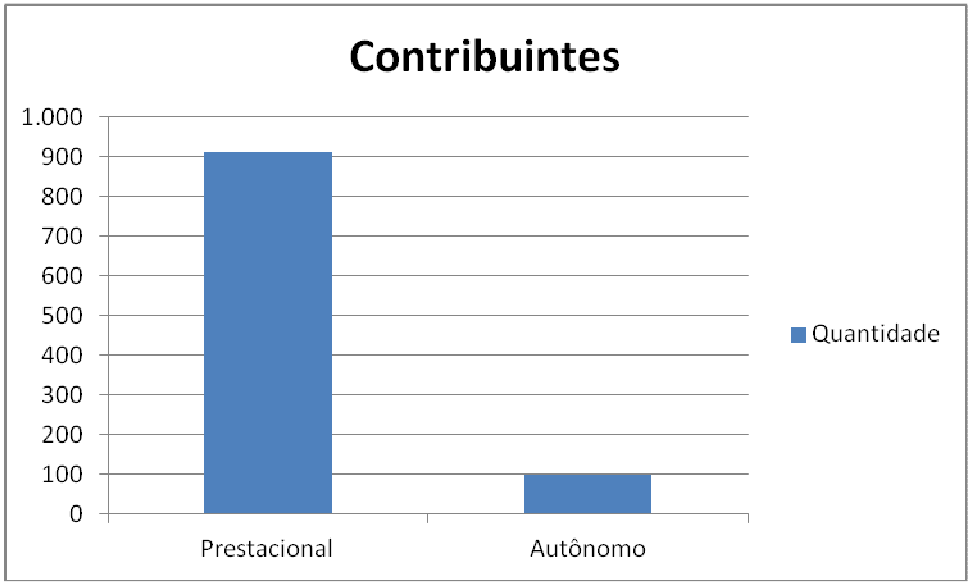


Figura 14 - Distribuição das modalidades de atuação do conjunto de dados final.

No processa da limpeza, o principal esforço foi realizado na retirada dos registros do conjunto de dados pertencentes a auditorias que não estavam finalizadas. O trabalho realizado pelo fiscal pode se estender por dias ou meses em determinado contribuinte, e nesse intervalo, informações são registradas no sistema de forma gradual e incompleta. Em outras situações, quando não constatada irregularidades, determinadas informações não são vinculadas de forma obrigatória, o que ocasionou a retirada dos registros que não estavam com informações que fossem relevantes para o objeto da pesquisa.

5.3.2 Identificação de relevância

Os campos que continham informações obsoletas, redundantes, que continham um baixo índice de preenchimento ou que não contribuia para a composição do perfil do contribuinte foram retirados da seleção inicial para não prejudicar a eficiência da tarefa de MD. Na primeira avaliação, foram observados 120 atributos que poderiam compor o conjunto de dados da mineração. Após sucessivas análises, observou-se que 26 atributos teriam informações relevantes, conforme ilustrado na figura 15.

Atributo	Descrição	Tipo
NUMR	Informação para distinguir dados de determinado contribuinte	Numérico
DT_IN	Data inicial do movimento fiscalizado	Data
DT_FIM	Data final do movimento fiscalizado	Data
VL_DEV	Valor da movimentação do contribuinte no ato da fiscalização	Numérico
INF_MIC	Informação de micro empresa	Categórico
DT_ABERT	Data de inicio de atividade	Data
NR_EMP	Número de empregados	Numérico
NT_JUR	Natureza jurídica	Categórico
INF_ESC	Informação se possui ou não escrita fiscal	Categórico
TP_TRIB	Contribui somente com o imposto do ISS ou demais	Categórico
TP_ISEN	Possui algum benefício de isenção	Categórico
TP_REC	Se o fator é baseado na movimentação ou estimado	Categórico
NR_ART_CTM	Informação dos artigos de fundamentação com base no CTM	Texto
NR_ART_RCTM	Informação dos artigos de fundamentação com base no RCTM	Texto
NUMR_ATV1	Atividade econômica principal	Numérico
NUMR_ATV2	Atividade econômica secundária	Numérico
TP_IRREG_1	Tipo de irregularidade constatada	Categórico
TP_IRREG_2	Tipo de irregularidade constatada	Categórico
TP_IRREG_3	Tipo de irregularidade constatada	Categórico
NR_SERV1	Serviço executado principal	Numérico
NR_SERV2	Serviço executado secundário	Numérico
NR_SERV3	Serviço executado secundário	Numérico
NR_BAIR	Código de localização	Numérico
INF_LIB	Informação de autônomo	Categórico
TP_EST	Tipo de estimativa de contribuição	Categórico
IRREG	Apresentação ou não de irregularidade	Categórico

Figura 15 – Composição dos atributos iniciais do conjunto de dados final.

A figura 15 apresenta os 26 atributos que originalmente teriam relevância para compor o conjunto de dados utilizado no processo de MD. O primeiro passo buscou a identificação de relevância dos atributos pela sua capacidade de personificar o perfil do contribuinte, tais como o número de empregados, atividade econômica, modalidade jurídica, tipo de tributação e tempo de atividade, os quais, após coletados, foram transformados em uma estrutura de dados em forma de matriz. Os atributos que não agregavam informação para personificar o perfil, como exemplo, dados do contador e identificação de sócios, foram considerados irrelevantes para compor o conjunto de dados do processo de MD. Posteriormente, os atributos com índice menor que 50% de preenchimento também foram retirados por não contribuírem efetivamente na predição do algoritmo ao criar um modelo. Os demais atributos visualizados estão apresentados no Anexo I.

5.4 TRANSFORMAÇÃO

Após definição de relevância, as atividades de transformação são necessárias para realizar desmembramento de campos, conversão, normalização numérica e agrupamentos.

5.4.1 Conversão

As informações relativas aos tipos de irregularidades constatadas foram coletadas em três campos: IRREG1, IRREG2 e IRREG3 com quatro[4] possíveis valores, representando as irregularidades com base em impostos ou irregularidades com base em desobrigações acessórias, normalmente relacionadas ao descumprimento de taxas. Tais informações foram convertidas em uma matriz no qual, um índice numérico

[4] (A) Recolhimento a menor de imposto, (B) Falta de recolhimento de imposto, (C) Recolhimento a menor de taxas e (D) Falta de recolhimento de taxas.

foi utilizado para tipificar as combinações de irregularidades que representavam 99% das combinações constatadas, convertendo-as para um único campo categórico, como apresentado na figura 16.

	Matriz Irregularidades		
	IRREG1	IRREG2	IRREG3
1	1	0	0
2	0	1	0
3	0	0	1
4	1	1	0
5	1	0	1
6	0	1	1
7	1	1	1

Figura 16 - Matriz de irregularidades constatadas.

Sendo:

IRREG1 → A falta de recolhimento de imposto devido.

IRREG2 → O recolhimento a menor de imposto devido.

IRREG3 → Descumprimento de desobrigação acessória.

As ocorrências correspondentes somente às infrações de obrigações acessórias totalizam um quantitativo de 15% das irregularidades constatadas. Relacionado ao retorno financeiro, as obrigações acessórias normalmente corresponde a taxas e possuem pequeno valor no quantitativo recuperado pelo departamento de auditoria. Tais ocorrências também foram inclusas no contexto deste trabalho porque os contribuintes registrados de alguma forma encontram-se irregulares na prestação de seus serviços.

A figura 17 ilustra o percentual das irregularidades após o processo de conversão para um único campo categórico.

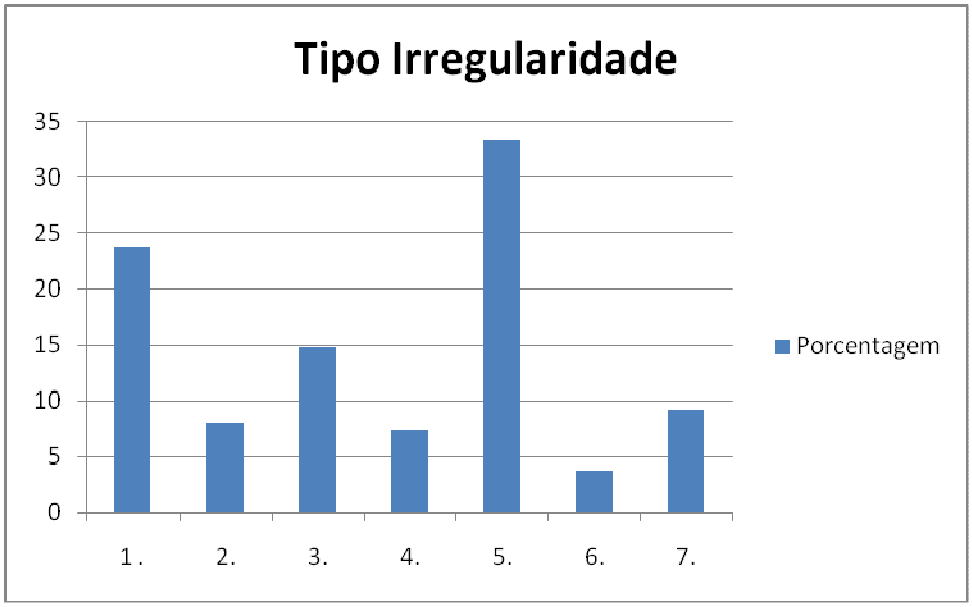


Figura 17 - Gráfico dos tipos de irregularidades visualizadas.

Os percentuais do gráfico, apresentados no gráfico da figura 17, correspondem às empresas catalogadas que apresentaram algum tipo de irregularidade. Após as atividades de pré-processamento e transformação, um novo atributo denominado IRREG foi inserido no conjunto de dados para distinguir os contribuintes que incidiram em alguma irregularidade sendo elas relacionadas à falta de recolhimento de impostos, recolhimento a menor, o descumprimento com obrigações acessórias ou a combinação deles, como mencionado na figura 16.

O atributo proposto IRREG é do tipo categórico e instituído com valor zero (0) para os contribuintes que estavam regulares em suas atividades e com valor um (1) para aqueles que apresentavam alguma não conformidade visualizada pelo fiscal. A tabela 4 demonstra o quantitativo de registros do conjunto de dados com respectivos percentuais dos contribuintes regulares e os que apresentaram alguma infração.

Tabela 4 - Quantidade de contribuintes regulares e irregulares.

Quantitativo de Contribuintes			
IRREGULARIDADE	DESCRIÇÃO	QUANTIDADE	PERCENTUAL
0	Contribuinte Regular	284	27,31%
1	Contribuinte Irregular	756	72,69%
Total de Contribuintes analisados		1040	100,00%

No sentido de categorizar os atributos que serão utilizados pelo algoritmo de classificação, uma segunda conversão foi realizada nos campos onde são informados a fundamentação legal para o trabalho da fiscalização realizada no contribuinte. Nesse caso, o atributo original é armazenado de forma descritiva em campo texto onde são informados os artigos, parágrafos e alíneas do respectivo CTM e RCTM referenciado para respaldar a atividade do fiscal. Tais informações foram desmembradas, retirando-se os dados que fossem identificados de forma categórica em novo atributo.

Na primeira etapa foi realizado um desmembramento para catalogar quais artigos foram informados no campo base CTM. Posteriormente uma matriz foi gerada para representar de forma categórica as ocorrências dos artigos registrados. Como o campo desmembrado é informado de forma facultativa, a matriz foi composta utilizando a combinação dos artigos que representavam 99% das combinações encontradas. Quando informado, um ou mais artigos são registrados. Assim, a matriz constituída foi composta através do percentual de combinações dos artigos informados, apresentadas em forma categórica com valores de 1 a 6, como apresentado na figura 18.

	Matriz CTM			
	x1	x2	x3	x4
1	1	0	0	0
2	0	1	0	0
3	0	0	1	0
4	0	0	0	1
5	1	0	1	0
6	1	0	0	1

Figura 18 - Matriz dos artigos CTM informados.

Sendo:

X1 → Artigos referentes à forma local e retenção do imposto devido.

X2 → Artigos referentes à determinação da instituição de contribuinte e suas respectivas atividades econômicas.

X3 → Artigos referentes à administração de notas fiscais, livros fiscais e sua utilização.

X4 → Artigos referentes à instituição, modo, prazos e pagamentos de taxas estabelecidas pelo município.

5.4.2 Normalização numérica

A normalização numérica é um recurso utilizado para reduzir disparidade dos dados dispersos de domínio numérico quando estes possuem um domínio com alta escala de variação, em destaque os que representam valores monetários. O campo valor base pode conter valores na casa de unidades de reais como valores na ordem de milhões.

Visando o equilíbrio da influência dos atributos utilizados pelo algoritmo de MD, a variável valor base foi transformada para uma escala com valores entre zero (0) e um (1). Segundo (BUSSAB *et al*., 1990) , a transformação é aplicada a variáveis

quantitativas[5], através da formula Y = (Xi – Xmin) / (Xmax – Xmin) onde: Y → O valor referente á variável após a transformação para escala entre 0 e 1; Xi → O valor original da variável a ser transformada; Xmin → O menor valor contido no conjunto original de dados referente à variável a ser transformada e Xmax → O maior valor contido no conjunto original de dados referente à variável a ser transformada.

5.4.3 Agrupamento

Os contribuintes, sendo eles pessoas físicas ou jurídicas de quaisquer atividades econômicas, precisam se cadastrar perante o município através do sistema CAE. Tal sistema relaciona as atividades regulares no município de Goiânia que obedecem a uma hierarquia imposta pelo Cadastro Nacional de Atividades Econômicas CNAE, mantido pela Receita Federal.

O CNAE[6] é formado por sete dígitos representando uma hierarquia formalizada em grupos, subgrupos e classes de atividades econômicas correlatas, como ilustra a tabela 5.

Tabela 5 - Hierarquia CNAE.

Hierarquiva CNAE					
Seção	Grupo	SubGrupo	Classe	CNAE	Descrição
Q					Saúde Humana e Serviços Sociais
	86				Atividade de Atenção a Saúde Humana
		863			Atividade de Atenção Ambulatorial Executadas por Médicos e Odontólogos
			8630-5		Atividade de Atenção Ambulatorial Executadas por Médicos e Odontólogos
				8630-5 / 04	Atividade Odontológica

Conforme descrito na tabela 5, à atividade econômica que representa o serviço de Odontologia é identificada pelo código CNAE **8630504,** o qual pertence à classe **8630-**

[5] Atributos que são representados por números

[6] Instrumento de padronização nacional dos códigos de atividades econômicas.

5, Atividade de Atenção Ambulatorial Executada por Médicos e Odontólogos, a qual pertence ao subgrupo **863**, Atividade de Atenção Ambulatorial Executadas por Médicos e Odontólogos, o qual pertence ao grupo **86**, Atividade de Atenção a Saúde Humana, o qual pertence à seção **Q**, Saúde Humana e Serviços Sociais.

Para compor o conjunto de dados desta pesquisa foram selecionadas as duas principais atividades econômicas dos contribuintes, visando o desempenho do algoritmo ao avaliar atividades que contem movimentação econômica, visto que o sistema CAE permite a inclusão das demais atividades. Tanto para a atividade primária como atividade secundária, o código de referência do CNAE foi convertido para seu respectivo subgrupo e grupo para facilitar a indução do algoritmo de classificação, como apresentado na tabela 6 e 7. A relação completa da hierarquia oficial dos CNAE está apresentada no Anexo II.

Tabela 6 - Agrupamento Grupo CNAE.

Grupos CNAE	
Grupos CNAE	Descrição
.....	
77	Alugueis não-imobiliários e Gestão de Ativos Intangíveis não-financeiros
78	Seleção, Agenciamento e Locação de mão-de-obra
79	Agência de Viagens, Operadores turísticos e Serviços de Reservas
80	Atividade de Vigilancia, Segurança e Investigação
81	Serviços para Edifícios e Atividades paisagisticas
82	Serviços de Escritório, de Apoio administrativo Prestados às Empresas
.....	

Tabela 7 - Agrupamento SubGrupo CNAE.

SubGrupos CNAE	
Sub CNAE	Descrição
.....	
771	Locação de Meios de Transporte sem Condutor
772	Aluguel de Objetos pessoais e Domésticos
773	Aluguel de Máquinas e Equipamentos sem Operador
774	Gestão de Ativos Intangíveis Não-financeiros
.....	

Aliado a estrutura do CNAE, o fiscal relata em sua vistoria os serviços que são executados pelo contribuinte em consonância com a estrutura disponibilizada no CTM. De forma similar as atividades econômicas informadas pelo contribuinte ao se regularizar no CAE, o fiscal relata os serviços executados pelo contribuinte, mas obedecendo a uma hierarquia própria estabelecida pelo CTM. Também nesse caso os dois principais serviços foram coletados, visto que o sistema de auditoria permite o registro das demais atividades realizadas.

Para os campos de serviços visualizados a conversão foi realizada ao seu respectivo grupo como apresentado na tabela 8, visto que a hierarquia disposta pelo CTM dispõe apenas de dois níveis de extensão. Por exemplo, o serviço **04.12** de Odontologia pertence ao grupo **04**, Serviços de saúde, assistência médicas e congêneres. A relação completa dos grupos e serviços homologados pelo CTM está disponível no Anexo III.

Tabela 8 - Lista Serviços CTM.

Grupos Serviços CTM	
Grupo	Descrição
01	Serviços de informática e congêneres
02	Serviços de pesquisas e desenvolvimento de qualquer natureza
03	Serviços prestados mediante locação, cessão de direito de uso e congêneres
04	Serviços de saúde, assistência médica e congêneres
05	Serviços de medicina e assistência veterinária e congêneres
06	Serviços de cuidados pessoais, estética, atividades físicas e congêneres
.....	

5.5 CONJUNTO DE DADOS PADRONIZADO

A seleção inicial identificou quais campos das diversas fontes de informação seriam relevantes para compor o conjunto de dados necessário ao processo de MD. Tais dados oriundos do sistema de atividades econômicas e do registro de auditorias foram coletados através de um programa criado especificamente para extração,

transformados em uma estrutura de dados e posteriormente repassados as etapas de pré-processamento e transformação. Tais etapas procederam às atividades de limpeza, análise de relevância e formatações com o intuito de estabelecer um padrão nos dados, tornando-os aptos a atividade de MD, em especial a classificação por intermédio de um algoritmo de árvore de decisão.

O escopo final do domínio dos contribuintes ficou reservado às modalidades não optantes do SIMPLES nacional e restrito aos que primariamente se dedicam a realização de serviços. Dentro desse conjunto há outra ótica de observação incidente sobre a natureza jurídica informada na constituição legal dos contribuintes, como demonstrado na figura 19.

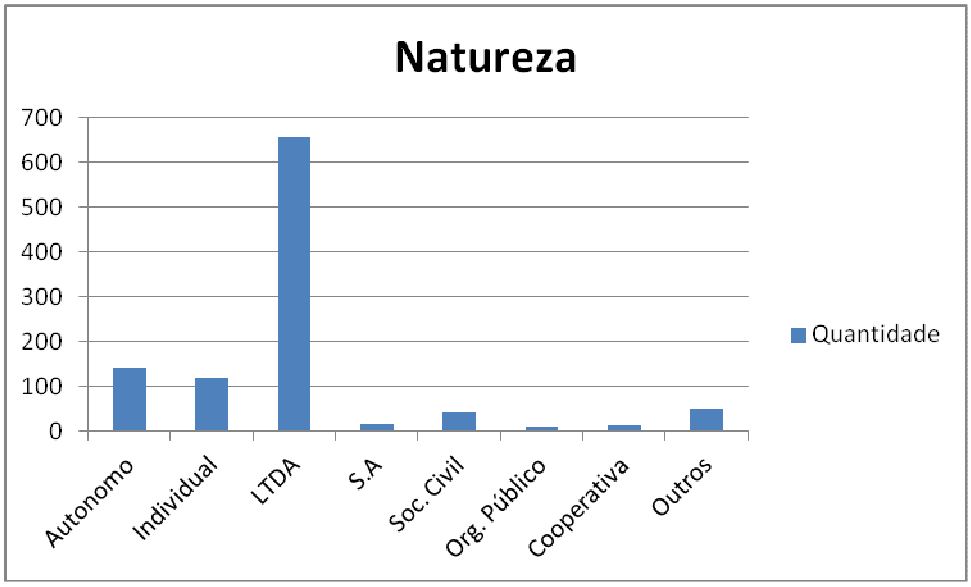


Figura 19 - Natureza dos contribuintes do conjunto de dados selecionado.

O conjunto final dos dados após desmembramentos, agrupamentos e conversões resultou em uma estrutura com atributos definidos e padronizados nos moldes de uma matriz com um quantitativo de 1040 registros que serão utilizados para o processo de MD.

5.6 CLASSIFICAÇÃO DOS CONTRIBUINTES

Após o estabelecimento do conjunto final dos dados, a próxima atividade envolve a definição do algoritmo de classificação. Nos trabalhos baseados na classificação de contribuintes com irregularidades correlatos, COVARLÃO (2009), BRAGA (2010) e ANDRADE (2009), a classificação foi realizada através de redes neurais e por intermédio de regressão logística, visto que nos dois casos o universo de atributos utilizados pelo algoritmo continha variáveis contínuas que representavam, em grande parte, a movimentação dos contribuintes. A utilização da rede neural foi necessária para permitir a utilização de atributo alvo composto de dois atributos, como exemplo, faturamento e região econômica para melhor definir o conjunto de dados selecionado.

Neste trabalho o algoritmo árvore de decisão foi utilizado com o objetivo de classificar os contribuintes perante a natureza dos dados disponíveis, como abordado no item 4.5, que englobam os dados socioeconômicos e de auditoria, visto que os contribuintes prestadores de serviço podem apresentar uma movimentação financeira contínua ou esporádica e não apresentam balanços anuais, os quais podem ser balizados com o movimento econômico e fiscal. Assim, o domínio dos dados engloba as informações socioeconômicas e do registro de irregularidades dos contribuintes de serviços.

Para permitir o aprendizado automático e favorecer o descobrimento de padrões é necessário estabelecer uma estrutura no conjunto de dados que permita a criação de um modelo classificador. Para a tarefa de classificação é necessário que o conjunto de dados contenha atributos preditivos e um atributo especial denominado atributo alvo que seja de caráter categórico, utilizado para determinar as classes dos registros que serão classificados. Visto essa particularidade, a variável IRREG foi criada como atributo alvo, possuindo valores [SIM, NÃO] e utilizada para montar o vetor inicial do

algoritmo de classificação por intermédio da árvore de decisão, como apresentado na figura 20.

.....	TM_ATIV	NR_EMP	NT_JUR	INF_ESC	TP_TIRB	TP_ISEN	TP_REC	IRREG
.....	9	0	0	2	1	0	0	SIM
.....	31	0	0	2	0	0	0	SIM
.....	19	0	0	2	1	0	0	SIM
.....	7	0	0	2	0	0	0	SIM
.....	8	0	0	2	1	0	0	SIM
.....	6	0	2	1	0	0	0	SIM
.....	16	4	2	2	0	0	0	SIM
.....	16	0	2	2	3	0	3	SIM
.....	25	5	2	1	0	0	4	SIM
.....	12	3	1	2	0	0	0	SIM
.....	10	3	2	2	0	0	4	SIM
.....	8	24	2	1	3	0	0	SIM
.....	11	0	1	2	0	0	0	NAO
.....	10	66	2	2	0	0	4	SIM
.....	16	0	0	2	1	0	0	NAO
.....	42	23	18	1	0	5	0	NAO
.....	1	0	1	2	0	0	0	SIM

Figura 20 - Exemplo de atributos preditivos e atributo alvo.

Algoritmos de MD de dados em especial, os classificadores são de caráter preditivo e nesses casos, desenvolvem aprendizado com uma rotina de treinamento para atingir o objetivo de classificação. Por essa característica peculiar o conjunto final de dados foi subdividido em um conjunto de treinamento e um conjunto de testes, como apresentado na tabela 9.

Tabela 9 - Subconjuntos gerados.

Subdivisão do Conjunto de Dados	
DESCRIÇÃO	QUANTIDADE
Dados de treinamento	500
Dados de teste	540
Total de Contribuintes do conjunto final	1040

Nesta etapa o esforço empreendido incide na elaboração do modelo para realizar a classificação dos contribuintes. Para realizar esta atividade e permitir a criação desta estrutura, a ferramenta *WEKA* será utilizada. A ferramenta necessita de um conjunto de dados formatado em um formato de arquivo específico *atributte-relation file format* (arff). Tal formatação é necessária para carregar o conjunto de dados para o formato específico utilizado nas execuções dos algoritmos classificadores da ferramenta como mostra a figura 21.

```
@relation dadosWeka-weka.filters.unsupervised.instance.RemovePerc

@attribute TMP_FIS numeric
@attribute CNAE_P1 numeric
@attribute CNAE_P2 numeric
@attribute CNAE_S1 numeric
@attribute TM_ATIV numeric
@attribute NR_EMP numeric
@attribute NT_JUR {0.0,2.0,1.0,4.0,18.0,' ',99.0,11.0,3.0,15.0,5.
@attribute INF_ESC numeric
@attribute TP_TIRB numeric
@attribute TP_ISEN numeric
@attribute TP_REC numeric
@attribute IRREG {SIM,NAO}

@data
5,85,851,?,6,0,2.0,2,13,0,0,SIM
5,66,662,43,15,0,2.0,2,3,0,0,NAO
```

Figura 21 - Estrutura do conjunto de dados para utilização no WEKA.

Após o carregamento do conjunto de dados, a primeira etapa da ferramenta é a etapa de pré-processamento. Nessa fase, além da formatação especifica para o carregamento, o conjunto de dados foi subdividido em dois novos conjuntos, de treinamento e de testes, como ilustrado na figura 22.

Na etapa de pré-processamento a divisão do conjunto original, contendo 1040 registros, foi dividida em um novo conjunto de teste contendo 51.9% dos registros e o percentual restante no conjunto que será utilizado para contemplar a etapa de treinamento, realizado de forma aleatória, como ilustrado na figura 22.

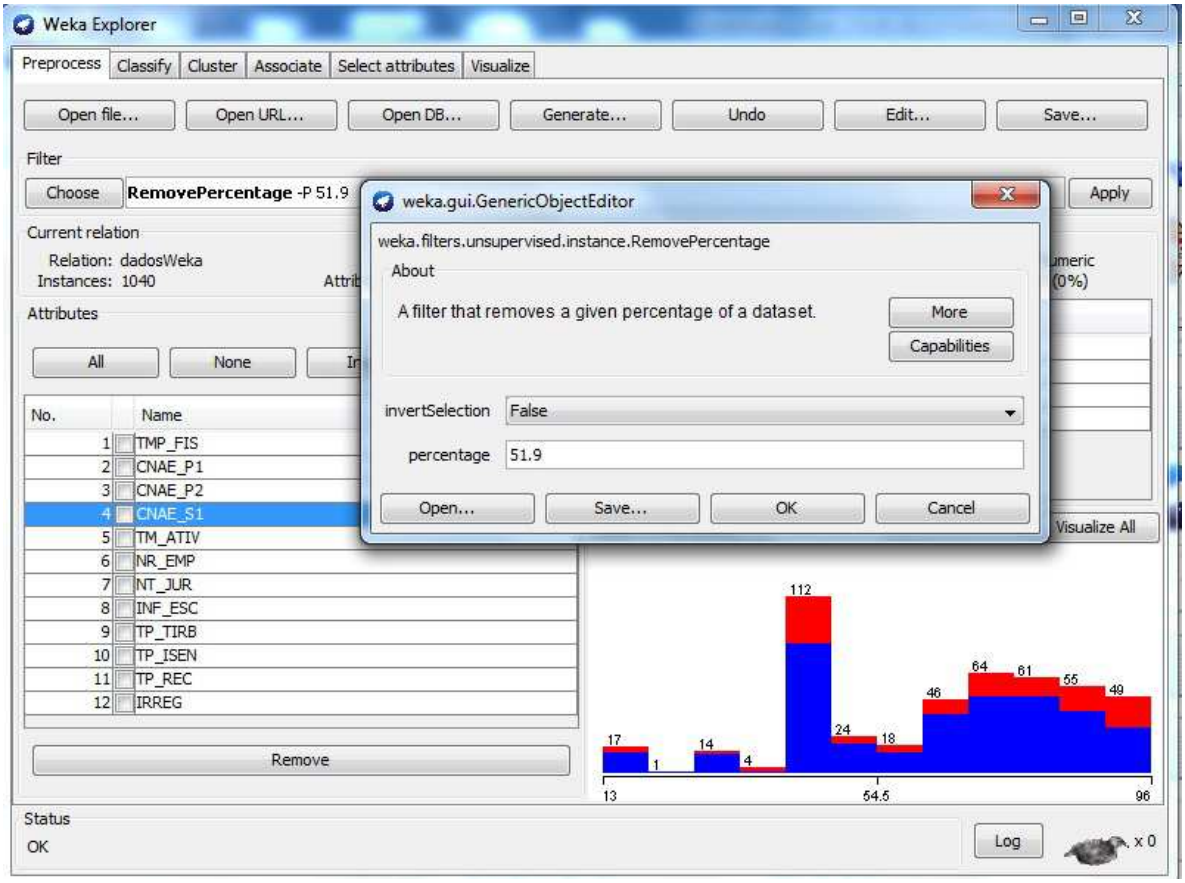


Figura 22 - Divisão do conjunto de dados em treinamento e teste.

Após a divisão do conjunto apresentado na figura 22 em dois novos conjuntos, de teste e treinamento, realizado pela etapa de pré-processamento da ferramenta WEKA, a aba de classificação foi selecionada para criar a estrutura do modelo classificador mediante a opção de árvore de decisão e com base no algoritmo J48 como apresentado na figura 23.

Figura 23 - Seleção do algoritmo de classificação.

O algoritmo J48, utilizado neste trabalho, disponível na ferramenta WEKA, tem capacidade de manipular atributos binários, ordinais, nominais e contínuos, mas para melhorar a indução dos atributos na construção árvore de decisão que representa o modelo, de preferência atributos categóricos e contínuos foram utilizados. Na etapa de pré-processamento algumas conversões foram realizadas para favorecer a legibilidade da árvore gerada pelo modelo, visto que na execução do algoritmo, ao criar um novo nível de ramificação da árvore, é realizado um produto cartesiano entre o novo atributo e o domínio de valores categóricos do mesmo. Na etapa de carregamento do conjunto de dados final para o WEKA novas padronizações foram necessárias, visto que por recomendação, o atributo que direciona os registros a uma determinada classe, é de preferência o último da lista.

5.7 SELEÇÃO DE ATRIBUTOS

A próxima análise tem por finalidade a seleção dos atributos que melhor representem o modelo de classificação. Esta etapa pode ser realizada diversas vezes, visto que a possibilidade de combinação dos atributos e o resultado final representam uma complexa definição sobre se o modelo proposto é satisfatório ou se precisa de novas amostras de atributos.

No primeiro experimento realizado, foram escolhidos os atributos que pertencem ao domínio socioeconômico dos contribuintes e que continham no mínimo 70% de preenchimento mediante a quantidade total de registros, visto que alguns são preenchidos de forma facultativa. Nesse experimento, foram selecionados 12 atributos, conforme ilustrado na tabela 10.

Tabela 10 - Relação dos atributos selecionados no primeiro experimento.

Atributo	Descrição
TMP_FISC	Tempo decorrido entre a atual auditoria e a última realizada
CNAE_P1	Grupo da atividade primária do contribuinte
CNAE_P2	SubGrupo da atividade primária do contribuinte
CNAE_S1	Grupo da atividade secundária do contribuinte
TM_ATIV	Tempo de atividade
NR_EMP	Número de empregados
NT_JUR	Natureza jurídica
INF_ESC	Informação se possui ou não escrita fiscal
TP_TRIB	Se envolve so o imposto do ISS ou demais
TP_ISEN	Se possui algum benefício de isenção
TP_REC	Se o fator e baseado na movimentação econômica ou estimativa
IRREG	Atributo alvo que direciona as classes

Após o treinamento realizado, foi selecionado o conjunto de testes contendo 540 registros para realizar a classificação. O resultado dessa combinação apresentou um percentual de acertos de 82,40%, representando 445 ocorrências e um percentual de 17,60% de erros, representando 95 ocorrências, como visualizado na figura 24.

Classifier output

```
Number of Leaves  :     30

Size of the tree :      45


Time taken to build model: 0.02 seconds

=== Evaluation on training set ===
=== Summary ===

Correctly Classified Instances         445               82.4074 %
Incorrectly Classified Instances        95               17.5926 %
Kappa statistic                          0.4645
Mean absolute error                      0.2802
Root mean squared error                  0.3739
Relative absolute error                 71.2662 %
Root relative squared error             84.3591 %
Total Number of Instances              540
```

Figura 24 - Resultado da classificação do primeiro experimento com 51,9% registros de teste e 48,1% para treinamento.

Observa-se também na figura 24, que a avaliação foi baseada mediante o conjunto de treinamentos, informação destacada na quarta linha.

A ferramenta permite ao carregar o conjunto de dados subdividi-lo de forma automática, no momento da classificação, com 66% dos registros para treinamento e 34% dos 1040 registros do conjunto final para teste. Nessa configuração, a avaliação foi baseada na subdivisão automática do conjunto total e o resultado obtido foi inferior ao resultado apresentado na divisão de dois conjuntos na etapa de pré-processamento, representando 51,9% para testes e 48,1% para treinamentos. O resultado apresentado foi de um percentual de acertos de 70,62%, representando um quantitativo de 250 acertos e um percentual de 29,38%, representando um quantitativo de 104 erros como apresentado na figura 25.

Classifier output

```
Number of Leaves  :     34

Size of the tree :      53


Time taken to build model: 0.07 seconds

=== Evaluation on test split ===
=== Summary ===

Correctly Classified Instances         250               70.6215 %
Incorrectly Classified Instances       104               29.3785 %
Kappa statistic                          0.168
Mean absolute error                      0.3619
Root mean squared error                  0.469
Relative absolute error                 90.7402 %
Root relative squared error            104.4629 %
Total Number of Instances              354
```

Figura 25 - Resultado da classificação do primeiro experimento com 34% registros de teste e 66% de treinamento.

Importante ressaltar a estrutura criada para representar o modelo de classificação. Nessa projeção o algoritmo criou um modelo em forma de árvore com um quantitativo

de 34 folhas e com tamanho total de 53 ramificações como visualizado na parte superior da figura 25. Devido à estrutura complexa idealizada pelo algoritmo, a figura 26 apresenta, como exemplo, parte da estrutura construída para representar o modelo criado para realizar a classificação.

A estrutura criada de 34 folhas e 53 ramificações representa a quantidade de perfis dos contribuintes com probabilidades de estarem cometendo alguma irregularidade. Um dos perfis apresentado pelo modelo criado na figura 26 representa as empresas de natureza jurídica limitada, com tempo de fiscalização maior que 1 ano, com tempo de atividade maior que 19 vezes, que não possuem escrita fiscal atuantes nas modalidades com código CNAE acima de 51.

Classifier output

```
|   |   TMP_FIS <= 1: SIM (12.0/4.0)
|   |   TMP_FIS > 1
|   |   |   CNAE_P1 <= 70: SIM (3.0)
|   |   |   CNAE_P1 > 70: NAO (10.0/2.0)
|   TMP_FIS > 2: SIM (117.0/33.0)
NT_JUR = 2.0
|   TMP_FIS <= 1
|   |   NR_EMP <= 1: SIM (43.0/15.0)
|   |   NR_EMP > 1: NAO (17.0/5.0)
|   TMP_FIS > 1
|   |   TM_ATIV <= 19: SIM (519.0/102.0)
|   |   TM_ATIV > 19
|   |   |   INF_ESC <= 1
|   |   |   |   TP_TIRB <= 1
|   |   |   |   |   CNAE_P1 <= 51: SIM (16.5/5.0)
|   |   |   |   |   CNAE_P1 > 51: NAO (16.5/4.5)
|   |   |   |   TP_TIRB > 1: SIM (10.0/2.0)
|   |   |   INF_ESC > 1
|   |   |   |   TP_TIRB <= 1: SIM (19.0/2.0)
|   |   |   |   TP_TIRB > 1
|   |   |   |   |   TM_ATIV <= 27: SIM (11.0/2.0)
|   |   |   |   |   TM_ATIV > 27: NAO (3.0)
NT_JUR = 1.0: SIM (120.0/16.0)
NT_JUR = 4.0
|   NR_EMP <= 0: SIM (29.0/7.0)
|   NR_EMP > 0: NAO (12.0/5.0)
NT_JUR = 18.0: NAO (1.0)
NT_JUR =  : NAO (1.0)
NT_JUR = 99.0
|   TP_ISEN <= 0
|   |   NR_EMP <= 4
|   |   |   CNAE_S1 <= 86: SIM (11.43/3.86)
```

Figura 26 - Exemplo da árvore construída para gerar o modelo classificador.

Os atributos remanescentes do contexto socioeconômico do contribuinte, não utilizados no primeiro experimento, continham informações facultativas, ocasionando ocorrências em que um determinado atributo não atingisse o percentual de 70% de preenchimento, em relação ao total de registro do conjunto de dados.

No segundo experimento, foram escolhidos todos os atributos que pertencem ao domínio socioeconômico do contribuinte, totalizando uma seleção com 16 atributos, como ilustrado na tabela 11.

Tabela 11 - Relação dos 16 atributos selecionados no segundo experimento.

Atributo	Descrição
TMP_FISC	Tempo decorrido entre a atual auditoria e a última realizada
CNAE_P1	Grupo da atividade primária do contribuinte
CNAE_P2	SubGrupo da atividade primária do contribuinte
CNAE_S1	Grupo da atividade secundária do contribuinte
TM_ATIV	Tempo de atividade
NR_EMP	Número de empregados
NT_JUR	Natureza jurídica
INF_ESC	Informação se possui ou não escrita fiscal
TP_TRIB	Se envolve so o imposto do ISS ou demais
TP_ISEN	Se possui algum benefício de isenção
TP_REC	Se o fator e baseado na movimentação econômica ou estimativa
INF_MIC	Se é micro empresa
INF_LIB	Se é profissional liberal
TP_EST	Se possui alguma estimativa no cálculo do imposto
VL_MOV	Valor movimentado pelas atividades selecionadas
IRREG	Atributo alvo que direciona as classes

O segundo experimento preservou o conjunto final de dados, utilizado no experimento anterior. Após o treinamento realizado, foi selecionado o conjunto de testes contendo 540 registros para realizar a aplicação do algoritmo de classificação. O resultado dessa combinação apresentou um percentual de acertos de 83,15%,

representando 449 ocorrências e um percentual de 16,85% de erros, representando 91 ocorrências, como visualizado na figura 27.

```
Number of Leaves  :     24

Size of the tree :      33


Time taken to build model: 0.01 seconds

=== Evaluation on training set ===
=== Summary ===

Correctly Classified Instances         449               83.1481 %
Incorrectly Classified Instances        91               16.8519 %
Kappa statistic                          0.522
Mean absolute error                      0.2287
Root mean squared error                  0.3379
Relative absolute error                 58.1671 %
Root relative squared error             76.243  %
Total Number of Instances              540
```

Figura 27 - Resultado da classificação do segundo experimento com 51,9% registros de teste e 48,1% para treinamento.

Nesse experimento, o algoritmo criou um modelo em forma de árvore com um quantitativo de 24 folhas e com tamanho total de 33 ramificações como visualizado na parte superior da figura 27. Com essa seleção de atributos, utilizando o mesmo conjunto de dados do experimento anterior, o algoritmo conseguiu criar uma estrutura menor, menos complexa que a estrutura criada no primeiro experimento, e como resultado, obteve uma melhora de 0,75% no índice de acerto.

A figura 27 demonstra que a avaliação do modelo criado, foi mediante o conjunto de treinamentos, informação destacada na quarta linha. Outro teste com esses atributos foi realizado, ao carregar o conjunto de dados e subdividi-lo de forma automática no momento da classificação, representando um conjunto de treinamento em 66% e um conjunto de testes em 34% dos 1040 registros totais. Nessa configuração o resultado obtido foi inferior ao resultado apresentado na divisão de dois

conjuntos na etapa de pré-processamento, representando 51,9% para testes e 48,1% para treinamentos. O resultado apresentado foi de um percentual de 75,70%, representando um quantitativo de 268 acertos e um percentual de 24,30%, representando um quantitativo de 86 erros como apresentado na figura 28.

```
Number of Leaves  :     38

Size of the tree :      61


Time taken to build model: 0.03 seconds

=== Evaluation on test split ===
=== Summary ===

Correctly Classified Instances         268               75.7062 %
Incorrectly Classified Instances        86               24.2938 %
Kappa statistic                          0.3444
Mean absolute error                      0.26
Root mean squared error                  0.4077
Relative absolute error                 65.2002 %
Root relative squared error             90.8093 %
Total Number of Instances              354
```

Figura 28 - Resultado da classificação do segundo experimento com 34% registros de teste e 66% de treinamento.

O resultado apresentado na figura 28 após a subdivisão automática do conjunto final dos dados demonstra um percentual de acertos inferior ao experimento realizado com a divisão dos conjuntos em teste e treinamento, que obteve um percentual de acertos de 83,15%. Importante ressaltar que nessa projeção, apesar do menor índice de acertos, o algoritmo criou um modelo em forma de árvore com um quantitativo de 38 folhas e com tamanho total de 61 ramificações, ou seja, uma piora na indução do modelo como visualizado na parte superior da figura 28.

Posteriormente, os atributos do contexto da auditoria foram adicionados, para geração de novos modelos. O experimento utilizou-se dos 26 atributos selecionados para o processo de MD, os quais estão descritos no item 5.3.2. Nesse caso, alguns

atributos da auditoria não puderam ser utilizados, visto que determinados atributos, como exemplo o atributo tipo de regularidade constatada, continha o mesmo direcionamento que o atributo alvo criado para direcionar os demais atributos às classes que indicam um contribuinte como irregular ou não. Nesse caso, o respectivo atributo pode ser utilizado para catalogar um novo processo dentre os contribuintes irregulares, dimensionando o grau de inconsistência de um respectivo contribuinte, como exemplo, irregularidade leve ou moderada.

Após diversos testes realizados com a adesão dos atributos da auditoria, um terceiro experimento foi elaborado com 20 atributos, como ilustrados na tabela 12.

Tabela 12 - Relação dos 20 atributos selecionados no terceiro experimento.

Atributo	Descrição
TMP_FISC	Tempo decorrido entre a atual auditoria e a última realizada
CNAE_P1	Grupo da atividade primária do contribuinte
CNAE_P2	SubGrupo da atividade primária do contribuinte
CNAE_S1	Grupo da atividade secundária do contribuinte
CNAE_S2	SubGrupo da atividade secundária do contribuinte
TM_ATIV	Tempo de atividade
NR_EMP	Número de empregados
NT_JUR	Natureza jurídica
INF_ESC	Informação se possui ou não escrita fiscal
TP_TRIB	Se envolve so o imposto do ISS ou demais
TP_ISEN	Se possui algum benefício de isenção
TP_REC	Se o fator e baseado na movimentação econômica ou estimativa
VL_DEV_N	Valor do movimento da atividade
INF_MIC	Se é micro empresa
INF_LIB	Se é profissional liberal
TP_EST	Se possui alguma estimativa no cálculo do imposto
SERV1_GR	Grupo da atividade econômica
NR_SERV1	SubGrupo da atividade econômica
TP_CTM	Fundamentação artigo CTM
IRREG	Atributo alvo que direciona as classes

Utilizando-se da mesma disposição de dados utilizados no experimento anterior, foi selecionado um conjunto de dados contendo 500 registros para realizar o

treinamento do algoritmo de classificação. O resultado dessa combinação apresentou um percentual de acertos de 92,03%, representando 497 ocorrências e um percentual de 7,97% de erros, representando 43 ocorrências, como visualizado na figura 29.

```
Number of Leaves  :     30

Size of the tree :      45


Time taken to build model: 0.01 seconds

=== Evaluation on training set ===
=== Summary ===

Correctly Classified Instances         497               92.037  %
Incorrectly Classified Instances        43                7.963  %
Kappa statistic                          0.807
Mean absolute error                      0.1239
Root mean squared error                  0.248
Relative absolute error                 31.5162 %
Root relative squared error             55.9486 %
Total Number of Instances              540
```

Figura 29 - Resultado da classificação do terceiro experimento com 51,9% registros de teste e 48,1% para treinamento.

Nesse experimento, o algoritmo criou um modelo em forma de árvore com um quantitativo de 30 folhas e com tamanho total de 45 ramificações como visualizado na parte superior da figura 29. A seleção final com atributos do contexto socioeconômico e de auditoria, utilizando o mesmo conjunto de dados dos experimentos anteriores, conseguiu criar um modelo de classificação com uma estrutura menor, equivalente ao primeiro experimento, e como resultado, obteve uma considerável melhora de 8,88% no índice de acerto.

A figura 29 demonstra que a avaliação do modelo criado, foi mediante o conjunto de treinamentos, informação destacada na quarta linha. Posteriormente, outro teste foi feito pela subdivisão automática realizada pela ferramenta no momento da

classificação, representando um conjunto de treinamento em 66% e um conjunto de testes em 34% da quantidade de registros total. Nessa configuração o resultado obtido foi inferior ao resultado apresentado na divisão de dois conjuntos na etapa de pré-processamento, representando 51,9% para testes e 48,1% para treinamentos. O resultado apresentado foi de um percentual de acertos de 89,55%, representando um quantitativo de 317 acertos e um percentual de 10,45%, representando um quantitativo de 37 erros como apresentado na figura 30.

```
Number of Leaves  :     32

Size of the tree :      49


Time taken to build model: 0.14 seconds

=== Evaluation on test split ===
=== Summary ===

Correctly Classified Instances         317               89.548  %
Incorrectly Classified Instances        37               10.452  %
Kappa statistic                          0.7578
Mean absolute error                      0.1478
Root mean squared error                  0.2929
Relative absolute error                 37.0534 %
Root relative squared error             65.2466 %
Total Number of Instances              354
```

Figura 30 - Resultado da classificação do terceiro experimento com 34% registros de teste e 66% de treinamento.

.

Com essa configuração, o algoritmo criou um modelo em forma de árvore com um quantitativo de 32 folhas e com tamanho total de 49 ramificações como visualizado na parte superior da figura 30.

5.8 AVALIAÇÃO DOS MODELOS

A avaliação do modelo proposto pelo processo de MD representa uma etapa complexa e desafiadora aos gestores, visto as possibilidades de observações que podem ser concatenadas e também, perante os métodos de avaliação utilizados. Pode envolver os fatores tempo, recursos financeiros, recursos humanos disponíveis e qualidade dos dados para determinar qual resultado pode ser considerado qualificado e eficaz. O modelo é avaliado como efetivo se responde aos objetivos tratados na fase de entendimento do problema (LAROSE, 2006).

Para realizar a avaliação dos modelos criados quanto à qualidade e eficácia e essencial a utilização de métodos estatísticos. Referente à técnica de classificação, a avaliação pode ser considerada efetivada pelos conceitos taxa de erros, falso positivo, falso negativo e custo de erros de ajustes, métodos estes utilizados no algoritmo de classificação C4.5 (QUINLAN, 2003), base do algoritmo J48, utilizado neste trabalho (WEKA, 2012).

A ferramenta *Weka* utiliza o método estatístico *Kappa* para realizar a avaliação do modelo criado. A estatística *Kappa* é utilizada para realizar a medida de concordância em escalas nominais fornecendo um cenário que apresenta o quanto as observações se afastam das observações esperadas, indicando assim o grau de legitimidade das interpretações feitas pelo modelo criado (THOMPSON, 2001). Tal informação está destacada nas figuras referentes aos resultados dos modelos criados neste trabalho, como ilustrado na figura 30. A estatística *Kappa* é calculada em três etapas. Primeiro calcula-se um índice que represente a concordância esperada pelo acaso. Em segundo, calcula-se a concordância observada e por último, a estatística é calculada pela divisão da diferença entre a concordância observada e a esperada, pela diferença entre a concordância absoluta e a esperada pelo acaso. O resultado busca a maior diferença possível entre a concordância observada e a esperada (THOMPSON, 2001).

A tabela 13 ilustra o índice utilizado pela *Kappa* que representa classificação da magnitude, com intervalo de confiança de 95%.

Tabela 13 - Índices Kappa.

Valor Kappa	Concordância
0	Pobre
0 - 0,20	Ligeira
0,21 - 0,40	Considerável
0,41 - 0,60	Moderad
0,61 - 0,80	Substâncial
0,81 - 1	Excelente

5.9 RESULTADOS ENCONTRADOS

Os experimentos realizados buscaram identificar o comportamento da combinação dos atributos do contexto socioeconômico dos contribuintes juntamente com os catalogados na auditoria. O primeiro experimento utilizou 12 atributos do contexto socioeconômico dos contribuintes que continham um mínimo de 70% de preenchimento. Nesse cenário, dois testes foram realizados como visualizado na figura 31.

	Porcentagem de acerto dos experimentos					
	1° Experimento		2° Experimento		3° Experimento	
	Acertos	Erros	Acertos	Erros	Acertos	Erros
1° Teste	82,40%	17,60%	83,15%	16,85%	92,03%	7,97%
2° Teste	70,62%	29,38%	75,70%	24,30%	89,55%	10,45%

1° teste - 59,01% dos registros de teste e 48,01% para treinamento.
2°teste - 34% dos registros de teste e 66% para treinamento.

Figura 31 - Porcentagem dos acertos do modelo dos experimentos realizados.

O primeiro teste foi realizado através da subdivisão do conjunto de dados em um conjunto de treinamento e outro de testes, que alcançou um índice de 82,40% de acertos e outro, através da subdivisão automática realizada pela ferramenta, que alcançou um índice de 70,62% de acertos, como ilustrado na figura 31.

O segundo experimento selecionou todos os atributos do contexto socioeconômico dos contribuintes, resultando numa seleção de 16 atributos. Nesse cenário, através da subdivisão do conjunto final de dados em um conjunto de treinamento com 48,1% dos registros e um conjunto de testes com 51,9%, o algoritmo de classificação criou um modelo que alcançou um índice de 83,15% de acertos. Novamente, a divisão automática realizada pela ferramenta indicou um índice menor de acerto, nesse caso, 75,10%, como visualizado na figura 31, observando o mesmo conjunto de 16 atributos.

O terceiro experimento adicionou os atributos da auditoria na composição do modelo. Após diversos testes realizados com a adesão dos atributos da auditoria, visualizou-se que alguns tinham mesma característica do atributo alvo criado para direcionar as demais variáveis ao modelo classificador, fato que inutilizava a utilização dos demais atributos preditivos, e assim, não poderiam ser utilizadas, como exemplo o atributo tipo de irregularidade. Assim, o terceiro experimento selecionou 20 atributos dentre os de auditoria e do contexto socioeconômico dos contribuintes, alcançando o melhor índice de acerto encontrado. O modelo construído nesse cenário indicou uma média de acertos em 92,03% dos registros testados, como visualizado na figura 31 através dos últimos indicadores.

VI CONCLUSÕES

O objetivo principal deste trabalho foi elaborar um modelo, baseada em uma metodologia especifica de MD, que auxiliasse o Departamento de Fiscalização na melhor identificação dos contribuintes do ISS do município de Goiânia com alguma espécie de irregularidade, baseando-se no denso volume de informações que o Departamento recebe e gerencia mensalmente.

No entendimento do problema da sonegação, reuniões foram realizadas com gestores do departamento de fiscalização e auditoria, departamento de arrecadação e do departamento de tecnologia responsável pelos controles desenvolvidos no município. Essa interação foi essencial para delimitar o problema, para definir modalidades de atuação que teriam relevância para o objetivo do trabalho, elucidar o domínio dos dados catalogados e, paralelamente, idealizar um mecanismo que fizesse a extração dos dados necessários à condução deste projeto de MD, observadas as orientações disposta na metodologia dos CRISP-DM.

O objetivo principal deste trabalho foi alcançado, ao passo que o modelo classificador dos contribuintes com irregularidades foi idealizado baseado na técnica de árvore de decisão, utilizando o algoritmo J48. Para permitir tal idealização, os objetivos específicos também foram alcançados para formalizar o conjunto de dados utilizado no processo de MD. A extração dos dados exigiu a criação de um mecanismo específico para essa finalidade, representando um marco, uma difícil etapa que viabilizou o prosseguimento das atividades do projeto de MD.

A decisão da definição da tarefa e da técnica necessária para realizar o processo de MD não é uma tarefa trivial, representando um importante fator de sucesso para o resultado do processo. Após a extração, formalização e padronização do conjunto de

dados foi possível visualizar o domínio do conjunto extraído, os objetivos perseguidos pelo trabalho e qual técnica seria eficiente para unificar tais necessidades. A tarefa de classificação é utilizada quando se conhece as classes, mas sem uma clara distinção entre os valores dos atributos que a compõem. Neste caso, o algoritmo de árvore de decisão foi selecionado para combinar e direcionar os atributos preditivos às classes previamente conhecidas, através de um atributo alvo que intitula um contribuinte como irregular ou não.

O processo de MD utilizou uma base composta de 1040 contribuintes que pertenciam primariamente à realização de serviços e que não tinham benefícios do SIMPLES Nacional. Esse conjunto de dados após as atividades de pré-processamento, limpeza e formatação, possibilitou a execução de diversos experimentos na busca de um modelo que melhor identificasse os contribuintes irregulares.

O primeiro experimento selecionou os atributos socioeconômicos com índice de preenchimento maior que 70%. Essa configuração composta de 12 atributos idealizou um modelo de classificação com índice de 82,40% de acertos. Posteriormente um segundo experimento utilizou todos os atributos do contexto socioeconômico do contribuinte. Esse experimento composto de 16 atributos atingiu uma melhoria aumentando a eficiência do modelo ao apresentar um índice de 83,15% de acerto.

O modelo encontrado que apresentou melhor resultado foi realizado com uma seleção composta de 20 atributos pertencente ao domínio socioeconômico e parte dos atributos catalogados na auditoria. Nesse cenário, o algoritmo escolhido conseguiu estabelecer um acerto aproximado de 92,03% de contribuintes regulares daqueles contribuintes que apresentaram alguma irregularidade.

Tal conjunção não finda o processo de MD, visto que outros algoritmos podem ser utilizados pra realizar a construção do modelo de classificação dos contribuintes

irregulares, juntamente com a utilização de demais métodos estatísticos que avalie a qualidade e eficácia do resultado encontrado.

Os resultados encontrados têm extrema importância para o Departamento de Fiscalização ao prover uma indicação do contexto dos contribuintes com irregularidades e em que tipo e frequência elas ocorrem. Além da possibilidade de planejamento de um trabalho de auditoria mais efetivo, tais indicadores podem fornecer subsídios para possibilitar o estabelecimento de estratégias de políticas públicas ao estudar o comportamento de alguns segmentos de contribuintes, a adesão de determinado segmento a regularidade, a mensuração de participação das atividades econômicas, o tipo de imposto recolhido por determinado segmento e o comportamento de benefícios concedidos aos contribuintes.

6.1 DIFICULDADES ENCONTRADAS

A principal dificuldade foi estabelecer o domínio dos dados necessários para realizar o projeto de MD, visto que os demais trabalhos publicados referentes à identificação dos contribuintes irregulares e detecção de evasão fiscal utilizaram informações das movimentações econômicas dos contribuintes, ao invés de informações socioeconômicas, domínio este utilizado neste trabalho. Pelo fato da pesquisa ser realizada em um Departamento Público e da forma que os dados estavam armazenados, a permissão, o entendimento e a extração demandaram autorizações e diversas reuniões realizadas com os responsáveis, tornando-se fator dificultador para a consecução das atividades realizadas neste trabalho.

Outro desafio encontrado foi à definição da técnica que realizasse o trabalho de classificação. A técnica de árvore de decisão foi escolhida com base no domínio das informações disponíveis e pela sua capacidade de trabalhar com variáveis categóricas

e contínuas, visto a sua aplicabilidade em diversos trabalhos realizados com o intuito de classificação.

6.2 INDICAÇÕES PARA FUTUROS TRABALHOS

Os modelos construídos neste processo de MD visaram classificar os contribuintes do ISS identificando o contexto dos que apresentaram alguma irregularidade. Uma segunda possibilidade a ser analisada refere-se à classificação do perfil do contribuinte com maiores índices de irregularidades constatadas através dos atributos que não puderam ser utilizados nesse processo de MD. O atributo tipo irregularidade poderia ser utilizado para dimensionar os registros de contribuintes irregulares a um novo cenário representando o grau de relevância das irregularidades cometidas.

REFERÊNCIAS

ADABAS. Data Management Without Limits, Disponivel em: <http://www.softwareag .com/br/products/adabas_2010/default.asp>. Acesso em: 11 nov. 2010.

AL-RADAIDEH, Q. A., NAGI, E. Al., Using Data Mining Techniques to Build a Classification Model for Predicting Employees Performance, "**International Journal of Advanced Computer Science and Application (IJACSA)"**, Vol 3, Fevereiro, 2012.

ANDRADE, H. S., 2009, **Um processo de Mineração de dados Aplicado à Sonegação Fiscal do ICMS,** Dissertação de M.Sc, MPCOMP/IFCE, Ceará,Brasil.

ANDRADE FILHO, E. O. **Auditoria de Impostos e Contribuições**. São Paulo, Editora Atlas, 2005.

ANGELINE D. M. D, JAMES S. P., **"**Association Rule Generation Using Apriori Mend Algorithm for Student's Placement**",**. **International Jounal of English Studies- IJES**. Int. J. Emerg. Sci ., 2(1), 78-86, Março 2012, ISSN 2222-4254.

BERRY, M. J. A.; LINOFF, G., **Data Mining Techniques for Marketing, Sales and Customer Support**. 2. ed. New York: John Wiley & Sons, 2004.

BONCHI, F. et al. **Using data mining techniques in fiscal fraud detection,** 1999, Disponível em: < http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.42.1403 >. Acesso em: 19 Nov. 2012.

BRAGA, C. V., 2010, **Rede Neural e Regressão Linear: Comparativo entre as Técnicas Aplicadas a Um Caso Prático na Receita Federal,** Dissertação de M.Sc, Programa de Pós-Graduação em Administração e Economia,IBMEC, Rio de janeiro ,Brasil.

BUSSAB, W. O.; MIAZAKI, E.S.; ANDRADE, D.F.. Introdução à Análise de Agrupamentos IN: Anais do 9º. Simpósio Nacional de Probabilidade e Estatístico. São Paulo: Associação Brasileira de Estatística /ABE, julho, 1990.

CABENA, P., HADJINIAN, P., STADLER, R., *et al*., **Discovering data mining**; from concept to implementation. Upper Saddle River, Prentice-Hall PTR, 1998.

CHIU, S., TAVELLA, D. **Data Mining And Market Intelligence For Optimal Marketing Returns**, Oxford, Butterworth-Heinemann, 2008.

CHIZZOTTI, A. **A pesquisa qualitativa em ciências humanas e sociais**. Petrópolis Vozes, 2008.

CLEARY, D., Predictive Analytics in the Public Sector, "**Eletronic Jornal of E-Government (EJEG)"**, Vol 9, n 2, p 132-140, Dezembro, 2011.

CORVALÃO, E. D., 2009, **Classificação de contribuintes: um modelo de duas fases**. Tese de D.Sc., Programa de Pós-Graduação em Engenharia de Produção, UFSC, Florianópolis,Brasil.

CUNHA, A. L., "Regra-matriz do ISSQN – Imposto Sobre Serviços de Qualquer Natureza", **Revista IMES – Direito**, ano VIII, n°13, pp 65, julho/dezembro de 2007.

DIGITAL. Programa Goiânia Digital, Disponivel em: < http://www.goiania.go.gov.br/html/estacaodigital/>. Acesso em: 02 abr. 2013.

DOMINGUES, M. A., 2004, **Generalização de Regras de Associação**. Dissertação de M.Sc, ICMC/ Universidade São Paulo, USP, São Paulo, Brasil.

FARIA, L. I. L, QUONIAN, L., Ferramentas para Estudos Prospectivos – Tutorial, "**3° Workshop Brasileiro de Inteligência competitiva e Gestão do Conhecimento**", São Paulo , SP, 16 a 18 de Setembro de 2002.

FAYYAD, U. M., PIATETSKY-SHAPIRO, G.; SMYTH, P., *et al*., **Advances in Knowledge Discovery and Data Mining**. AAAI/MIT Press, 1996.

FAYYAD, U., PIATETSKY-SHAPIRO, G., SMYTH, P., "From Data Mining To Knowledge Discovery in Databases**", AI Magazine**, vol. 17, p. 37-54, 1996.

FIGUEIRA, V. C., 2006, **Modelos de Regressão Logística**. Dissertação M.Sc., Instituto de Matemática, UFRS, Rio Grande do Sul, Brasil.

FUTEMA, F. **Sonegação fiscal cresce e atinge quase 30% das empresas, diz IBPT.** Folha de São Paulo, São Paulo, 18 ago. 2005. Caderno dinheiro.

GANTZ, J., REINSEL, D., **Extracting Value From Chao,** IDC Analyze the future, USA, 2011, Disponível em: < http://www.emc.com/collateral/analyst-reports/idc-extracting-value-from-chaos-ar.pdf>, Acesso em 18 de Janeiro de 2013.

GIUDICI, P., **Applied Data Mining: Statistical methods for business and industry.** London, John Willey & Sons, 2003.

GIL, A. C., **Estudo de Caso**. São Paulo, Atlas, 2009.

GIRIOLI, L. S., 2010, **Análise do uso de medidas de desempenho de empresas presente na pesquisa em contabilidade no Brasil**. Dissertação de M.Sc, Faculdade de Economia, Administração e Contabilidade da Universidade de São Paulo, SP, Brasil.

HAN, J., KAMBER, M., **Data Mining Concepts and Techniques**, 2ª Edição, San Francisco, Elsevir. 2006, Disponível em: <http://books.google.sk/books?id=AfL0t-YzOrEC&printsec=frontcover&hl=sk&source=gbs_ge_summary_r&cad=0#v= onepage&q&f=true >, Acesso em: 17 de agosto 2012.

HARRISON, T.H., **Intranet data warehouse,** São Paulo, Editora Berkeley, 1998.

HAYKIN, S. S., **Neural Networks** : A comprehensive Foundation, 2ª Edição, Upper Saddle River, Prentice-Hall PTR ,1999.

IBGE. **Contagem da população 2007,** Rio de Janeiro, 2007, Disponível em:< http://www.ibge.gov.br/home/estatistica/populacao/contagem2007/contagem.pdf >. Acesso em: 17 de Agosto 2012.

INMON, W.H.; HACKATHORN, R. D., **Como usar o Data Warehouse.** Rio de Janeiro, Infobook, 1997**.**

JACKSON, J., "Data Mining: A Conceptual Overview." **Communications of the Association for Information Systems**, v. 8, p. 267-296, 2002.

LAROSE, D. T., **Discovdery Knowledge in Data. An Introducing to DATA MNING,** New Jersey, John Wiley & Sons, 2006.

Lei complementar n° 123, de 14 de Dezembro de 2006. Institui o Estatuto Nacional da Microempresa e da Empresa de Pequeno Porte em: < http://www.receita.fazenda.gov.br/Legislacao/LeisComplementares/2006/leicp123.htm >. Acesso em: 20 jan. 2012.

MEIRELLES, H. L., **Direito Administrativo brasileiro**. São Paulo, Editora Atual, 2004.

MICCI-BARRECA, D., RAMACHANDRAN, S., **Improving Tax Administration with Data Mining.** Analytics Elite, 2006, Disponível em: <http://www.spss.ch/upload/1122641565_Improving%20tax%20administration%20with%20data%20mining.pdf >. Acesso em: 20 Jan. 2012.

MIT, Emerging Technology that will change the word, **Technology Review,** 2001, Disponível em: <http://www.technologyreview.com/featured-story/400868/emerging-technologies-that-will-change-the-world/>, Acesso em 17 de Agosto de 2012.

MITRA, S.; ACHARYA, T. **Data mining: multimedia, soft computing, and bioinformatics,** 2003**,** Disponível em: <http://books.google.com/books?id=VPeOaKNfDlgC&pg=PA4&dq=kdd&hl=pt-BR#v=onepage&q=kdd&f=false>. Acesso em: 17 de agosto 2012.

NASCIMENTO, J. R., 2010, **Um estudo sobre a influência das regras e procedimentos de controle fiscal via internet nos resultados da arrecadação tributaria de Municípios do Estado e São Paulo.** Dissertação de M.Sc, Faculdade de Economia, Administração e Contabilidade da Universidade de São Paulo, SP, Brasil.

NETO, M. A. S., VILLWOC, R, SCHEER S, *et al.*, "Técnicas de Mineração Visual de Dados Aplicadas aos Dados de Instrumentação da Barragem de Itaipu", **Revista Gestão e Produção**, São Carlos, v.17, n. 4, p 721-734, 2010.

OLSON, D. L., DELEN, D., **Advanced Data Mining Techniques,** Berlim:,Springer, 2008.

PENTAHO. **Introducing the Pentaho BI Suite Community Edition. Pentaho Open Source Business Inteligence**, 2009. Disponivel em: <http://wiki.pentaho.com/download/attachments/12386846/community_user_guide.pdf ?version=1>. Acesso em: 11 nov. 2011.

PIZZIRANI, F., 2003, **Otimização Topológica de Estruturas Utilizando Algoritmos Geneticos**.. Dissertação de M.Sc, Faculdade de Engenharia Mecânica, Unicamp, São Paulo, Brasil.

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

REZENDE, S. O., **Sistemas inteligentes: fundamentos e aplicações**. São Paulo, Manole, 2005, Disponível em: < http://books.google.com.br/books?id=UsJe_PlbnWcC&pg=PA8&lpg=PA8&dq=sistemas+inteligentes+fundamentos+e+aplica%C3%A7%C3%B5es&source=bl&ots=EkSSSLWBWt&sig=1f2gStkR7XFURWZi1J9VFGCNJR4&hl=pt-BR&sa=X&ei=SjxsULb-L5Pi9gSowoCoBw&sqi=2&ved=0CCgQ6AEwAQ>. Acesso em: 17de agosto 2012.

ROCHA, F. G., 2003, **Contribuição de Modelos de Séries Temporais para a Previsão da Arrecadação de ISS,** Dissertação de M.Sc, Universidade Estadual de Campinas, Unicamp, São Paulo.

ROGER, R. J.; GEATZ, M. W. **Data Mining: A Tutorial-Based primer.** Boston, Addilson Wesley, 2003**.**

RUD, O. P., **Data mining cookbook:** modeling data for marketing, risk, and customer relationship management, New York: John Wiley & Sons, 2001.

SAS. SAS Enterprise Miner Reference, Disponivel em: < http://www.sas.com/technologies/analytics/datamining/miner/index.html>. Acesso em: 02 abr. 2013.

SEMMA. SEMMA Metodology, Disponivel em: < http://www.sas.com/offices/europe/uk/technologies/analytics/datamining/miner/semma.html>. Acesso em: 02 abr. 2013.

SFERRA, H. H., CORRÊA, Â. M. C. J, “Conceitos e Aplicações de *Data Mining*”. **Revista de Ciência & Tecnologia**, v.1, n. 22, p. 19-34. jul./dez. 2003.

SHEARER, C., The CRISP-DM Model: the new blueprint for data mining, “**Journal of Data Mining”**, v. 5, n. 4, pp. 13-22, 2000.

SILVER, D. L., **Knowledge discovery and data mining**, Technical Report MBA6522, CogNova Technologies London Health Science Center, 1996.

SIQUEIRA, M. L.; RAMOS, F. S. “A economia da sonegação: teorias e evidências empíricas”, **Revista de Economia Contemporânea**, v. 9, n. 3, p. 555-581, Set./Dez. 2005. Rio de Janeiro.

SUMATHI, S., SIVANANDAM, S.N., **Introduction to Data Mining and its Applications,** Berlim, Springer, 2006.

THOMPSON. J. R., **“**Estimation equations for kappa statistics**”,**. **Statistics in Medicine**, Volume 20, Edição 19, 2895 - 2906, Outubro 2001.

TOMÉ, F. D. P., **A prova no direito tributário**. São Paulo, Noeses, 2005.

TAN, P., STEINBACH, M., KUMAR V., **Introduçao ao DATA MINING Mineração de Dados**, Rio de Janeiro, Editora Ciência Moderna, 2009.

TURRIONE, J. B.; MELLO, C. H. P., **Metodologia de Pesquisa em Engenharia de Produção**. Minas Gerais, Universidade Federal de Itajubá, UNIFEI, 2011.

VENKATESWAR RAO, G.T., CHITOOR, B., REDDY K, V., VENUGOPAL RAO, G., “ Intelligent Data Mining for Detecting Tax Evasion: Using an Innovative Approach for Name Search”. **International Conference on Management of Data COMAD 2005b**, India, 20 a 22 de Dez. 2005.

VIGLIONI, G. M. C., 2007, **Metodologia para previsão de demanda ferroviária utilizando data mining**. 2007. Dissertação de M.Sc, Instituto Militar de Engenharia, IME. Rio de Janeiro, Brasil.

WEKA. Data Mining Software in Java, Disponivel em: < http://www.cs.waikato.ac.nz/ml/weka/>. Acesso em: 19 Fev. 2013.

WITTEN, I. H.; FRANK, E. **Data mining: practical machine learning tools and techniques**, 2005. Disponível em: <http://books.google.com/books?id=QTnOcZJzlUoC&printsec=frontcover&dq=data+mining&hl=pt-BR#v=onepage&q=&f=false>. Acesso em: 17de agosto de 2012.

YAMASHITA, D. **Elisão e Evasão de tributos**. São Paulo, Lex Editora, 2005.

YODA, E. M., 2003, **Inteligência Computacional no Projeto Automático de Redes Neurais Híbridas e Redes NeuroFuzzy Hererogêneas**. Dissertação de M.Sc, Faculdade de Engenharia Eletríca e Computação, Unicamp, São Paulo.

YIN, R. K., **Applications of case study research**. Thousand Oaks, California, Sage 2003.

YU, F., QIN, Z., JIA, X., “Data Mining Application Issues in Fraudulent Tax Declaration Detection”. **Machine Learning and Cybernetics, 2003 International Conference on,** Vo. 4. Pag. 2202 a 2206, China. 02 a 05 de Nov. 2003

ZORNETZER, S. F., MCKENNA, T. M., LAU C., *et al,* **An Introduction to Neural and Electronic Networks**. 2a. Ed. London, Academic Press, 1994, Disponível em: <http://books.google.com.br/books/about/An_Introduction_to_Neural_and_Electronic.html?id=ocP4j8LJSl4C&redir_esc=y>. Acesso em: 17 de agosto 2012.

ANEXO I

Tabela dos 120 atributos selecionados com relevância para o processo de MD após a extração, pré-processamento e formatação dos dados dos contribuintes.

Atributo	Descrição
NUMR	Informação para distringuir dados de determinado contribuinte
DT_IN	Data inicial do movimento fiscalizado
DT_FIM	Data final do movimento fiscalizado
VL_DEV	Valor averiguado da movimentaçao do contribuinte no ato da fiscalização
INF_MIC	Informação de micro empresa
DT_ABERT	Data de inicio de atividade
NR_EMP	Número de empregados
NT_JUR	Natureza jurídica
INF_ESC	Informação se possui ou não escrita fiscal
TP_TRIB	Se contribui somento com o imposto do ISS ou demais
TP_ISEN	Se possui algum benefício de isenção
TP_REC	Se o fator e baseado na movimentação econômica ou estimativa
NR_ART_CTM	Informação dos artigos de fundamentaçao com base no CTM
NR_ART_RCTM	Informação dos artigos de fundamentaçao com base no RCTM
NUMR_ATV1	Atividade economica principal
NUMR_ATV2	Atividade economica secundária
TP_IRREG_1	Tipo de irregularidade constadada
TP_IRREG_2	Tipo de irregularidade constadada
TP_IRREG_3	Tipo de irregularidade constadada
NR_SERV1	Serviço executado principal
NR_SERV2	Serviço executado secundário
NR_SERV3	Serviço executado secundário
NR_BAIR	Código de localização
INF_LIB	Informação de autônomo
TP_EST	Tipo de estimativa de contribuição
IRREG	Apresentaçao ou não de irregularidade
NM_RAZAO	Razão Social
NM_FANTASIA	Nome Fantasia
INF_OPER	Indica a situação da atividade
INF_MOT	Motivo de alteraçao da atividade
DT_EVEN	Data evento da atividade
INF_ORIG	Origem da atividade
INF_FOROPER	Forma de operação da atividade
INF_END	Endereço da atividade econômica
CD_LOG	Código logradouro
NM_LOG	Nome logradouro
NR_LOG	Número logradouro

END_QD	Quadra
END_LOT	Lote
END_COMPL	Complemento
CD_BAIRRO	Código bairro
INF_DDD	Ddd do telefone da atividade
TEL	Telefone da atividade
DT_MOV	Data de entrada da atividade
DT_BAIXA	Data da baixa da atividade
DT_INC	Data de inclusão da atividade
CD_ATIV	Atividade
DT_ALTER	Data de alteração da atividade
INF_ISE	Tipo de isenção
DT_ISE	Data da isenção
NR_INSP	Número de inspeção
NR_INSC_E	Inscrição estadual
NR_INSC_C	Inscrição comercial
TP_REG	Tipo registro
NR_SOC	Número de sócios
NR_SOC_L	Número de sócios liberais
DT_NASC	Data nascimento contribuinte
NR_IDEN	Número identidade contribuinte
ORG_EX	Orgão expedidor do contribuinte
REG_CLAS	Registro de classe
SG_CLAS	Sigla do registro de classe do contribuinte
NM_MAE	Nome mãe do contribuinte
END_CONT	Endereço do contato
CD_LOG	Código logradouro do contato
NM_LOG	Nome logradouro do contato
NR_LOG	Número logradouro do contato
END_QD_C	Quadra do contato
END_LOT_C	Lote do contato
END_COMPL_C	Complemento do contato
CD_BAIRRO_C	Código bairro do contato
INF_DDD_C	Ddd do telefone do contato
TEL_C	Telefone do contato
CD_MUN	Município do contato
NR_CEP	Cep do contato
NR_CPF_S	Número de CPF dos sócios
NR_CPF_R	Número de CPF do responsável
NR_CIM	Número do cim do responsável
DT_VIS	Data vistoria
NR_VIS	Vistoriador
NR_PROC	Processo
TP_DEV	Tipo devolução

DT_PROC	Data processamento
NR_OPE	Operador
HR_PROC	Hora processamento
NR_FIC	Número fic
INF_DEB	Gerar débitos
NR_CAE	Número de inscrição
NR_CGC	Número cgc ou cpf
NR_CIM_C	Número cim atividade
CD_FON	Código de fonetização
CD_FON_E	Código de fonetização do endereço
CD_FON_L	Código de fonetização do logradouro
INF_RED	Redução atividade
DT_REV	Data de revisao da atividade
DT_INE	Data de início de estimativa
DT_FINE	Data final de estimativa
VL_EST	Valor estimativa
NR_IPTU	Número IPTU
NR_INF_CONT	Contador
INF_SIM	Atividade simples
TP_DOC	Tipo documento atividade
CD_GEO	Código georeferenciamento
NM_CONT	Nome contador
DT_RED	Data redução atividade
CD_GRUPO	Código do grupo da atividade
CD_EST	Estimativa contribuinte
CD_ESTAT	Código de estatística
CD_BAI_C	Bairro contador
CD_FON_F	Código de fonetização da fantasia
INF_OBS	Observação atividade
NR_CNPJ_C	Cpnj do contador
ST_EST	Estação digital
DT_EST	Data Estação digital
INF_SIM	Informação simples
DT_SIM	Data do simples
INF_LOGP	Logradouro público atividade
INF_INSCP	Inscrição principal
INF_REG_A	Regime de apuração
INF_SEX	Sexo do contribuinte
NR_EST_V	Estabelecimento virtual

ANEXO II

Tabela da relação completa das atividades econômicas de abrangência nacional mantida pela Comissão Nacional de Classificação.

Grupo	Descrição
01	AGRICULTURA, PECUÁRIA E SERVIÇOS RELACIONADOS
02	PRODUÇÃO FLORESTAL
03	PESCA E AQÜICULTURA
05	EXTRAÇÃO DE CARVÃO MINERAL
06	EXTRAÇÃO DE PETRÓLEO E GÁS NATURAL
07	EXTRAÇÃO DE MINERAIS METÁLICOS
08	EXTRAÇÃO DE MINERAIS NÃO-METÁLICOS
09	ATIVIDADES DE APOIO À EXTRAÇÃO DE MINERAIS
10	FABRICAÇÃO DE PRODUTOS ALIMENTÍCIOS
11	FABRICAÇÃO DE BEBIDAS
12	FABRICAÇÃO DE PRODUTOS DO FUMO
13	FABRICAÇÃO DE PRODUTOS TÊXTEIS
14	CONFECÇÃO DE ARTIGOS DO VESTUÁRIO E ACESSÓRIOS
15	PREPARAÇÃO DE COUROS E FABRICAÇÃO DE ARTEFATOS DE COURO, ARTIGOS PARA VIAGEM E CALÇADOS
16	FABRICAÇÃO DE PRODUTOS DE MADEIRA
17	FABRICAÇÃO DE CELULOSE, PAPEL E PRODUTOS DE PAPEL
18	IMPRESSÃO E REPRODUÇÃO DE GRAVAÇÕES
19	FABRICAÇÃO DE COQUE, DE PRODUTOS DERIVADOS DO PETRÓLEO E DE BIOCOMBUSTÍVEIS
20	FABRICAÇÃO DE PRODUTOS QUÍMICOS
21	FABRICAÇÃO DE PRODUTOS FARMOQUÍMICOS E FARMACÊUTICOS
22	FABRICAÇÃO DE PRODUTOS DE BORRACHA E DE MATERIAL PLÁSTICO
23	FABRICAÇÃO DE PRODUTOS DE MINERAIS NÃO-METÁLICOS
24	METALURGIA
25	FABRICAÇÃO DE PRODUTOS DE METAL, EXCETO MÁQUINAS E EQUIPAMENTOS
26	FABRICAÇÃO DE EQUIPAMENTOS DE INFORMÁTICA, PRODUTOS ELETRÔNICOS E ÓPTICOS
27	FABRICAÇÃO DE MÁQUINAS, APARELHOS E MATERIAIS ELÉTRICOS
28	FABRICAÇÃO DE MÁQUINAS E EQUIPAMENTOS
29	FABRICAÇÃO DE VEÍCULOS AUTOMOTORES, REBOQUES E CARROCERIAS
30	FABRICAÇÃO DE OUTROS EQUIPAMENTOS DE TRANSPORTE, EXCETO VEÍCULOS AUTOMOTORES
31	FABRICAÇÃO DE MÓVEIS
32	FABRICAÇÃO DE PRODUTOS DIVERSOS

33	MANUTENÇÃO, REPARAÇÃO E INSTALAÇÃO DE MÁQUINAS E EQUIPAMENTOS
35	ELETRICIDADE, GÁS E OUTRAS UTILIDADES
36	CAPTAÇÃO, TRATAMENTO E DISTRIBUIÇÃO DE ÁGUA
37	ESGOTO E ATIVIDADES RELACIONADAS
38	COLETA, TRATAMENTO E DISPOSIÇÃO DE RESÍDUOS; RECUPERAÇÃO DE MATERIAIS
39	DESCONTAMINAÇÃO E OUTROS SERVIÇOS DE GESTÃO DE RESÍDUOS
41	CONSTRUÇÃO DE EDIFÍCIOS
42	OBRAS DE INFRA-ESTRUTURA
43	SERVIÇOS ESPECIALIZADOS PARA CONSTRUÇÃO
45	COMÉRCIO E REPARAÇÃO DE VEÍCULOS AUTOMOTORES E MOTOCICLETAS
46	COMÉRCIO POR ATACADO, EXCETO VEÍCULOS AUTOMOTORES E MOTOCICLETAS
47	COMÉRCIO VAREJISTA
49	TRANSPORTE TERRESTRE
50	TRANSPORTE AQUAVIÁRIO
51	TRANSPORTE AÉREO
52	ARMAZENAMENTO E ATIVIDADES AUXILIARES DOS TRANSPORTES
53	CORREIO E OUTRAS ATIVIDADES DE ENTREGA
55	ALOJAMENTO
56	ALIMENTAÇÃO
58	EDIÇÃO E EDIÇÃO INTEGRADA À IMPRESSÃO
59	ATIVIDADES CINEMATOGRÁFICAS, PRODUÇÃO DE VÍDEOS E DE PROGRAMAS DE TELEVISÃO; GRAVAÇÃO DE SOM E EDIÇÃO DE MÚSICA;
60	ATIVIDADES DE RÁDIO E DE TELEVISÃO
61	TELECOMUNICAÇÕES
62	ATIVIDADES DOS SERVIÇOS DE TECNOLOGIA DA INFORMAÇÃO
63	ATIVIDADES DE PRESTAÇÃO DE SERVIÇOS DE INFORMAÇÃO
64	ATIVIDADES DE SERVIÇOS FINANCEIROS
65	SEGUROS, RESSEGUROS, PREVIDÊNCIA COMPLEMENTAR E PLANOS DE SAÚDE
66	ATIVIDADES AUXILIARES DOS SERVIÇOS FINANCEIROS, SEGUROS, PREVIDÊNCIA COMPLEMENTAR E PLANOS DE SAÚDE
68	ATIVIDADES IMOBILIÁRIAS
69	ATIVIDADES JURÍDICAS, DE CONTABILIDADE E DE AUDITORIA
70	ATIVIDADES DE SEDES DE EMPRESAS E DE CONSULTORIA EM GESTÃO EMPRESARIAL
71	SERVIÇOS DE ARQUITETURA E ENGENHARIA; TESTES E ANÁLISES TÉCNICAS
72	PESQUISA E DESENVOLVIMENTO CIENTÍFICO
73	PUBLICIDADE E PESQUISA DE MERCADO
74	OUTRAS ATIVIDADES PROFISSIONAIS, CIENTÍFICAS E TÉCNICAS
75	ATIVIDADES VETERINÁRIAS
77	ALUGUÉIS NÃO-IMOBILIÁRIOS E GESTÃO DE ATIVOS INTANGÍVEIS

	NÃO-FINANCEIROS
78	SELEÇÃO, AGENCIAMENTO E LOCAÇÃO DE MÃO-DE-OBRA
79	AGÊNCIAS DE VIAGENS, OPERADORES TURÍSTICOS E SERVIÇOS DE RESERVAS
80	ATIVIDADES DE VIGILÂNCIA, SEGURANÇA E INVESTIGAÇÃO
81	SERVIÇOS PARA EDIFÍCIOS E ATIVIDADES PAISAGÍSTICAS
82	SERVIÇOS DE ESCRITÓRIO, DE APOIO ADMINISTRATIVO E OUTROS SERVIÇOS PRESTADOS ÀS EMPRESAS
84	ADMINISTRAÇÃO PÚBLICA, DEFESA E SEGURIDADE SOCIAL
85	EDUCAÇÃO
86	ATIVIDADES DE ATENÇÃO À SAÚDE HUMANA
87	ATIVIDADES DE ATENÇÃO À SAÚDE HUMANA INTEGRADAS COM ASSISTÊNCIA SOCIAL, PRESTADAS EM RESIDÊNCIAS COLETIVAS E PARTICULARES
88	SERVIÇOS DE ASSISTÊNCIA SOCIAL SEM ALOJAMENTO
90	ATIVIDADES ARTÍSTICAS, CRIATIVAS E DE ESPETÁCULOS
91	ATIVIDADES LIGADAS AO PATRIMÔNIO CULTURAL E AMBIENTAL
92	ATIVIDADES DE EXPLORAÇÃO DE JOGOS DE AZAR E APOSTAS
93	ATIVIDADES ESPORTIVAS E DE RECREAÇÃO E LAZER
94	ATIVIDADES DE ORGANIZAÇÕES ASSOCIATIVAS
95	REPARAÇÃO E MANUTENÇÃO DE EQUIPAMENTOS DE INFORMÁTICA E COMUNICAÇÃO E DE OBJETOS PESSOAIS E DOMÉSTICOS
96	OUTRAS ATIVIDADES DE SERVIÇOS PESSOAIS
97	SERVIÇOS DOMÉSTICOS
99	ORGANISMOS INTERNACIONAIS E OUTRAS INSTITUIÇÕES EXTRATERRITORIAIS

ANEXO III

Apresenta lista de Serviços de acordo com o art. 52 do CTM do Município de Goiânia.

1 – Serviços de informática e congêneres.
1.01 – Análise e desenvolvimento de sistemas.
1.02 – Programação.
1.03 – Processamento de dados e congêneres.
1.04 – Elaboração de programas de computadores, inclusive de jogos eletrônicos.
1.05 – Licenciamento ou cessão de direito de uso de programas de computação.
1.06 – Assessoria e consultoria em informática.
1.07 – Suporte técnico em informática, inclusive instalação, configuração e manutenção de programas de computação e bancos de dados.
1.08 – Planejamento, confecção, manutenção e atualização de páginas eletrônicas.
2 – Serviços de pesquisas e desenvolvimento de qualquer natureza.
2.01 – Serviços de pesquisas e desenvolvimento de qualquer natureza.
3 – Serviços prestados mediante locação, cessão de direito de uso e congêneres.
3.01 – Cessão de direito de uso de marcas e de sinais de propaganda.
3.02 – Exploração de salões de festas, centro de convenções, escritórios virtuais, stands, quadras esportivas, estádios, ginásios, auditórios, casas de espetáculos, parques de diversões, canchas e congêneres, para realização de eventos ou negócios de qualquer natureza.
3.03 – Locação, sublocação, arrendamento, direito de passagem ou permissão de uso, compartilhado ou não, de ferrovia, rodovia, postes, cabos, dutos e condutos de qualquer natureza.
3.04 – Cessão de andaimes, palcos, coberturas e outras estruturas de uso
4 – Serviços de saúde, assistência médica e congêneres.
4.01 – Medicina e biomedicina.
4.02 – Análises clínicas, patologia, eletricidade médica, radioterapia, quimioterapia, ultrasonografia, ressonância magnética, radiologia, tomografia e congêneres.
4.03 – Hospitais, clínicas, laboratórios, sanatórios, manicômios, casas de saúde, prontossocorros, ambulatórios e congêneres.
4.04 – Instrumentação cirúrgica.
4.05 – Acupuntura.
4.06 – Enfermagem, inclusive serviços auxiliares.
4.07 – Serviços farmacêuticos.
4.08 – Terapia ocupacional, fisioterapia e fonoaudiologia.
4.09 – Terapias de qualquer espécie destinadas ao tratamento físico, orgânico e mental.
4.10 – Nutrição.
4.11 – Obstetrícia.
4.12 – Odontologia.
4.13 – Ortóptica.
4.14 – Próteses sob encomenda.
4.15 – Psicanálise.
4.16 – Psicologia.
4.17 – Casas de repouso e de recuperação, creches, asilos e congêneres.
4.18 – Inseminação artificial, fertilização *in vitro* e congêneres.
4.19 – Bancos de sangue, leite, pele, olhos, óvulos, sêmen e congêneres.
4.20 – Coleta de sangue, leite, tecidos, sêmen, órgãos e materiais biológicos de qualquer espécie.
4.21 – Unidade de atendimento, assistência ou tratamento móvel e congêneres.
4.22 – Planos de medicina de grupo ou individual e convênios para prestação de assistência médica, hospitalar, odontológica e congêneres.
4.23 – Outros planos de saúde que se cumpram através de serviços de terceiros contratados, credenciados, cooperados ou apenas pagos pelo operador do plano mediante indicação do beneficiário.

5 – Serviços de medicina e assistência veterinária e congêneres.
5.01 – Medicina veterinária e zootecnia.
5.02 – Hospitais, clínicas, ambulatórios, prontos-socorros e congêneres, na área veterinária.
5.03 – Laboratórios de análise na área veterinária.
5.04 – Inseminação artificial, fertilização *in vitro* e congêneres.
5.05 – Bancos de sangue e de órgãos e congêneres.
5.06 – Coleta de sangue, leite, tecidos, sêmen, órgãos e materiais biológicos de qualquer espécie.
5.07 – Unidade de atendimento, assistência ou tratamento móvel e congêneres.
5.08 – Guarda, tratamento, amestramento, embelezamento, alojamento e congêneres.
5.09 – Planos de atendimento e assistência médico-veterinária.
6 – Serviços de cuidados pessoais, estética, atividades físicas e congêneres.
6.01 – Barbearia, cabeleireiros, manicuros, pedicuros e congêneres.
6.02 – Esteticistas, tratamento de pele, depilação e congêneres.
6.03 – Banhos, duchas, sauna, massagens e congêneres.
6.04 – Ginástica, dança, esportes, natação, artes marciais e demais atividades físicas.
6.05 – Centros de emagrecimento, spa e congêneres.
7 – Serviços relativos a engenharia, arquitetura, geologia, urbanismo, construção civil, manutenção, limpeza, meio ambiente, saneamento e congêneres.
7.01 – Engenharia, agronomia, agrimensura, arquitetura, geologia, urbanismo, paisagismo e congêneres.
7.02 – Execução, por administração, empreitada ou subempreitada, de obras de construção civil, hidráulica ou elétrica e de outras obras semelhantes, inclusive sondagem, perfuração de poços, escavação, drenagem e irrigação, terraplanagem, pavimentação, concretagem e a instalação e montagem de produtos, peças e equipamentos (exceto o fornecimento de mercadorias produzidas pelo prestador de serviços fora do local da prestação dos serviços, que fica sujeito ao ICMS).
7.03 – Elaboração de planos diretores, estudos de viabilidade, estudos organizacionais e outros, relacionados com obras e serviços de engenharia, elaboração de anteprojetos, projetos básicos e projetos executivos para trabalhos de engenharia.
7.04 – Demolição.
7.05 – Reparação, conservação e reforma de edifícios, estradas, pontes, portos e congêneres (exceto o fornecimento de mercadorias produzidas pelo prestador dos serviços, fora do local da prestação dos serviços, que fica sujeito ao ICMS).
7.06 – Colocação e instalação de tapetes, carpetes, assoalhos, cortinas, revestimentos de parede, vidros, divisórias, placas de gesso e congêneres, com material fornecido pelo tomador do serviço.
7.07 – Recuperação, raspagem, polimento e lustração de pisos e congêneres.
7.08 – Calafetação.
7.09 – Varrição, coleta, remoção, incineração, tratamento, reciclagem, separação e destinação final de lixo, rejeitos e outros resíduos quaisquer.
7.10 – Limpeza, manutenção e conservação de vias e logradouros públicos, imóveis, chaminés, piscinas, parques, jardins e congêneres.
7.11 – Decoração e jardinagem, inclusive corte e poda de árvores.
7.12 – Controle e tratamento de efluentes de qualquer natureza e de agentes físicos, químicos e biológicos.
7.13 – Dedetização, desinfecção, desinsetização, imunização, higienização, desratização, pulverização e congêneres.
7.14 – Florestamento, reflorestamento, semeadura, adubação e congêneres.
7.15 – Escoramento, contenção de encostas e serviços congêneres.
7.16 – Limpeza e dragagem de rios, portos, canais, baias, lagos, lagoas, represas, açudes e congêneres.
7.17 – Acompanhamento e fiscalização da execução de obras de engenharia, arquitetura e urbanismo.
7.18 – Aerofotogrametria (inclusive interpretação), cartografia, mapeamento, levantamentos topográficos, batimétricos, geográficos, geodésicos, geológicos, geofísicos e congêneres.
7.19 – Pesquisa, perfuração, cimentação, mergulho, perfilagem, concretação, testamunhagem, pescaria, estimulação e outros serviços relacionados com a exploração e exploração de petróleo, gás natural e de outros recursos minerais.

7.20 – Nucleação e bombardeamento de nuvens e congêneres.
8 – Serviços de educação, ensino, orientação pedagógica e educacional, instrução, treinamento e avaliação pessoal de qualquer grau ou natureza.
8.01 – Ensino regular pré-escolar, fundamental, médio e superior.
8.02 – Instrução, treinamento, orientação pedagógica e educacional, avaliação de conhecimentos de qualquer natureza.
9 – Serviços relativos a hospedagem, turismo, viagens e congêneres.
9.01 – Hospedagem de qualquer natureza em hotéis, *apart-service* condominiais, *flat*, *aparthotéis*, hotéis residência, *residence-service*, *suíte-service*, hotelaria marítima, motéis, pensões e congêneres; ocupação por temporada com fornecimento de serviço (o valor da alimentação e gorjeta, quando incluído no preço da diária, fica sujeito ao Imposto Sobre Serviços).
9.02 – Agenciamento, organização, promoção, intermediação e execução de programas de turismo, passeios, viagens, excursões, hospedagens e congêneres.
9.03 – Guias de turismo.
10 – Serviços de intermediação e congêneres.
10.01 – Agenciamento, corretagem ou intermediação de câmbio, de seguros, de cartões de crédito, de planos de saúde e de planos de previdência privada.
10.02 – Agenciamento, corretagem ou intermediação de títulos em geral, valores mobiliários e contratos quaisquer.
10.03 – Agenciamento, corretagem ou intermediação de direitos de propriedade industrial, artística ou literária.
10.04 – Agenciamento, corretagem ou intermediação de contratos de arrendamento mercantil (*leasing*), de franquia (*franchising*) e de faturização (*factoring*).
10.05 – Agenciamento, corretagem ou intermediação de bens móveis ou imóveis, não abrangidos em outros itens ou subitens, inclusive aqueles realizados no âmbito de Bolsas de Mercadorias e Futuros, por quaisquer meios.
10.06 – Agenciamento marítimo.
10.07 – Agenciamento de notícias.
10.08 – Agenciamento de publicidade e propaganda, inclusive o agenciamento de veiculação por quaisquer meios.
10.09 – Representação de qualquer natureza, inclusive comercial.
10.10 – Distribuição de bens de terceiros.
11 – Serviços de guarda, estacionamento, armazenamento, vigilância e congêneres.
11.01 – Guarda e estacionamento de veículos terrestres automotores, de aeronaves e de embarcações.
11.02 – Vigilância, segurança ou monitoramento de bens e pessoas.
11.03 – Escolta, inclusive de veículos e cargas.
11.04 – Armazenamento, depósito, carga, descarga, arrumação e guarda de bens de qualquer espécie.
12 – Serviços de diversões, lazer, entretenimento e congêneres.
12.01 – Espetáculos teatrais.
12.02 – Exibições cinematográficas.
12.03 – Espetáculos circenses.
12.04 – Programas de auditório.
12.05 – Parques de diversões, centros de lazer e congêneres.
12.06 – Boates, *taxi-dancing* e congêneres.
12.07 – *Shows*, *ballet*, danças, desfiles, bailes, óperas, concertos, recitais, festivais e congêneres.
12.08 – Feiras, exposições, congressos e congêneres.
12.09 – Bilhares, boliches e diversões eletrônicas ou não.
12.10 – Corridas e competições de animais.
12.11 – Competições esportivas ou de destreza física ou intelectual, com ou sem a participação do espectador.
12.12 – Execução de música.
12.13 – Produção, mediante ou sem encomenda prévia, de eventos, espetáculos, entrevistas, *shows*, *ballet*, danças, desfiles, bailes, teatros, óperas, concertos, recitais, festivais e congêneres.

12.14 – Fornecimento de música para ambientes fechados ou não, mediante transmissão por qualquer processo.
12.15 – Desfiles de blocos carnavalescos ou folclóricos, trios elétricos e congêneres.
12.16 – Exibição de filmes, entrevistas, musicais, espetáculos, *shows*, concertos, desfiles, óperas, competições esportivas, de destreza intelectual ou congêneres.
12.17 – Recreação e animação, inclusive em festas e eventos de qualquer natureza.
13 – Serviços relativos a fonografia, fotografia, cinematografia e reprografia.
13.01 – Fonografia ou gravação de sons, inclusive trucagem, dublagem, mixagem e congêneres.
13.02 – Fotografia e cinematografia, inclusive revelação, ampliação, cópia, reprodução, trucagem e congêneres.
13.03 – Reprografia, microfilmagem e digitalização.
13.04 – Composição gráfica, fotocomposição, clicheria, zincografia, litografia, fotolitografia.
14 – Serviços relativos a bens de terceiros.
14.01 – Lubrificação, limpeza, lustração, revisão, carga e recarga, conserto, restauração, blindagem, manutenção e conservação de máquinas, veículos, aparelhos, equipamentos, motores, elevadores ou de qualquer objeto (exceto peças e partes empregadas, que ficam sujeitas ao ICMS).
14.02 – Assistência técnica.
14.03 – Recondicionamento de motores (exceto peças e partes empregadas, que ficam sujeitas ao ICMS).
14.04 – Recauchutagem ou regeneração de pneus.
14.05 – Restauração, recondicionamento, acondicionamento, pintura, beneficiamento, lavagem, secagem, tingimento, galvanoplastia, anodização, corte, recorte, polimento, plastificação e congêneres, de objetos quaisquer.
14.06 – Instalação e montagem de aparelhos, máquinas e equipamentos, inclusive montagem industrial, prestados ao usuário final, exclusivamente com material por ele fornecido.
14.07 – Colocação de molduras e congêneres.
14.08 – Encadernação, gravação e douração de livros, revistas e congêneres.
14.09 – Alfaiataria e costura, quando o material for fornecido pelo usuário final, exceto aviamento.
14.10 – Tinturaria e lavanderia.
14.11 – Tapeçaria e reforma de estofamentos em geral.
14.12 – Funilaria e lanternagem.
14.13 – Carpintaria e serralheria.
15 – Serviços relacionados ao setor bancário ou financeiro, inclusive aqueles prestados por instituições financeiras autorizadas a funcionar pela União ou por quem de direito.
15.01 – Administração de fundos quaisquer, de consórcio, de cartão de crédito ou débito e congêneres, de carteira de clientes, de cheques pré-datados e congêneres.
15.02 – Abertura de contas em geral, inclusive conta-corrente, conta de investimentos e aplicação e caderneta de poupança, no País e no exterior, bem como a manutenção das referidas contas ativas e inativas.
15.03 – Locação e manutenção de cofres particulares, de terminais eletrônicos, de terminais de atendimento e de bens e equipamentos em geral.
15.04 – Fornecimento ou emissão de atestados em geral, inclusive atestado de idoneidade, atestado de capacidade financeira e congêneres.
15.05 – Cadastro, elaboração de ficha cadastral, renovação cadastral e congêneres, inclusão ou exclusão no Cadastro de Emitentes de Cheques sem Fundos – CCF ou em quaisquer outros bancos cadastrais.
15.06 – Emissão, reemissão e fornecimento de avisos, comprovantes e documentos em geral; abono de firmas; coleta e entrega de documentos, bens e valores; comunicação com outra agência ou com a administração central; licenciamento eletrônico de veículos; transferência de veículos; agenciamento fiduciário ou depositário; devolução de bens em custódia.
15.07 – Acesso, movimentação, atendimento e consulta a contas em geral, por qualquer meio ou processo, inclusive por telefone, *facsímile*, *internet* e telex, acesso a terminais de atendimento, inclusive vinte e quatro horas; acesso a outro banco e a rede compartilhada; fornecimento de saldo, extrato e demais informações relativas a contas em geral, por qualquer meio ou processo.

15.08 – Emissão, reemissão, alteração, cessão, substituição, cancelamento e registro de contrato de crédito; estudo, análise e avaliação de operações de crédito; missão, concessão, alteração ou contratação de aval, fiança, anuência e congêneres; serviços relativos a abertura de crédito, para quaisquer fins.
15.09 – Arrendamento mercantil (*leasing*) de quaisquer bens, inclusive cessão de direitos e obrigações, substituição de garantia, alteração, cancelamento e registro de contrato, e demais serviços relacionados ao arrendamento mercantil (*leasing*).
15.10 – Serviços relacionados a cobranças, recebimentos ou pagamentos em geral, de títulos quaisquer, de contas ou carnês, de câmbio, de tributos e por conta de terceiros, inclusive os efetuados por meio eletrônico, automático ou por máquinas de atendimento; fornecimento de posição de cobrança, recebimento ou pagamento; emissão de carnês, fichas de compensação, impressos e documentos em geral.
15.11 – Devolução de títulos, protesto de títulos, sustação de protesto, manutenção de títulos, reapresentação de títulos, e demais serviços a eles relacionados.
15.12 – Custódia em geral, inclusive de títulos e valores mobiliários.
15.13 – Serviços relacionados a operações de câmbio em geral, edição, alteração, prorrogação, cancelamento e baixa de contrato de câmbio; emissão de registro de exportação ou de crédito; cobrança ou depósito no exterior; emissão, fornecimento e cancelamento de cheques de viagem; fornecimento, transferência, cancelamento e demais serviços relativos a carta de crédito de importação, exportação e garantias recebidas; envio e recebimento de mensagens em geral relacionadas a operações de câmbio.
15.14 – Fornecimento, emissão, reemissão, renovação e manutenção de cartão magnético, cartão de crédito, cartão de débito, cartão salário e congêneres.
15.15 – Compensação de cheques e títulos quaisquer; serviços relacionados a depósito, inclusive depósito identificado, a saque de contas quaisquer, por qualquer meio ou processo, inclusive em terminais eletrônicos e de atendimento.
15.16 – Emissão, reemissão, liquidação, alteração, cancelamento e baixa de ordens de pagamento, ordens de crédito e similares, por qualquer meio ou processo; serviços relacionados à transferência de valores, dados, fundos, pagamentos e similares, inclusive entre contas em geral.
15.17 – Emissão, fornecimento, devolução, sustação, cancelamento e oposição de cheques quaisquer, avulso ou por talão.
15.18 – Serviços relacionados a crédito imobiliário, avaliação e vistoria de imóvel ou obra, análise técnica e jurídica, emissão, reemissão, alteração, transferência e renegociação de contrato, emissão e reemissão do termo de quitação e demais serviços relacionados a crédito imobiliário.
16 – Serviços de transporte de natureza municipal.
16.01 – Serviços de transporte de natureza municipal.
17 – Serviços de apoio técnico, administrativo, jurídico, contábil, comercial e congêneres.
17.01 – Assessoria ou consultoria de qualquer natureza, não contida em outros itens desta lista; análise, exame, pesquisa, coleta, compilação e fornecimento de dados e informações de qualquer natureza, inclusive cadastro e similares.
17.02 – Datilografia, digitação, estenografia, expediente, secretaria em geral, resposta audível, redação, edição, interpretação, revisão, tradução, apoio e infra-estrutura administrativa e congêneres.
17.03 – Planejamento, coordenação, programação ou organização técnica, financeira ou administrativa.
17.04 – Recrutamento, agenciamento, seleção e colocação de mão-de-obra.
17.05 – Fornecimento de mão-de-obra, mesmo em caráter temporário, inclusive de empregados ou trabalhadores, avulsos ou temporários, contratados pelo prestador de serviço.
17.06 – Propaganda e publicidade, inclusive promoção de vendas, planejamento de campanhas ou sistemas de publicidade, elaboração de desenhos, textos e demais materiais publicitários.
17.07 – Franquia (*franchising*).
17.08 – Perícias, laudos, exames técnicos e análises técnicas.
17.09 – Planejamento, organização e administração de feiras, exposições, congressos e congêneres.
17.10 – Organização de festas e recepções; bufê (exceto o fornecimento de alimentação e bebidas, que fica sujeito ao ICMS).

17.11 – Administração em geral, inclusive de bens e negócios de terceiros.
17.12 – Leilão e congêneres.
17.13 – Advocacia.
17.14 – Arbitragem de qualquer espécie, inclusive jurídica.
17.15 – Auditoria.
17.16 – Análise de Organização e Métodos.
17.17 – Atuária e cálculos técnicos de qualquer natureza.
17.18 – Contabilidade, inclusive serviços técnicos e auxiliares.
17.19 – Consultoria e assessoria econômica ou financeira.
17.20 – Estatística.
17.21 – Cobrança em geral.
17.22 – Assessoria, análise, avaliação, atendimento, consulta, cadastro, seleção, gerenciamento de informações, administração de contas a receber ou a pagar e em geral, relacionados a operações de faturização (*factoring*).
17.23 – Apresentação de palestras, conferências, seminários e congêneres.
18 – Serviços de regulação de sinistros vinculados a contratos de seguros; inspeção e avaliação de riscos para cobertura de contratos de seguros; prevenção e gerência de riscos seguráveis e congêneres.
18.01 – Serviços de regulação de sinistros vinculados a contratos de seguros; inspeção e avaliação de riscos para cobertura de contratos de seguros; prevenção e gerência de riscos seguráveis e congêneres.
19 – Serviços de distribuição e venda de bilhetes e demais produtos de loteria, bingos, cartões, pules ou cupons de apostas, sorteios, prêmios, inclusive os decorrentes de títulos de capitalização e congêneres.
19.01 – Serviços de distribuição e venda de bilhetes e demais produtos de loteria, bingos, cartões, pules ou cupons de apostas, sorteios, prêmios, inclusive os decorrentes de títulos de capitalização e congêneres.
20 – Serviços portuários, aeroportuários, ferroportuários, de terminais rodoviários, ferroviários e metroviários.
20.01 – Serviços portuários, ferroportuários, utilização de porto, movimentação de passageiros, reboque de embarcações, rebocador escoteiro, atração, desatração, serviços de praticagem, capatazia, armazenagem de qualquer natureza, serviços acessórios, movimentação de mercadorias, serviços de apoio marítimo, de movimentação ao largo, serviços de armadores, estiva, conferência, logística e congêneres.
20.02 – Serviços aeroportuários, utilização de aeroporto, movimentação de passageiros, armazenagem de qualquer natureza, capatazia, movimentação de aeronaves, serviços de apoio aeroportuários, serviços acessórios, movimentação de mercadorias, logística e congêneres.
20.03 – Serviços de terminais rodoviários, ferroviários, metroviários, movimentação de passageiros, mercadorias, inclusive suas operações, logística e congêneres.
21 – Serviços de registros públicos, cartórios e notariais.
21.01 – Serviços de registros públicos, cartorários e notariais.
22 – Serviços de exploração de rodovia.
22.01 – Serviços de exploração de rodovia mediante cobrança de preço ou pedágio dos usuários, envolvendo execução de serviços de conservação, manutenção, melhoramentos para adequação de capacidade e segurança de trânsito, operação, monitoração, assistência aos usuários e outros serviços definidos em contratos, atos de concessão ou de permissão ou em normas oficiais.
23 – Serviços de programação e comunicação visual, desenho industrial e congêneres.
23.01 – Serviços de programação e comunicação visual, desenho industrial e congêneres.
24 – Serviços de chaveiros, confecção de carimbos, placas, sinalização visual, *banners*, adesivos e congêneres.
24.01 – Serviços de chaveiros, confecção de carimbos, placas, sinalização visual, *banners*, adesivos e congêneres.
25 – Serviços funerários.
25.01 – Funerais, inclusive fornecimento de caixão, urna ou esquifes; aluguel de capela; transporte do corpo cadavérico; fornecimento de flores, coroas e outros paramentos; desembaraço de certidão de óbito; fornecimento de véu, essa e outros adornos; embalsamento, embelezamento, conservação ou restauração de cadáveres.

25.02 – Cremação de corpos e partes de corpos cadavéricos.
25.03 – Planos ou convênio funerários.
25.04 – Manutenção e conservação de jazigos e cemitérios.
26 – Serviços de coleta, remessa ou entrega de correspondências, documentos, objetos, bens ou valores, inclusive pelos correios e suas agências franqueadas; *courrier* e congêneres.
26.01 – Serviços de coleta, remessa ou entrega de correspondências, documentos, objetos, bens ou valores, inclusive pelos correios e suas agências franqueadas; *courrier* e congêneres.
27 – Serviços de assistência social.
27.01 – Serviços de assistência social.
28 – Serviços de avaliação de bens e serviços de qualquer natureza.
28.01 – Serviços de avaliação de bens e serviços de qualquer natureza.
29 – Serviços de biblioteconomia.
29.01 – Serviços de biblioteconomia.
30 – Serviços de biologia, biotecnologia e química.
30.01 – Serviços de biologia, biotecnologia e química.
31 – Serviços técnicos em edificações, eletrônica, eletrotécnica, mecânica, telecomunicações e congêneres.
31.01 – Serviços técnicos em edificações, eletrônica, eletrotécnica, mecânica, telecomunicações e congêneres.
32 – Serviços de desenhos técnicos.
32.01 – Serviços de desenhos técnicos.
33 – Serviços de desembaraço aduaneiro, comissários, despachantes e congêneres.
33.01 – Serviços de desembaraço aduaneiro, comissários, despachantes e congêneres.
34 – Serviços de investigações particulares, detetives e congêneres.
34.01 – Serviços de investigações particulares, detetives e congêneres.
35 – Serviços de reportagem, assessoria de imprensa, jornalismo e relações públicas.
35.01 – Serviços de reportagem, assessoria de imprensa, jornalismo e relações públicas.
36 – Serviços de meteorologia.
36.01 – Serviços de meteorologia.
37 – Serviços de artistas, atletas, modelos e manequins.
37.01 – Serviços de artistas, atletas, modelos e manequins.
38 – Serviços de museologia.
38.01 – Serviços de museologia.
39 – Serviços de ourivesaria e lapidação.
39.01 – Serviços de ourivesaria e lapidação (quando o material for fornecido pelo tomador do serviço).