Business Intelligence

Ricardo Holt Pereira

APLICAÇÃO DE DATAMINING COMO EFICAZ MECANISMO DE CONTROLE NA EVASÃO DE RECEITAS FISCAIS E NA DETECÇÃO DE FRAUDES

Monografia de Final de Curso

04/08/2009

Monografia apresentada ao Departamento de Engenharia Elétrica da PUC - Rio como parte dos requisitos para a obtenção do título de Especialização em Business Intelligence

Orientadores:

Marley Maria B. R. Velasco

Marco Aurélio Cavalcante Pacheco

Dedicatória

À minha mãe, por tudo.

Agradecimentos

A elaboração deste trabalho só foi possível graças ao incentivo e orientação que recebi do Corpo Docente do curso BI Master – Business Intelligence da PUC-RJ.

A todos que com tamanha generosidade e disponibilidade me acompanharam nesta jornada, dedico o meu agradecimento mais sincero pela experiência e conhecimento que me transmitiram.

RESUMO

Esta monografia é fruto de longas pesquisas realizadas no decorrer de anos de trabalho "in loco" no campo onde se desenrolam verdadeiras batalhas no enfrentamento de um dos mais graves e complexos problemas com os quais se defronta o Poder Público, o controle da evasão de receitas fiscais e a detecção de fraudes praticadas pelos contribuintes.

O que aqui se pretende demonstrar é a viabilidade de uma solução alternativa para tão grave e antigo problema, através da utilização de Dataminig e redes neurais artificiais.

A eficácia desta ferramenta como opção viável pode ser medida e comprovada diante da constatação da existência, nos dias de hoje, de comportamentos fraudulentos não passíveis de identificação pela análise humana.

Assim é que, na atualidade, todos os métodos convencionais de detecção de fraude têm se mostrado de sucesso absolutamente limitado.

Como exemplo prático, a opção utilizada foi a análise do desempenho da arrecadação do Imposto sobre circulação de Mercadorias e Serviços (ICMS) em unidade X (real) da Federação.

ABSTRACT

This paper is the result of research carried out during long years of work "in situ" in the field where real battles take place in the face of one of the most serious and complex problems which are facing the State, the control of evasion of income and detection of tax frauds perpetrated by taxpayers.

What is sought here is to demonstrate feasibility of an alternative to so serious and old problem through the use of artificial neural networks and Dataminig.

The effectiveness of this tool as a viable option can be measured and verified before the finding of the existence, today, of fraudulent conduct non-identifyed by human analysis.

Thus, in actuality, all conventional methods of detection of fraud have had quite limited success.

As a practical example, the option was used to analyze the performance of the collection of tax on the movement of goods and services (ICMS) in unit X (real) of the Federation.

SUMÁRIO

1. INTRODUÇÃO

A expressão *Business Intelligence* (BI), que como se sabe, comporta tradução como "inteligência de negócios", abrange todo o procedimento de coleta, organização, análise, compartilhamento e monitoramento de informações que propiciam eficaz suporte na gestão de negócios e na tomada de decisão.

Em conseqüência, não seria de todo equivocado afirmar que BI é uma verdadeira fonte de conhecimento que se nutre por intermédio de criterioso acervo de informações primárias que, decodificadas geram novas informações, produzindo exato conhecimento que propicia a otimização do processo de gestão e as tomadas de decisão.

A aplicação do conjunto de ferramentas de BI não admite limites. Todo e qualquer ambiente empresarial comporta a utilização de BI.

Na iniciativa privada são incontáveis as hipóteses e a excelência dos resultados obtidos como é já de conhecimento geral.

No presente trabalho, aproveitando a experiência de alguns anos no Serviço Público, a opção será focar a abordagem somente na esfera estatal, mais especificamente no tema evasão de receitas tributárias em decorrência de fraudes praticadas pelos contribuintes.

O problema é tão antigo quanto à tributação em si e, ainda hoje é objeto de incessantes buscas de solução em virtude da complexidade do tema que abrange, dentre inúmeras outras, questões de natureza ética e sociológicas.

A experiência adquirida em nove anos de atividades junto ao órgão responsável pela arrecadação de um dos mais importantes estados da Federação demonstra que pelo menos um dos fatores que contribuem para gerar o comportamento sonegador é a ineficiência da máquina fiscal.

Registre-se que não cabe aqui atribuir o insucesso dos procedimentos de arrecadação ao trabalho desenvolvido pelos profissionais do Fisco.

Isto porque a experiência comprova que nem mesmo o esforço, competência e dedicação de tais especialistas são suficientes para gerar um índice de sucesso mais significativo na identificação de comportamentos fraudulentos por parte dos contribuintes.

Tomando-se como exemplo o ICMS (Imposto sobre a Circulação de Mercadorias e Serviços) chega-se a conclusão de que na avaliação do comportamento da receita por ele gerada, todos os métodos convencionais de detecção de fraudes, sejam eles detalhados com a análise dos pagamentos efetuados ou globais através de comparações entre valores da mesma receita, apresentam resultados insatisfatórios

A razão do insucesso e inadequação atual de tais métodos é relativamente singela, vale dizer, existem comportamentos fraudulentos cuja identificação não se faz possível través da análise humana

Dai vem a proposta desta monografia que é evidenciar o sucesso da utilização de uma das ferramentas de BI, ou seja, o Datamining, para identificar os comportamentos fraudulentos praticados pelos contribuintes e que escapam da análise humana.

Os algoritmos de datamining propiciarão a identificação de situações e a solução de problemas tomando como ponto de partida a observação de uma série de parâmetros não lineares e complexos, sobretudo onde o ambiente seja dinâmico.

O exemplo escolhido e sobre o qual se pretende discorrer é, mais uma vez, o da evolução do ICMS em determinado estado brasileiro, a partir de agora denominado Estado X.

Convêm ressaltar que a base de dados utilizada é real, porém, em respeito aos princípios da ética e sigilo fiscal, todas as informações sobre identificação dos contribuintes (inscrição estadual) e valores arrecadados foram alterados.

Com base no histórico de pagamentos efetuados nos últimos cinco anos a título de ICMS pelos maiores contribuintes da rede varejista e de supermercados, será apresentada uma solução que propõe uma forma alternativa de “separar” os

comportamentos de fraude nos pagamentos daqueles que, a princípio, aparentam estar corretos.

Com base em um histórico de autos de infração, nesta solução a mineração de dados será capaz classificar contribuintes que deverão migrar para a “malha fina” da fiscalização.

1.1 MOTIVAÇÃO

Maior eficácia dos modelos de fiscalização do pagamento do Imposto sobre Circulação de Mercadorias e Serviços (ICMS) dos contribuintes do Estado X, utilizando a mineração de dados (Data Mining), através da utilização de Redes Neurais artificiais MLP, algoritmos BI1, J48e Nayve Baines.

Estes algoritmos serão capazes de identificar a ocorrência ou não de possível fraude no pagamento de ICMS, por parte de um específico contribuinte em determinado período mês/ano de arrecadação.

1.2 OBJETIVOS DO TRABALHO

Estudar o comportamento histórico dos pagamentos mensais dos maiores contribuintes de ICMS da rede varejista e de supermercados do Estado X, relativos aos últimos cinco anos (2008, 2007, 2006, 2005, 2004), para que se possa classificá-los de forma a tornar possível a identificação de padrões e comportamentos fraudulentos.

1.3 DESCRIÇÃO DO TRABALHO

Apresentação da Base de Dados do Data warehouse de Arrecadação da Secretaria de Fazenda do Estado X

Nesta etapa será apresentado o universo de informações contidas no Data warehouse da Secretaria, mostrando os atributos desse modelo, mais especificamente no datamart de arrecadação dos contribuintes de ICMS.

Seleção de Variáveis Necessárias para a mineração de dados

Analise e coleta de atributos pertinentes para a modelagem da rede neural, tendo como base atributos qualitativos e quantitativos do datamart de arrecadação

Extração das informações do DW (preservando o sigilo fiscal)

As informações extraídas serão correspondentes aos últimos cinco anos de arrecadação (de 2008 até 2004), agregadas mensalmente, relativas aos maiores contribuintes de ICMS da rede varejista e de supermercados do Estado X .

Limpeza das Variáveis

São mostrados estudos com os atributos escolhidos para modelagem onde serão feitos a mineração dos dados. Esse estudo vai ser feito via seleção de registros do banco de dados, e posteriormente inseridos no software Weka, onde se procurará as melhores variáveis para a analise da detecção de fraudes.

Modelagem da Solução

Toda a modelagem da rede será feita no software Weka, com base nas informações anteriormente analisadas. Nesta fase todos os atributos serão tratados e convertidos de modo a facilitar o treinamento dos algoritmos para que sejam feitas as melhores classificações possíveis.

Seleção de Variáveis para cada tipo de modelo

A seleção de variáveis para cada modelo será feita objetivando analisar os componentes principais e ranquear os atributos de melhores resultados no modelo proposto. Para isso serão utilizados diversos métodos tais como LSE, SIE etc....

Análise dos Resultados

Os resultados das classificações serão analisados para assim se conseguir entender os padrões de fraude no pagamento do ICMS.

1.4. ORGANIZAÇÃO DA MONOGRAFIA

Esta monografia esta organizada nos seguintes capítulos:

O capítulo 2 apresenta a descrição do problema proposto para a classificação dos contribuintes “fraudadores”.

O capítulo 3 apresenta as principais metodologias utilizadas no desenvolvimento da modelagem da solução de dataminig.

O capítulo 4 apresenta a arquitetura da Rede Neural e os outros algoritmos de classificação, assim como o detalhamento dos procedimentos adotados para a sua modelagem.

O capítulo 5 detalha os resultados obtidos durante o trabalho, identificando as melhores soluções de classificação.

Finalmente, o capítulo 6 mostra as conclusões e identifica possíveis trabalhos futuros.

2. DESCRIÇÃO DO PROBLEMA

Classificar padrões de comportamentos fraudulentos no pagamento de ICMS, por parte dos maiores contribuintes da rede varejista e de supermercados do Estado X.

Para se descobrir isso será feita uma rede neural, com treinamento supervisionado, que irá procurar classificar os contribuintes infratores, com base em um histórico de infrações cometidas.

As informações analisadas terão como base os seguintes itens abaixo:

- Informações de cadastro dos Contribuintes
- Receitas recolhidas
- Períodos de recolhimento (mês/ano)
- Código nacional de atividade fiscal (CNAEF)
- Valores totais recolhidos (agregado daquele mês/ano)
- Dados sobre infrações dos contribuintes

Estas informações serão agregadas por valor total arrecadado no período correspondente ao mês/ano, dos últimos cinco anos, o que equivale a +ou- 30.000 registros.

3. REDES NEURAIS ARTIFICIAIS E O MODELO DE CLASSIFICAÇÃO DOS CONTRIBUINTES DO ESTADO X

Nesta parte é mostrada a metodologia empregada no trabalho, além dos critérios e medidas usadas para a elaboração dos resultados.

Este trabalho usou os dados dos contribuintes de ICMS do Estado X. Estes dados foram extraídos do Data warehouse do estado de fazenda do Estado X, referentes aos anos de 2004, 2005, 2006, 2006, 2007. Estas informações estão no escopo dos maiores contribuintes arrecadadores do setor varejista e de supermercados.

3.1. MÉTODO PROPOSTO PARA CLASSIFICAÇÃO

A seguir serão mostradas todas as fases da pesquisa, descrevendo cada passo do processo de aprendizagem da rede neural artificial e os outros algoritmos de classificação que serão mostrados no capitulo seguinte.

Fase 1 – Modelagem das Variáveis de Avaliação

Nesta fase são definidas todas as variáveis de entrada da rede neural. As seguintes variáveis foram selecionadas como entrada:

Variáveis de entrada :

- Inscrição (Criptografado),
- Tipo de Contribuinte,
- CNAEF Principal,
- Código de Receita,

- Situação de Cadastro do Contribuinte,
- Município,
- Inspetoria,
- Mês_Ano de Recolhimento,
- Valor Total Arrecadado.

Variável de Saída :

- Indicador de Auto de Infração.

As variáveis acima serão melhores explicitadas no capitulo seguinte (Arquitetura).

Fase 2 – Definição da Topologia e do Algoritmo de Aprendizado da Rede Neural Artificial / Algoritmos de Classificação.

A finalidade da fase dois é selecionar a topologia e estabelecer os índices de desempenho desta rede.

Para este trabalho optou-se pelas redes neurais artificiais com treinamento Supervisionado, onde o objeto do trabalho é a classificação e a análise dos comportamentos dos pagamentos de ICMS, tendo em vista que se sabe de ocorrência ou não da infração naquele determinado mês.

Sendo mais específico optou-se para este trabalho os algoritmos de Classificação IB1, J48, Nayve Baines e a rede neural MLP.

O programa Weka foi escolhido por reunir grande quantidade de variáveis de entrada, baixo custo e uma amigável interface gráfica.

3.2. MÉTODO PARA DESENVOLVIMENTO DE APLICAÇÕES DE DATAMINING COM TREINAMENTO SUPERVISIONADO

Etapa 1 – Coleta dos Dados.

Etapa 2 – Modelagem do Sistema (banco de dados e codificação das variáveis).

Etapa 3 – Resultados Obtidos.

Na etapa um são coletados os dados de arrecadação dos contribuintes do Estado X, dentro do universo dos maiores contribuintes de ICMS do setor varejista e de supermercados dos últimos cinco anos.

Na etapa dois são analisadas as variáveis existentes sobre estes pagamentos, assim como as características dos contribuintes.

Na etapa três é feita uma análise sobre os resultados de cada método usado.

Na etapa quatro é feita uma conclusão sobre o melhor resultado obtido na classificação dentre os algoritmos IB1, J48, Nayve Baines e a rede MLP.

4. MODELO DE REDE NEURAL ARTIFICIAL/ALGORITMOS DE DATAMINING PROPOSTOS

Neste capitulo será apresentada a arquitetura usada, assim como os algoritmos de aprendizado.

Entre os numerosos algoritmos de aprendizado propostos para a aplicação de classificação, resolvi aplicar quatro algoritmos muito eficazes, que são : IB1, J48 e Nayve Baines, além das redes Multi-Layer Perceptrons.

O IB1 utiliza o algoritmo nearest-neighbor cuja técnica principal se baseia em classificar uma nova instância considerando-se a classe de sua vizinha mais próxima. O algoritmo IB1 armazena todas as instâncias de treinamento, que são processadas incrementalmente.

O algoritmo J48 é uma implementação do algoritmo C4.5 release 8 que, gera árvore de decisão e é considerado o mais popular algoritmo da Weka. O J48 constrói um modelo de árvore de decisão baseado num conjunto de dados de treinamento, sendo que esse modelo é utilizado para classificar as instâncias do conjunto de teste.

O algoritmo Naïve Bayes é um algoritmo de para uso em modelagem de previsão. O nome Naive Bayes foi atribuído pelo fato de o algoritmo usar o teorema de Bayes, mas não considerar dependências que possam existir. Portanto, suas suposições são consideradas ingênuas. Esse algoritmo é computacionalmente menos intenso de que outros algoritmos da Microsoft e, portanto, é útil para gerar modelos de mineração rapidamente para descobrir as relações entre as colunas de entrada e as colunas previsíveis.

O algoritmo da rede Multi-Layer Perceptron são redes de múltiplas camadas que podem resolver funções lineares e não-lineares, onde deve-se minimizar o erro de todos os processadores da camada de saída, para todos os padrões. Foi demonstrado que a Multi-Layer perceptron é um aproximador universal, isto é, pode representar qualquer função.

4.1. ARQUIVOS DE DADOS GERADOS

Considerando que os dados necessários a este estudo se encontravam inicialmente disponíveis apenas em um sistema gerenciador de banco de dados e linguagem de programação (Oracle e PL/SQL) não adequada ao tratamento estatístico de dados, tornou-se necessário iniciar o processamento desses dados nesse mesmo ambiente, viabilizando a geração de um arquivo com a finalidade de armazenar os registros a serem utilizados.

Este arquivo foi formatado de modo que pudesse ser transferido, sem maiores esforços, para plataformas menores (PC-Windows), onde existe maior disponibilidade de softwares aplicativos. Esta premissa orientou a modelagem de dados realizada e o planejamento das rotinas necessárias à geração desse banco de dados. O arquivo de dados foi transferido e transformado em uma tabela, para que pudesse ser trabalhado em uma instância Oracle Desktop.

O arquivo de dados foi gerado compondo tantos registros de pagamentos dos maiores contribuintes dos últimos cinco anos. Dentre as varáveis disponíveis no banco de dados que serão empregadas na modelagem e aprendizado da rede neural, fazendo parte de um conjunto de blocos, foram escolhidas aquelas julgadas mais significativas para medição do desempenho dos contribuintes de ICMS, separando-as em três grupos distintos:

- DADOS CADASTRAIS DO CONTRIBUINTE

1. *Inscrição (Criptografado, usado somente para a extração dos dados)* - Inscrição estadual do contribuinte de ICMS do Estado X. Para não violar o sigilo fiscal este número foi criptografado.

2. *Tipo de Contribuinte* - Identifica se o contribuinte é matriz ou filial

3. *Situação de Cadastro do Contribuinte* - Indica situação de cadastro atual do contribuinte, isto é, se ele está paralisado ou habilitado.

4. *Município* - Identifica o Município de localização do contribuinte, dentro do Estado X.

5. *Inspetoria* - Identifica a Inspetoria de fiscalização daquele contribuinte

- DADOS ECONÔMICO-FISCAIS

6. *CNAEF Principal* - Identifica a atividade econômica preponderante desenvolvida pelo contribuinte, classificando-a quanto à atividade, produto e público-alvo.

7. *Código de Receita* - Identifica o tipo de receita pago pelo contribuinte, referente aquele tributo de ICMS.

8. *Mês Ano de Recolhimento* - Mostra o período de referência de arrecadação daquele documento.

9. *Valor Total Arrecadado* - Identifica o total arrecadado (agregado pelos itens acima) naquele mês/ano.

- DADOS DE AUTO DE INFRAÇÕES

10. *Indicador de Auto de Infração* - Indica se ouve ou não infração para o imposto dentro daquele mês/ano.

Deste entendimento, verifica-se que o banco de dados que será adotado para a Rede Neural Artificial será uma matriz composta de +ou- 30.000 linhas (representando os registros agregados de pagamento dos últimos cinco anos dos maiores contribuintes de ICMS do setor varejista e de supermercados do Estado X) e nove colunas (cinco representando os dados cadastrais, quatro representando os dados econômicos fiscais e uma coluna indicando a se houve ou não infração). Essas nove colunas são identificadoras das variáveis de entrada e saída na Rede Neural, devendo existir, tantos neurônios quantos forem às variáveis de entrada/saída.

5. RESULTADOS

5.1 MODELAGEM DA SOLUÇÃO

1 – VARIÁVEIS DE ENTRADA

MUNICIPIO	ENTRADA
ANGRA DOS REIS	1
ARARUAMA	2
BARRA MANSA	3
BELFORD ROXO	4
CAMPOS DOS GOYTACAZES	5
DUQUE DE CAXIAS	6
ITABORAI	7
ITAGUAI	8
ITATIAIA	9
NILOPOLIS	10
NITEROI	11
NOVA IGUACU	12
PARAIBA DO SUL	13
PETROPOLIS	14
QUEIMADOS	15
RESENDE	16
RIO DAS OSTRAS	17
RIO DE JANEIRO	18
SAO GONCALO	19
SAO JOAO DE MERITI	20
SAO PEDRO DA ALDEIA	21
SAQUAREMA	22
TERESOPOLIS	23
TRES RIOS	24
VALENCA	25
VOLTA REDONDA	26
NÃO IDENTIFICADO	99

TIPO ESTABELECIMENTO	ENTRADA
ÚNICO	1
PRINCIPAL	2
NÃO IDENTIFICADO	3

SITUAÇÃO CONTRIBUINTE	ENTRADA
HABILITADO REGULAR	HR
IMPEDIDO	IP
PARALISADO	PL
SUSPENSO	SP

CÓDIGO NACIONAL DE ATIVIDADE ECONÔMICA E FISCAL	ENTRADA
ALUGUEL DE FITAS DE VÍDEO, DVDS E SIMILARES	1
COMERCIO VAREJISTA DE ARTIGOS DE ARMARINHO	2
COMÉRCIO ATACADISTA DE MERCADORIAS EM GERAL, SEM PREDOMINÂNCIA DE ALIMENTOS OU DE INSUMOS AGROPECUÁRIOS	3
COMÉRCIO ATACADISTA ESPECIALIZADO EM OUTROS PRODUTOS ALIMENTÍCIOS NÃO ESPECIFICADOS ANTERIORMENTE	4
COMÉRCIO VAREJISTA DE ARTIGOS DE JOALHERIA	5
COMÉRCIO VAREJISTA DE ARTIGOS DO VESTUÁRIO E ACESSÓRIOS	6
COMÉRCIO VAREJISTA DE CALÇADOS	7
COMÉRCIO VAREJISTA DE HORTIFRUTIGRANJEIROS	8
COMÉRCIO VAREJISTA DE MATERIAIS DE CONSTRUÇÃO EM GERAL	9
COMÉRCIO VAREJISTA DE MATERIAIS DE CONSTRUÇÃO NÃO ESPECIFICADOS ANTERIORMENTE	10
COMÉRCIO VAREJISTA DE MERCADORIAS EM GERAL, COM PREDOMINÂNCIA DE PRODUTOS ALIMENTÍCIOS – HIPERMERCADOS	11
COMÉRCIO VAREJISTA DE MERCADORIAS EM GERAL, COM PREDOMINÂNCIA DE PRODUTOS ALIMENTÍCIOS – MINIMERCADOS, MERCEARIAS E ARMAZÉNS	12
COMÉRCIO VAREJISTA DE MERCADORIAS EM GERAL, COM PREDOMINÂNCIA DE PRODUTOS ALIMENTÍCIOS – SUPERMERCADOS	13
COMÉRCIO VAREJISTA DE MÓVEIS	14
COMÉRCIO VAREJISTA DE OUTROS PRODUTOS NÃO ESPECIFICADOS ANTERIORMENTE	15
COMÉRCIO VAREJISTA DE TECIDOS	16

COMÉRCIO VAREJISTA ESPECIALIZADO DE ELETRODOMÉSTICOS E EQUIPAMENTOS DE ÁUDIO E VÍDEO	17
CONFECÇÃO DE PEÇAS DO VESTUÁRIO, EXCETO ROUPAS ÍNTIMAS E AS CONFECCIONADAS SOB MEDIDA	18
FORNECIMENTO DE ALIMENTOS PREPARADOS PREPONDERANTEMENTE PARA EMPRESAS	19
LANCHONETES, CASAS DE CHÁ, DE SUCOS E SIMILARES	20
LOJAS DE DEPARTAMENTOS OU MAGAZINES	21
NÃO IDENTIFICADO	99

TIPO DE RECEITA PAGO	ENTRADA
ICMS NORMAL	213
ICMS ESTIMATIVA	221
ICMS SUBSTITUICAO TRIBUTARIA	230
ICMS IMPORTACAO	248
ICMS AQS. AT. FIXO OU MAT. FORA ESTADO	272
ICMS PARCELAMENTO	280
ICMS AUTO DE INFRACAO	302
ICMS SERVICOS DE TRANSPORTE	361
ICMS OUTROS	370
ICMS AUTO DE INFRACAO PARCELAMENTO	396
ICMS AUTO DE INFRAÇÃO ANISTIA	469
ICMS AUTO DE INFRAÇAO PARCELAMENTO ANISTIA	477
TAXA DE SERVIÇOS ESTADUAIS - NATUREZA FAZENDÁRIA	2003
TAXA DE SERVIÇOS ESTADUAIS/NATUREZA Ñ FAZENDÁRIA	2011
TAXA DE SERVIÇOS ELETRÔNICOS - NATUREZA FAZENDÁRIA	2020
DIVIDA ATIVA ICM	5002
DIVIDA ATIVA ICMS	5029
DIVIDA ATIVA ICMS PARCELAMENTO	5037
DIVIDA ATIVA OUTROS	5070
DÍVIDA ATIVA NÃO TRIBUTÁRIA	5070
DIVIDA ATIVA OUTROS PARCELAMENTO	5088
DÍVIDA ATIVA NÃO TRIBUTÁRIA - PARCELAMENTO	5088
DÍVIDA ATIVA TRIBUTÁRIA - OUTROS	5207
DÍVIDA ATIVA TRIBUTÁRIA - OUTROS - PARCELAMENTO	5215
DÍVIDA ATIVA IPVA	5223

DIVIDA ATIVA ICMS ANISTIA	5320
DIVIDA ATIVA ICMS – RESOL. SEF 6490/02	5371
MULTA FORMAL ICMS PARCELAMENTO	5398
MULTA FORMAL ICMS	5517
DÍVIDA ATIVA ICMS FECP	5720
DÍVIDA ATIVA ICMS FECP PARCELAMENTO	5738
ICMS FECP	7501
ICMS FECP PARCELAMENTO	7510
ICMS FECP AUTO DE INFRAÇÃO	7528
DEPOSITO RECURSAL	9113
OUTRAS RECEITAS	9997

MÊS	ENTRADA
JANEIRO	1
FEVEREIRO	2
MARÇO	3
ABRIL	4
MAIO	5
JUNHO	6
JULHO	7
AGOSTO	8
SETEMBRO	9
OUTUBRO	10
NOVEMBRO	11
DEZEMBRO	12

ANO	ENTRADA
2004	1
2005	2
2006	3
2007	4
2008	5

2 – VARIÁVEL DE SAÍDA

INDICADOR DE INFRAÇÃO	ENTRADA
NÃO COMETEU INFRAÇÃO	0
COMETEU INFRAÇÃO	1

5.2 DISTRIBUIÇÃO DOS VALORES

Município		
No.	Label	Count
1	1	690
2	2	46
3	3	206
4	4	448
5	5	1151
6	6	1769
7	7	202
8	8	276
9	9	45
10	10	320
11	11	1815
12	12	1224
13	13	194
14	14	550
15	15	206
16	16	169
17	17	77
18	18	16697
19	19	1303
20	20	934
21	21	117
22	22	89
23	23	168
24	24	54
25	25	349

26	26	1728
27	27	0

Tipoestab		
No.	Label	Count
1	1	2005
2	2	28822
3	3	0

Situação		
No.	Label	Count
1	HR	29900
2	IP	84
3	PL	570
4	SP	273

Cnaef		
No.	Label	Count
1	1	140
2	2	180
3	3	763
4	4	178
5	5	121
6	6	3290
7	7	693
8	8	919
9	9	462
10	10	653
11	11	798
12	12	115
13	13	9776
14	14	188
15	15	149
16	16	107
17	17	2745
18	18	169

19	19	388
20	20	3812
21	21	5181
22	99	0

Receita

No.	Label	Count
1	1	10132
2	2	3
3	3	2847
4	4	104
5	5	3900
6	6	453
7	7	221
8	8	64
9	9	12
10	10	148
11	11	4
12	12	5
13	13	3095
14	14	75
15	15	377
16	16	1
17	17	3
18	18	6
19	19	5
20	20	53
21	21	4
22	22	18
23	23	1
24	24	47
25	25	59
26	26	41
27	27	411
28	28	1
29	29	3

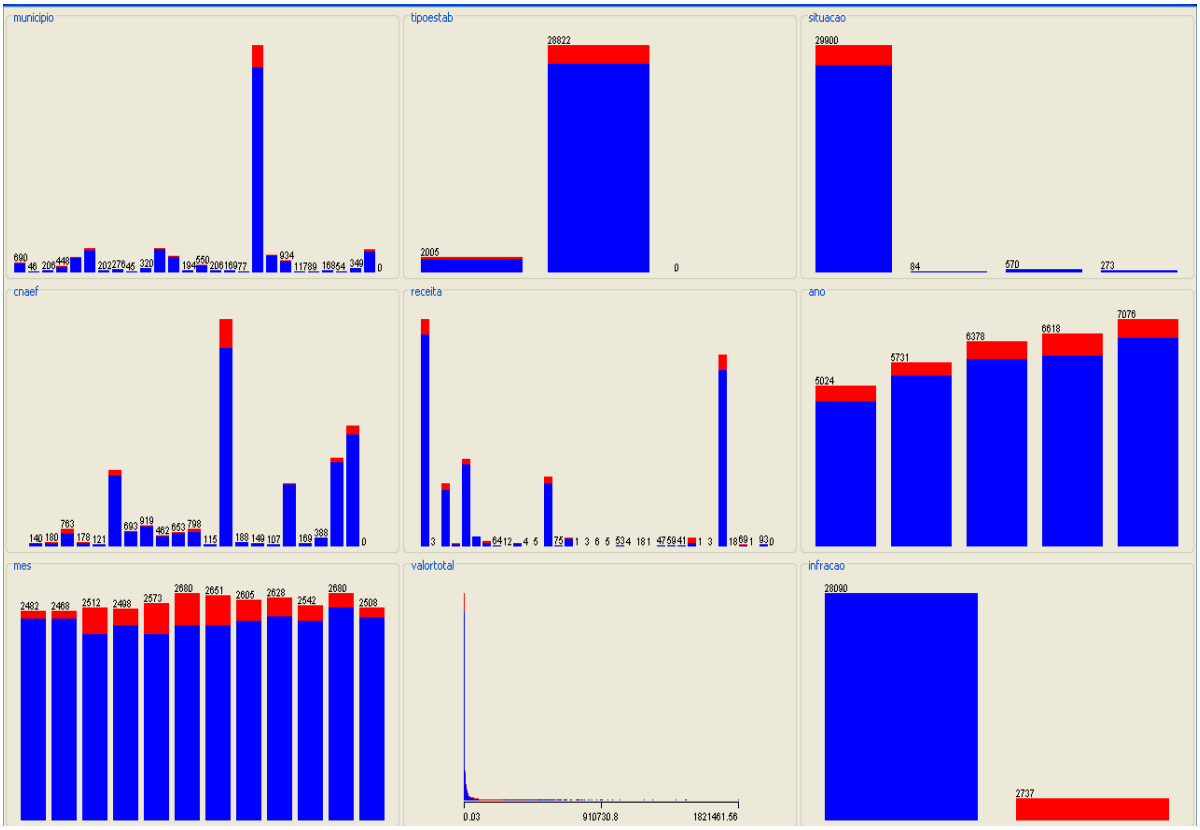
30	30	8553
31	31	18
32	32	69
33	33	1
34	34	93
35	99	0

Ano		
No.	Label	Count
1	1	5024
2	2	5731
3	3	6378
4	4	6618
5	5	7076

Mês		
No.	Label	Count
1	1	2482
2	2	2468
3	3	2512
4	4	2498
5	5	2573
6	6	2680
7	7	2651
8	8	2605
9	9	2628
10	10	2542
11	11	2680
12	12	2508

Valor total	
Statistic	Value
Minimum	0,03
Maximum	1821461,56
Mean	15411,94
StdDev	55432,57

Infração		
No.	Label	Count
1	0	28090
2	1	2737



5.3 SELEÇÃO DE VARIÁVEIS

Rank	GainRatioAttributeEval	InfoGainAttributeEval	OneRAttributeEval	ReliefAttributeEval
1	0.010121 5 receita	0.026434 5 receita	91.462 5 receita	0.21652 7 mes
2	0.007222 2 tipoestab	0.020965 4 cnaef	91.404 8 valortotal	0.21525 4 cnaef
3	0.006653 4 cnaef	0.013004 7 mes	91.121 3 situacao	0.14037 6 ano
4	0.003628 7 mes	0.005718 1 municipio	91.121 1 municipio	0.12825 1 municipio
5	0.002639 3 situacao	0.002507 2 tipoestab	91.121 2 tipoestab	0.04959 5 receita
6	0.002037 1 municipio	0.001043 6 ano	91.121 7 mes	0.00938 3 situacao
7	0.001784 8 valortotal	0.001033 8 valortotal	91.121 4 cnaef	0.00791 2 tipoestab
8	0.000451 6 ano	0.000614 3 situacao	91.121 6 ano	0.00129 8 valortotal

5.4 RESULTADOS DOS TREINAMENTOS

MultilayerPerceptron

Scheme: weka.classifiers.functions.MultilayerPerceptron -L 0.3 -M 0.2 -N 500 -V 0 -S 0 -E 20 -H a

Relation: HMEQ_Treinamento_Monog

Instances: 30827

Attributes: 9

municipio

tipoestab

situacao

cnaef

receita

ano

mes

valortotal

infracao

Test mode: evaluate on training data

Time taken to build model: 3303.36 seconds

=== Evaluation on training set ===

=== Summary ===

Correctly Classified Instances	**28809**	**93.4536 %**
Incorrectly Classified Instances	**2018**	**6.5464 %**
Kappa statistic	**0.4321**	
Mean absolute error	**0.0697**	
Root mean squared error	**0.2472**	
Relative absolute error	**42.1188 %**	
Root relative squared error	**85.9219 %**	
Total Number of Instances	**30827**	

=== Detailed Accuracy By Class ===

	TP Rate	**FP Rate**	**Precision**	**Recall**	**F-Measure**	**ROC Area**	**Class**
	0.998	**0.697**	**0.935**	**0.998**	**0.965**	**0.831**	**0**
	0.303	**0.002**	**0.933**	**0.303**	**0.457**	**0.831**	**1**
Weighted Avg.	**0.935**	**0.634**	**0.934**	**0.935**	**0.919**	**0.831**	

=== Confusion Matrix ===

a	**b**	**<-- classified as**
27835	**66**	**a = 0**
1952	**974**	**b = 1**

NaiveBayes

Scheme: weka.classifiers.bayes.NaiveBayes

Relation: HMEQ_Treinamento_Monog

Instances: 30827

Attributes: 9

municipio

tipoestab

situacao

cnaef

receita

ano

mes

valortotal

infracao

Test mode: evaluate on training data

=== Evaluation on training set ===

=== Summary ===

Correctly Classified Instances	**28115**	**91.2025 %**
Incorrectly Classified Instances	**2712**	**8.7975 %**
Kappa statistic	**0.1671**	

Mean absolute error **0.1479**

Root mean squared error **0.2718**

Relative absolute error **91.3657 %**

Root relative squared error **95.5669 %**

Total Number of Instances **30827**

=== Detailed Accuracy By Class ===

	TP Rate	FP Rate	Precision	Recall	F-Measure	ROC Area	Class
	0.989	0.88	0.92	0.989	0.953	0.751	0
	0.12	0.011	0.52	0.12	0.195	0.751	1
Weighted Avg.	0.912	0.803	0.885	0.912	0.886	0.751	

=== Confusion Matrix ===

a	b	<-- classified as
27787	303	a = 0
2409	328	b = 1

IB1

Scheme: **weka.classifiers.lazy.IB1**

Relation: **HMEQ_Treinamento_Monog**

Instances: **30827**

Attributes: **9**

municipio

tipoestab

situacao

cnaef

receita

ano

mes

valortotal

infracao

Test mode: evaluate on training data

=== Classifier model (full training set) ===

IB1 classifier

Time taken to build model: 0.06 seconds

=== Evaluation on training set ===

=== Summary ===

Correctly Classified Instances	**30771**	**99.8183 %**
Incorrectly Classified Instances	**56**	**0.1817 %**
Kappa statistic	**0.9888**	
Mean absolute error	**0.0018**	
Root mean squared error	**0.0426**	
Relative absolute error	**1.1225 %**	

Root relative squared error **14.9846 %**

Total Number of Instances **30827**

=== Detailed Accuracy By Class ===

	TP Rate	FP Rate	Precision	Recall	F-Measure	ROC Area	Class
	0.999	0.01	0.999	0.999	0.999	0.994	0
	0.99	0.001	0.99	0.99	0.99	0.994	1
Weighted Avg.	0.998	0.009	0.998	0.998	0.998	0.994	

=== Confusion Matrix ===

a	b	<-- classified as
28062	28	a = 0
28	2709	b = 1

J48

Scheme: **weka.classifiers.trees.J48 -C 0.25 -M 2**

Relation: **HMEQ_Treinamento_Monog**

Instances: **30827**

Attributes: **9**

municipio

tipoestab

situacao

cnaef

receita

ano

mes

valortotal

infracao

Test mode: evaluate on training data

=== Evaluation on training set ===

=== Summary ===

Correctly Classified Instances	**28383**	**92.0719 %**
Incorrectly Classified Instances	**2444**	**7.9281 %**
Kappa statistic	**0.1927**	
Mean absolute error	**0.1442**	
Root mean squared error	**0.2685**	
Relative absolute error	**89.0834 %**	
Root relative squared error	**94.3904 %**	
Total Number of Instances	**30827**	

=== Detailed Accuracy By Class ===

TP Rate	FP Rate	Precision	Recall	F-Measure	ROC Area	Class
0.999	0.882	0.921	0.999	0.958	0.623	0
0.118	0.001	0.915	0.118	0.209	0.623	1

Weighted Avg. 0.921 0.804 0.92 0.921 0.892 0.623

=== Confusion Matrix ===

a b <-- classified as

28060 30 | a = 0

2414 323 | b = 1

6. CONCLUSÕES E TRABALHOS FUTUROS

Para o caso em estudo, o algoritmo IB1 se mostrou mais eficiente e com menor índice de erros. Os outros algoritmos se mostraram também eficazes, porem menos eficientes no tempo de resultados.

Obtendo-se novas informações sobre características sobre os contribuintes, assim como novas informações tributárias, futuramente será possível identificar novos padrões de classificação com a ajuda de novos algoritmos e a adoção de uma nova rede neural.

REFERÊNCIAS BIBLIOGRÁFICAS

1. CONTRERAS, J.C.S. *Previsão de arrecadação do ICMS através de redes neurais no Brasil*. Acessado em 20/04/2009. Disponível em http://www.gprocura.com.br/dp/_23361/Previsao-de-arrecadacao-do-ICMS-atraves-de-redes-neurais-no-Brasil.html

2. SISNANDO, S. R. A. & FREITAS, M. A. S. *Previsão e Avaliação do Desempenho dos Contribuintes do ICMS do Estado do Ceará Utilizando as Redes Neurais Artificiais. Acessado em 20/04/2009. Disponível em* http://www.bnb.gov.br/content/aplicacao/Publicacoes/REN-Numeros_Publicados/docs/ren2006_v37_n1_a8.pdf

3. MILAGRE, J. A. Fraudes dificultam perícia em notas fiscais eletrônicas. Acessado em 20/04/2009. Disponível em http://webinsider.uol.com.br/index.php/2008/11/18/fraudes-dificultam-pericia-em-nota-fiscais-eletronicas-nf-e/

4. SANTOS, R. A. F. *Uso de Redes Neurais Artificiais na Detecção de Fraudes*. Acessado em 20/04/2009. Disponível em http://www.serasa.com.br/revista/revista5.htm

5. ICMS - IMPOSTO SOBRE CIRCULAÇÃO DE MERCADORIAS E PRESTAÇÃO DE SERVIÇOS. Acessado em 20/04/2009. Disponível em http://www.portaltributario.com.br/tributario/legislacao_icms.htm

6. VELASCO, Marley. *Apostila curso BI-Master PUC-Rio*. Rio de Janeiro: PUC-Rio,

7. KIMBALL, R. & ROSS, M. *The Data Warehouse Toolkit: The Complete Guide to Dimensional Modeling* (Second Edition ed.). New York: Wiley, 2002.

8. PACHECO, M. A. & VELASCO, M. *Sistemas Inteligentes de apoio a Decisão*. Rio de Janeiro: PUC-Rio, 2007.

9. INMOM, W.H. & TERDEMAN, R. H. *Exploration Warehousing: Turning Business Information into Business Opportunity*. New York: John Wiley and Sons, 2000.