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

# MINERAÇÃO DE DADOS APLICADA À CLASSIFICAÇÃO DOS CONTRIBUINTES DE ICMS DA SEFAZ-GO

**SANTIAGO MEIRELES ROCHA**

**2017**

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

# MINERAÇÃO DE DADOS APLICADA À CLASSIFICAÇÃO DOS CONTRIBUINTES DE ICMS DA SEFAZ-GO

SANTIAGO MEIRELES ROCHA

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.

Orientadora: Profa. Maria José Pereira Dantas, *Doutora.*

**Goiânia**
**Agosto, 2017**

Dados Internacionais de Catalogação da Publicação (CIP)
(Sistema de Bibliotecas PUC Goiás)

R672 Rocha, Santiago Meireles
Mineração de dados aplicados à classificação dos contribuintes da SEFAZ-GO
de ICMS[ manuscrito]/ Santiago Meireles Rocha.-- 2017.
76 f.; il. 30 cm

Texto em português com resumo em inglês
Dissertação (mestrado) - Pontifícia Universidade Católica de Goiás, Programa de Pós-Graduação Stricto Sensu em Engenharia de Produção e Sistemas, Goiânia, 2017

Inclui referências f. 71-76

1. Sonegação fiscal. 2. Imposto sobre circulação de mercadorias e serviços. 3. Mineração de dados (Computação). I.Dantas, Maria José Pereira. II.Pontifícia Universidade Católica de Goiás. III. Título.

CDU: 004.45:336.226

**MINERAÇÃO DE DADOS APLICADA A CLASSIFICAÇÃO DOS CONTRIBUINTES DO ICMS DA SEFAZ- GO**

**SANTIAGO MEIRELES ROCHA**

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 agosto de 2017.

Prof. Marcos Lajovic Carneiro, Dr.
Coordenador do Programa de Pós-Graduação em
Engenharia de Produção e Sistemas

Banca examinadora:

Profa. Maria José Pereira Dantas, Dra.
Orientadora

Prof. Thyago Carvalho Marques, Dr.

Prof. José Elmo de Menezes, Dr.

Prof. Marcos Lajovic Carneiro, Dr.

Goiânia - Goiás
Agosto 2017

*Dedico esse trabalho à Deus por ter me dado forças para eu alcançar o que tanto sonhei.*

## AGRADECIMENTOS

Aos professores do Curso de Mestrado em Engenharia de Produção e Sistemas, pela dedicação e ensinamentos que serviram para a construção do meu conhecimento.

À minha mãe, Maria Delcy Meireles Rocha, por ser a maior motivadora e nunca deixar de acreditar na realização desse projeto.

À minha esposa Daniela e às minhas filhas Bruna e Gabriela, tão queridas, que alegram e dão sentido à minha vida e me dão forças para seguir em frente.

Aos meus colegas do Curso de Mestrado em Engenharia de Produção e Sistemas, alegrias, angústias, dúvidas, incertezas e conquistas compartilhadas.

À Secretaria de Fazenda do Estado de Goiás, em especial aos Auditores Fiscais Adonidio Neto Vieira Junior e Olímpio de Oliveira Junior e ao Gestor de Tecnologia da Informação, Antônio Henrique Pereira pelo pronto atendimento em fornecer os dados que serviram de subsídio à esta pesquisa, e por dirimir dúvidas a respeito de Auditorias de ICMS.

Ao Prof. Dr. Sibélius Vieira Lelis que conduziu as orientações no início deste trabalho e se afastou por motivos de saúde.

À Profa. Dra. Maria José Pereira Dantas, minha orientadora, que soube, com maestria conduzir as orientações para que pudéssemos alcançar o objetivo.

*"Quem me dera ao menos uma vez, fazer com que o mundo saiba que Seu nome está em tudo e mesmo assim ninguém Lhe diz ao menos obrigado"*
(Renato Russo)

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 À CLASSIFICAÇÃO DOS CONTRIBUINTES DE ICMS DA SEFAZ-GO

Santiago Meireles Rocha
Agosto, 2017.

Orientadora: Profª. Maria José Pereira Dantas, Dra.

Com o aumento exponencial do volume de dados armazenados e o alto potencial de conhecimento oculto nesses dados que pode auxiliar nas estratégias e nas tomadas de decisão das organizações, muito vem se investido em tecnologia da informação e telecomunicação. A presente dissertação teve como objetivo aplicar o processo de Descoberta do Conhecimento em Base de Dados (DCBD) a fim de classificar os contribuintes de ICMS da SEFAZ-GO em Alto Sonegador e Baixo Sonegador, por meio da tarefa de mineração de dados Classificação Supervisionada, implementada pelo algoritmo J48, na plataforma computacional WEKA. Foram realizados 3 experimentos com uma amostra de dados de contribuintes de ICMS do setor atacadista do município de Goiânia-GO, com atributos selecionados a partir do Código do Tributário do Estado de Goiás. Durante os experimentos foram aplicados os algoritmos *AttributeSelection* e *Discretize*, para a redução de atributos e transformação das variáveis contínuas em discretas, respectivamente. Os índices estatísticos Matriz de Confusão e Coeficiente de *Kappa* foram utilizados como métricas de validação do modelo proposto. Após cada experimento, regras de classificação foram extraídas formando assim o modelo preditivo proposto de classificação. Obteve-se, no melhor cenário, uma taxa de classificação correta de 84% de acerto. A mineração de dados é uma realidade dentro de muitas organizações e pode ser uma forte aliada no cumprimento da, nada trivial, tarefa de descoberta de conhecimento nas bases de dados corporativas.

**Palavras chave:** Sonegação Fiscal, Árvore de Decisão, DCBD, WEKA

# ABSTRACT

With the exponential increase in the volume of data stored and the high potential for hidden knowledge in these data that can aid in the strategies and decision making of organizations, much has been invested in information technology and telecommunication. The purpose of this dissertation was to apply the Knowledge Discovery in Database (DCBD) process in order to classify the taxpayers of SEFAZ-GO ICMS in High Eviction and Low Eviction, through the task of data mining Supervised Classification, Implemented by the algorithm J48, on the WEKA computing platform. Three experiments were carried out with a sample of ICMS taxpayers data from the wholesale sector of the city of Goiânia-GO, with attributes selected from the Tax Code of the State of Goiás. During the experiments, the *AttributeSelection* and *Discretize* algorithms were applied. Reduction of attributes and transformation of the continuous variables into discrete ones, respectively. The statistical indices Confusion Matrix and *Kappa* Coefficient were used as validation metrics of the proposed model. After each experiment, classification rules were extracted, thus forming the proposed predictive model of classification. In the best scenario, a correct classification rate of 84% accuracy was obtained. Data mining is a reality within many organizations and can be a strong ally in fulfilling the, trivial, task of knowledge discovery in corporate databases.



**Keywords**: Tax evasion, Decision tree, KDD, WEKA

# SUMÁRIO

## LISTA DE FIGURAS

## LISTA DE QUADROS

## LISTA DE SIGLAS

DCBD – Descoberta do Conhecimento em Banco de Dados

EFD – Escrituração Fiscal Digital

GETI – Gerência de Tecnologia da Informação

GPL – GNU General Public License (Licença Pública Geral)

IBGE – Instituto Brasileiro de Geografia e Estatística

ICMS – Imposto sobre a Circulação de Mercadorias e Serviços

IPVA – Imposto sobre a Propriedade de Veículos Automotores

ITCD – Imposto sobre Transmissão Causa Mortis e Doação de Quaisquer Bens ou Direitos

KDD – Knowledge Discovery Data in Databases

PIB – Produto Interno Bruto

SEFAZ – Secretaria de Estado de Fazenda

SPED – Sistema Público de Escrituração Digital

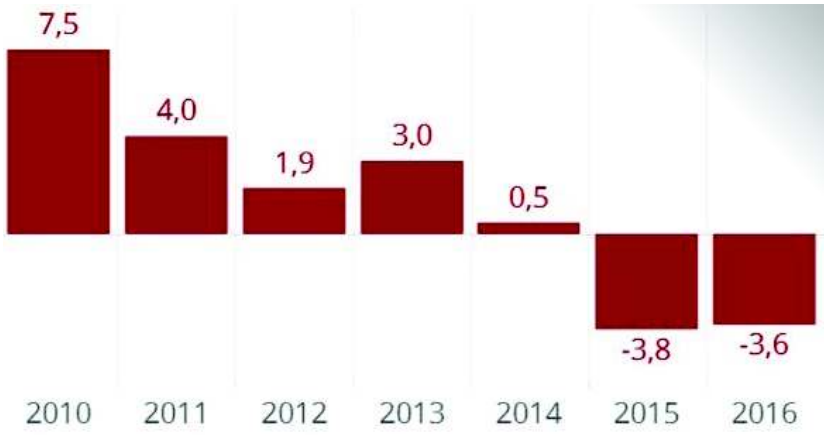
SRE – Superintendência da Receita Estadual

WEKA – Waikato Enviroment for Knowledge Analysis

# 1. INTRODUÇÃO

O Brasil vem atravessando por uma das piores crises econômico-financeira da sua história, gerando milhões de desempregados e uma forte queda no Produto Interna Bruto (PIB). Segundo dados do Instituto Brasileiro de Geografia e Estatística (IBGE), por meio da Pesquisa Nacional por Amostra de Domicílios Contínua, o desemprego alcançou os 12%, cerca de 12,3 milhões de desempregados, no último trimestre de 2016, fechando o ano com a taxa média e desocupação de 11,5%, a maior, desde 2012, início da série histórica do indicador (G1, 2017). Ainda segundo o IBGE, o PIB brasileiro caiu de em 2016, pelo segundo ano consecutivo e as projeções para 2017 foram revisadas para baixo com estimativa de crescimento de 1,0% para 0,5% (VALOR, 2017). A Figura 1 apresenta da evolução do PIB brasileiro nos últimos sete anos.

Figura 1 – Evolução do PIB, ano a ano em porcentagem (%)



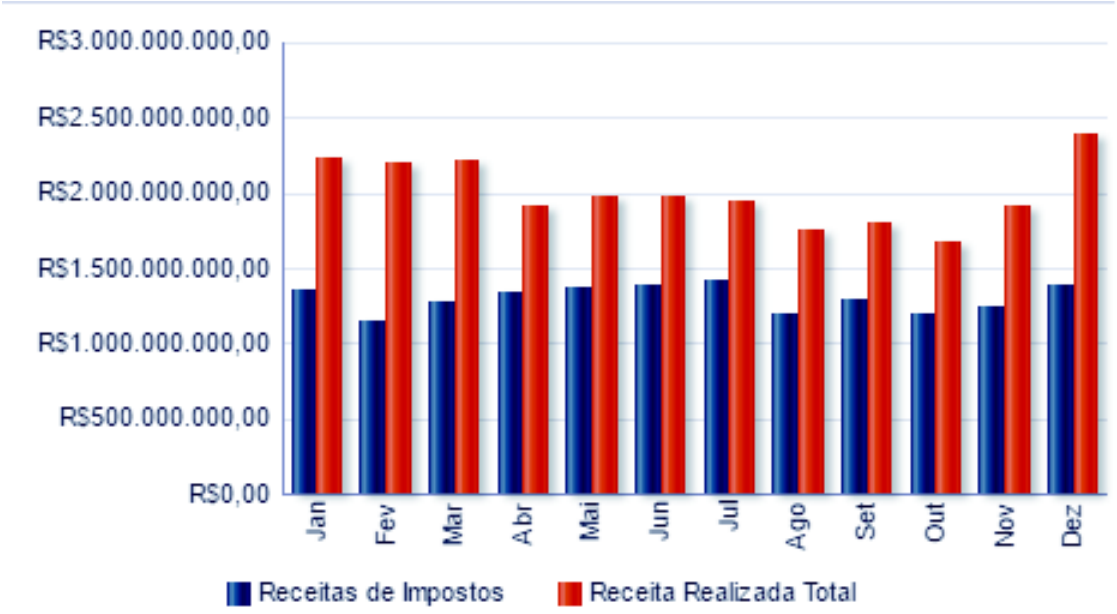

Fonte: Adaptado IBGE (2017)

Como consequência da crise, percebe-se uma queda no consumo da população e uma consequente queda da arrecadação de tributos e impostos dos órgãos governamentais de arrecadação.

Muitos são os desafios da administração pública em prover à sociedade serviços fundamentais para desenvolvimento humano, tais como saúde, educação e segurança, dentre outros, cuja manutenção depende diretamente das receitas obtidas pelos governos, advindos da arrecadação de tais tributos e impostos.

As receitas tributárias estaduais têm uma expressiva representação no valor total arrecadado pelos estados. Segundo o Portal da Transparência do Estado de Goiás, em 2016, 69,28% da arrecadação foram provenientes do recolhimento dos impostos, conforme mostra a Figura 2, que apresenta, mensalmente, a comparação entre a receita total e a receita arrecadada com impostos (TRANSPARÊNCIA, 2017), o que ratifica a importância dos tributos na arrecadação do Estado;

Figura 2 – Representação da Receita de Impostos na Receita Total – 2016.

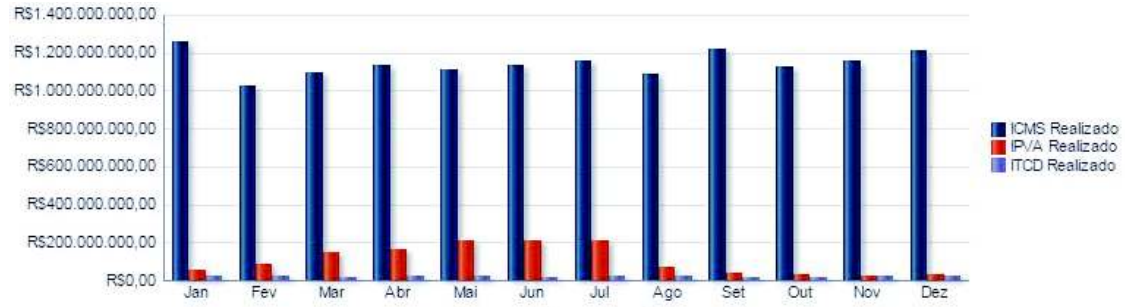



Fonte: Transparência (2017)

Atualmente, o principal imposto de competência dos Estados no Brasil é o Imposto sobre a Circulação de Mercadorias e Serviços (ICMS) (ANDRADE, 2009). Em Goiás, o ICMS representou, em 2016, 60,87% da arrecadação bruta e 87,86% da arrecadação dos

impostos geridos pela Secretaria da Fazenda do Estado de Goiás (SEFAZ-GO), tais como Imposto sobre a Propriedade de Veículos Automotores (IPVA), Imposto sobre Transmissão Causa Mortis e Doação de Quaisquer Bens ou Direitos (ITCD), além, do próprio ICMS (GOIÁS, 2017). A Figura 3 mostra a representatividade de cada imposto por mês, no ano de 2016.

Figura 3 – Representatividade por imposto, em 2016.




Fonte: Transparência (2017)

A Constituição Federal de 1988 estabeleceu a competência da legislação do ICMS aos Estados, respeitando-se algumas regras constitucionais. O recolhimento do imposto aos cofres públicos é obrigação tributária principal do sujeito passivo e cabe às administrações tributárias a fiscalização e apuração de tal recolhimento, a fim de combater a sonegação.

Pelilizanni (1990) define que a sonegação é todo ato que, de forma legal ou ilegal, leva ao não pagamento ou pagamento menor do imposto devido, seja essa atitude tomada consciente ou inconscientemente pelo contribuinte. A sonegação pode ser praticada de forma lícita, por meio da elisão ou de forma ilícita quando se dá a evasão. A elisão é praticada dentro da lei, com planejamentos fiscais, resultando no não recolhimento de impostos e cabendo aos Fiscos apenas ajustar a legislação a fim de impossibilitar tais práticas. Já a evasão se caracteriza pelo dolo em infringir a legislação a fim de evitar o pagamento do imposto

## 1.1. Problematização

Em tempos de crise, pode ser considerada natural uma perceptível queda na arrecadação de impostos, o que aumenta ainda mais a complexidade da gestão pública dos recursos financeiros, bem como os desafios do Fisco em responder aos questionamentos: a arrecadação caiu devido à crise? a crise está motivando o aumento da sonegação? ou a ocorrência de ambas questões concomitantemente?

O fato é que a estrutura de fiscalização dos Fiscos é infinitamente menor que o universo de contribuintes que precisa ser auditado, fazendo com que investimentos em recursos de Tecnologia da Informação e Telecomunicação sejam cada vez mais utilizados, estrategicamente, para equilibrar essa balança. Com a informatização das empresas e das estruturas organizacionais, um grande volume de dados é gerado diariamente, contendo, em suas minúcias, conhecimento que pode ser de grande valor estratégico.

Com projetos como a Lei de Responsabilidade Fiscal e o Sistema Público de Escrituração Digital (SPED), as empresas de vários seguimentos ficaram obrigadas a enviar seus livros fiscais para os Fiscos, em formato digital (GOIÁS, 2016). Isso fez com que as Secretarias de Fazenda Estaduais armazenassem um imenso volume de dados, com enorme potencial de exploração do conhecimento, mas sua utilização não acompanha a curva do crescimento do volume de dados. A mineração de dados é parte de um processo maior chamado Descoberta de Conhecimento em Base de Dados (DCBD), que busca por padrões nos milhares de dados armazenados e mantidos pelas organizações (FAYYAD *et al.*,1996). Em geral, as técnicas de mineração de dados executam as tarefas de classificação e agrupamento dos dados e descoberta de regras de associação entre os dados (STEINER et al, 2006). Segundo Balamurugan et al. (2008), a mineração de dados atrai muitos pesquisadores e analista de bancos de dados pela diversidade de aplicações

que possui, como na área de estratégias de mercado e finanças, auxiliando a extrair informações importantes dos bancos de dados das empresas.

De acordo com Digiampietri et al. (2008), a mineração de dados e a estatísticas estão sendo aplicadas para tentar identificar e detectar operações fraudulentas em transações de cartão de crédito, telecomunicações, detecção de terrorismo e detecção de crime financeiro.

## 1.2. Justificativa

As organizações vêm aprimorando seus recursos tecnológicos em busca da exploração e descoberta de conhecimento em suas imensas bases de dados que lhes possam trazer vantagens competitivas e respostas mais rápidas às decisões impostas pelos mercados.

O Setor Público não foge à essa tendência e investimentos em Tecnologia da Informação e Telecomunicação possuem dotações cativas em seus orçamentos anuais e planejamentos plurianuais.

Liu *et al*. (2012) afirmam que de posse das informações dos impostos pagos pelos contribuintes e utilizando um algoritmo de mineração de dados no sistema de administração tributária, é possível identificar padrões de comportamentos operacionais anormais dos contribuintes, podendo determinar, também, se há presença de fraude fiscal.

Muitos estudos têm sido desenvolvidos com o objetivo de prover assertividade na escolha dos contribuintes a serem fiscalizados, como em Piccirilli (2013), Andrade (2009) e Solveira (2001), obtendo resultados satisfatórios na seleção de contribuintes a serem fiscalizados. A partir dos contribuintes selecionados, também é valoroso para o planejamento das atividades de fiscalização, o agrupamento destes por risco potencial de

sonegação, conforme apresentado por Oliveira (2009), na Secretaria da Fazenda do Estado da Bahia (SEFAZ-BA), indicando prioridades na execução das auditorias, resultando na recuperação da evasão dos impostos e consequente aumento na arrecadação tributária.

## 1.3. Objetivos

De forma a alcançar as metas propostas, seguem o objetivo geral do trabalho bem como seus objetivos específicos.

### 1.3.1. Objetivo Geral

O objetivo geral da presente pesquisa foi propor um modelo preditivo de classificação de contribuintes de ICMS, a partir da análise dos dados mantidos pela Secretaria da Fazenda do Estado de Goiás (SEFAZ-GO), relativos à contribuição de ICMS das empresas ativas do setor atacadista, situadas no município de Goiânia-GO, a fim de obter indicações de sonegação, por meio de modelos de mineração de dados baseados na classificação supervisionada por árvore de decisão.

### 1.3.2. Objetivos Específicos

A partir do objetivo geral, decorrem os seguintes objetivos específicos:

– Identificar e selecionar atributos relevantes dos contribuintes para a classificação e agrupamento, com base no Código Tributário do Estado de Goiás.

– Coletar dados dos contribuintes referentes aos atributos selecionados por meio da Escrituração Fiscal Eletrônica, Nota Fiscal Eletrônica, Sistema de Autos de Infração e Sistema de Cadastro de Contribuintes;

– Realizar o pré-processamento, limpeza e transformação dos dados coletados;

– Aplicar aos dados coletados para treinamento, a técnica de mineração de dados Classificação Supervisionada, utilizando a ferramenta WEKA.

– Analisar a robustez do modelo proposto por meio de índices computacionais e estatísticos.

## 1.4. Estrutura do trabalho

O presente trabalho está estruturado em cinco capítulos.

O capítulo 2 é composto pela revisão bibliográfica de diferentes autores que envolvem conceitos e definições relacionados ao tema estudado.

O capítulo 3 aborda a metodologia utilizada para a elaboração do trabalho, que envolve a classificação da pesquisa, a descrição da empresa estudada e o detalhamento das etapas a serem seguidas.

O capítulo 4 apresenta um modelo de mineração de dados baseado na classificação supervisionada e agrupamento de contribuintes do ICMS da SEFAZ-GO.

O capítulo 5 abrangerá as conclusões do estudo, verificando o alcance dos objetivos propostos, a eficácia e eficiência do modelo, além de apresentar sugestões para trabalhos futuros.

O trabalho ainda contém as referências bibliográficas.

# 2. REFERENCIAL TEÓRICO

Este capítulo visa apresentar conceitos e definições que embasaram a realização da presente pesquisa.

## 2.1. Sonegação fiscal e ICMS

O pagamento dos impostos é importante para o desenvolvimento da humanidade e sua sonegação pode trazer danos à sociedade. Conforme Silva (2014), a sonegação fiscal é conceituada como as práticas ilegais que se enquadram em crimes contra a ordem tributária.

A administração pública estadual tem, por meio das auditorias fiscais, uma de suas principais atribuições, a recuperação de créditos tributários não declarados ou não pagos, gerando assim a sonegação fiscal. Segundo aspectos relevantes da Lei nº 8.137/90, sonegação fiscal é a ocultação dolosa, mediante fraude, astúcia ou habilidade, do recolhimento de tributo devido ao Poder Público.

O ICMS é o tributo que incide sobre a operação de circulação de mercadoria e sobre as prestações de serviço de transporte interestadual e intermunicipal e de comunicação, ainda que a operação e as prestações se iniciem no exterior (GOIÁS, 2016). A Constituição Federal de 1988 atribuiu aos Estados e ao Distrito Federal a competência para instituir o ICMS, que é cobrado nas várias etapas da cadeia produtiva, incidindo sobre o valor agregado nas operações sucessivas, até à venda ao consumidor final (MENDONÇA, 2000). A obrigação da apuração e lançamento do ICMS é do próprio sujeito passivo, ou seja, é a pessoa física ou jurídica a encarregada de recolher os impostos aos cofres públicos (VIEIRA, 2014).

## 2.2. SPED e EFD

O Sistema Público de Escrituração Fiscal (SPED) foi instituído por meio do Decreto 6.022/2007 e é o instrumento que consolida as atividades de recepção, validação, armazenamento e autenticação de livros e documentos que integram as escriturações contábil e fiscal dos contribuintes, inclusive imunes ou isentos, por meio de transmissão eletrônica de informações. Trata-se de uma solução tecnológica que normatiza e padroniza o formato digital no qual os livros e documentos que integram a escrituração contábil e fiscal devem ser enviados às receitas estaduais e federal (SPED, 2017).

EFD é um arquivo em formato digital que contém as escriturações de documentos fiscais e demais informações de interesse dos fiscos estaduais e Secretaria da Receita Federal do Brasil, além de registros de apuração de impostos referentes às operações e prestações praticadas pelo contribuinte (GOIÁS, 2016).

A EFD deve ser gerada, assinada digitalmente e transmitida via internet pelo contribuinte ao ambiente SPED, seguindo o leiaute definido no Ato Cotepe/ICMS nº 09/2008, e contempla os seguintes livros fiscais: Registro de Entradas, Registro de Saídas, Registro de Apuração do ICMS, Registro de Apuração do IPI, Registro do Inventário e do Documento de Controle de Crédito de ICMS do Ativo Imobilizado – CIAP (GOIÁS, 2016).

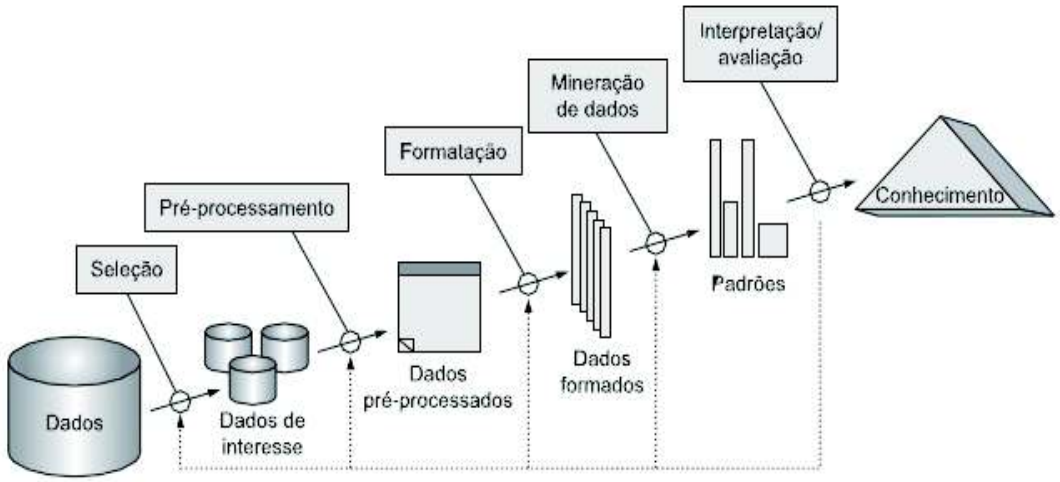
Desde 1º de janeiro de 2012, os contribuintes de ICMS do Estado de Goiás ficaram obrigados a aderirem ao projeto de EFD, conforme inciso II do Art. 4º-A da Instrução Normativa nº 1.020/10-GSF, de 27 de dezembro de 2010, excetuando as pequenas e as micros empresas optantes pelo Simples Nacional (GOIÁS, 2016).

## 2.3. Descoberta do conhecimento em base de dados - DCBD

O baixo custo de armazenagem, associados às disponibilidades de computadores de alta performance a um custo razoável e, principalmente, a necessidade de informações efetivas para a gestão empresarial vêm fazendo com que os bancos de dados das organizações acumulem um enorme volume de dados, ocultando conhecimentos de grande valor para as tomadas de decisão. Porém, esse grande volume de dados mantidos pelas empresas não significa que ela obtém e utiliza todas as informações intrínsecas nos dados, devido a incapacidade humana e inabilidade técnica na sua captação e interpretação. (CARDOSO e MACHADO, 2008). Goldschmidt (2011) acrescenta que tal coleta e interpretação das informações fica inviável sem o auxílio de ferramentas computacionais apropriadas. Conhecidas como ferramentas de mineração de dados, seu principal objetivo é encontrar conhecimento novo, que agregue valor e auxilie os gestores na tomada de decisão.

A Descoberta de conhecimento em base de dados é acrônimo do termo inglês *Knowledge Discovery in Databases* (KDD). Fayyad *et al.* (1996) definem KDD como um complexo processo de cinco etapas, interativo e iterativo, que busca a identificação de padrões compreensíveis, válidos, novos, potencialmente úteis, por meio da análise de um grande volume de dados. É a aplicação de técnicas que objetivam a transformação dos dados armazenados em conhecimento útil para decisões assertivas. A Figura 4 apresenta as etapas que compõem a descoberta do conhecimento em base de dados.

Figura 4– Etapas do processo DCBD




Fonte: Adaptado de Fayyad *et al.* (1996)

– Etapa 1 – Seleção: Nesta etapa, deve ser definido quais dados serão selecionados para a descoberta do conhecimento. Antes, porém, deve ser entendido junto às partes interessadas, quais são os objetivos a serem alcançados com a mineração de dados. Sem o completo entendimento dos objetivos, dados mal selecionados podem levar à resultados frustrantes;

– Etapa 2 – Pré-processamento: Nesta etapa, ocorre a limpeza dos dados coletados, descartando os dados incompletos, inconsistentes, fora do padrão, duplicados, entre outras situações que venham contaminar a massa de dados selecionada;

– Etapa 3 – Formatação: Nesta etapa, ocorre a transformação dos dados originais em um formato que é melhor utilizado nas etapas seguintes, porém sem perdas nas propriedades. São comuns nessa etapa atividades como redução da dimensão, normalização e categorização, transformação dos dados não estruturados em estruturados;

– Etapa 4 – Mineração de dados: É a principal etapa do DCBD, tanto que muitos autores tratam os termos como sinônimos. Diferentes técnicas podem ser utilizadas para

extrair o conhecimento e revelar os padrões, estruturas e tendências nos dados selecionados.

– Etapa 5 – Interpretação: Nesta etapa, os resultados alcançados são analisados, preferencialmente, com a participação das partes interessadas que conhecem o segmento de negócio analisado, podendo retornar a qualquer um dos passos anteriores para a equalização das técnicas;

### 2.4. Mineração de dados – MD

Conforme apresentado na Etapa 4 do item 2.3, a mineração de dados ou *Data Mining* é a etapa mais importante do processo DCBD e refere-se à aplicação de técnicas para a mineração do conhecimento oculto nos grandes volumes de dados mantidos pelas organizações. É a etapa que, de fato, analisa as informações em um grande volume de dados em busca de correlações e padrões relevantes para os usuários tomadores de decisão. Tais descobertas significativas podem propiciar em algum tipo de vantagem, normalmente econômica (WITTEN e FRANK, 2005).

Giudic (2013) define MD como o processo de seleção, exploração e modelagem de grandes volumes de dados a fim de descobrir padrões ou relações que, em primeira análise, são desconhecidos, com o objetivo de apresentar resultados úteis para os gestores da organização.

MD busca a construção de modelos computacionais a partir da descoberta de padrões e relacionamentos não triviais ocultos nas bases de dados analisadas, predição, promovendo a descoberta de anomalias, conhecimento e inúmeras oportunidades de aplicação, focada, para tal, em técnicas estatísticas e de inteligência artificial (RODRIGUES e AMARAL, 2012).

Em geral, as técnicas de mineração de dados executam as tarefas de classificação e agrupamento dos dados e descoberta de regras de associação entre os dados (STEINER *et al.*, 2006). Larose (2006) ainda destaca as tarefas de descrição, regressão e predição. O quadro 1 apresenta um resumo das principais tarefas de mineração de dados.

Quadro 1- Quadro-Resumo das principais tarefas de mineração de dados

| Tarefa | Características | Exemplos de Aplicação |
|---|---|---|
| **Descrição** | Utilizada para descrever padrões e tendências em dados. | Geralmente utilizada em conjunto com técnicas de análise exploratória de dados; |
| **Classificação** | Uma das tarefas mais comuns da MD, busca associar um conjunto de atributos à um atributo objetivo, categórico, chamado de atributo chave. | Determinar quando uma transação de cartão de crédito pode ser uma fraude; |
| **Estimativa ou Regressão** | Tarefa similar à classificação, porém, é utilizada quando o conjunto de atributos em questão é identificado por um valor numérico e não categórico. | Estimar a pressão ideal de um paciente baseando-se na idade, sexo e massa corporal; |
| **Predição** | Tarefa similar às tarefas de classificação e regressão, porém, esta visa prever o valor de um terminado atributo. | Predizer o valor de uma ação três meses adiante; |
| **Agrupamento** | Tarefa que não tem atribuição de classificação, regressão ou predição, mas sim a de agrupar as instâncias dos dados conforme os valores de seus atributos, não necessitando a definição de um atributo alvo, ou seja, as instâncias não são categorizadas, como na classificação. | Para auditoria, separando comportamentos suspeitos; |
| **Associação** | Busca identificar a relação entre os atributos, apresentando-se na forma SE ocorre A, ENTÃO ocorre B. | Identificar os usuários de planos que respondem bem a oferta de novos serviços; |

Fonte: Adaptado Camilo e Silva (2009)

Steiner *et al.* (2006) destacam que, dentre os métodos capazes de realizar o reconhecimento de padrões por meio da classificação, método utilizado nessa pesquisa, estão as populares Árvores de decisão, as Máquinas de Suporte de Vetores (*Support Vector Machines*, SVM), os Métodos Estatísticos, as Redes Neurais, os Algoritmos Genéticos e as Meta-Heurísticas. O Quadro 2 apresenta um resumo dos principais métodos de mineração de dados.

Quadro 2 - Quadro-Resumo dos principais métodos de mineração de dados

| Método | Características |
|---|---|
| **Árvores de Decisão** | Fluxograma top-down em forma de árvore onde cada nó (atributo) indica um teste a ser feito sobre um valor. Cada nó inferior está ligado e representa um possível valor do nó superior. As folhas indicam qual classe a instância pertence. A partir da estrutura da árvore, extraem-se as regras. |
| **SVM** | Técnica permite gerar modelos lineares e não-lineares, podem ser utilizadas para tarefas de classificação e predição. |
| **Classificação Bayesiana** | Técnica estatística baseada no teorema de Thomas Bayes que diz ser possível encontrar a probabilidade de um certo evento ocorrer, dada a probabilidade de um outro evento que já ocorreu: Probabilidade (B dado A) = Probabilidade(A e B)/Probabilidade(A). O algoritmo considera a inexistência de relação de dependência entre os atributos que, nem sempre isto é possível. |
| **Redes Neurais** | Com origem na psicologia e na neurobiologia, esta técnica simula o comportamento dos neurônios humanos. A rede possui um conjunto de entradas, às quais são aplicados pesos gerando saídas. Ao logo do processo de aprendizado, os pesos são ajustados a fim de aumentar a taxa de acerto de classificações corretas. |
| **Algoritmo Genético** | Seguindo a teoria da evolução onde o mais forte prevalece, o algoritmo, a partir de um estado inicial, passa por inúmeras iterações, simulando a seleção natural pelas melhores soluções. |

Fonte: Adaptado Camilo e Silva (2009)

## 2.5. Classificação e árvore de decisão

Classificação é uma das mais populares tarefas de mineração de dados, cujo propósito é identificar relacionamentos entre os atributos que caracterizam uma instância, chamados de atributos preditivos, e o alvo, chamado de atributo classe, a fim de identificar uma uniformidade e poder ser utilizado como predição para novas instâncias.

A classificação é implementada por um algoritmo de aprendizagem de máquina que, a partir de um conjunto de dados de treinamento previamente classificados, relaciona-os de acordo com suas características, produzindo um modelo que é validado a partir de um conjunto de dados teste.

Como exemplo de classificação, considere um gerente de *marketing* que possui um banco de dados contendo as seguintes informações sobre clientes: nome, idade, renda mensal, profissão e se o cliente comprou ou não produtos eletrônicos na hipotética loja. O Quadro 3 mostra exemplos de registros do banco de dados.

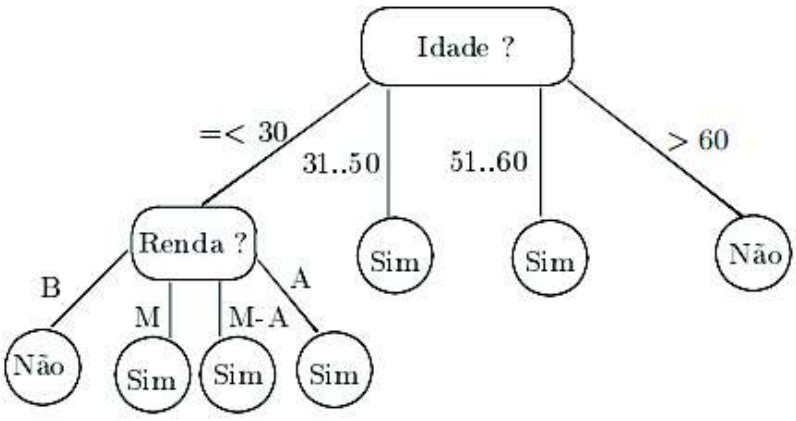
Quadro 3 - Exemplo de registros do banco de dados

| Nome | Idade | Renda | Profissão | ClasseProdEletr |
|---|---|---|---|---|
| Daniel | =< 30 | Média | Estudante | Sim |
| João | 31..50 | Média-Alta | Professor | Sim |
| Carlos | 31..50 | Média-Alta | Engenheiro | Sim |
| Maria | 31..50 | Baixa | Vendedora | Não |
| Paulo | =< 30 | Baixa | Porteiro | Não |
| Otávio | > 60 | Média-Alta | Aposentado | Não |

Fonte: Amo (2004)

A fim de evitar despesas desnecessárias enviando propaganda com lançamento de produtos eletrônicos para clientes que não possuem esse perfil, decidiu-se pela criação de um modelo de classificação de potenciais clientes que pudesse ser utilizado para classificar quais novos clientes receberão tal campanha. Então, divide-se o banco de dados em atributos preditivos, que são idade, renda e profissão, e atributo classe, que indica se o cliente compra ou não compra produtos eletrônicos. A Figura 5 apresenta uma possível árvore de decisão gerada a partir do banco de dados.

Figura 5– Exemplo de árvore de decisão.




Fonte: Amo (2004)

A partir de então, são geradas as regras de classificação de clientes, tornando possível a classificação de novos clientes de forma preditiva. Um exemplo de regra de classificação seria: clientes com idade inferior ou igual a 30 anos e a renda mensal é Alta, então Sim, potencial comprador de produtos eletrônicos.

Árvore de decisão é um modelo preditivo com uma estrutura hierárquica usado comumente como método de classificação. A principal vantagem de se utilizar a árvore de decisão é que a técnica fornece uma forma significativa de representar o conhecimento adquirido, por meio de regras de classificação SE-ENTÃO (LIN *et al.*, 2015).

A árvore de decisão é representada graficamente como uma árvore, com nós e ramos, mas no sentido invertido, sendo que cada nó contém um teste e seus resultados vão formando os demais ramos. Nas extremidades da árvore estão os nós folhas, que representam os valores de predição para a variável independente ou atributo classe. Quando a variável independente ou atributo classe é categórica, a árvore de decisão pode ser chamada de árvore de classificação, ou pode ser chamada de árvore de regressão, quando a variável independente ou atributo classe é numérica (MEIRA *et. al,* 2008).

De acordo com Chen *et al.* (2011), as árvores de decisão podem ser usadas para resolver problemas de classificação. Ela consiste em uma estrutura de árvore possuindo nós internos e externos. Cada nó terminal possui uma etiqueta que indica a classe prevista de um determinado vetor de características. Ela também é chamada de árvore de classificação ou modelo de previsão e, segundo Agrawal e Agrawal (2015), existem dois métodos para a construção, construção em declive e construção em poda de baixo para cima. Devido sua simplicidade e aplicabilidade em diferentes áreas de interesses, a árvore de decisão tornou-se um dos métodos mais utilizados. Normalmente, os algoritmos

utilizados na construção das árvores se diferem pelas estratégias, particionamento de nós e poda da árvore (GONZÁLEZ E VELÁSQUEZ, 2013).

### 2.6. Algoritmos de árvores de decisão

*Hunt* foi um dos primeiros algoritmos de indução de árvore de decisão e serviu como referência para o surgimento de algoritmos relacionados (TAN et al., 2009). *Classification and Regression Trees* (CART), Chi-Squared *Automatic Induction* (CHAID), Quest, SLIQ, SPRINT, *Induction Decision Tree* (ID-3), C4.5 e J48 são exemplos desses algoritmos.

Considerado por muitos pesquisadores como o criador das árvores de decisão, o professor da Universidade de Sydney – Austrália, Ross Quinlan, desenvolveu, em meados da década de 70, o algoritmo ID3, que tinha como requisito básico para a criação das árvores de decisão, a discretização dos atributos preditivos, ou seja, a redução do domínio valores que um atributo pode assumir à um conjunto discreto de dados (QUINLAN, 1993). Quinlan ainda propôs os algoritmos ID4, ID6, C4.5 e See 5 (LEMES*, et al*., 2005), destacando a capacidade do algoritmo C4.5 trabalhar com atributos nominais, ordinais, numéricos e ausentes. O algoritmo ainda descarta os atributos ou ramos da árvore, que não agregam valor ao processo decisório.

Uma das decisões que o algoritmo tem de realizar é escolha dos atributos preditivos para os nós das árvores. Existem diferentes tipos de critérios de seleção, sendo este o que pode diferenciar cada um dos algoritmos de indução por árvores de decisão. A maioria dos algoritmos de que implementam classificação por árvore de decisão, subdivide a árvore com base em um único atributo (BASGALUPP, 2010). O atributo mais importante é apresentado na árvore de decisão como o primeiro nó e os atributos menos relevantes

são apresentados nos nós subsequentes (Lemos *et al.,* 2005). As árvores são divididas segundo os critérios impureza, distância e dependência. Muitos algoritmos de indução dividem um nó em nós inferiores considerando o menor grau de impureza dos atributos. Tem-se uma impureza nula quando todos os exemplos de uma amostra pertencerem à mesma classe e tem-se uma impureza máxima quando a quantidade de exemplos da amostra for a mesma para cada uma das classes possíveis.

O algoritmo ID3 implementa a medida baseada em impureza Ganho de Informação, que utiliza a entropia para determinar o quão relevante é uma condição de teste. Isso se dá com a comparação da entropia do pai, antes da divisão da árvore, com o grau de entropia dos nós filhos, após a divisão, sendo escolhido como condição teste na árvore, o atributo que gerar a maior diferença. O cálculo do Ganho de Informação é dado pela equação 1.

$$ganho = entropia(pai) - \sum_{j=1}^{n} \frac{N(v_j)}{N} entropia(v_j) \quad (1)$$

n = número de valores do atributo;

N = número total de exemplos do nó pai

N($v_\gamma$) = número de exemplos associados ao nó filho $v_\gamma$

O cálculo da entropia é dado pela equação 2.

$$entropia(t) = -\sum_{i=1}^{k} p(i|t) log_2 p(i|t) \quad (2)$$

p($i|t$) = fração de exemplos pertencentes à classe $i$, no nó $t$

$k$ = número de classes0

O problema de se usar o ganho de informação é que o método prioriza os atributos com de maior domínio, ou seja, com o maior número de valores possível. Para resolver tal problema, Quinlan (1993) propôs o *Gain Ratio* que é a utilização do ganho da informação ponderado, calculado pela equação 3.

$$GainRatio(t) = \frac{ganho}{entropia(t)} \tag{3}$$

O algoritmo C4.5 utiliza a medida Gain Ratio para escolher o atributo que melhor divide as instâncias, gerando árvores mais precisa e menos complexas. A Figura 6 apresenta o pseudo-código do algoritmo C4.5.

Figura 6 - Pseudo-código do algoritmo C4.5

```
Algoritmo C4.5
- repetir várias vezes (aproximadamente 10)
        CONSTRUIR
        Escolher conjunto de trabalho do conjunto de treinamento
        REPETIR
                formar árvore para conjunto de trabalho
                        SE critério de parada satisfeito
                                escolher melhor classe
                        SENÃO
                                escolher melhor teste de atributo
                                dividir conjunto de treinamento em concordância
                                formar árvore nos sub-conjuntos
                testar no resto do conjunto de treinamento
                adicionar itens mal classificados ao conjunto de treinamento
        ATÉ não haver melhorias
        PODAR
        ENQUANTO a árvore de decisão contiver sub-árvores complexas e com pouco benefício
                Substituir sub-árvores por folhas
- selecionar a árvore podada mais promissora
```

Fonte: Souza (2011)

Segundo Balamurugan et al. (2008), o algoritmo J48 implementa uma árvore de decisão que cria uma árvore binária. O algoritmo J48 é a versão escrita em Java, implementada na ferramenta computacional WEKA, do algoritmo C4.5, versão 8. Os

algoritmos C4.5, C5.0 e J48 estão entre os mais populares e poderosos classificadores de árvores de decisão (AL-RADAIDEH *et al.*, 2011).

O algoritmo CART cuja implementação no WEKA tem o nome de SimpleCart, proposto por Breiman em 1984, implementa uma técnica não paramétrica capaz de induzir árvores de classificação, quando o atributo classe é categórico, e árvores de regressão, quando o atributo classe é contínuo. As árvores geradas por esse algoritmo são sempre binárias. Segundo seus criadores, o algoritmo CART possui uma eficiente técnica de poda, o que possibilita a produção de árvores mais simples, precisas e com alta capacidade de generalização (BASGALUPP, 2010).

O algoritmo NBTree é um modelo de indução hibrido que implementa classificação por árvore de decisão e *naive Bayes*. Os nós divisores utilizam condições baseadas em um único atributo, como nas árvores de decisão, mas os nós folhas possuem classificadores *Naive Bayes.* O algoritmo ADTree (*Alternating Decision Tree*) proposto por Yoav Freund e Llew Mason, é uma generalização de árvores de decisão e *decision stump*. O algoritmo é restrito para problemas de duas classes. O algoritmo LMT adapta a ideia de regressão linear de predição de atributos contínuos para a predição de atributos categóricos por meio da regressão logística, tendo como vantagem a possibilidade de se obter a estimativa da probabilidade da classe, ao invés de apenas uma classe. BFTree é um algoritmo baseado na heurística *best-first*, que implementa dois métodos de poda, a pré poda, *best-first-based pre-pruning* e pós poda, *best-first-based post-pruning.*

### 2.7. Métricas de validação

É fundamental que os modelos de classificação baseados em árvores de decisão sejam avaliados e validados. Para a construção de uma árvore de decisão são utilizados

os dados de aprendizagem, indução, onde o algoritmo vai construir a árvore e, em seguida, são utilizados dados de testes, avaliação, onde a capacidade de predição da mesma será avaliada. *Overfitting* é a capacidade do algoritmo super aprender com os dados de testes gerando modelos de baixa generalização, o que pode ser evitado com a escolha da amostra de dados e configuração adequada da ferramenta de mineração (BASGALUPP, 2010).

A indução e avaliação das árvores de decisão podem utilizar os métodos de re-substituição e reamostragem. O primeiro método utiliza o mesmo conjunto de dados utilizado para treinamento para teste, implicando em resultados otimistas, já que o modelo não garante uma boa generalização para novos exemplares. Já o método de reamostragem, parte do princípio de se utilizar um conjunto de dados para aprendizado e outro para testes. Para amostras grandes, pode-se utilizar método *hold-out* que utiliza 2/3 da amostra para treinamento e 1/3 para testes. Para amostras não grandes o bastante, são sugeridos os métodos de reamostragem, tais como *random subampling* e validação cruzada com k partições. (BASGALUPP, 2010).

O método validação cruzada com k partições consiste na divisão aleatória da amostra em k - 1 grupos, que são executados pelo algoritmo para treinamento e 1 grupo é utilizado como base para o teste. A execução repete-se por k vezes, garantindo que cada grupo seja utilizado uma vez como base para o teste. Por fim, a correção total é calculada pela média dos resultados obtidos em cada execução, chegando à uma estimativa de qualidade do modelo de conhecimento gerado, permitindo, assim, as análises estatísticas (SANTOS; et. al, 2009). Segundo Tan *et al.* (2005), um bom estimador para validação cruzada é utilizar 10 partições, ou seja, k=10.

A Matriz de Confusão representa a qualidade do modelo comparando a classificação da instância pelo modelo com sua classificação real, explicitando da base de

teste os verdadeiros-positivos (VP), verdadeiros-negativos (VN), falsos-positivos (FP) e falsos-negativos (FN). A Tabela 1 apresenta um exemplo de matriz de confusão para um exemplo com duas classes.

Tabela 1 – Exemplo matriz de confusão para um exemplo com duas classes

| | Classe predita | |
|---|---|---|
| Classe real | Positiva | Negativa |
| Positiva | VP | FN |
| Negativa | FP | VN |

Fonte: (BASGALUPP, 2010).

- Verdadeiros-Positivos: instâncias classificadas corretamente, pertencentes à classe Positiva;
- Falsos-Positivos: instâncias classificadas erradamente como Positiva, mas que pertencem à classe Negativa;
- Verdadeiros-Negativos: instâncias classificadas corretamente, pertencentes à classe Negativa;
- Falsos-Negativos: instâncias classificadas erradamente como Negativa, mas que pertencem à classe Positiva;

A partir da matriz de confusão, uma das variáveis que podem ser quantificadas é o Coeficiente de *Kappa*, que, segundo (PERROCA e GAIDZINSKI, 2003), retrata o grau de concordância dos dados, gerando, assim, um aspecto de confiabilidade e precisão dos dados classificados. O resultado obtido por esse coeficiente varia entre 0 e 1, e quanto mais próximo de 1, melhor a qualidade dos dados classificados.

Segundo Thompson (2011), o Coeficiente de *Kappa* é calculado em três etapas a seguir: (i) calcula-se o índice que representa a concordância esperada pelo acaso; (ii) calcula-se a concordância observada; (iii) calcula-se a estatística pela divisão da diferença

entre a concordância esperada e observada 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.

A Tabela 2 apresenta o índice utilizado pelo Coeficiente de *Kappa* que representa classificação da magnitude com intervalo de confiança de 95%.

Tabela 2 – Índices do Coeficiente de *Kappa*

| Valor *Kappa* | Concordância |
|---|---|
| <0 | Inexistente |
| 0 | Pobre |
| 0 a 0,20 | Ligeira |
| 0,21 a 0,40 | Considerável |
| 0,41 a 0,60 | Moderada |
| 0,61 a 0,80 | Substancial |
| 0,81 a 1 | Excelente |

Fonte: Cavalcante (2014)

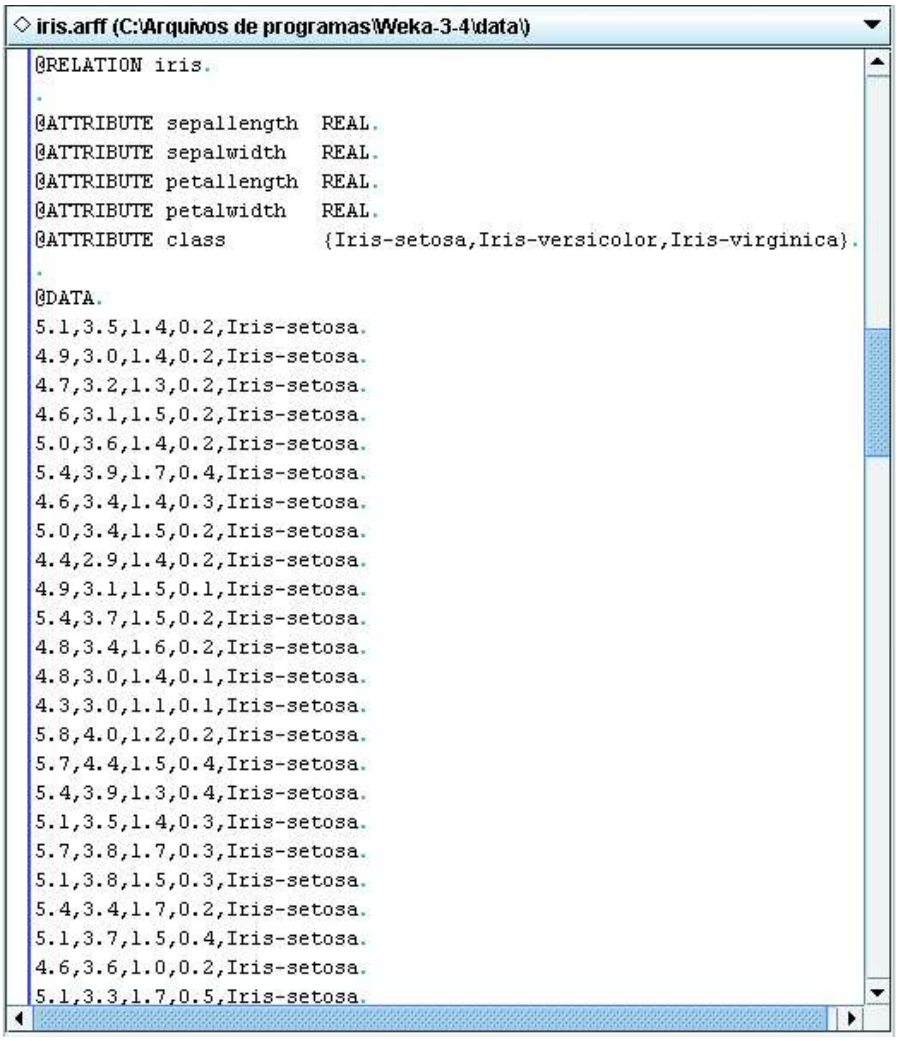
## 2.8. Waikato Enviroment for Knowledge Analysis - WEKA

WEKA é uma plataforma formada por um conjunto de implementações de algoritmos de diversas técnicas de Mineração de Dados. O WEKA é um software livre, ou seja, está sob domínio da licença GPL - *General Public License* (Licença Pública Geral) e está disponível em http://www.cs.waikato.ac.nz/ml/weka (WEKA, 2016).

A ferramenta WEKA, além dos métodos de classificação já citados no item 2.4, possui ainda implementados Regras de Aprendizagem, os algoritmos *Naive Bayes*, Tabelas de decisão, Regressão local de pesos, Aprendizado baseado em instância, Regressão lógica, *Perceptron*, *Perceptron* multicamada e Comitê de *perceptrons*. Ressalta-se que a ferramenta conta ainda com métodos para Predição Numérica, Agrupamento e Associação (DAMACENO, 2016).

O arquivo para a carga dos dados no software WEKA deve estar no Formato de Arquivo Atributo-Relação (.arff), que contém duas sessões, uma para o cabeçalho e outra para os dados. O cabeçalho contém um nome para a base de dados definido imediatamente à anotação @RELATION, uma lista de variáveis e seus respectivos tipos de dados definidos com a anotação @ATTRIBUTE. A anotação @DATE inicia a segunda sessão que se refere aos dados que serão minerados. O WEKA suporta 4 tipos de dados: nominal, numérico, *string* (valores de texto arbitrário) e data (VIEIRA, 2014). A Figura 7 apresenta o arquivo iris.arff, disponível como exemplo da plataforma WEKA.

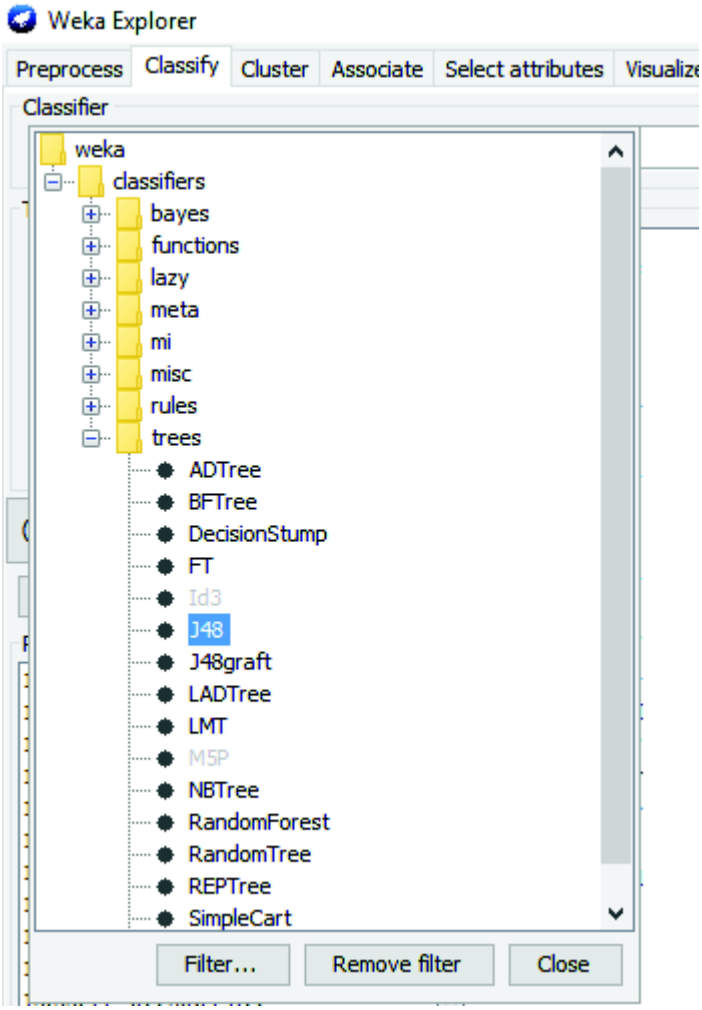
Figura 7 – Arquivo iris.arff disponivel como exemplo na plataforma WEKA.




Fonte: WEKA (2016)

A ferramenta WEKA implementa vários algoritmos classificadores para árvore de decisão, além do J48. A Figura 8 mostra a lista de algoritmos disponíveis pelo *software*.

Figura 8 - Algoritmos de classificação por árvore de decisão implementados no WEKA




Fonte: Autor (2017)

## 2.9. Trabalhos relacionados

Muitos trabalhados em mineração de dados têm sido realizados em busca do conhecimento oculto nos grandes volumes de dados mantidos pelas organizações, nas mais diversas áreas do conhecimento, utilizando as várias técnicas e métodos, tanto no Brasil, quanto no exterior.

Meira *et al.* (2008) desenvolveram uma árvore de decisão com o objetivo de auxiliar na compreensão de manifestações epidêmicas da ferrugem do cafeeiro. A árvore de decisão foi treinada com 364 exemplos preparados a partir de dados coletados em lavouras de café em produção, classificando corretamente 78% do conjunto de treinamento e a sua acurácia estimada em 73% para a classificação de novos exemplos.

Lima *et al.* (2010) utilizaram a mineração de dados e o método árvore de decisão na exploração do banco de dados de uma empresa incubadora de ovos, do setor de Avicultura. O método árvore de decisão foi aplicado para a geração de regras de classificação que identificaram padrões nas aves fêmeas indesejáveis pela empresa, dando suporte às tomadas de decisões e redução de desperdícios.

A literatura apresenta também muitos trabalhos em mineração de dados aplicados no suporte à gestão fazendária, combate à sonegação e evasão fiscal. Souza (2002) aplicou algoritmo de mapas auto-organizáveis das redes neurais artificiais, para classificar potenciais contribuintes sonegadores de ICMS, na Sefaz-GO, a partir da análise do banco de dados da extinta Declaração Periódica de Informações – DPI, enviada mensalmente pelos contribuintes e, atualmente, substituída pela Escrituração Fiscal Digital – EFD.

Andrade (2009) utilizou algoritmos das redes neurais artificiais para agrupamento, seleção de atributos e classificação de contribuinte potenciais sonegadores de ICMS, na Sefaz-BA. Levergger (2013) aplicou o método árvore de decisão com o objetivo de classificar os contribuintes de ISS nas categorias regular e irregular, a partir da análise do banco de dados da Secretaria de Finanças do município de Goiânia, com um índice de acertos de 92,03%.

González e Velásquez (2013) apresentam em seu trabalho, a aplicação e comparação dos métodos de redes neurais artificiais, mapas auto-organizáveis, gas e multilayer perceptron, além das árvores de decisão, na caracterização e detecção de fraudes de micro e pequena empresas chilenas que fazem uso de notas fiscais falsas para forjar aquisições, aumentando seus créditos fiscais e reduzindo assim os impostos a serem pagos.

Além desses, o Quadro 4 apresenta um resumo dos trabalhos correlatos utilizados como referência para a presente pesquisa.

Quadro 4 - Quadro-Resumo dos trabalhos correlatos

| **Autor/Ano** | **Metodologia** | **Objetivo** | **Resultado** |
|---|---|---|---|
| Habibi *et al.* (2015) | Usou dados de indivíduas que passaram por uma triagem e aplicou a Weka aplicando a Arvore de decisão e o algoritmo "J48" | Examinar um modelo preditivo utilizando funcionalidades relacionadas com a diabetes tipo 2 fatores de risco | A curva ROC indicou alta capacidade do modelo, especialmente na identificação da saúde pessoas |
| Goumagias *et al.* (2012) | Utiliza um modelo dinâmico de suporte à decisão baseado em Markov o qual captura as características do sistema de imposto grego | Descrever um modelo de suporte à decisão que incorpora os principais recursos do sistema tributária grego | O apoio à decisão escolhido para explorar a problemática da evasão fiscal sugeriu mudanças na forma de cobranças dos impostos |
| Wu *et al.* (2012) | Filtrar possíveis relatórios de impostos como medida paramétrica para resultados da mineração de dados de associação de DBMiner que foi utilizado separadamente em Data Cube 1 e Data Cube 2 para obter regras de associação. | Aplicar uma técnica de mineração de dados para aumentar o desempenho na detecção de evasão de impostos. | A técnica de mineração de dados proposta reduziu as perdas decorrentes da evasão fiscal |
| Liu *et al.* (2012) | Com uma análise de agrupamento aplicou um algoritmo de mineração de dados baseado em cluster para descobrir impostos com dados anormais | Resolver problemas das análises de impostos ficais por meio da mineração de dados. | Provou a eficácia do algoritmo |
| Amini *et al.* (2011) | Comparação dos algoritmos e de especificidade com base no critério C4.5 e KNN na mineração de dados WEKA | Prever a incidência de AVC | O algoritmo da árvore de decisão C4.5 e o k-vizinho mais próximo, puderam ser usados para prever o AVC em grupos de alto risco. |
| Chen *et al.* (2011) | Usa várias aplicações de redes neurais, incluindo o Multi-*Layer Perceptrons* | Desenvolver um modelo automático para criação e revisão dos | A ferramenta automatizada mostrou ser viável para detectar |

| | | | |
|---|---|---|---|
| | (MLPs), *Learning Vector* Quantização (LVQ), árvore de decisão e métodos da Rede Neural Hiper-Retangular (HRCNN) | relatórios de impostos (fiscais) do norte de Taiwan. | relatórios fiscais com erros. Independente da rede neural utilizada a taxa de reconhecimento foi satisfatória. |
| Digiampietri *et al.* (2008) | Utilizou um sistema chamado CARANCHO para destacar operações aduaneiras suspeitas | Criar ferramenta que auxilia funcionários aduaneiros a identificar sonegações de impostos | Apresentou algumas informações de inteligência artificial usadas na detecção de fraude aduaneira no Brasil |
| Balamurugan *et al.* 2008 | Classificou os email com o uso, por exemplo de rede neural, arvore de decisão. Usou Weka com base em diferentes tamanhos de dados de emails suspeitos que são detectados. | Detectar e-mails de fraude | Mostrou que o classificador ID3, pela aplicação de uma árvore binária, dará uma taxa de detecção satisfatória |

Fonte: Autor, 2017

# 3. METODOLOGIA

Para se ter acesso aos referidos dados, o pesquisador formalizou um pedido junto Superintendência da Receita Estadual (SRE), pertencente à SEFAZ-GO, por meio do processo 201600004019252, firmando um acordo que o modelo produzido pela presente pesquisa ficará disponível para sua plena utilização pela instituição.

A pesquisa limitou-se em analisar os dados das médias e grandes empresas do segmento atacadista, do município de Goiânia, auditadas via auditoria de ICMS, nos exercícios de 2013 a 2016, e apresentaram irregularidades, gerando, assim, a lavratura de um ou mais autos de infração.

## 3.1.Classificação da Pesquisa

A presente pesquisa está classificada quanto à sua natureza, quanto os objetivos, quanto à abordagem do problema e quanto aos procedimentos técnica, conforme apresentado a seguir.

Quanto aos objetivos, Jung (2003) afirma que uma pesquisar pode ser classificada como básica, que consiste na aquisição do conhecimento sem finalidades práticas ou imediatas, ou a pesquisa pode ser classificada como aplicada, que consiste na aplicação do conhecimento da pesquisa básica e da tecnologia para se obter aplicações práticas. A presente pesquisa é classificada como aplicada, pois, se utiliza-se de técnicas e ferramentas disponíveis na literatura e no mercado, a fim de propor um modelo preditivo de mineração de dados.

Jung (2003) define que, quanto aos objetivos, a pesquisa pode ser classificada em exploratória, descritiva ou explicativa. A pesquisa descritiva, tem por finalidade observar, registrar e analisar os fenômenos sem entrar no mérito do seu conteúdo, não vendo assim

interferência do pesquisador. A pesquisa explicativa tem como foco a classificação, análise e interpretação dos fatores que determinam ou que contribuem para a ocorrência dos fatos. A presente pesquisa está classificada como exploratória, que, segundo Lakatos e Marconi (2001), trata-se de uma investigação de pesquisa empírica cujo objetivo é a formulação de questões, com a finalidades de desenvolver hipóteses, aumentar a familiaridade do pesquisador o fato ou fenômeno, para a realização de uma pesquisa futura mais precisa ou modificar e clarear conceitos.

Quanto à sua abordagem, uma pesquisa pode ser classificada como qualitativa, quantitativa ou ambas. Segundo Gerhardt e Silveira (2009), a pesquisa qualitativa não se preocupa com representatividade numérica, mas sim, com o aprofundamento da compreensão de um grupo social ou de uma organização. Já a pesquisa quantitativa, tem por objetivo traduzir as opiniões e informações coletadas, em números, a fim de classificá-las e analisá-las, utilizando-se de recursos e técnicas estatísticas. A presente pesquisa tem a abordagem quantitativa, por analisar os dados numéricos advindos dos sistemas de informação mentidos pela SEFAZ-GO e traduzi-los em regras de classificação por árvore de decisão.

Por fim, quanto aos procedimentos técnicos, a presente pesquisa pode ser classificada em estudo de caso, pesquisa de campo e pesquisa bibliográfica. A pesquisa de campo é pautada na observação dos fatos na coleta de dados e no registro de variáveis que se presumem relevantes, para analisá-los. A pesquisa bibliográfica abrange a bibliografia já tornada pública em relação ao tema de estudo e sua finalidade é ter contato direto com o que foi publicado sobre determinado assunto (LAKATOS e MARCONI, 2001).

### 3.2.Etapas da pesquisa

Para a evolução da pesquisa, foram seguidos os passos da DCBD, cuja primeira etapa é a seleção dos dados de interesse, dentro de um universo de dados, que, no presente estudo, foi a base de dados manipulada pelos sistemas de informação mantidos pela Gerência de Tecnologia da Informação - GETI, da SEFAZ-GO. Para a seleção dos dados, utilizou-se como referência o Código Tributário do Estado de Goiás, e esses foram coletados a partir dos sistemas de Escrituração Digital Fiscal - EFD, Nota Fiscal Eletrônica - NFE, Auto de Infração Eletrônico - AIE e Cadastro de Contribuintes do Estado de Goiás - CCE. Os atributos previamente selecionados para a pesquisa foram agrupados em 5 conjuntos: Dados Cadastrais do Contribuinte Atacadista; Registro de apuração de ICMS; Registro de Controle de Crédito de ICMS; Registro de Inventário; Resultado da auditoria de ICMS.

A fim de ratificar a relevância dos atributos pré-selecionados e a possível identificação de outros atributos relevantes para a predição de potenciais sonegadores, foi utilizada a técnica de consulta aos especialistas, na qual foram realizadas reuniões periódicas e informais com Auditores Fiscais da Receita Estadual de Goiás que trabalham diretamente com Auditoria de ICMS desde, pelo menos, os últimos 5 anos. Ao longo da pesquisa foram realizadas cinco reuniões. Por meio da experiência e conhecimento dos auditores, conseguiu-se uma contribuição fundamental para o fechamento dos atributos analisados, dentre outras decisões destacadas adiante.

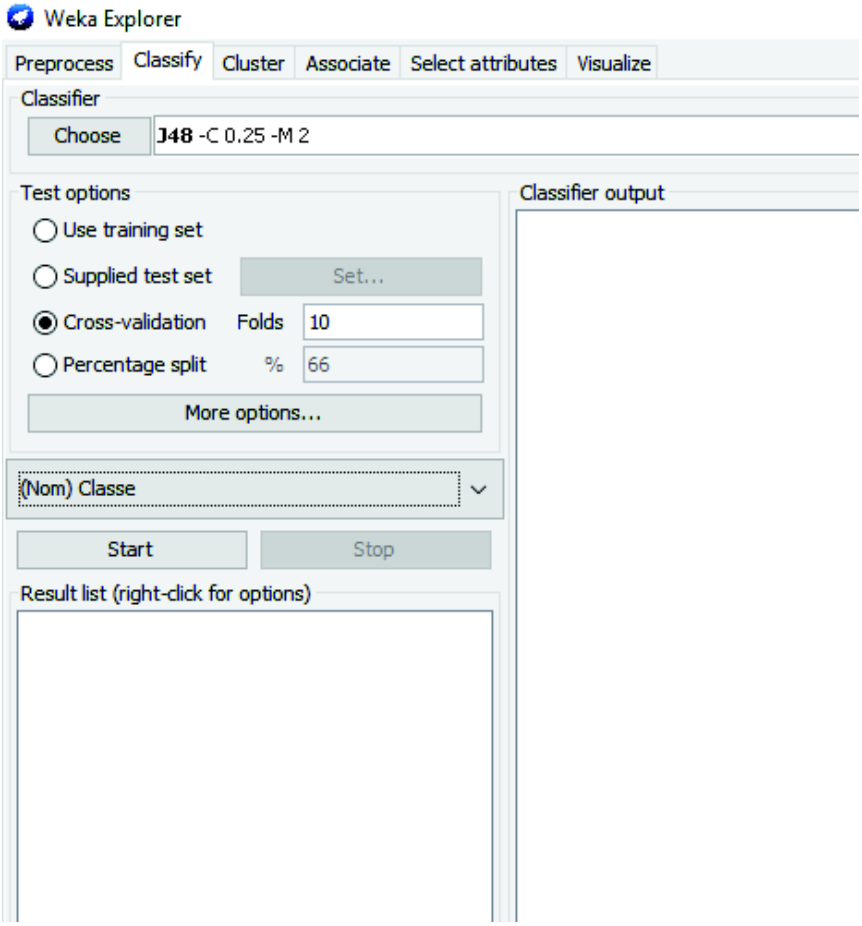
Como resultado da primeira etapa, foi gerado um arquivo com os dados dos contribuintes, conforme os limites da pesquisa.

Após a coleta dos dados, foi realizada a fase de pré-processamento, a fim de remover ruídos, erros, exemplos fora do padrão, dados incompletos, entre outras características que poderiam contaminar a massa de dados selecionada para a mineração. Em seguida, foi realizada a transformação dos dados para um arquivo. ARFF e a atribuição de cada registro à uma classe dentro do domínio Alto Valor de Sonegação (ALTO) e Baixo Valor de Sonegação (BAIXO), com base no atributo Soma_valor_original_debito_por_ano, que foi utilizado somente na preparação dos dados para realizar a classificação prévia dos dados para treinamento. Foram considerados contribuintes com Alto valor de sonegação, aqueles cujo somatório dos valores originais dos autos de infração superou R$ 50.000,00 e considerados como baixo para contribuintes cujo somatório dos valores originais dos autos de infração foi inferior a R$ 50.000,00, ambos no ano analisado. As decisões de se utilizar apenas duas classes para o atributo alvo, bem como o limite de R$ 50.000,00 para se classificar as instâncias foram tomadas em conjunto com os especialistas, baseada em suas experiências.

Após sua geração, o arquivo . ARFF, passou-se para a etapa de mineração de dados. O arquivo foi carregado na ferramenta computacional WEKA e, em seguida, foi selecionado o algoritmo J48 para a classificação, com a opção *Cross-validation* selecionada, com o valor 10 no campo *Folds*. A opção *Cross-validation* minimiza o efeito do *overfitting*, que é quando o modelo estatístico se ajusta demasiadamente ao conjunto de dados da amostra. Conforme já apresentado, o ambiente WEKA disponibiliza vários algoritmos de classificação por árvore de decisão e optou-se J48, por ser amplamente referenciado na literatura e, segundo Habibi *et al.* (2015), é o algoritmo mais importante para desenvolver o modelo de predição, dentro do ambiente da WEKA.

A Figura 9 apresenta tela de execução *Classify* do *software* WEKA com a configuração especificada e o algoritmo J48 selecionado.

Figura 9– Tela de execução Classify do software WEKA




Fonte: Autor (2017)

A mineração de dados foi dividida em 3 experimentos, com o intuito de testar a relevância de cada conjunto de atributos na mineração, mesclando os dados conforme descrito a seguir:

- Experimento 1
    - Dados do Cadastro de Contribuintes
    - Registro de Apuração do ICMS
    - Resultado da auditoria de ICMS

- Experimento 2
    - Dados do Cadastro de Contribuintes
    - Registro de Apuração do ICMS
    - Registro de Controle de Crédito de ICMS
    - Resultado da auditoria de ICMS
- Experimento 3
    - Dados do Cadastro de Contribuintes
    - Registro de Apuração do ICMS
    - Registro de Inventário
    - Resultado da auditoria de ICMS

Realizados os testes de eficiência e eficácia a partir dos indicadores computacionais, o modelo foi avaliado a partir de indicadores estatísticos, visando sua avaliação quanto à precisão preditiva e com relação à robustez. Para tal, foram analisados os percentuais de instâncias classificadas corretamente (taxa de acertos), além das análises estatísticas a partir da Matriz de Confusão e Coeficente de *Kappa*.

### 3.3. Recursos utilizados

Para a realização do presente estudo foram utilizados os softwares *Microsoft Access* e WEKA. *Access* é um gerenciador de banco de dados pertencente ao pacote *Office Professional* e foi utilizado para nas etapas de preparação e limpeza dos dados. O software WEKA foi utilizado para a mineração dos dados. O hardware utilizado foi um notebook *Acer* com processador Intel *Core* i3-6100U (2.3 GHz, 3MB L3 *Cache*), Memória 4GB DDR3 L e 1000 GB HDD.

# 4. RESULTADOS E DISCUSSÕES

## 4.1. Experimento 1

Para o Experimento 1, foram utilizados os conjuntos de atributos descritos a seguir:

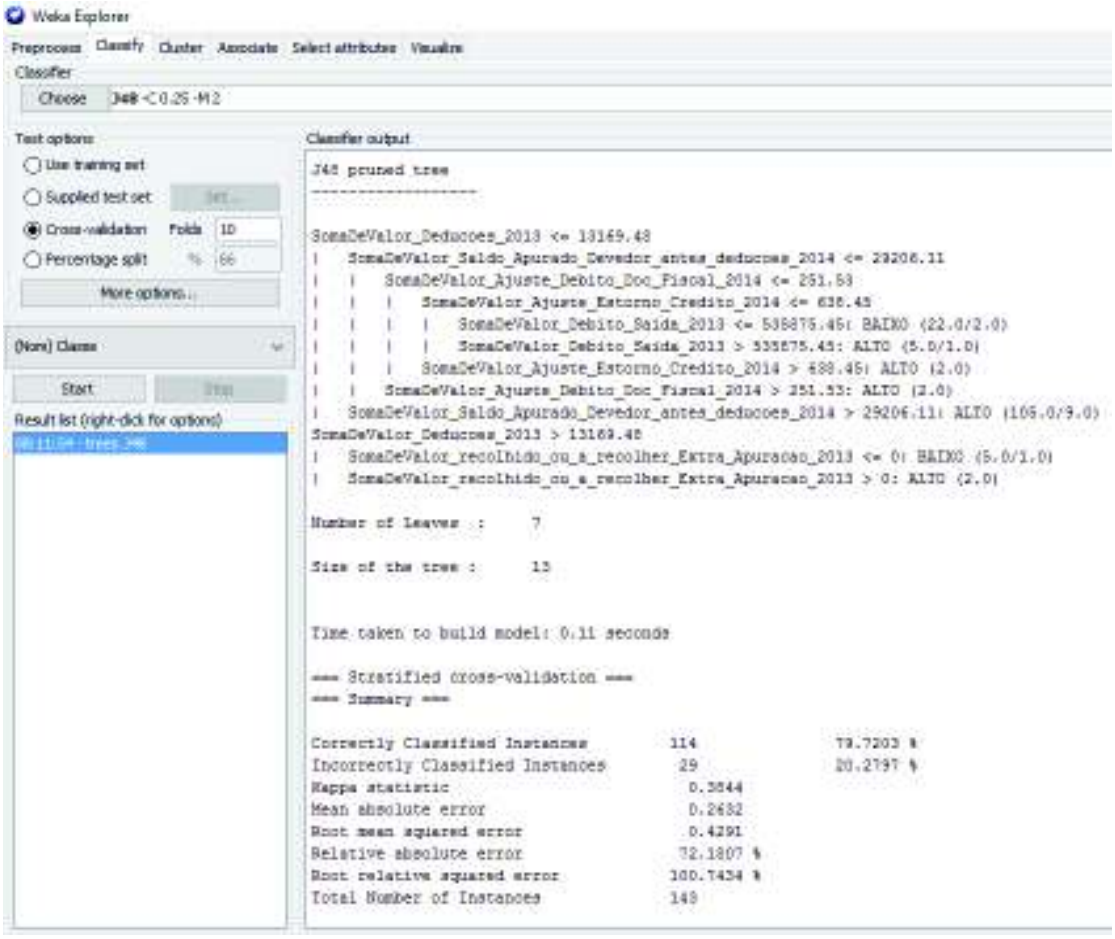
- Dados Cadastrais do contribuinte: Porte, Natureza_Juridica e Classe_Atividade_Economica;
- Registros de apuração de ICMS: Valor_Credito_Entrada_por_ano, Valor_Debito_Saida_por_ano, Valor_Ajuste_Debito_Doc_Fiscal_por_ano, Valor_Ajuste_Debito_por_ano, Valor_Ajuste_Estorno_Debito_por_ano, Valor_Ajuste_Credito_Doc_Fiscal_por_ano, Valor_Ajuste_Credito_por_ano, Valor_Ajuste_Estorno_Credito_por_ano, Valor_Saldo_Credor_Periodo_Anterior_por_ano, Valor_Saldo_Apurado_Devedor_antes_deducoes_por_ano, Valor_Deducoes_por_ano, Valor_ICMS_Recolher_por_ano, Valor_Saldo_Credor_Transp_Periodo_Seguinte_por_ano e Valor_recolhido_ou_a_recolher_Extra_Apuracao_por_ano;
- Resultado da auditoria de ICMS: Soma_valor_original_debito_por_ano;

Após a fase de pré-processamento para o experimento 1, produziu-se um arquivo com os dados de 143 contribuintes, sendo 109 classificados como ALTO por terem sofrido autos de infração cuja a soma do valor original foi superior a R$ 50.000,00. Os demais 34 contribuintes foram considerados BAIXO. Os resultados obtidos nessa etapa foram comparados com os próprios dados utilizados no aprendizado da mineração, utilizando o algoritmo J48, na opção validação cruzada, método de particionamento por reamostragem, dentre as quatro opções que a ferramenta disponibiliza. A cada execução

do algoritmo, foram analisados os percentuais de acerto e erro da classificação, o Coeficiente de *Kappa*, as regras extraídas da árvore de decisão gerada e a identificação dos atributos não utilizados para a classificação.

Para a primeira execução do algoritmo sobre os dados, foram utilizados 32 atributos, sendo que os dados Registros de Apuração de ICMS dos anos 2013 e 2014 foram analisados em conjunto, além do atributo classe, com os valores pré-determinados ALTO e BAIXO. O resultado apresentado nessa classificação foi abaixo dos 70% de acerto. Para a segunda execução, foram retirados os atributos Natureza_Juridica e Classe_Atividade_Economica e o resultado apresentou uma melhora significativa, aumentando para 79,72% o percentual de classificações corretas. Esse aumento no percentual de acerto é um indicativo que tais atributos não são relevantes para a classificar o contribuinte sonegador. A Figura 10 apresenta os resultados estratificados apresentados pelo WEKA.

Figura 10 – Resultado da segunda execução, com 30 atributos, 2013 e 2014




Fonte: Autor (2017)

A Figura 10 apresenta o resultado da execução do algoritmo sobre a massa de dados contendo 143 instâncias, cujas 114 instâncias foram classificadas corretamente. O Coeficiente de *Kappa* apresentou um índice considerável de concordância, segundo a Tabela 2 – Índices de Coeficiente de *Kappa*, de 0,3844. Percebe-se, ainda, que do conjunto 30 atributos utilizados para a mineração de dados, somente 06 atributos foram selecionados pelo WEKA para classificar os contribuintes. Os demais atributos foram considerados irrelevantes para a tarefa de classificação. Analisando a árvore acima, pode-se extrair as seguintes regras:

**Regra 1**

SomaDeValor_Deducoes_2013 <= 13169.48 e

SomaDeValor_Saldo_Apurado_Devedor_antes_deducoes_2014 <= 29206.11 e

SomaDeValor_Ajuste_Debito_Doc_Fiscal_2014 <= 251.53 e

SomaDeValor_Ajuste_Estorno_Credito_2014 <= 638.45 e

SomaDeValor_Debito_Saida_2013 <= 535875.45 então BAIXO

**Regra 2**

SomaDeValor_Deducoes_2013 <= 13169.48 e

SomaDeValor_Saldo_Apurado_Devedor_antes_deducoes_2014 <= 29206.11 e

SomaDeValor_Ajuste_Debito_Doc_Fiscal_2014 <= 251.53 e

SomaDeValor_Ajuste_Estorno_Credito_2014 <= 638.45 e

SomaDeValor_Debito_Saida_2013 > 535875.45 então ALTO

**Regra 3**

SomaDeValor_Deducoes_2013 <= 13169.48 e

SomaDeValor_Saldo_Apurado_Devedor_antes_deducoes_2014 <= 29206.11 e

SomaDeValor_Ajuste_Debito_Doc_Fiscal_2014 <= 251.53 e

SomaDeValor_Ajuste_Estorno_Credito_2014 > 638.45 então ALTO

**Regra 4**

SomaDeValor_Deducoes_2013 <= 13169.48 e

SomaDeValor_Saldo_Apurado_Devedor_antes_deducoes_2014 <= 29206.11 e

SomaDeValor_Ajuste_Debito_Doc_Fiscal_2014 > 251.53 então ALTO

**Regra 5**

SomaDeValor_Deducoes_2013 <= 13169.48 e

SomaDeValor_Saldo_Apurado_Devedor_antes_deducoes_2014 > 29206.11 então ALTO

**Regra 6**

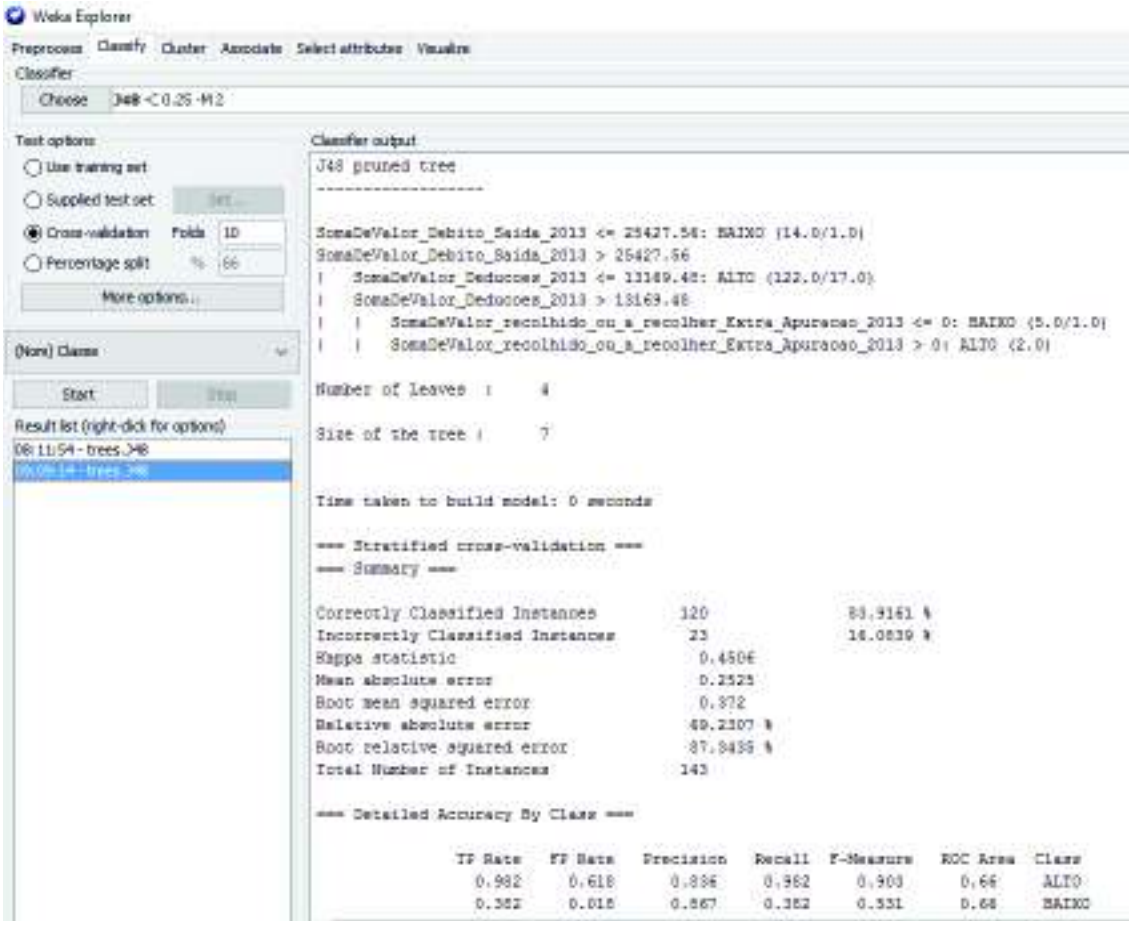
SomaDeValor_Deducoes_2013 > 13169.48 e
SomaDeValor_recolhido_ou_a_recolher_Extra_Apuracao_2013 <= 0 então BAIXO

**Regra 7**

SomaDeValor_Deducoes_2013 > 13169.48 e
SomaDeValor_recolhido_ou_a_recolher_Extra_Apuracao_2013 > 0 então ALTO

Em busca de melhoria no desempenho do algoritmo J48 na classificação dos contribuintes, foram realizadas mais duas execuções com essa mesma massa de dados, porém, na terceira execução foram analisados somente os dados de 2013 e na quarta execução foram analisados somente os dados de 2014. As Figuras 11 e 12 apresentam os resultados para os dados de 2013 e 2014, respectivamente.

Figura 9 – Resultado da terceira execução, somente dados de 2013




Fonte: Autor (2017)

A Figura 11 apresenta o resultado da execução do algoritmo J48 sobre a massa de dados contendo 143 instâncias, mas com atributos referentes ao ano de 2013. A performance do algoritmo melhorou e alcançou 83,91% de eficiência, classificando corretamente 120 instâncias. O Coeficiente de *Kappa* apresentou o valor 0,4506, o que, segundo a Tabela 2, indica um índice moderado de concordância de 0,4506. Percebe-se também que com a redução dos atributos selecionados para a execução do algoritmo, somente 03 atributos foram considerados relevantes pela WEKA. Para a árvore gerada nessa execução, pode-se extrair as seguintes regras:

**Regra 1**

SomaDeValor_Debito_Saida_2013 <= 25427.56 então BAIXO

**Regra 2**

SomaDeValor_Debito_Saida_2013 > 25427.56 e
SomaDeValor_Deducoes_2013 <= 13169.48 então ALTO

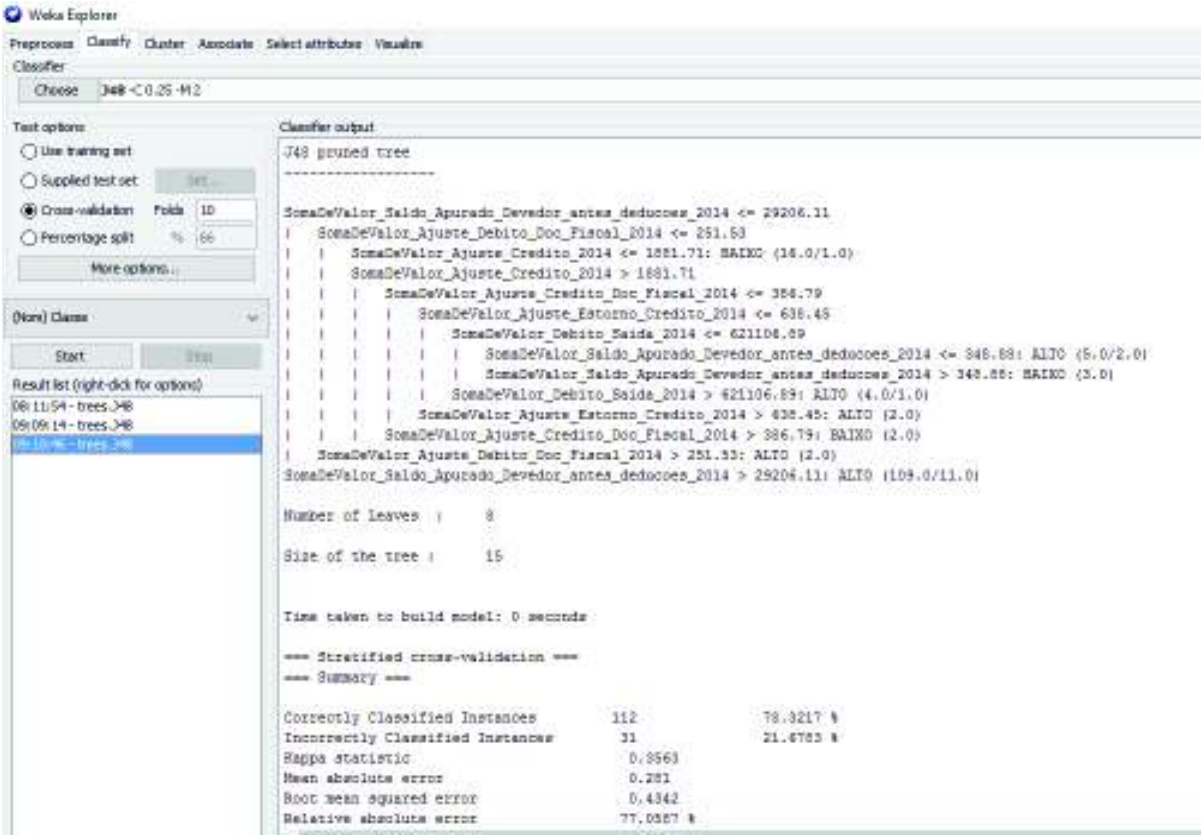
**Regra 3**

SomaDeValor_Debito_Saida_2013 > 25427.56 e
SomaDeValor_Deducoes_2013 > 13169.48 e
SomaDeValor_recolhido_ou_a_recolher_Extra_Apuracao_2013 <= 0 então BAIXO

**Regra 4**

SomaDeValor_Debito_Saida_2013 > 25427.56 e
SomaDeValor_Deducoes_2013 > 13169.48
SomaDeValor_recolhido_ou_a_recolher_Extra_Apuracao_2013 > 0 então ALTO

Figura 10– Resultado da quarta execução, somente atributos de 2014




Fonte: Autor (2017)

A Figura 12 apresenta o resultado da execução do algoritmo J48 sobre a massa de dados contendo 143 instâncias, mas com atributos referentes ao ano de 2014. A performance do algoritmo não foi tão eficiente alcançando 78,32% de êxito, classificando corretamente 112 instâncias. O Coeficiente de *Kappa* apresentou um índice considerável de concordância de 0,3563. Percebe-se que o WEKA considerou 06 atributos relevantes para a tarefa de classificação. Para a árvore gerada nessa execução, pode-se extrair as seguintes regras:

**Regra 1**

SomaDeValor_Saldo_Apurado_Devedor_antes_deducoes_2014 <= 29206.11 e
SomaDeValor_Ajuste_Debito_Doc_Fiscal_2014 <= 251.53 e
SomaDeValor_Ajuste_Credito_2014 <= 1881.71 então BAIXO

**Regra 2**

SomaDeValor_Saldo_Apurado_Devedor_antes_deducoes_2014 <= 29206.11 e
SomaDeValor_Ajuste_Debito_Doc_Fiscal_2014 <= 251.53 e
SomaDeValor_Ajuste_Credito_2014 > 1881.71 e
SomaDeValor_Ajuste_Credito_Doc_Fiscal_2014 <= 386.79 e
SomaDeValor_Ajuste_Estorno_Credito_2014 <= 638.45 e
SomaDeValor_Debito_Saida_2014 <= 621106.89 e
SomaDeValor_Saldo_Apurado_Devedor_antes_deducoes_2014 <= 348.88 então ALTO

**Regra 3**

SomaDeValor_Saldo_Apurado_Devedor_antes_deducoes_2014 <= 29206.11 e
SomaDeValor_Ajuste_Debito_Doc_Fiscal_2014 <= 251.53 e
SomaDeValor_Ajuste_Credito_2014 > 1881.71 e
SomaDeValor_Ajuste_Credito_Doc_Fiscal_2014 <= 386.79 e
SomaDeValor_Ajuste_Estorno_Credito_2014 <= 638.45 e
SomaDeValor_Debito_Saida_2014 <= 621106.89 e
SomaDeValor_Saldo_Apurado_Devedor_antes_deducoes_2014 > 348.88 então BAIXO

**Regra 4**

SomaDeValor_Saldo_Apurado_Devedor_antes_deducoes_2014 <= 29206.11 e
SomaDeValor_Ajuste_Debito_Doc_Fiscal_2014 <= 251.53 e
SomaDeValor_Ajuste_Credito_2014 > 1881.71 e
SomaDeValor_Ajuste_Credito_Doc_Fiscal_2014 <= 386.79 e
SomaDeValor_Ajuste_Estorno_Credito_2014 <= 638.45 e
SomaDeValor_Debito_Saida_2014 > 621106.89 então ALTO

**Regra 5**

SomaDeValor_Saldo_Apurado_Devedor_antes_deducoes_2014 <= 29206.11 e
SomaDeValor_Ajuste_Debito_Doc_Fiscal_2014 <= 251.53 e
SomaDeValor_Ajuste_Credito_2014 > 1881.71 e
SomaDeValor_Ajuste_Credito_Doc_Fiscal_2014 <= 386.79 e
SomaDeValor_Ajuste_Estorno_Credito_2014 > 638.45 então ALTO

**Regra 6**

SomaDeValor_Saldo_Apurado_Devedor_antes_deducoes_2014 <= 29206.11 e
SomaDeValor_Ajuste_Debito_Doc_Fiscal_2014 <= 251.53 e
SomaDeValor_Ajuste_Credito_2014 > 1881.71 e
SomaDeValor_Ajuste_Credito_Doc_Fiscal_2014 > 386.79 então BAIXO

**Regra 7**

SomaDeValor_Saldo_Apurado_Devedor_antes_deducoes_2014 <= 29206.11 e
SomaDeValor_Ajuste_Debito_Doc_Fiscal_2014 > 251.53 então ALTO

**Regra 8**

SomaDeValor_Saldo_Apurado_Devedor_antes_deducoes_2014 > 29206.11 então ALTO

Os resultados obtidos pelo Experimento 1 foram indícios de que a técnica de mineração de dados Classificação Supervisionada, onde tem-se o atributo algo assumindo os valores BAIXO SONEGADOR e ALTO SONEGADOR, pode se apresentar como uma ferramenta de apoio à predição, gerando uma boa expectativa para os experimentos 2 e 3 realizados em sequência.

### 4.2. Experimento 2

Para o Experimento 2, foram adicionados à amostra de dados, os atributos do conjunto de atributos Registro de Controle de Crédito de ICMS, conforme descrito a seguir:

- Dados Cadastrais do contribuinte;
- Registros de apuração de ICMS;

- Registro de Controle de Crédito de ICMS: Valor_Credito_Anterior; Valor_Credito_Apropriado; Valor_Crédito_Recebido; Valor_Credito_Utilizado; Valor_Credito_a_transportar;
- Resultado da auditoria de ICMS;

Para a etapa de mineração de dados, foi utilizado um arquivo com 115 instâncias, sendo 51 classificados como ALTO e 64 contribuintes foram considerados BAIXO. O critério para determinação da classe foi o mesmo para os três experimentos, ou seja, alto para contribuintes cuja soma do valor original dos autos de infração sofridos foi superior a R$ 50.000,00 e baixo cuja soma seja inferior à esse valor. Inicialmente, 23 atributos, incluindo o atributo classe, foi carregado na ferramenta computacional WEKA. Em seguida, foi selecionado o algoritmo J48 para a classificação, com a opção *Cross-validation* selecionada, com o valor 10 no campo *Folds*, conforme definido na metodologia.

Ao processar o arquivo nas configurações ressaltadas, o resultado obtido foi uma taxa de acerto de 55,65%, o que, para fins de predição de indícios de sonegação fiscal como forma de seleção de contribuintes para futuras auditorias presenciais, pode não ser razoável. Para esse cenário, a matriz de confusão apresentou que o modelo classificou corretamente como ALTO somente 22 contribuintes de 51 e classificou corretamente como BAIXO 42 de um total de 64, gerando um Coeficiente de *Kappa* de 0,0889, índice que indica apenas com ligeira concordância.

Para a segunda execução sobre a amostra de dados, optou-se por remover os atributos Porte, Natureza_Juridica e Classe_Atividade_Econômica do conjunto de atributos Dados Cadastrais do Contribuinte, por considerar que, por serem todos os contribuintes do setor atacadista do município de Goiânia-GO, não haveria perda na

qualidade na amostra minerada, conforme já havia-se feito no experimento 1. Feito isso, a quantidade de atributos reduziu de 23 para 20. O resultado apresentado foi um aumento da taxa de acerto para 63,47% e uma melhoria no desempenho do modelo, com a matriz de confusão classificando corretamente como ALTO 27 contribuintes de 51 e corretamente como BAIXO 49 de 64 contribuintes, elevando o coeficiente de *Kappa* para 0,2420, índice classificado como considerável.

Para a terceira execução, optou-se por aplicar um filtro de redução de atributos para que o próprio algoritmo escolhesse, computacionalmente, quais seriam os atributos mais relevantes dentro da amostra de dados. Na guia de pré-processamento, a ferramenta WEKA disponibiliza vários algoritmos de filtro para classificação, tanto supervisionada como não supervisionada e, para a redução de atributos, foi utilizado o algoritmo *AttributeSelection*. Após a aplicação do filtro, houve uma redução de 20 para 6 atributos e percebeu-se que o conjunto de atributos Registro de Controle de Crédito de ICMS foi removido da amostra. O resultado foi insatisfatório, com a taxa de acerto reduzindo para 57,39%

Para a quarta execução, retornou-se o estado da amostra para 20 atributos que, com exceção do atributo classe, são todos do tipo de dados numérico, ou seja, valores contínuos. A estratégia utilizada nessa execução foi a discretização dos atributos numéricos transformando-os em um pequeno número de intervalos distintos. Para tanto, foi aplicado o algoritmo *Discretize* em todos os atributos numéricos da amostra. Dos 20 atributos utilizados nessa etapa, a transformação de valores contínuos para discreto foi realizada nos 5 atributos a seguir, fazendo com que os mesmos pudessem assumir duas faixas de valor: SomaDeValor_Credito_Entrada, SomaDeValor_Debito_Saida, SomaDeValor_Ajuste_Credito, SomaDeValor_ICMS_Recolher e SomaDeValor_Saldo_Apurado_Devedoir_antes_deducoes. Para os demais atributos, o

algoritmo *Discretize* os agrupou em um único grupo, fazendo com estes fossem descartados na construção da árvore de decisão, pois não agregaria valor ao processo de classificação. O resultado foi um aumento da taxa de acertos para 69,56%, com melhorias significantes na matriz de confusão, com as relações de acerto de 37/51 e 43/64, e no coeficiente de *kappa*, elevando-se para 0,3919, índice ainda classificado como considerável, segundo a Tabela 2. Esse foi o melhor cenário obtido para esse experimento. A Figura 13 apresenta os resultados gerados pela ferramenta Weka.

Figura 11– Resultados gerados pela ferramenta WEKA, no melhor cenário

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

Correctly Classified Instances          80               69.5652 %
Incorrectly Classified Instances        35               30.4348 %
Kappa statistic                          0.3919
Mean absolute error                      0.4118
Root mean squared error                  0.473
Relative absolute error                 83.3832 %
Root relative squared error             95.1765 %
Total Number of Instances              115

=== Detailed Accuracy By Class ===

               TP Rate   FP Rate   Precision   Recall  F-Measure   ROC Area  Class
                 0.725     0.328      0.638     0.725     0.679      0.628    ALTO
                 0.672     0.275      0.754     0.672     0.711      0.628    BAIXO
Weighted Avg.    0.696     0.298      0.703     0.696     0.697      0.628

=== Confusion Matrix ===

  a  b   <-- classified as
 37 14 |  a = ALTO
 21 43 |  b = BAIXO
```

Fonte: Autores (2017)

A Figura 14 mostra as informações sobre o processamento da classificação supervisionada, descrevendo os atributos considerados e, em seguida, a árvore de decisão gerada.

Figura 12 – Informações geradas pelo WEKA sobre a classificação supervisionada.

```
Scheme:weka.classifiers.trees.J48 -C 0.25 -M 2
Relation:     contribuinte-weka.filters.unsupervised.attribute.Remove-R1-3-weka.filters.
Instances:    115
Attributes:   20
              SomaDeValor_Credito_Entrada
              SomaDeValor_Debito_Saida
              SomaDeValor_Ajuste_Debito_Doc_Fiscal
              SomaDeValor_Ajuste_Debito
              SomaDeValor_Ajuste_Estorno_Debito
              SomaDeValor_Ajuste_Credito_Doc_Fiscal
              SomaDeValor_Ajuste_Credito
              SomaDeValor_Ajuste_Estorno_Credito
              SomaDeValor_Saldo_Credor_Periodo_Anterior
              SomaDeValor_Saldo_Apurado_Devedor_antes_deducoes
              SomaDeValor_Deducoes
              SomaDeValor_ICMS_Recolher
              SomaDeValor_Saldo_Credor_Transp_Periodo_Seguinte
              SomaDeValor_recolhido_ou_a_recolher_Extra_Apuracao
              SomaDeValor_Credito_Anterior
              SomaDeValor_Credito_Apropriado
              SomaDeValor_Crédito_Recebido
              SomaDeValor_Credito_Utilizado
              SomaDeValor_Credito_a_transportar
              Classe
Test mode:10-fold cross-validation
=== Classifier model (full training set) ===

J48 pruned tree
------------------

SomaDeValor_Debito_Saida = '(-inf-873093.38]': BAIXO (55.0/12.0)
SomaDeValor_Debito_Saida = '(873093.38-inf)'
|   SomaDeValor_Credito_Entrada = '(-inf-420227.02]': BAIXO (4.0/1.0)
|   SomaDeValor_Credito_Entrada = '(420227.02-inf)'
|   |   SomaDeValor_Ajuste_Credito = '(-inf-21594.34]': BAIXO (6.0/2.0)
|   |   SomaDeValor_Ajuste_Credito = '(21594.34-inf)': ALTO (50.0/14.0)

Number of Leaves  :     4

Size of the tree :      7
```

Fonte: Autores (2017)

As regras geradas pela árvore de decisão são descritas a seguir. Percebe-se que número de regras é sempre igual ao número de folhas da árvore (*Number of Leaves*):

**Regra 1:**

SomaDeValor_Debito_Saida = (-inf:-873.093,38), então BAIXO;

**Regra 2:**

SomaDeValor_Debito_Saida = (873.093,38-inf) e

SomaDeValor_Credito_Entrada = (-inf:-420.227,02) então BAIXO;

**Regra 3:**

SomaDeValor_Debito_Saida = (873.093,38-inf) e

SomaDeValor_Credito_Entrada = (420.227,02-inf) e

SomaDeValor_Ajuste_Credito = (-inf:-21.594,34) então BAIXO;

**Regra 4:**

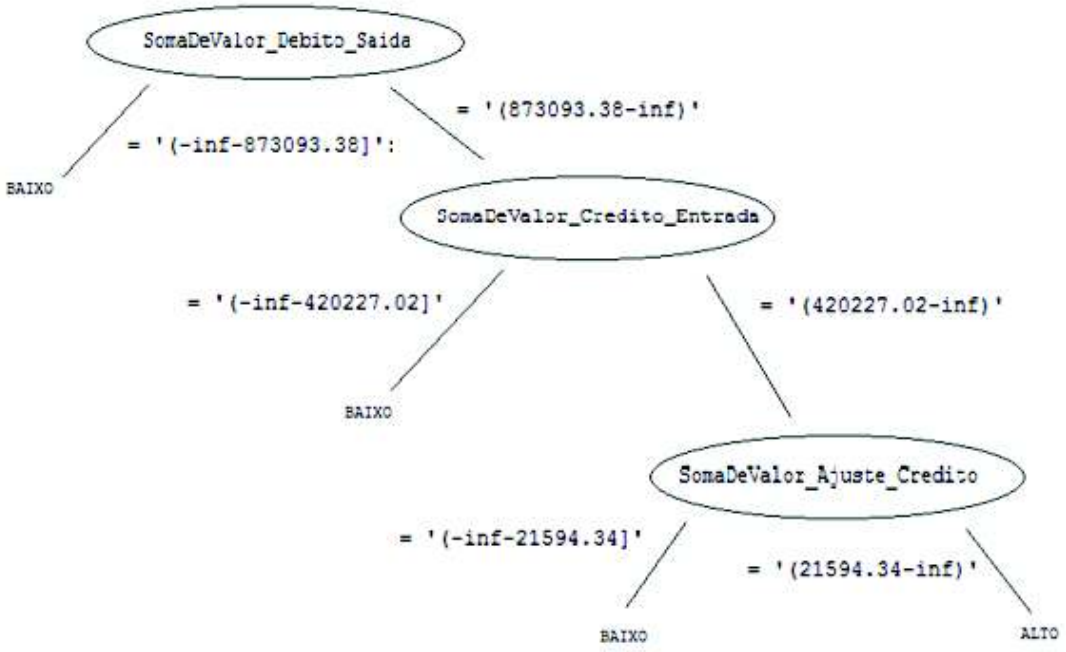
SomaDeValor_Debito_Saida = (873.093,38-inf) e

SomaDeValor_Credito_Entrada = (420.227,02-inf) e

SomaDeValor_Ajuste_Credito = (21.594,34-inf) então ALTO;

A Figura 15 mostra a representação gráfica da árvore de decisão gerada.

Figura 13 - Arvore de decisão




Fonte: Autores (2017)

## 4.3. EXPERIMENTO 3

Para o Experimento 3, foram utilizados os conjuntos de atributos descritos a seguir:

- Dados Cadastrais do contribuinte;
- Registros de apuração de ICMS;
- Registro de Inventário: Vl_Unitário;
- Resultado da auditoria de ICMS;

Para este experimento, foi utilizado uma amostra de dados com 91 contribuintes. A justificativa para a redução do tamanho da amostra a cada experimento, é que, ao adicionar novos atributos à amostra, foram desconsideradas as instâncias que não tinham valor para tais atributos. O resultado apresentado nesse experimento foi o de menor relevância nesse estudo, alcançando uma taxa de acerto de 68,13%, sem a utilização de filtros. Foram utilizados os filtros *AttributeSelection e Discretize* como no experimento 2*,* porém não houve melhora no desempenho do algoritmo. O Coeficiente *Kappa* foi de 0,295 mostrando-se um índice apenas considerável. A Figura 16 apresenta os resultados gerados pela ferramenta WEKA.

Figura 14- Resultados gerados pela ferramenta WEKA

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

Correctly Classified Instances          62               68.1319 %
Incorrectly Classified Instances        29               31.8681 %
Kappa statistic                          0.295
Mean absolute error                      0.369
Root mean squared error                  0.5029
Relative absolute error                 75.7381 %
Root relative squared error            101.8782 %
Total Number of Instances               91

=== Detailed Accuracy By Class ===

               TP Rate   FP Rate   Precision   Recall  F-Measure   ROC Area  Class
                 0.368     0.094      0.737     0.368     0.491      0.637    ALTO
                 0.906     0.632      0.667     0.906     0.768      0.637    BAIXO
Weighted Avg.    0.681     0.407      0.696     0.681     0.652      0.637

=== Confusion Matrix ===

  a  b   <-- classified as
 14 24 |  a = ALTO
  5 48 |  b = BAIXO
```

Fonte: Autor (2017)

A Figura 17 apresenta a árvore de decisão gerada no melhor cenário do experimento 3, pelo *software* WEKA.

Figura 15– Árvore de decisão gerada no melhor cenário do experimento 3, pelo software WEKA

```
SomaDeValor_Saldo_Apurado_Devedor_antes_deducoes <= 1843314.05
|   Vl_Unitário <= 154.68: ALTO (5.0)
|   Vl_Unitário > 154.68
|   |   SomaDeValor_Saldo_Apurado_Devedor_antes_deducoes <= 87258.33: BAIXO (23.0/1.0)
|   |   SomaDeValor_Saldo_Apurado_Devedor_antes_deducoes > 87258.33
|   |   |   SomaDeValor_recolhido_ou_a_recolher_Extra_Apuracao <= 846.5
|   |   |   |   Vl_Unitário <= 4796.87: ALTO (5.0)
|   |   |   |   Vl_Unitário > 4796.87
|   |   |   |   |   SomaDeValor_Ajuste_Credito <= 965144.52
|   |   |   |   |   |   SomaDeValor_Ajuste_Estorno_Debito <= 75
|   |   |   |   |   |   |   SomaDeValor_Credito_Entrada <= 973008.84
|   |   |   |   |   |   |   |   Vl_Unitário <= 17490.9: BAIXO (5.0)
|   |   |   |   |   |   |   |   Vl_Unitário > 17490.9
|   |   |   |   |   |   |   |   |   SomaDeValor_Credito_Entrada <= 53175.6: BAIXO (4.0)
|   |   |   |   |   |   |   |   |   SomaDeValor_Credito_Entrada > 53175.6
|   |   |   |   |   |   |   |   |   |   SomaDeValor_Saldo_Apurado_Devedor_antes_deducoes <= 279979.09: ALTO (9.0/1.0)
|   |   |   |   |   |   |   |   |   |   SomaDeValor_Saldo_Apurado_Devedor_antes_deducoes > 279979.09
|   |   |   |   |   |   |   |   |   |   |   SomaDeValor_Debito_Saida <= 1482846.27: BAIXO (3.0)
|   |   |   |   |   |   |   |   |   |   |   SomaDeValor_Debito_Saida > 1482846.27: ALTO (2.0)
|   |   |   |   |   |   |   SomaDeValor_Credito_Entrada > 973008.84: BAIXO (7.0)
|   |   |   |   |   |   SomaDeValor_Ajuste_Estorno_Debito > 75: ALTO (2.0/1.0)
|   |   |   |   |   SomaDeValor_Ajuste_Credito > 965144.52: ALTO (4.0)
|   |   |   SomaDeValor_recolhido_ou_a_recolher_Extra_Apuracao > 846.5: BAIXO (11.0/1.0)
SomaDeValor_Saldo_Apurado_Devedor_antes_deducoes > 1843314.05: ALTO (11.0)

Number of Leaves  :     13

Size of the tree :      25
```

Fonte: Autor (2017)

Conforme pode se observar na Figura 17, o número de folhas (*leaves*) é igual a 13 que equivale ao número de regras extraídas descritas a seguir:

**Regra 1**

SomaDeValor_Saldo_Apurado_Devedor_antes_deducoes <= 1843314.05 e
Vl_Unitário <= 154.68 então ALTO

**Regra 2**

SomaDeValor_Saldo_Apurado_Devedor_antes_deducoes <= 1843314.05 e
Vl_Unitário > 154.68 e
SomaDeValor_Saldo_Apurado_Devedor_antes_deducoes <= 87258.33 então BAIXO

**Regra 3**

SomaDeValor_Saldo_Apurado_Devedor_antes_deducoes <= 1843314.05 e
Vl_Unitário > 154.68 e
SomaDeValor_Saldo_Apurado_Devedor_antes_deducoes > 87258.33 e
SomaDeValor_recolhido_ou_a_recolher_Extra_Apuracao <= 846.5 e
Vl_Unitário <= 4796.87 então ALTO

**Regra 4**

SomaDeValor_Saldo_Apurado_Devedor_antes_deducoes <= 1843314.05 e
Vl_Unitário > 154.68 e
SomaDeValor_Saldo_Apurado_Devedor_antes_deducoes > 87258.33 e
SomaDeValor_recolhido_ou_a_recolher_Extra_Apuracao <= 846.5 e
Vl_Unitário > 4796.87 e
SomaDeValor_Ajuste_Credito <= 965144.52 e
SomaDeValor_Ajuste_Estorno_Debito <= 75 e
SomaDeValor_Credito_Entrada <= 973008.84 e

Vl_Unitário <= 17490.9 então BAIXO

**Regra 5**

SomaDeValor_Saldo_Apurado_Devedor_antes_deducoes <= 1843314.05 e
Vl_Unitário > 154.68 e
SomaDeValor_Saldo_Apurado_Devedor_antes_deducoes > 87258.33 e
SomaDeValor_recolhido_ou_a_recolher_Extra_Apuracao <= 846.5 e
Vl_Unitário > 4796.87 e
SomaDeValor_Ajuste_Credito <= 965144.52 e
SomaDeValor_Ajuste_Estorno_Debito <= 75 e
SomaDeValor_Credito_Entrada <= 973008.84 e
Vl_Unitário > 17490.9 e
SomaDeValor_Credito_Entrada <= 53175.6 então BAIXO

**Regra 6**

SomaDeValor_Saldo_Apurado_Devedor_antes_deducoes <= 1843314.05 e
Vl_Unitário > 154.68 e
SomaDeValor_Saldo_Apurado_Devedor_antes_deducoes > 87258.33 e
SomaDeValor_recolhido_ou_a_recolher_Extra_Apuracao <= 846.5 e
Vl_Unitário > 4796.87 e
SomaDeValor_Ajuste_Credito <= 965144.52 e
SomaDeValor_Ajuste_Estorno_Debito <= 75 e
SomaDeValor_Credito_Entrada <= 973008.84 e
Vl_Unitário > 17490.9 e
SomaDeValor_Credito_Entrada > 53175.6 e
SomaDeValor_Saldo_Apurado_Devedor_antes_deducoes <= 279979.09 então ALTO

**Regra 7**

SomaDeValor_Saldo_Apurado_Devedor_antes_deducoes <= 1843314.05 e
Vl_Unitário > 154.68 e

SomaDeValor_Saldo_Apurado_Devedor_antes_deducoes > 87258.33 e
SomaDeValor_recolhido_ou_a_recolher_Extra_Apuracao <= 846.5 e
Vl_Unitário > 4796.87 e
SomaDeValor_Ajuste_Credito <= 965144.52 e
SomaDeValor_Ajuste_Estorno_Debito <= 75 e
SomaDeValor_Credito_Entrada <= 973008.84 e
Vl_Unitário > 17490.9 e
SomaDeValor_Credito_Entrada > 53175.6 e
SomaDeValor_Saldo_Apurado_Devedor_antes_deducoes > 279979.09 e
SomaDeValor_Debito_Saida <= 1482846.27 então BAIXO

**Regra 8**

SomaDeValor_Saldo_Apurado_Devedor_antes_deducoes <= 1843314.05 e
Vl_Unitário > 154.68 e
SomaDeValor_Saldo_Apurado_Devedor_antes_deducoes > 87258.33 e
SomaDeValor_recolhido_ou_a_recolher_Extra_Apuracao <= 846.5 e
Vl_Unitário > 4796.87 e
SomaDeValor_Ajuste_Credito <= 965144.52 e
SomaDeValor_Ajuste_Estorno_Debito <= 75 e
SomaDeValor_Credito_Entrada <= 973008.84 e
Vl_Unitário > 17490.9 e
SomaDeValor_Credito_Entrada > 53175.6 e
SomaDeValor_Saldo_Apurado_Devedor_antes_deducoes > 279979.09
SomaDeValor_Debito_Saida > 1482846.27 então ALTO

**Regra 9**

SomaDeValor_Saldo_Apurado_Devedor_antes_deducoes <= 1843314.05 e
Vl_Unitário > 154.68 e
SomaDeValor_Saldo_Apurado_Devedor_antes_deducoes > 87258.33 e
SomaDeValor_recolhido_ou_a_recolher_Extra_Apuracao <= 846.5 e
Vl_Unitário > 4796.87 e

SomaDeValor_Ajuste_Credito <= 965144.52 e

SomaDeValor_Ajuste_Estorno_Debito <= 75 e

SomaDeValor_Credito_Entrada <= 973008.84 e

Vl_Unitário > 17490.9 e

SomaDeValor_Credito_Entrada > 973008.84 então BAIXO

**Regra 10**

SomaDeValor_Saldo_Apurado_Devedor_antes_deducoes <= 1843314.05 e

Vl_Unitário > 154.68 e

SomaDeValor_Saldo_Apurado_Devedor_antes_deducoes > 87258.33 e

SomaDeValor_recolhido_ou_a_recolher_Extra_Apuracao <= 846.5 e

Vl_Unitário > 4796.87 e

SomaDeValor_Ajuste_Credito <= 965144.52 e

SomaDeValor_Ajuste_Estorno_Debito > 75: ALTO

**Regra 11**

SomaDeValor_Saldo_Apurado_Devedor_antes_deducoes <= 1843314.05 e

Vl_Unitário > 154.68 e

SomaDeValor_Saldo_Apurado_Devedor_antes_deducoes > 87258.33 e

SomaDeValor_recolhido_ou_a_recolher_Extra_Apuracao <= 846.5 e

Vl_Unitário > 4796.87 e

SomaDeValor_Ajuste_Credito > 965144.52 então ALTO

**Regra 12**

SomaDeValor_Saldo_Apurado_Devedor_antes_deducoes <= 1843314.05 e

Vl_Unitário > 154.68 e

SomaDeValor_Saldo_Apurado_Devedor_antes_deducoes > 87258.33 e

SomaDeValor_recolhido_ou_a_recolher_Extra_Apuracao > 846.5 então BAIXO

**Regra 13**

SomaDeValor_Saldo_Apurado_Devedor_antes_deducoes > 1843314.05 então ALTO

### 4.4. Comparação dos modelos

O Quadro 5 apresenta um resumo, conforme características de cada experimento, comparando os resultados do melhor cenário de cada um.

Quadro 5– Resumo dos resultados dos experimentos

| Exp. | Tam. Amostra | Atributos Discretizados? (Discreitize) | Tam. Árvore | Nº de Folhas | Instâncias Classificadas Corretamente | Instâncias Classificadas Incorretamente | Coific. Kappa | Número Atributos Utilizados |
|---|---|---|---|---|---|---|---|---|
| 1 | 143 | Não | 7 | 4 | 83,91% | 16,09% | 0,4506 | 3 |
| 2 | 115 | Sim | 7 | 4 | 69,56% | 30,44% | 0,3919 | 3 |
| 3 | 91 | Não | 25 | 13 | 68,13% | 31,87% | 0,2950 | 7 |

Fonte: Autor (2007)

### 4.5. Comparação do algoritmo j48 com outros algoritmos implementados no weka

Após a compilação dos resultados da mineração de dados dos três experimentos realizados, optou-se então por refazer os experimentos utilizando outros algoritmos de árvore de decisão, disponíveis no ambiente WEKA. O Quadro 6 mostra o comparativo do resultado do experimento 1 executado pelo J48 com os resultados obtidos pela execução da mesma amostra de dados com alguns dos outros algoritmos disponíveis.

Quadro 6 - Comparativo dos resultados do experimento 1

| Algoritmo | Tam. Árvore | Nº de Folhas | Instâncias Classificadas Corretamente | Instâncias Classificadas Incorretamente | Coific. Kappa | Número Atributos Utilizados |
|---|---|---|---|---|---|---|
| J48 | 7 | 4 | 83,91% | 16,09% | 0,4506 | 3 |
| ADTree | 31 | 21 | 79,72% | 20,28% | 0,3558 | 8 |
| BFTree | 3 | 2 | 81,82% | 18,18% | 0,4015 | 1 |
| LMT | 3 | 2 | 83,22% | 16,78% | 0,4195 | 1 |
| NBTree | 3 | 2 | 83,91% | 16,09% | 0,4506 | 1 |
| SimpleCart | 3 | 2 | 84,62% | 15,38% | 0,4679 | 1 |

Fonte: Autor (2017)

Como no experimento 2 os dados foram discretizados, foi possível testar a amostra com o algoritmo ID3. O Quadro 7 mostra o comparativo do resultado do experimento 2 executado pelo J48 com os resultados obtidos pela execução da mesma amostra de dados com alguns dos outros algoritmos disponíveis.

Quadro 7 - Comparativo dos resultados do experimento 2

| **Algoritmo** | **Tam. Árvore** | **Nº de Folhas** | **Instâncias Classificadas Corretamente** | **Instâncias Classificadas Incorretamente** | **Coific. Kappa** | **Número Atributos Utilizados** |
|---|---|---|---|---|---|---|
| J48 | 7 | 4 | 69,56% | 30,44% | 0,3919 | 3 |
| ID3 | - | 10 | 68,70% | 31,30% | 0,3733 | 5 |
| ADTree | 25 | 17 | 72,18% | 27,82% | 0,4385 | 5 |
| BFTree | 19 | 10 | 69,56% | 30,44% | 0,3895 | 5 |
| LMT | 5 | 3 | 73,04% | 26,96% | 0,4593 | 2 |
| NBTree | 1 | 1 | 73,91% | 26,09% | 0,4736 | 1 |
| SimpleCart | 9 | 5 | 69,56% | 30,44% | 0,3895 | 4 |

Fonte: Autor (2017)

O Quadro 8 mostra o comparativo do resultado do experimento 3 executado pelo J48 com os resultados obtidos pela execução da mesma amostra de dados com alguns dos outros algoritmos disponíveis.

Quadro 8 - Comparativo dos resultados do experimento 3

| **Algoritmo** | **Tam. Árvore** | **Nº de Folhas** | **Instâncias Classificadas Corretamente** | **Instâncias Classificadas Incorretamente** | **Coific. Kappa** | **Número Atributos Utilizados** |
|---|---|---|---|---|---|---|
| J48 | 25 | 13 | 68,13% | 31,87% | 0,2950 | 7 |
| ADTree | 31 | 21 | 61,54% | 38,46% | 0,2063 | 5 |
| BFTree | 5 | 3 | 58,24% | 42,76% | 0,1219 | 2 |
| LMT | 1 | 1 | 65,93% | 34,07% | 0,2340 | 1 |
| NBTree | 3 | 2 | 73,63% | 26,37% | 0,4280 | 1 |
| SimpleCart | 3 | 2 | 58,24% | 42,76% | 0,1014 | 1 |

Fonte: Autor (2017)

# 5. CONCLUSÃO

A presente pesquisa analisou os dados mantidos pela SEFAZ-GO relativos à contribuição de ICMS das empresas ativas do setor atacadista, situadas no município de Goiânia-GO, a fim de propor um modelo preditivo de classificação de contribuintes de ICMS, em busca de indicações de sonegação, por meio de técnicas de mineração de dados baseadas na classificação dos contribuintes. O software utilizado nesse estudo foi o WEKA, que possui um conjunto de métodos e algoritmos para identificação de regras de classificação e padrões em grandes volumes de dados. O algoritmo utilizado foi o J48 e foram realizadas 3 execuções sobre a massa de dados contribuintes de ICMS do setor atacadista do município de Goiânia-GO que sofreram autos de infração entre 2013 e 2016. Concluí-se que, todas as etapas do processo DCBD são fundamentais para o êxito do projeto e que as etapas de pré-processamento, que consumiu aproximadamente 70% do tempo da presente pesquisa, podem ser fundamentais para que os resultados apresentados pela mineração de dados sejam relevantes. Os atributos selecionados foram baseados no Código Tributário de Estado de Goiás relacionados ao recolhimento do ICMS, com auxílio de opiniões de especialistas com vasta experiência na execução de auditores de ICMS . Apesar de o modelo gerado pelo presente estudo ter alcançado, em seu melhor cenário, a taxa de classificações corretas em torno de 84%, é potencialmente possível que a seleção de outros conjuntos atributos contidos na EFD e não analisados neste estudo, em conjunto com os aqui minerados, possam elevar a efetividade do modelo de predição gerado por meio da execução do algoritmo J48, na ferramenta computacional WEKA, e essa investigação é fortemente recomendada para trabalhos futuros. Conclui-se ainda, que a robustez de um modelo deve ser comprovada por índices estatísticos, que, no presente estudo foram utilizados a Matriz de Confusão e o Coeficiente de *Kappa*, que alcançou

0,4506, em seu melhor cenário, podendo ser estendido a outros indicadores, para que, de fato, possam gerar cenários preditivos capazes de contribuir para decisões mais assertivas por parte dos gestores.

# REFERÊNCIAS

Al-Radaideh, Q. A.; Al-Ananbeh, A.; Al-Shawakfa, E. M., A Classification Model For Predicting The Suitable Study Track For School Students, Department of Computer Information Systems, Faculty of Information Technology, 2011, Jordan.

AMO, S. Técnicas de mineração de dados. Universidade Federal de Uberlândia, Faculdade de Computação, disponível em: <http://www.deamo.prof.ufu.br/>, 2004. Acesso em 30 de Junho de 2017.

AGRAWAL, S.; AGRAWAL, J. Survey on Anomaly Detection using Data Mining Techniques. Procedia - Procedia Computer Science, 2015. v. 60, p. 708–713. Disponível em: <http://dx.doi.org/10.1016/j.procs.2015.08.220>.

ANDRADE, Helder da Silva. Um Processo de Mineração de Dados Aplicado ao Combate à Sonegação Fiscal do ICMS, Dissertação (Mestrado), Universidade Estadual do Ceará, 2009.

BASGALUPP, Márcio Porto. LEGAL-Tree: Um algoritmo genético multi-objetivo lexicográfico para indução de árvores de decisão. Tese de D.Sc., ICMC/USP São Carlos, São Paulo, 2010. Disponível em: < http://www.teses.usp.br/teses/disponiveis/55/55134/tde-12052010-165344/pt-br.php > Acesso em: 09 de Agosto de 2017.

CAMILO, C. O.; SILVA, J. C. Mineração de Dados: Conceitos, Tarefas, Métodos e Ferramentas, RT-INF_001-09, UFG, 2009.

CARDOSO, Olinda N. P.; MACHADO, Rosa T. M.. Gestão do Conhecimento usando data mining: estudo de caso na Universidade Federal de Lavras. Revista de Administração Pública, v. 42, n. 3, p. 495-528, jun. 2008. Disponível em: <http://www.scielo.br/pdf/rap/v42n3/a04v42n3.pdf> Acesso em: 01 de Fevereiro de 2017.

CASTELLÓN, P.; VELÁSQUEZ, J. D. Expert Systems with Applications Characterization and detection of taxpayers with false invoices using data mining techniques. Expert Systems with Applications, 2013. v. 40, p. 1427–1436.

Cavalcante, Renata de Souza Alves Paula. Descoberta de conhecimento na plataforma lattes: um estudo de caso no instituto federal goiás – Goiânia: PUC-Goiás/MEPROS, 2014.

CHEN, J. et al. Automation in Construction Application of neural networks for detecting erroneous tax reports from construction companies. Automation in Construction, 2011. v. 20, n. 7, p. 935–939. Disponível em: <http://dx.doi.org/10.1016/j.autcon.2011.03.011>.

DAMACENO, Marcelo. Introdução à Mineração de Dados Utilizando o Weka, Disponível em < http://connepi.ifal.edu.br/ocs/index.php/connepi/CONNEPI2010/paper/viewFile/258/207 >. Acesso em: 03 mai 2016.

DIGIAMPIETRI, L. A., ROMAN, N. T., MEIRA, L. A., FERREIRA, C. D., KONDO, A. A., CONSTANTINO, E. R., LANNA, A Uses of Artificial Intelligence in the Brazilian Customs Fraud Detection System. Digital Government Society of North America, 2008. p. 181–187.

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

G1. O Portal de Notícias da Globo. Disponível em < http://g1.globo.com/economia/noticia/desemprego-fica-em-12-no-4-trimestre-de-2016.ghtml>. Acesso em: 24 mar.2017.

GERHARDT, T. E. SILVEIRA, D. T. Métodos de pesquisa. Coordenado pela Universidade Aberta do Brasil – UAB/UFRGS e pelo Curso de Graduação Tecnológica – Planejamento e Gestão para o Desenvolvimento Rural da SEAD/UFRGS. – Porto Alegre: Editora da UFRGS, 2009.

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

GOIÁS**.** Portal EFD Goiás. 2016. Disponível em : <http://www.efd.go.gov.br/>. Acesso em: 02 mai 2016.

GOIÁS. Secretaria da Fazenda do Estado de Goiás. 2016. Disponível em: <http://www.sefaz.go.gov.br/LTE/Lte_ver_40_3_htm/Rcte/RCTE.htm > Acesso em: 03 mai.2016.

GOLDSCHMIDT, R. R. Tópicos Especiais em Inteligência Computacional. Instituto Superior de Tecnologia do Rio de Janeiro – Série Livros Didáticos Digitais Informática para todos. Rio de Janeiro: IST – Rio, 2011.

GOUMAGIAS, N. D.; SARAIDARIS, A. A decision support model for tax revenue collection in Greece. Decision Support Systems, 2012. v. 53, n. 1, p. 76–96. Disponível em: <http://dx.doi.org/10.1016/j.dss.2011.12.006>.

HABIBI, S.; AHMADI, M.; ALIZADEH, S. Type 2 Diabetes Mellitus Screening and Risk Factors Using Decision Tree : Results of Data Mining. 2015. v. 7, n. 5, p. 304–310.

JUNG, C. F. Metodologia Científica: Ênfase em Pesquisa Tecnológica. 3ª Ed. Revisada e Ampliada – 2003/I.

LAKATOS, E. M. MARCONI, M. A. Fundamentos de metodologia científica. 5 Ed. São Paulo: Atlas, 2003.

LAROSE, D. T. Discovering Knowledge in Data: An Introduction to Data Mining. John Wiley and Sons, 2006.

LEVERGGER Piccirilli, Tiago. Mineração de Dados Aplicada à Classificação dos Contribuintes do ISS, Dissertação (Mestrado), Pontifícia Universidade Católica de Goiás, 2013.

LIN, C.; Chiu A., Huang, S. Y.; Yen, D. C. Detecting the financial statement fraud: The analysis of the differences between data mining techniques and experts' judgments, Knowledge-Based Systems, 2015, Permanent link to this document: http://dx.doi.org/10.1016/j.knosys.2015.08.011

LIU, B. et al. Outlier Detection Data Mining of Tax Based on Cluster 1. 2012. v. 33, p. 1689–1694. Disponível em: <http://dx.doi.org/10.1016/j.phpro.2012.05.272>.

MEIRA, Carlos A.A.; Rodrigues, Luiz H.A.; Moraes, Sérgio A. Análise da epidemia da ferrugem do cafeeiro com árvore de decisão, Trop. plant pathol. vol.33 no.2 Brasília Mar./Apr. 2008. Permanent link to this document: http://dx.doi.org/10.1590/S1982-56762008000200005

MENDONÇA, Marcelo Amaral Gonçalves de. Utilização de Redes Neurais Artificiais e Séries Temporais: Análise da Arrecadação do ICMS do Estado de São Paulo. Mestrado (Dissertação) – Universidade Federal do Rio de Janeiro/COPPEAD, 2000.

OLIVEIRA, Francisco Nobre de. Estratégias para Aperfeiçoar o Processo de Recuperação de Receitas Tributárias no Estado da Bahia: Um Modelo para o ICMS Baseado em Redes Neurais Artificiais, Dissertação (Mestrado), Universidade Federal da Bahia, 2011.

PELLIZZARI, Deoni. A Grande Farsa da Tributação e da Sonegação. Petrópolis: Editora Vozes, 1990.

PERROCA, M. G.; GAIDZINSKI, R. R. Avaliando a confiabilidade interavaliadores de um instrumento para classificação de pacientes - coeficiente Kappa. Rev. Esc. Enferm. USP, 2003; 37 (1): 72-80. Disponível em: <http://www.scielo.br/pdf/reeusp/v37n1/09.pdf>. Acesso em 03 Julho de 2017.

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

RODRIGUES, Fabrício Alves; AMARAL, Laurence Rodrigues. Aplicação de Métodos Computacionais de Mineração de Dados na Classificação e Seleção de Oncogenes Medidos por Microarray, Revista Brasileira de Cancerologia 2012; 58(2): 241-249, 2012. Disponível em <http://www1.inca.gov.br/rbc/n_58/v02/pdf/14_artigo_aplicacao_metodos_computacionais_mineracao_dados_classificacao_selecao_oncogenes_medidos_microarray.pdf >, Acesso em 01.jul.2017

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

SANTOS, Luciano Drosda M.,; et al., Procedimentos de Validação Cruzada em Mineração de Dados para ambiente de Computação Paralela. Dep. Acad. Informática, UTFPR, 2009; Disponível em < http://www.lbd.dcc.ufmg.br/colecoes/erad/2009/047.pdf> Acesso em 04.de Julho de 2017.

SILVEIRA, Marcos Renato Moreira. Sistema Neural para Quantificação e Qualificação da Sonegação Fiscal de ICMS em Empresas do Tipo Débito/Crédito, Dissertação (Mestrado), Universidade Estadual do Norte Fluminense, 2001.

SOUZA, Américo José Vasconcelos de. O Uso de Mapas Auto-Organizáveis para Classificar Contribuintes do ICMS, Dissertação (Mestrado), Universidade Federal de Goiás, 2002.

SOUZA, Primavera Botelho de. Uma estratégia baseada em algoritmos de mineração de dados para validar plano de operação de voo a partir de predições de estados dos satélites do INPE / Primavera Botelho de Souza. – São José dos Campos : INPE, 2011.xxii+149 p. ; (sid.inpe.br/mtc-m19/2011/04.15.19.12-TDI)

STEINER, M. T. A.; Soma N. Y.; Shimizu, T.; Nievola, J. C.; Steiner Neto, P. J. Abordagem de um Problema Médico por Meio do Processo de KDD com Ênfase à análise exploratória dos dados, Gestão & Produção, v.13, n2, p.335-337, maio-ago. 2006.

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

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

TRANSPARÊNCIA, Portal da Transparência do Estado de Goiás. Disponível em <http://www.transparencia.go.gov.br/portaldatransparencia/receitas/receita-estadual> Acesso em 24.mar.2017.

UNIVERSITY OF WAIKATO. Weka 3 – Machine Learning Software in Java. Disponível no site da University of Waikato (2016). URL: http://www.cs.waikato.ac.nz/ml/weka

VALOR. Valor Econômico. Disponível em <http://www.valor.com.br/brasil/4910646/estimativa-da-fazenda-para-expansao-do-pib-neste-ano-recua-de-1-para-05> Acesso em 24 mar.2017.

VIEIRA, Mário Henrique Paes. Aplicação de técnicas de mineração em um programa de concessão de benefícios ao consumidor: o caso do Programa Nota Legal do Distrito Federal, Dissertação (Mestrado), Universidade de Brasília, 2014.

WITTEN, Ian. H.; FRANK, Eibe. Data Mining: Practical Machine Learning Tools and Techniques. Elsevier, Second Edition, 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: 05 de março de 2017.

WU, R. et al. Expert Systems with Applications Using data mining technique to enhance tax evasion detection performance. 2012. v. 39, p. 8769–8777.