Universidade de Brasília
Instituto de Ciências Exatas
Departamento de Ciência da Computação

Mineração de dados como suporte à detecção de lavagem de dinheiro

Ebberth Lopes de Paula

Dissertação apresentada como requisito parcial para conclusão do
Mestrado Profissional em Computação Aplicada

Orientador
Prof. Dr. Marcelo Ladeira

Brasília
2016

Ficha catalográfica elaborada automaticamente,
com os dados fornecidos pelo(a) autor(a)

PP324m

Paula, Ebberth Lopes de
Mineração de dados como suporte à detecção de lavagem de dinheiro nas exportações / Ebberth Lopes de Paula; orientador Marcelo Ladeira. -- Brasília, 2016.
114 p.

Dissertação (Mestrado - Mestrado Profissional em Computação Aplicada) -- Universidade de Brasília, 2016.

1. Aprendizagem supervisionada. 2. Autoencoder. 3. Redes neurais profundas. 4. Combate à lavagem de dinheiro. 5. Exportações. I. Ladeira, Marcelo, orient. II. Título.

Universidade de Brasília
Instituto de Ciências Exatas
Departamento de Ciência da Computação

Mineração de dados como suporte à detecção de lavagem de dinheiro

Ebberth Lopes de Paula

Dissertação apresentada como requisito parcial para conclusão do
Mestrado Profissional em Computação Aplicada

Prof. Dr. Marcelo Ladeira (Orientador)
CIC/UnB

Prof. Dr. Rommel Novaes Carvalho
CIC/UnB

Dr. Igor Assis Braga
Big Data Assessoria Empresarial

Prof. Dr. Marcelo Ladeira
Coordenador do Programa de Pós-graduação em Computação Aplicada

Brasília, 15 de dezembro de 2016

Dedicatória

Se nos alegramos por aquilo que nos dá potência, então não haveriam outros a quem dedicar este trabalho.

Se amamos o que desejamos e se amamos também o que se faz presente no tempo presente, então não haveriam outros a quem dedicar este trabalho.

Se nos sacrificamos pelo que nos é sacro, então também não haveriam outros a quem dedicar este trabalho.

Este trabalho é dedicado à Juliana, Bruno, Victor, Maria Eduarda e às *two the single ladys*.

Agradecimentos

São tantos aqueles que tornaram este trabalho possível que não conseguiria agradecer a todos nominalmente. Escolho nomear alguns, na esperança que a gratidão a estes aqui expressa transcenda aos demais.

Agradeço a todos amigos e colegas da RFB que de alguma forma colaboraram com este trabalho, em especial os colegas da Dipes e da Coana; ao Marcelo Lingerfelt pelas correções, revisões, paciência e incentivo, e ao Leon Solon sem o qual este trabalho definitivamente não existiria.

Aos colegas de mestrado, em especial aos mineradores, por todos momentos intensos, maravilhosos e inesquecíveis que vivemos juntos.

À Universidade de Brasília e a todos seus professores, funcionários e alunos, em especial ao meu orientador, professor Marcelo Ladeira pela sua paciência, dedicação a este trabalho e eterno bom humor; e aos professores do PPCA, pela forma como exercem sua profissão dignificando aqueles que se dignam a aprender.

A todos que trabalham em prol de uma sociedade melhor, em especial ao coordenador do PPCA, professor Marcelo Ladeira, e ao professor Rommel Novaes que de forma silenciosa têm introduzido no serviço público brasileiro a semente revolucionária da mineração de dados, cujos frutos certamente já estamos colhendo.

A minha família, em especial a minha esposa, companheira, mulher e amada, Juliana, pelo apoio incondicional, pelos incentivos nos momentos difíceis, pela paciência e amor demonstrados ao longo deste trabalho; a meus filhos, Bruno e Victor, por aquilo que é inominável no aprendizado do dia a dia e que certamente está presente nessas linhas; a minha filha, Maria Eduarda, por ter surgido no decorrer deste trabalho trazendo mais alegria e com isso mais potência para o trabalho; ao caríssimo Warton Monteiro, pelo apoio, por dividir suas histórias acadêmicas, e pelas projeções que em mim representa e que neste trabalho se fizeram presentes.

Resumo

Este trabalho apresenta o uso de técnicas de mineração de dados para detecção de empresas exportadoras brasileiras suspeitas de operarem exportações fictícias e consequente incorrência no crime de lavagem de dinheiro. A partir de estudos de aprendizagem de máquina com algoritmos supervisionados, foi desenvolvido um modelo capaz de classificar empresas suspeitas de operarem exportações fictícias. Em paralelo, foram desenvolvidos ainda estudos não supervisionados com *Deep Learning Autoencoder* e identificado um padrão de relacionamento entre os atributos numéricos representativos dos dados econômicos, mercantis, tributários e sociais das empresas que permitem a identificação de anomalias em dados de outras empresas. As empresas identificadas a partir do modelo supervisionado proposto neste trabalho foram submetidas à área específica de fiscalização aduaneira dentro da RFB e julgadas aptas a integrarem a programação de seleção para fiscalizações no ano de 2017. A metodologia desenvolvida, seus resultados e sua aplicabilidade foram divulgadas a todos escritórios de pesquisa e investigação da RFB por meio de Informação de Pesquisa e Investigação (IPEI). Um estudo de caso apresentando a metodologia aqui desenvolvida está previsto para ocorrer no 1º Encontro Nacional da RedeLab de 2017. Melhorias futuras a este trabalho incluem a detecção de anomalias e classificação de suspeição na exportação com maior granularidade dos dados, permitindo a sua identificação independente da empresa: por exemplo, a partir de transações, por rotas de produtos ou por tipo de mercadoria.

Palavras-chave: Aprendizagem supervisionada, Autoencoder, Rede neurais profundas, Combate à lavagem de dinheiro, Exportações

Abstract

This research presents the use of data mining techniques to detect brazilian exporting companies suspected of operating dummy exports and consequently incurring the crime of money laundering. Based on studies involving supervised analyzes, a model was developed capable of classifying companies suspected of operating dummy exports. Based on studies with *Deep Learning Autoencoder*, a pattern of relationship was identified between the numerical attributes representative of the economic and tax data of the companies. From this pattern, is possible to identify anomalies in data of another companies. The companies identified in this study were submitted to the specific area of customs supervision and found fit to integrate the selection schedule for inspections in the year 2017. The technique developed was disclosed to all investigation offices of the RFB through a document called IPEI. A case study presenting the methodology developed is expected to take place at the first national meeting of RedeLab 2017. Future improvements to this work include detection of anomalies and classification of export suspicious with greater granularity of the data, allowing them to be identified independently of the company: for example from transactions, product routes and by commodity type.

Keywords: Supervised learning, Deep Learning Autoencoder, Anti-money laundering, Exports

Sumário

[1] *Distributed Random Forest* é o nome dado à implementação da técnica *Random Forest* na plataforma *H2O*

Lista de Figuras

Lista de Tabelas

Lista de Abreviaturas e Siglas

AIRE Anomaly Index using Rank and Entropy.

AUC Área sob a curva ROC.

BArr Base Arrecadação.

BC Banco Central do Brasil.

BCad Base Cadastros.

BCE Base Comércio Exterior.

BCTBF Base Contribuições, Tributos e Benefícios Fiscais.

BEmp Base Empregados.

BMF Base Movimentações Financeiras.

BN Rede Bayesiana.

BNFe Base Notas Fiscais Eletrônicas.

BRIF Base Retenções de Impostos na Fonte.

COAF Conselho de Controle de Atividades Financeiras.

CRISP-DM Cross Industry Standard Process for Data Mining.

CTN Código Tributário Nacional.

Dacon Demonstrativo de Apuração de Contribuições Sociais.

DARF Documento de Arrecadação de Receitas Federais.

DBF Declaração de Benefícios Fiscais.

DBN Dynamic Bayesian Network.

DCTF Declaração de Contribuições Federais.

DE Declaração de Exportação.

DI Declaração de Importação.

DIMOF Declarações de Informações sobre Movimentação Financeira.

DIRF Declaração do Imposto de Renda Retido na Fonte.

DLA Deep Learning Autoencoder.

DRF Distributed Random Forest.

EART Euclidean Adaptive Resonance Theory.

fnr Taxa de Falsos Negativos.

fpr Taxa de Falsos Positivos.

GBM Gradient Boosting Machines.

GLM Generalized Linear Models.

GPS Guia da Previdência Social.

HPB Hierarchical Pattern Bayes.

Lab-LD Laboratório de Tecnologia Contra a Lavagem de Dinheiro.

LD Lavagem de Dinheiro.

mcc Matthews Correlation Coefficient.

MDIC Ministério da Indústria, Comércio Exterior e Serviços.

MJ Ministério da Justiça.

MSE Mean Square Error.

NCM Nomenclatura Comum do Mercosul.

NFe Notas Fiscais Eletrônicas.

Rede-LAB Rede Nacional de Laboratórios contra Lavagem de Dinheiro.

RFB Receita Federal do Brasil.

SARDBN Suspicious Activity Reporting using Dynamic Bayesian Network.

SERPRO Serviço Federal de Processamento de Dados.

Siscomex Sistema Integrado de Comércio Exterior Brasileiro.

SPED Sistema Público de Escrituração Digital.

SWRL Semantic Web Rule Language.

tnr Taxa de Verdadeiros Negativos.

tpr Taxa de Verdadeiros Positivos.

Capítulo 1

Introdução

Este capítulo introduz o problema que se pretende abordar como tema da pesquisa de mestrado. Inicialmente é apresentada a definição do problema e, a seguir, são detalhadas a justificativa do tema, os objetivos que se pretende alcançar e as contribuições esperadas.

1.1 Definição do problema

Pretende-se desenvolver um modelo de mineração de dados para apoio à seleção de exportadores pessoas jurídicas suspeitos de Lavagem de Dinheiro (LD), isto é, empresas de dentro do Brasil que atuam no comércio exterior promovendo exportações fictícias.

Serão analisados os dados fiscais e econômicos das empresas brasileiras exportadoras de bens e mercadorias de quaisquer espécies e origens que realizaram diretamente operações no comércio exterior durante o ano calendário de 2014 e 2015 (parcial).

Serão utilizadas as bases de dados disponíveis na Receita Federal do Brasil (RFB) relativas ao comércio exterior (sistema *Siscomex*) e comércio interno (sistema *SPED*); bases cadastrais (*CPF* e *CNPJ*); bases dos tributos internos administrados pela RFB; e bases de dados provenientes das informações fiscais declaradas por terceiros.

1.2 Justificativa do tema

O presente tema justifica-se sob diversos aspectos a seguir apresentados:

Inviabilidade do tratamento manual As exportações brasileiras são anualmente direcionadas a quase 200 países. Milhares de notas fiscais com suspensão de impostos nas mercadorias destinadas à exportação são diariamente emitidas. Cerca de 20.000 pessoas jurídicas operaram anualmente, direta ou indiretamente, no envio de bens e mercadorias ao exterior. A Nomenclatura Comum do Mercosul (NCM), utilizada para a classificação

fiscal das mercadorias, tem a capacidade de distinguir entre quase 10.000 tipos de bens e mercadorias, cada um deles sujeito potencialmente a uma legislação específica. Neste contexto de alta cardinalidade de atributos nominais, é desejável que a detecção de suspeição de atividades ilícitas se dê com auxílio de algum método automatizado.

Impacto nas relações econômicas De acordo com o Ministério da Indústria, Comércio Exterior e Serviços (MDIC), nos anos de 2014[1] e 2015[2], as exportações brasileiras somaram US$ 416,2 bilhões. Se aplicados ao Brasil, os percentuais de estimativas de lavagem de dinheiro mundial apresentadas na Seção 3.1, algo entre US$ 9 bilhões e US$ 22 bilhões seriam provenientes de dinheiro sujo.

Impacto na arrecadação de tributos A exportação fictícia, quando oriunda de bens e serviços efetivamente produzidos, implica na destinação do produto ao mercado interno informal (sem o pagamento de impostos) e, sendo o produto industrializado, no aproveitamento do crédito tributário da cadeia produtiva para dedução de outros impostos da indústria. Tal prática tem o potencial de reduzir significativamente os tributos arrecadados pelas administrações tributárias federal e estadual.

Ineditismo Dentro da estrutura orgânica da inteligência financeira nacional existem diversos esforços empreendidos pelo governo com vistas ao combate à lavagem de dinheiro [3]. No entanto, ainda não se realizou um trabalho de mineração de dados no domínio das exportações. Soma-se ainda que a RFB possui posição singular neste contexto, pois, além de deter informações do comércio exterior da parte da sua competência aduaneira, detém informações relativas a todos tributos internos de competência federal, tanto das pessoas físicas quanto das pessoas jurídicas.

1.3 Objetivos e contribuição esperada

Esta seção apresenta os objetivos do presente trabalho e a contribuição esperada com seu desenvolvimento.

[1] `http://www.mdic.gov.br/comercio-exterior/estatisticas-de-comercio-exterior/balanca-comercial-brasileira-mensal-2/2-uncategorised/1184-balanca-comercial-janeiro-dezembro-2014` - Acessado em 25/12/2016

[2] `http://www.mdic.gov.br/comercio-exterior/estatisticas-de-comercio-exterior/balanca-comercial-brasileira-mensal-2/2-uncategorised/1185-balanca-comercial-janeiro-dezembro-2015` - Acessado em 25/12/2016

[3] Para detalhes sobre esses esforços, sugere-se a visita ao site do Conselho de Controle de Atividades Financeiras (COAF) em `www.coaf.fazenda.gov.br`

Objetivo geral

Propor um modelo que, a partir da aplicação de técnicas de mineração de dados, classifique os contribuintes que operaram no comércio exterior em dois grupos: com e sem suspeita de lavagem de dinheiro.

Objetivos específicos

- Para os contribuintes suspeitos de lavagem de dinheiro na exportação, propor um índice que indique uma ordem de prioridade para a investigação pela RFB.
- Identificar os atributos mais relevantes para explicar a exportação fictícia.
- Analisar a sensibilidade de cada atributo preditivo do índice proposto para a seleção dos contribuintes suspeitos.

Contribuições esperadas

Conforme exposto na Seção 2.1, não foram identificados trabalhos que se utilizam de técnicas de mineração de dados para detecção de lavagem de dinheiro no comércio exterior. Desta forma, seguem duas contribuições tecnológicas que se espera com o presente trabalho.

- Desenvolvimento, com uso de técnicas de mineração de dados, de modelo preditivo inédito de identificação de exportadores suspeitos de operarem lavagem de dinheiro no comércio exterior.
- Desenvolvimento de índice que, a partir do modelo preditivo citado, indique uma ordem de prioridade para investigação e fiscalização pela RFB.

1.4 Estrutura deste documento

O Capítulo 2 traz, além de uma revisão do estado da arte dos trabalhos de mineração de dados no domínio das finanças e da detecção de fraudes, a fundamentação teórica das técnicas de mineração de dados aqui empregadas. O Capítulo 3 objetiva contextualizar a lavagem de dinheiro dentro de um panorama do comércio exterior e caracteriza-la tanto pelo aspecto legal quanto doutrinário. O Capítulo 4 traz a metodologia aplicada para a busca do atingimento dos objetivos propostos. Os Capítulos 5 a 8 apresentam o desenvolvimento do referencial metodológico *Cross Industry Standard Process for Data Mining* (CRISP-DM) aplicado ao presente trabalho, quais sejam, *entendimento do negócio*, *entendimento e preparação dos dados*, *indução do modelo e análise dos resultados*, *avaliação*

do modelo. Por fim, o Capítulo 9 apresenta as conclusões, os resultados alcançados e os trabalhos futuros. Nos Apêndices A e B é possível encontrar todo o código de programação utilizado no presente trabalho. O Apêndice C traz cópia do artigo submetido pelo autor, dentre outros, ao 15º *IEEE International Conference on Machine Learning and Applications*.

Capítulo 2

Fundamentação Teórica

Esse capítulo inicia-se com o levantamento do estado da arte das técnicas de mineração de dados no domínio do combate à lavagem de dinheiro e fraudes financeiras. A seguir são descritos os conceitos fundamentais do modelo de referência metodológica *Cross Industry Standard Process for Data Mining* (CRISP-DM) e das técnicas de mineração de dados *Gradient Boosting Machine* (GBM), *Random Forest* e *Deep Learning Autoencoder* (DLA).

2.1 Estado da arte

A partir dos trabalhos que têm sido feitos nas aplicações de técnicas de mineração de dados no domínio do combate à lavagem de dinheiro e fraudes, foram selecionadas aqueles que mais se aproximam do problema apresentado na Seção 1.1. Como se verá a seguir, a escassez de trabalhos específicos para mineração de dados visando a identificação de lavagem de dinheiro nas exportações levou a uma necessidade de ampliação da busca os seguintes temas correlatos:

1. Identificação de lavagem de dinheiro no comércio exterior;
2. Identificação de lavagem de dinheiro em geral;
3. Identificação de fraudes;
4. Mineração de dados no comércio exterior brasileiro;
5. Técnicas de detecção de anomalias em dados.

Identificação de lavagem de dinheiro no comércio exterior Esforços que envolvam o uso de mineração de dados em dados do comércio exterior na detecção de lavagem de dinheiro não foram identificados em artigos científicos publicados, mesmo quando amplia-se a pesquisa para mais de dez anos. Ressalta-se porém, que há referências do uso

de inteligência artificial para esse fim pelo *Financial Crimes Enforcement Network* [1] em 1995 [1] e 1998 [2]. A análise desses artigos mostra que eles não apresentam um nível de detalhamento relevante para esta pesquisa.

Identificação de lavagem de dinheiro em geral Ao se ampliar a busca por produções científicas que envolvem a identificação de lavagem de dinheiro em geral com o uso de mineração de dados, percebe-se que a análise de transações financeiras detêm a unanimidade das publicações. Assim, Larik e Haider [3] enfrentam o problema da entrada de dinheiro sujo no sistema financeiro com uma abordagem híbrida de detecção de anomalias nas transações financeiras. Esta abordagem emprega clusteres não supervisionados para encontrar padrões de comportamentos normais para os clientes, conjugado com o uso de técnicas estatísticas para identificar o desvio de uma transação particular do correspondente comportamento esperado no seu agrupamento. É sugerida uma variante do *Euclidean Adaptive Resonance Theory* (EART) [4] para agrupar os clientes em diferentes clusteres. A perspectiva dos autores, diferentemente da que será abordada neste trabalho, é a de uma instituição financeira com foco nas transações. Porém a abordagem híbrida se aplicaria à RFB pois se espera que grupos de contribuintes tenham valores de movimentações financeiras agregadas baseados em variáveis que os identifiquem em um grupo particular. Daí, assim como no trabalho ora exposto, anomalias em relação ao grupo a que pertencem podem ser usadas como indicadores de suspeição do contribuinte.

Khan et al. [5] apresentam uma abordagem de rede bayesiana (BN) [6] para analisar as transações de clientes de uma instituição financeira a fim de detectar padrões suspeitos. Baseado no histórico de transações, a abordagem proposta atribui um patamar a partir da qual a transação se torna suspeita. O problema dessa abordagem quando transposta para o domínio do problema da RFB é a ausência de um período histórico relevante. Outra questão, não abordada pelos autores, foi o possível excesso de falsos positivos. Via de regra cabe à instituição financeira apenas informar ao órgão governamental responsável pelos crimes financeiros a transação suspeita, a análise e o diferimento desta transação cabe ao órgão governamental incumbido da prevenção e combate à lavagem de dinheiro.

Raza e Haider [7] agregam as duas abordagens citadas acima para criar o que eles chamaram de *Suspicious Activity Reporting using Dynamic Bayesian Network* (SARDBN), uma combinação de clusterização com *dynamic Bayesian network* (DBN) [8] para identificar anomalias em sequencias de transações. Os autores criaram ainda um índice chamado de *Anomaly Index using Rank and Entropy* (AIRE) que mede o grau de anomalia em uma operação e compara com um valor limiar pré-definido para marcar a transação como

[1] Escritório do Departamento do Tesouro dos Estados Unidos que coleta e analisa informações sobre as transações financeiras, a fim de combater nacional e internacionalmente a lavagem de dinheiro, financiamento do terrorismo e outros crimes financeiros.

normal ou suspeito. Essa abordagem por índice, assemelha-se ao patamar proposto por Khan et al. [5], contudo, esta divisão em duas fases aparenta sofrer menos dos problemas apontados no item anterior, pois a clusterização avalia primeiramente a totalidade dos clientes e o AIRE avalia as transações de um dado cliente de forma individual.

Rajput et al. [9] abordam o problema propondo uma ontologia de bases e regras escritas em *Semantic Web Rule Language* (SWRL) [10]. Tal abordagem exigiria menos computação e permitiria o reuso da base de conhecimento em domínios similares.

Rohit e Patel [11], Tabela 2.1 [2], mostram a diversidade de abordagens no domínio das finanças para tratar a identificação de transações suspeitas de lavagem de dinheiro.

Tópico	**Formulação do Problema**	**Tecnologia/ Algoritmo/ Método**	**Conjunto de Dados e parâmetro de avaliação**
Research on Money Laundering Detection based on Improved Minimum Spanning Tree Clustering and Its Application [12]	Uma nova métrica de dissimilaridade foi proposta e um novo modelo algoritmo de detecção de lavagem de dinheiro baseado em *Improved Minimum Spanning Tree clustering*	*Minimum spanning tree* [13] *; outliers* [14]; *clustering analysis* [15]	Obtidas no mundo real e em tempo real
Application of Data Mining for Anti-Money Laundering Detection: A Case Study [16]	Estudo de caso de aplicação de uma solução que combina mineração de dados e técnicas de computação natural é apresentado para detectar padrões de lavagem de dinheiro.	*Clustering (K-mean* [17] *),Neural networks* [18], *heuristics* [19], *genetics algorithm* [13]	Dados de transações bancárias e em tempo real
An Improved Support-Vector Network Model for Anti-Money Laundering [20]	Proposição de um *support vector machine* [21] melhorado, usando a função de validação cruzada para obter a melhor escolha parâmetros.	*Improved support vector machine* [20]	Dados reais de transações bancárias e acurácia
Research on Anti-Money Laundering Based on Core Decision Tree Algorithm [22]	Apresenta um algoritmo de árvore de decisão para identificar atividades de lavagem de dinheiro combinado com algoritmos de agrupamento.	*Clustering (K-mean [17], BIRCH* [23]*), decision tree algorithm* [24]	Dados sintéticos e eficiência
Money Laundering Detection Using TFA system [25]	Apresenta um sistema baseado em fluxo de transações para detecção de lavagem de dinheiro.	*Clustering (K-mean [17]), Frequent pattern Mining (SM, BIDE)* [26]	Dados reais de transações bancárias e acurácia
Applying Data Mining in Money Laundering Detection Vietnamese Banking Industry [27]	Propõe uma abordagem de detecção de lavagem de dinheiro usando técnicas de clusterização	*CLOPE algorithm* [28]	Dados reais de transações bancárias e Tempo de processamento; acurácia

Tabela 2.1: Abordagens para identificação de lavagem de dinheiro no domínio das finanças

Analisando as motivações dos artigos para a detecção da lavagem de dinheiro no domínio das finanças, percebe-se que elas não visam diretamente o combate ao crime, mas tão somente um gerador de alertas para comunicação aos órgãos governamentais por obrigação legal. Para tal, pode-se supor que, desde que a sensibilidade seja alta, qualquer algoritmo serve, pois o ônus da análise dos falsos-positivos recairá sobre o órgão governamental receptor dos alertas. É uma situação peculiar e derivada da legislação.

Identificação de fraudes Fora do domínio da lavagem de dinheiro, Sharma e Panigrahi [29] mostram que as técnicas de mineração de dados como modelos logísticos [30], redes neurais [18], redes bayesianas [6] e árvores de decisão [24] têm sido aplicadas extensivamente para fornecer soluções para os problemas inerentes à detecção e classificação de dados fraudulentos. A partir do estudo de quarenta e cinco artigos publicados entre 1995 e 2011 em periódicos diversos sobre fraudes no sistema financeiro, os autores apresentam os quatro grupos de abordagens em mineração de dados mais usados. A Tabela 2.2 apresenta uma síntese desse levantamento.

[2] Adaptação da tabela originalmente apresentada no artigo

Método	% de artigos
Redes Neurais	31%
Modelos de Regressão	40%
Logica Fuzzy	16%
Algoritmos genéticos e sistemas especialistas	13%

Tabela 2.2: Abordagens para detecção de fraudes no domínio das finanças

Mineração de dados no comércio exterior brasileiro Jambeiro ([31] e [32]) ao analisar o uso de métodos bayesianos aplicados a bases de importação e NCM em um problema de classificação de padrões de interesse prático para a Receita Federal do Brasil levantou uma importante questão quanto aos dados: a alta cardinalidade dos atributos e suas interações não lineares.

Mostrou também que

> empiricamente as estratégias bayesianas mais avançadas para tratamento de atributos de alta cardinalidade, como pré-processamento para redução de cardinalidade e substituição de tabelas de probabilidades condicionais (CPTs) de redes bayesianas (BNs) por tabelas default (DFs), árvores de decisão (DTs) e grafos de decisão (DGs) embora tragam benefícios pontuais não resultam em ganho de desempenho geral em nosso domínio alvo.

Seu trabalho se voltou então para um novo método bayesiano de classificação, chamado de *Hierarchical Pattern Bayes* (HPB). "O tempo de execução do HPB é exponencial no número de atributos, mas independe de sua cardinalidade. Assim, em domínios onde os atributos são poucos, mas possuem alta cardinalidade, ele é muito mais rápido" que algoritmos tradicionais.

Técnicas de detecção de anomalias em dados A detecção de anomalias tem sido tema de várias pesquisas, artigos de revisão e livros. Chandola et al. [33] apresentam uma visão estruturada da extensa pesquisa sobre técnicas de detecção de anomalias abrangendo várias áreas de pesquisa e domínios de aplicação.

> Hodge e Austin [34] forneceram uma extensa pesquisa de técnicas de detecção de anomalias desenvolvidas em domínios de aprendizagem mecânica e estatística. Uma ampla revisão das técnicas de detecção de anomalias para dados numéricos e simbólicos foi apresentada por Agyemang et al. [35]. Markou e Singh [36] [37] apresentaram uma revisão extensiva das técnicas de detecção de novidade utilizando redes neurais e abordagens estatísticas. Patcha e Park [38] apresentam um levantamento das técnicas de detecção de anomalias utilizadas especificamente para detecção de intrusão cibernética. Uma grande quantidade de pesquisas sobre a detecção de valores atípicos foi feita em estatísticas e tem sido revista em vários livros [39], [40], [14], bem como outros artigos de pesquisa [41] e [42].

Goodfellow et al. [43] apresentam o uso de *Deep Learning* como técnica de detecção de anomalias quando da sua configuração como *Autoencoder*. LeCun et al. [44] mostram que *Deep Learning* é um método de representação de aprendizagem com vários níveis de abstração obtidos através da composição de módulos simples, mas não lineares, que transformam cada representação em um nível (começando com a entrada bruta) em uma representação de nível superior, ligeiramente mais abstrato. Com as composições suficientes de tais transformações, funções muito complexas podem ser aprendidas. *Deep Learnig* tem se mostrado como um algoritmo capaz de atingir o estado da arte para vários domínios onde não há linearidade entre os atributos preditivos e que apresentam alta cardinalidade nos atributos nominais.

2.2 Modelo de referência CRISP-DM

De acordo com Nisbet et al.[45], *Cross Industry Standard Process for Data Mining* (CRISP-DM) é o mais completo modelo de processo para expressar a mineração de dados. De acordo com pesquisas realizadas pelo site *KDnuggets*[3] em 2007 e 2014, CRISP-DM é a metodologia mais utilizada por cientistas de dados (ver Figura 2.1). CRISP-DM foi criado a partir de um consórcio entre as empresas NCR[4], SPSS[5] e Daimler-Benz[6].

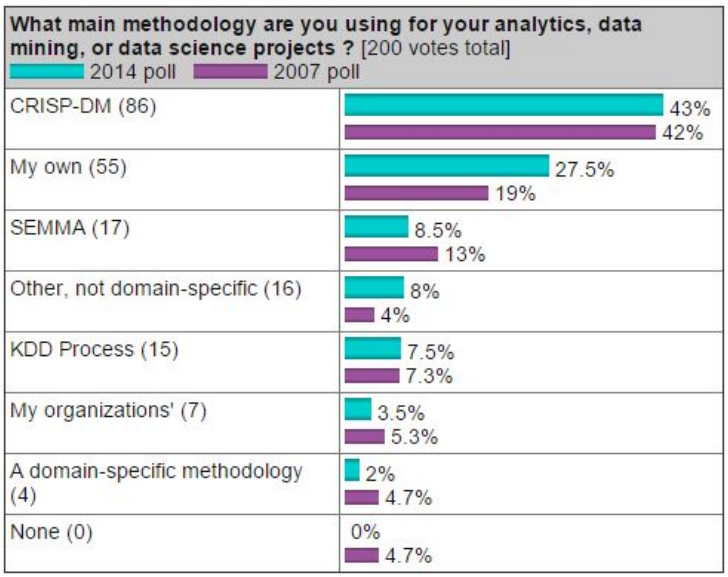

Figura 2.1: Pesquisas realizadas em 2007 e 2014 pelo sítio *KDnuggets*[3].

[3] `http://www.kdnuggets.com/2014/10/crisp-dm-top-methodology-analytics-data-mining-data-science-projects.html`

[4] `https://www.ncr.com/`

[5] `http://www.spss.com.hk/corpinfo/index.htm` - A SPSS foi adquirida pela IBM em 2010

[6] `http://www.daimler.com/`

O processo CRISP-DM define uma hierarquia que consiste em fases principais, tarefas genéricas, tarefas especializadas e instâncias de processo. As fases principais estão relacionadas na Figura 2.2 [7].

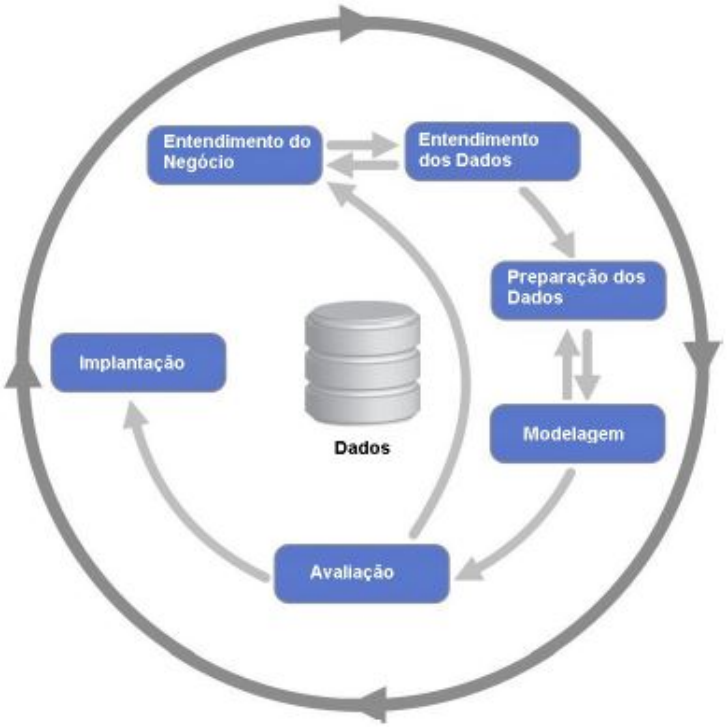

Figura 2.2: Fases do CRISP-DM..

De acordo com Chapman et al.[46], cada fase do processo consiste em um número de atividades genéricas de segundo nível, cada uma com várias operações especializadas. Um nível adicional de tarefas específicas do domínio deve ser definido em termos do problema empresarial específico a ser resolvido no contexto dos dados utilizados para resolvê-lo. Assim, a organização deste processo pode ser vista da seguinte forma hierárquica:

Fases da mineração de dados
Atividades
Operações
Tarefas

O detalhamento das fases que se seguem foram baseados no guia de mineração de dados do CRISP-DM [46].

[7]Imagem adaptada para o português a partir de [46].

Entendimento do negócio Esta fase objetiva ter um claro entendimento do que se pretende a partir da mineração de dados e como os resultados alcançados se parecerão em termos dos processos de negócios que serão beneficiados.

Entendimento dos dados Esta fase parte de uma coleta inicial dos dados seguida de atividades que possibilitem a familiarização com seu conjunto, a identificação de problemas de qualidade e a descoberta de *insights* dentro dos dados que permitam a formulação de hipóteses para informações que não estejam aparentes.

Preparação dos dados A fase de preparação de dados é constituída de atividades que visam a construção, a partir dos dados brutos iniciais, do conjunto de dados final. Tarefas de preparação de dados não possuem uma ordem prescrita e são susceptíveis de serem realizadas várias vezes. Essas tarefas incluem a seleção de tabela, registro e atributo, bem como transformação e limpeza de dados para ferramentas de modelagem.

Modelagem Nesta fase, várias técnicas de modelagem são selecionadas e aplicadas, e seus parâmetros são calibrados para valores ótimos. Normalmente, existem várias técnicas para o mesmo tipo de problema de mineração de dados. Algumas técnicas têm requisitos específicos sobre a forma dos dados. Portanto, voltar à fase de preparação de dados é muitas vezes necessário.

Avaliação Esta fase do projeto se dá após a construção de um modelo (ou modelos) que, a partir de uma perspectiva de análise de dados, aparenta ter alta qualidade. Para se ter certeza de que o modelo atinge adequadamente os objetivos de negócios, antes de proceder à implantação final do modelo, é importante avaliá-lo cuidadosamente e rever as etapas executadas para criá-lo. Um dos principais objetivos aqui é o de determinar se existe alguma questão de negócio importante que não tenha sido suficientemente considerada. No final desta fase, uma decisão sobre o uso dos resultados de mineração de dados deve ser alcançada.

Implantação Essa fase geralmente envolve a aplicação de modelos 'ao vivo' dentro dos processos de tomada de decisão de uma organização. A complexidade desta fase depende dos requisitos, podendo ser tão simples quanto gerar um relatório ou tão complexa como implementar um processo de mineração de dados repetível em toda a empresa. Em muitos casos esta fase não é executada pelo analista de dados. No entanto, mesmo se o analista realizar o esforço de implantação, é importante que o cliente compreenda as ações que precisam ser realizadas para realmente usar os modelos criados.

2.3 *Gradient Boosting Machines* (GBM)

Gradient Boosting Machines é um algoritmo de aprendizagem automática proposto por Friedman [47] [48] que pode tanto ser usado para tarefas de regressão, quanto para tarefas de classificação. O princípio básico de funcionamento deste algoritmo é bastante simples: dada uma função de perda e dado um estimador fraco o algoritmo procura um modelo que minimiza essa função de perda.

O algoritmo é inicializado com um palpite sobre a melhor resposta e é feito o cálculo do gradiente da função de perda. Então o modelo é ajustado para minimizar essa função. Esse novo modelo é então adicionado ao modelo anterior e nova interação ocorre até que se atinja um limite estipulado pelo usuário.

De acordo com Kuhn e Johnson [49], a princípio, qualquer estimador parametrizável pode ser escolhido como um estimador fraco a fim de atender à exigência do algoritmo. Porém, a escolha de árvores de decisão [24] como estimadores fracos é particularmente interessante, pois possuem a flexibilidade de se enfraquecerem à medida que se restringe sua profundidade. Acresce ainda que árvores distintas podem ser facilmente adicionadas e sua criação é extremamente rápida o que beneficia o processo de modelagem aditiva.

Friedman et al. [50] pontuam duas desvantagens deste algoritmo: sua estratégia gulosa que escolhe, a cada estágio, a solução ótima sem se importar em encontrar um ótimo global; e sua suscetibilidade a *over-fitting* na base de treinamento.

A degradação da capacidade de generalização deste algoritmo por *over-fitting* pode ser combatida a partir de várias técnicas:

- Redução do número de iterações;
- Redução da taxa de aprendizagem;
- Reamostragem da base de treinamento a cada interação;
- Penalização da complexidade da árvore.

Segundo Ridgeway [51], quando utilizam-se técnicas de adição em estimadores fracos para se obter um estimador mais forte (*boosting*) a importância de cada variável é dada em função da redução da função de perda que a sua adição dentro de cada árvore proporciona. Desta forma, a importância global de uma variável será dada pela média das suas contribuições em todas interações.

2.4 *Distributed Random Forest*[8] (DRF)

De forma similar ao *Gradient Boosting Machine*, *Ramdon Forest* aproveita-se das propriedades de *bagging*[9] propostas por Breiman [52] e se utiliza de um conjunto de modelos e árvores como base do seu aprendizado. Porém, a forma como esse conjunto é construído difere substancialmente em cada técnica: em *Ramdon Forest*, diferentemente de *Gradient Boosting Machine*, cada árvore é criada independentemente, possui com um limite máximo de profundidade e contribui de forma equanime com a formação do modelo final. Apesar dessas diferenças, de acordo com Kuhn e Johnson [49], ambas, *Gradient Boosting Machine* e *Ramdon Forest* oferecem performances preditivas competitivas entre si.

Segundo Rossini et al. [53], a criação independente das árvores fornece uma vantagem em termos de processamento à *Ramdon Forest*, pois favorece o paralelismo.

O algoritmo 1, adaptado de [49], apresenta o algoritmo básico de uma *Random Forest*. A partir dele podemos extrair as principais informações necessárias à sua construção: o número m de árvores a serem construídas, seu tamanho máximo (não há poda nas árvores) e o número de atributos que serão aleatoriamente selecionados.

Algorithm 1 Algoritmo básico de uma *Random Forest*. Adaptado de [49].

```
Selecione a quantidade de modelos a ser construído, m
for i = 1 to m do
    Gere uma amostragem por bootstrap dos dados originais
    Treine um modelo de árvore de decisão nesta amostra
    for cada split da árvore do
        Selecione randomicamente k preditores
        Selecione o melhor preditor entre os k preditores e particione os dados
    end for
    Use um critério de paragem para determinar quando a árvore está completa (não use
    poda)
end for
```

Breiman [54] provou que, diferentemente de *Gradient Boosting Machine*, *Ramdon Forest* não está sujeita a *over-fitting* de modo que a técnica não é afetada negativamente se usada com um grande número de árvores. A fim de não incorrer em excesso de carga camputacional, Kuhn e Johnson [49] sugerem que o modelo seja iniciado com 1.000 árvores e apenas caso a performance em *Cross-Validation* ainda apresente melhoras com essa quantidade, sejam incorporadas mais árvores até que o nível de performance pare de crescer.

[8] *Distributed Random Forest* é o nome dado à implementação da técnica *Random Forest* na plataforma *H2O*

[9] Preditores usando *bagging* foram propostos por Breiman [52] em 1996. Segundo o pesquisador, qualquer preditor que possa produzir uma alta variância com baixo enviesamento, por exemplo árvores de decisão, tem sua performance preditiva melhorada pela redução da variância do preditor

Importante notar que, de acordo com Strobl et al. [55] o cálculo de importância das variáveis na predição do modelo é fortemente impactado pela existência de variáveis correlacionadas no conjunto de treinamento e pelo número de atributos selecionados de forma aleatória. Um dos principais impactos é a diluição da importância dos preditores principais.

2.5 *Deep Learning Autoencoder* (DLA)

De acordo com Goodfellow et al. [43], um *Autoencoder* é uma rede neural treinada para tentar copiar sua entrada para sua saída. A rede pode ser vista como consistindo de duas partes: uma função de codificação e uma de decodificação que reconstrói os valores da entrada. Os *Autoencoders* podem ser vistos como um caso especial de redes *feedforward* [56] e podem ser treinados com as mesmas técnicas: normalmente gradientes descendentes calculados por *backpropagation* [56].

Porém, um *Autoencoder* que tenha sucesso em copiar os valores de entrada na sua saída não será útil. Normalmente o modelo da rede será feito de forma que a cópia seja apenas aproximada. Como o modelo é forçado a priorizar quais aspectos da entrada devem ser copiados, este muitas vezes acaba por aprender propriedades úteis dos dados.

Para melhor entendimento do processo, a Figura 2.3 (retirada de [43]) apresenta a estrutura genérica de um *Autoencoder*, a qual mapeia uma entrada x para uma saída (reconstrução) r através de uma representação ou código h (camadas ocultas). As funções f e g representam suas funções componentes: codificação (*encode*) de x para h e decodificação (*decode*) de h para r. Uma maneira de obter informações úteis dessa rede é restringir as dimensões de h de forma que este seja de menor dimensão que x.

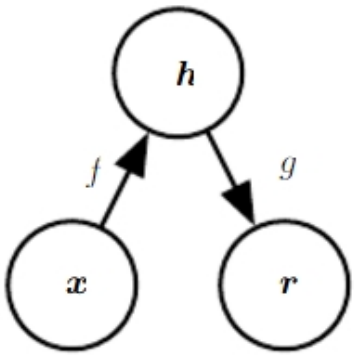

Figura 2.3: Estrutura genérica de um *Autoencoder* .

O foco de atenção do aprendizado do *Autoencoder* não reside portanto na sua saída, mas sim na diferença entre os valores da entrada e da saída. O processo de aprendizagem pode ser então descrito como a minimização da função

$$L(x, g(f(x))) \tag{2.1}$$

onde L pode ser uma função de perda como *mean squared error* (MSE) que funciona penalizando a diferença entre x e $g(f(x))$. O aprendizado neste tipo de rede força que o *Autoencoder* capture as características mais importantes dos dados [57].

2.5.1 Detecção de anomalias em *Autoencoders*

Uma das principais informações a serem extraídas da redução de dimensão ocorrida em h é a detecção de anomalias [43]. O pressuposto desta informação é que em um espaço dimensional reduzido, dados regulares e anômalos aparecem significativamente diferentes.

Assim, segundo Goodfellow et al. [43], dado um conjunto de treinamento como o conjunto a seguir, $\{x(1), x(2), ..., x(m)\}$, assume-se que cada $x(i) \in \mathbb{R}^D$ é representado por um vetor de D variáveis diferentes (Ver Figura 2.4 retirada de [43]). Durante a fase *encode*, os dados são comprimidos em um subconjunto menor para, a seguir, serem reconstruídos como $\{\widehat{x}(1), \widehat{x}(2), ..., \widehat{x}(m)\}$ de forma que o somatório do erro de reconstrução na Equação 2.2 para cada $x(i)$ seja o menor possível.

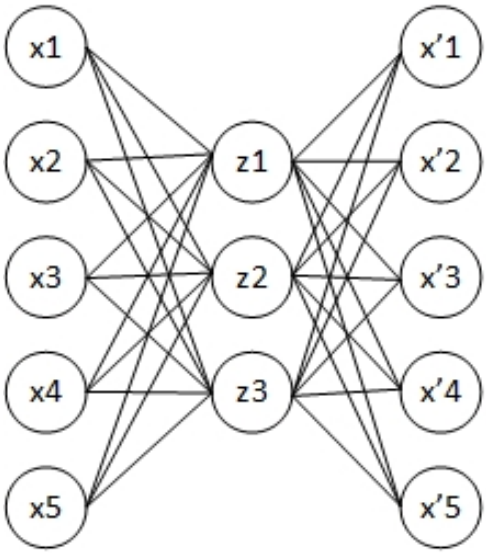

Figura 2.4: *Autoencoder*.

O cálculo do erro de reconstrução apresentado na Equação 2.2 é então utilizado como um índice de anomalias: o valor do erro terá valores baixos se $x(i)$ for um vetor que

satisfaz as relações do modelo encontrado na fase de treinamento. Em sentido contrário, o valor do erro é maior com vetores anômalos.

$$Err(i) = \sqrt{\sum_{j=1}^{D}(x_j(i) - \widehat{x}_j(i))^2} \tag{2.2}$$

2.6 Métricas de avaliação

Para podermos avaliar os modelos de classificação, são necessários parâmetros que nos permitam relativizar um modelo com ele mesmo (quando estamos procurando o ajuste de melhor resultado) ou com outro modelo (quando estamos procurando o modelo com a melhor performance para um dado problema). Existem diversas métricas aceitas pela comunidade de mineração de dados que se prestam a esta tarefa [58]. A seguir apresentamos aquelas que foram utilizadas neste trabalho, todas retirados de [58].

Matriz de confusão A matriz de confusão não é propriamente uma métrica, porém é através dos valores extraídos dela que muitas métricas são calculadas. Podemos dizer que a matriz de confusão é um tipo específico de tabela que permite comparar os resultados de um classificador em função dos valores reais. Na Figura 2.5 vemos nas colunas da matriz os valores sabidamente verdadeiros e nas linhas, os valores classificados por um algoritmo classificador.

		Valor Verdadeiro	
		positivos	negativos
Valor Predito	positivos	**VP** Verdadeiro Positivo	**FP** Falso positivo
	negativos	**FN** Falso negativo	**VN** Verdadeiro Negativo

Figura 2.5: Matriz de Confusão.

O preenchimento da matriz se dá como se segue:

- O quantitativo de valores preditos classificados como *positivo* e que se sabe serem realmente positivos é incluído na célula VP: são chamados de *verdadeiros positivos*.

- O quantitativo de valores preditos classificados como *negativo* e que se sabe serem realmente negativos é incluído na célula VN: são chamados de *verdadeiros negativos*.
- O quantitativo de valores preditos classificados como *positivo* e que se sabe serem realmente negativos é incluído na célula FP: são chamados de *falsos positivos*.
- O quantitativo de valores preditos classificados como *negativo* e que se sabe serem realmente positivos é incluído na célula FN: são chamados de *falsos negativos*.

Acurácia É uma métrica bastante simples. Seu cálculo é feito baseado no número de acertos em função do total de amostras. Deve ser usada quando se tem classes balanceadas. Para classes desbalanceadas, ela causa uma falsa impressão de bom desempenho.

Seu cálculo é dado pela seguinte equação:

$$Acurácia = \frac{VP + VN}{P + N} \tag{2.3}$$

onde VP representa o total de *verdadeiros positivos*, VN o total de *verdadeiros negativos*, P o total de amostras sabidamente positivas e N o total de amostras sabidamente negativas.

Recall Esta métrica é calculada por classes a partir da razão existente entre a quantidade de classificações corretas naquela classe e o total de itens verdadeiramente pertencentes àquela classe. Caso tenhamos uma classificação binária, a métrica seria calculada pela razão entre os verdadeiros positivos (VP) e o total de positivos P, conforme a equação a seguir.

$$Recall = \frac{VP}{P} \tag{2.4}$$

Precisão A *precisão* é calculada a partir da razão entre a quantidade de verdadeiros positivos de uma determinada classe e a soma desses com o quantitativo de dados classificados errados para essa classe, os falsos positivos (FP). A métrica *precisão* é dada pela seguinte fórmula:

$$Precisão = \frac{VP}{VP + FP} \tag{2.5}$$

F1 *measure* Trata-se da média harmônica entre as métricas *precisão* e *recall*. É calculada da seguinte forma:

$$F1 = \frac{2 * precisão * recall}{precisão + recall} \tag{2.6}$$

Uma das suas principais vantagens é ser uma métrica que sofre pouco com classes desbalanceadas.

Área sob a curva ROC (AUC) A curva ROC é formada pela relação gráfica bidimensional entre a taxa de verdadeiros positivos (tpr) no eixo Y e a taxa de falsos positivos (fpr) no eixo x. A área sob essa curva será tanto maior quanto forem maiores as taxas de verdadeiros positivos em comparação com as taxas de falsos positivos.

Nas situações de tpr com crescimento idêntico ou inferior à fpr temos que a AUC será igual a 50% da área total do gráfico ou menor. Nestes casos temos que a performance do algoritmo é igual à do arremesso de uma moeda não viciada, ou pior.

Em projetos onde se quer apenas identificar as classes, sem a observação da probabilidade de sua ocorrência, ela é um excelente indicador do melhor ponto de corte. A possibilidade de escolha de ponto de corte na curva dá a essa métrica a possibilidade de maximizar uma determinada característica que se busca atender no problema de negócio. Outra vantagem dessa métrica é sua capacidade de trabalhar bem com *datasets* que possuam classes desproporcionais.

Logloss Para classificações binárias, a fórmula para o cálculo de *logloss* é:

$$Logloss = -\frac{1}{N}\sum_{i=1}^{N}(y_i log(p_i) + (1 - y_i) log(1 - p_i)) \tag{2.7}$$

sendo que p é a probabilidade do exemplo pertencer a uma determinada classe e y é o valor real da variável dependente.

Uma de suas principais vantagens é punir previsões incorretas classificadas a partir de uma probabilidade alta. Apesar da equação apresentada ser para classificações binárias, ela pode ser usada em problemas de múltiplas classes. Em classes desbalanceadas, o *logloss* pode tender a apresentar valores melhores para modelos que favoreçam a classe de maior tamanho.

É uma métrica que deve ser escolhida quando a percepção da probabilidade de uma classe for mais importante que a simples classificação.

***Mean Squared Erro* (MSE)** Essa métrica é muito utilizada em modelos com resultados numéricos para mensurar a média da diferença entre o valor obtido pelo modelo e o valor esperado. A elevação ao quadrado dessa diferença antes da realização do somatório visa eliminar os valores negativos de erro antes de se efetuar a soma.

A fórmula para o cálculo do MSE é:

$$MSE = \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{n} \tag{2.8}$$

***Root Mean Squared Erro* (RMSE)** É uma variação do MSE que apresenta seu resultado na mesma unidade de medida dos dados, pois é simplesmente a raiz quadrada de MSE.

***Matthews correlation coefficient* (mcc)** Essa métrica é específica para mensuração da qualidade de classificadores binários e pode ser calculada diretamente a partir da matriz de confusão pela seguinte fórmula:

$$mcc = \frac{VP.VN - FP.FN}{\sqrt{(VP + FP).(VP + FN).(VN + FP).(VN + FN)}} \tag{2.9}$$

De acordo com seu propositor [59] o mcc mensura a correlação entre os dados observados e os preditos nas classificações binárias. Os valores de retorno do coeficiente estão entre -1 e $+1$, sendo que $+1$ indica uma predição perfeita, 0 indica que o modelo não é melhor que uma predição randômica e -1 indica uma total discordância entre previsão e observação.

Capítulo 3

Contextualização

O presente capítulo apresenta a lavagem de dinheiro dentro de um panorama mundial e nacional. Apresenta a posição da RFB dentro da estrutura orgânica da inteligência financeira no Brasil e o seu papel no combate à LD. São apresentados também a ligação da LD com o comércio exterior e com as chamadas exportações fictícias. Ao final, são apresentados os pressupostos indicativos da ocorrência de LD a partir da legislação nacional vigente.

3.1 Lavagem de dinheiro - Panorama mundial e nacional

De acordo com o Egmont Group[1] [60],

> lavagem de dinheiro é o processo pelo qual o criminoso transforma recursos oriundos de atividades ilegais em ativos com origem aparentemente legal. Essa prática geralmente envolve múltiplas transações, para ocultar a origem dos ativos financeiros e permitir que eles sejam utilizados sem comprometer os criminosos. A dissimulação é, portanto, a base para toda operação de lavagem que envolva dinheiro proveniente de um crime antecedente.

O Escritório das Nações Unidas sobre Drogas e Crime [61] afirmou que

> por trás da lavagem de dinheiro está o crime organizado transnacional, o tráfico de drogas, o tráfico de armas, o tráfico de pessoas e a corrupção. Este é um crime que aparenta não ter vítimas ... a lavagem de dinheiro permite aos criminosos desfrutar de suas riquezas ilegais e empreender novos negócios ilícitos. O valor estimado de dinheiro lavado anualmente no mundo está entre 2% e 5% do PIB mundial, ou seja, algo entre US$ 800 bilhões e US$ 2 trilhões.

[1] Grupo internacional criado para promover em âmbito mundial o tratamento de comunicações suspeitas relacionadas à LD. `http://www.egmontgroup.org/`

Tais declarações evidenciam a dimensão sócio-econômica do problema, trazendo a questão do crime que antecede a LD, aquele crime que teve resultados financeiros os quais se pretende reinserir de forma 'lícita', lavada, na economia.

No Brasil, em 2012, a Lei nº 9.613 de 1998, alterada pela Lei nº 12.683 de 2012 [62], trouxe importantes avanços para a prevenção e combate à lavagem de dinheiro como:

- a extinção do rol taxativo de delitos criminais antecedentes, admitindo-se agora como crime antecedente da LD qualquer infração penal.
- a imputação explícita do crime de lavagem de dinheiro àqueles que fraudam as exportações.

Esta lei estabelece ainda uma estrutura de combate aos crimes de lavagem ou ocultação de bens, direitos e valores, apresentada na Figura 3.1 extraída de [63], na qual se insere a RFB como uma instituição de controle atuando na inteligência financeira.

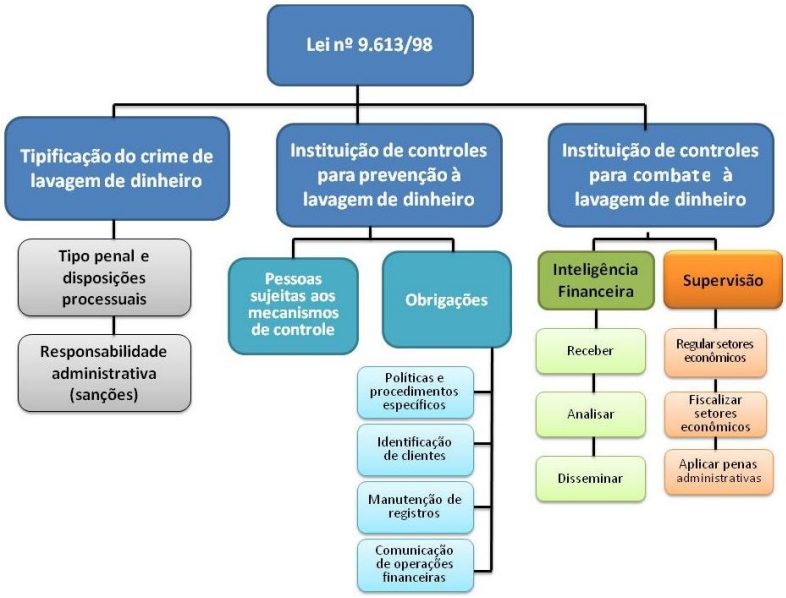

Figura 3.1: Estrutura orgânica da inteligência financeira no Brasil.

3.2 A Receita Federal do Brasil e o combate à lavagem de dinheiro

No contexto apresentado, a RFB é responsável, dentre outras atribuições correlatas, por "planejar, coordenar e executar as atividades de inteligência fiscal no combate à lavagem e

ocultação de bens, direitos e valores" [64]. A partir de mecanismos diversos como denúncias, fiscalizações, demandas judiciais, cruzamento de dados, dentre outros, os casos que podem se enquadrar nos crimes de lavagem de dinheiro são selecionados para investigação. A qualidade dessa seleção é determinada pela experiência do analista de inteligência[2] alocado à tarefa.

As bases de dados da RFB utilizadas nesta atividade são, em sua maioria, originárias de declarações prestadas por contribuintes ou por terceiros detentores de informações econômico-fiscais e cadastrais de interesse do fisco[3].

A RFB conta ainda com um *Laboratório de Tecnologia Contra a Lavagem de Dinheiro* (Lab-LD) integrante da *Rede Nacional de Laboratórios contra Lavagem de Dinheiro* (Rede-LAB) do Ministério da Justiça (MJ) instalado em abril de 2014 [65]. Este laboratório possui diversas ferramentas para análises estatísticas e mineração em grandes volumes de dados.

3.3 Fases da lavagem de dinheiro

De acordo com o Conselho de Controle de Atividades Financeiras (COAF) [66]

> para disfarçar lucros ilícitos sem comprometer os envolvidos, a lavagem de dinheiro realiza-se por meio de um processo dinâmico que requer: primeiro, o distanciamento dos fundos de sua origem, evitando uma associação direta deles com o crime; segundo, o disfarce de suas várias movimentações para dificultar o rastreamento desses recursos; e terceiro, a disponibilização do dinheiro novamente para os criminosos depois de ter sido suficientemente movimentado no ciclo de lavagem e poder ser considerado 'limpo'.

Ainda de acordo com o COAF [66],

> os mecanismos mais utilizados no processo de lavagem de dinheiro envolvem teoricamente essas três etapas independentes que, com frequência, ocorrem simultaneamente.
>
> 1. Colocação – a primeira etapa do processo é a colocação do dinheiro no sistema econômico. Objetivando ocultar sua origem, o criminoso procura movimentar o dinheiro em países com regras mais permissivas e naqueles que possuem um sistema financeiro liberal. A colocação se efetua por meio de depósitos, compra de instrumentos negociáveis ou compra de bens. Para dificultar a identificação

[2]Servidor da RFB encarregado da investigação de ilícitos tributários.

[3]Atualmente, 32 declarações diferentes são transmitidas à RFB em diversas periodicidades de acordo com as legislações específicas que regem as obrigações acessórias de cada tributo administrado pela União. Acresce-se a Escrituração Contábil Digital, Escrituração Fiscal Digital e as Notas Fiscais Eletrônicas (NFe) de todas empresas brasileiras tributadas pelo lucro real. Há ainda o intercâmbio de informações entre entidades por meio de convênios e, quando se trata de investigações de crimes tributários, as apreensões de bases de dados durante diligências judiciais e o uso de fontes abertas na Internet.

da procedência do dinheiro, os criminosos aplicam técnicas sofisticadas e cada vez mais dinâmicas, tais como o fracionamento dos valores que transitam pelo sistema financeiro e a utilização de estabelecimentos comerciais que usualmente trabalham com dinheiro em espécie.

2. Ocultação – a segunda etapa do processo consiste em dificultar o rastreamento contábil dos recursos ilícitos. O objetivo é quebrar a cadeia de evidências ante a possibilidade da realização de investigações sobre a origem do dinheiro. Os criminosos buscam movimentá-lo de forma eletrônica, transferindo os ativos para contas anônimas – preferencialmente, em países amparados por lei de sigilo bancário – ou realizando depósitos em contas abertas em nome de "laranjas" ou utilizando empresas fictícias ou de fachada.
3. Integração – nesta última etapa, os ativos são incorporados formalmente ao sistema econômico. As organizações criminosas buscam investir em empreendimentos que facilitem suas atividades – podendo tais sociedades prestarem serviços entre si. Uma vez formada a cadeia, torna-se cada vez mais fácil legitimar o dinheiro ilegal.

A Figura 3.2, retirada do site da KYCMap [4] ilustra essas três fases.

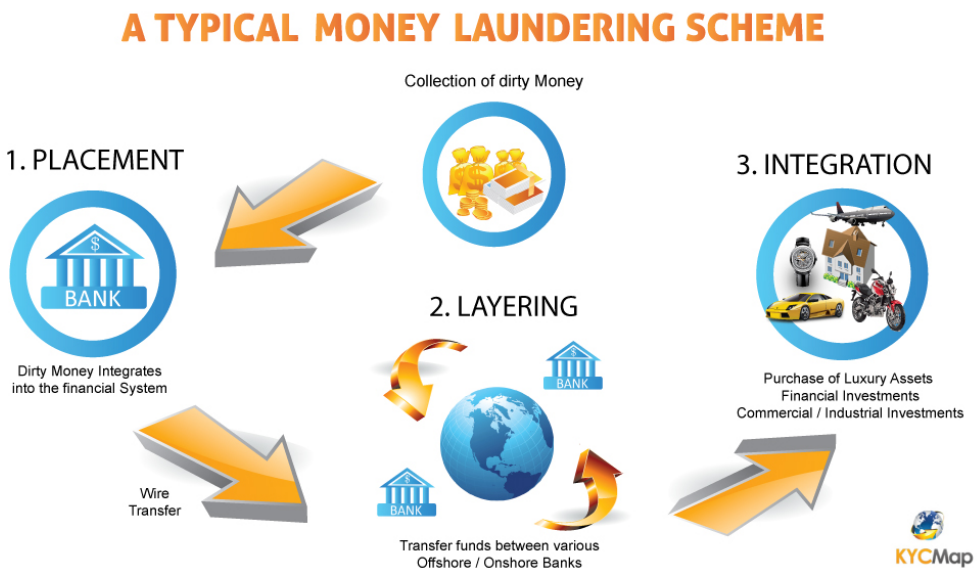

Figura 3.2: Fases da lavagem de dinheiro.

[4]KYC é o acrônimo de *Know Your Client*. KYCMap é uma empresa americana especializada em fornecer informações detalhadas à indústria de investimentos sobre a tolerância ao risco dos seus clientes e sobre regras mundiais de combate à lavagem de dinheiro. `http://kycmap.com/what-is-money-laundering/`

3.4 Comércio exterior e a exportação fictícia como instrumento da lavagem de dinheiro

Diversos autores ([67], [60], [68] e [69]) apontam casos de lavagem de dinheiro com o uso do comércio exterior em algumas de suas fases, pois, aproveitando-se das dificuldades dos países em trocar massivamente informações, operam a 'limpeza' do dinheiro. Olivia Greene [70] cita que as fraudes no comércio exterior representam

> um sistema de remessa financeira que permite às organizações ilegais a oportunidade de mover e armazenar receitas disfarçadas de comércio legítimo. O valor pode ser movido neste processo por falsa-faturação, sobre-faturação e sub-faturação de mercadorias que são importadas ou exportadas.

As exportações fictícias são aqui entendidas como as operações de comércio exterior em que há remessa de capital ao Brasil a partir de transação comercial internacional entre empresas sem contudo haver o efetivo envio da mercadoria. Este capital enviado é proveniente de crime cometido no exterior, ou até mesmo no Brasil, que retorna 'lavado' e 'legal' ao território nacional.

Via de regra, no Brasil, assim como em grande parte do mundo, as exportações recebem incentivos fiscais sendo pouco ou nada tributadas. De maneira inversa, a regra na importação é a tributação.

3.5 Lavagem de dinheiro nas exportações - pressupostos indicativos da ocorrência do crime

Toda fraude à exportação é forte indício de lavagem de dinheiro. Tal afirmação decorre do inciso III, parágrafo 1º do artigo 1º da lei 9.613/98 [71] de 1998 que, ao tratar da tipificação penal do crime de lavagem de dinheiro, é explícito quanto ao seu cometimento por aqueles que fraudam as importações e exportações. Trata-se de uma presunção legal, específica para o comércio exterior, e facilitadora da identificação da materialidade e autoria do crime.

Assim, cabe à RFB noticiar à autoridade policial e ao Ministério Público o indício de fraude no comércio exterior para que se dê início ao inquérito nos termos do Título II do Código de Processo Penal Brasileiro [72]. Decorre assim que o indício de fraude na exportação é pressuposto da suspeição de lavagem de dinheiro e suficiente para a apresentação de notícia crime.

A identificação da exportação fictícia compete às áreas aduaneiras de fiscalização, repressão e investigação da RFB que, a partir da experiência acumulada ao longo dos

anos, adotam diversos indicadores para selecionar um contribuinte para atuação do fisco federal. Esses indicadores encontram-se manualizados com grau de sigilo *reservado* e são de acesso restrito a apenas os servidores da RFB que deles necessitam.

Evidentemente, há outros pressupostos a serem observados no tratamento de casos de lavagem de dinheiro no comércio exterior envolvendo exportações fictícias, porém, todos eles são de competência da autoridade policial judiciária e fogem portanto das atividades regulares da RFB. Assim, esses pressupostos não são levados em conta neste trabalho.

Apesar do exposto, entende-se que o tratamento por mineração de dados não é suficiente para um encaminhamento automático às autoridades judiciárias dos achados suspeitos. Uma fase manual, posterior ao tratamento por mineração, é necessária para que uma equipe de especialistas da RFB agregue informações para melhor subsidiar as ações penais e tributárias posteriores.

Capítulo 4

Metodologia de Pesquisa

Este capítulo apresenta o método de pesquisa que será utilizado para se atingir os objetivos propostos. Por se tratar de um modelo de referência de mineração de dados já consolidado no mercado, sempre que aplicável, as fases, atividades, operações e tarefas do CRISP-DM descritas na Seção **??** foram incorporadas à presente metodologia.

4.1 Etapa 1: levantamentos preliminares

Esta etapa compreende os passos percorridos para a definição dos objetivos. As etapas posteriores são dependentes de seus resultados. Os Capítulos 1, 2, 3 e 5 são resultantes desta etapa.

Entendimento do Negócio Concentra-se em entender os objetivos e requisitos do projeto a partir da perspectiva de negócios e, posteriormente, em converter esse conhecimento em uma definição do problema de mineração de dados.

Revisão Bibliográfica Este passo concentra-se numa leitura crítica sobre trabalhos científicos relacionados ao problema de mineração de dados levantado no passo anterior. Busca-se tanto trabalhos clássicos quanto fontes mais recentes sobre o assunto da pesquisa.

4.2 Etapa 2: aquisição dos dados

Esta etapa vai desde a busca pelos dados que tenham relação com a definição do problema até a entrega do dado à ferramenta de modelagem. O Capítulo 6 é resultante desta etapa.

Entendimento dos dados O entendimento dos dados começa com a coleta de dados inicial e prossegue com atividades que permitem:

- identificação junto a especialistas dos atributos mais relevantes;
- familiarização com os dados;
- identificação de problemas de qualidade de dados;
- descoberta dos primeiros *insights* sobre os dados.

Preparação dos dados Abrange as atividades necessárias para construir o conjunto de dados final a partir dos dados em estado bruto. Esses dados serão usados para alimentar a ferramenta de modelagem.

Essa fase é suscetível de ser realizada várias vezes e não possui uma ordem prescrita. As tarefas incluem o planilhamento, registro e seleção de atributos, bem como a transformação e limpeza de dados.

4.3 Etapa 3: indução do modelo e análise de resultados

Esta etapa compreende a aplicação de diversas técnicas de modelagem de mineração de dados seguidas de testes com vistas à determinação de qual técnica alcança adequadamente os objetivos propostos. É comum e quase sempre necessário o retorno à fase de preparação de dados (Seção 4.2). O Capítulo 7 é resultante desta etapa.

Modelagem Nesta fase, várias técnicas de modelagem são selecionadas e aplicadas. Seus parâmetros são calibrados para os valores ótimos.

Tipicamente, existem várias técnicas para o mesmo tipo de problema de mineração de dados e algumas técnicas têm requisitos específicos sobre o formato dos dados.

Análise dos modelos Nesta fase os modelos, ou modelo, construídos na fase anterior e que se mostraram promissores a partir de uma perspectiva de análise de dados serão cuidadosamente verificados com relação à consecução dos objetivos propostos.

No final desta fase, uma decisão sobre a utilização dos resultados de mineração de dados deve ser alcançada.

4.4 Etapa 4: validação

Aqui o modelo deverá ser submetido a dados completamente novos, preferencialmente dados recentes. Os resultados deverão ser submetidos à avaliação de terceiros quanto à sua eficácia para a solução do problema definido. O Capítulo 8 é resultante desta etapa.

Aplicação do modelo em novas bases Nesta fase o modelo deverá ser aplicado sobre novas bases não utilizadas na etapa de *Busca do Modelo*. Estes novos dados deverão ser preparados sob os mesmos procedimentos da Seção 4.2 - *Preparação dos dados*.

O resultado deverá ser avaliado a partir de métricas objetivas e aceitas na comunidade de mineração de dados e que se adequem aos dados e à modelagem adotada.

Capítulo 5

Entendimento do Negócio

Boa parte do entendimento do negócio foi abordado no Capítulo 3, quando, para melhor entendimento do problema, foi definido o conceito de lavagem de dinheiro e de suas fases; foi apresentada a atuação da RFB no seu combate; e foi exposto o uso da exportação fictícia como seu instrumento, bem como os pressupostos indicativos de sua ocorrência.

Assim, acresce-se aqui outras informações, mais pormenorizadas, e obtidas em sua maioria junto a terceiros especialistas: servidores da RFB que atuam ou atuaram nas áreas de investigação de crimes de lavagem de dinheiro e de fiscalização aduaneira nas atividades alfandegárias de *zona primária*[1], *zona secundária*[2] ou de coordenação de trabalhos.

Assim, este capítulo objetiva apresentar aspectos do problema definido no Capítulo 1 sob a ótica da mineração de dados. As seções a seguir são baseadas em tarefas do CRISP-DM para a fase de Entendimento do Negócio. A Seção 5.1 traz o levantamento da situação atual do problema e suas perspectivas. A Seção 5.2 apresenta os recursos disponíveis dentro da RFB para a realização deste trabalho. A Seção 5.3 traz as restrições legais impostas à condução deste trabalho. A Seção 5.4 apresenta critérios de aceitação da mineração de dados que se propõe neste estudo.

5.1 Abordagem atual do problema e perspectivas

Pelo regimento interno da RFB [74] encontramos quatro formas a partir das quais o fisco federal atua no combate às exportações fictícias: controle aduaneiro de fronteira, portos e aeroportos; repressão aduaneira; fiscalização aduaneira; e investigação de crimes tributários.

[1]De acordo com o Decreto-Lei 37 [73] de 1966, a zona primaria é constituída por pontos de fronteira e áreas terrestres ou aquáticas nos portos e aeroportos alfandegados assim demarcadas pela autoridade aduaneira.

[2]Zona secundária compreende toda a área do território nacional não declarada como zona primária.

Repressão e controle aduaneiros de fronteira, portos e aeroportos atuam diretamente sobre a transação comercial no momento em que esta está ocorrendo. Seja a partir de denúncias ou a partir de um planejamento estratégico, servidores da RFB interceptam a mercadoria nas estradas ou nos pontos de embarque e desembarque para verificação da sua conformidade documental e física. Muitas exportações fictícias são assim identificadas, especialmente a falsa declaração de conteúdo.

Fiscalização aduaneira tem uma atuação com viés documental e contábil. Esta tem prazo para ocorrer até a prescrição do fato gerador do tributo, podendo portanto levar anos. Sua origem é a seleção de contribuintes por cruzamento de dados ou por denúncias. Os contribuintes selecionados são incluídos na programação fiscal dos anos subsequentes para que seja realizada a fiscalização. Para o foco deste trabalho, observa-se que, quando a fiscalização aduaneira atua na exportação, atua após o trânsito da mercadoria o que dificulta a materialidade do ilícito porventura detectado.

Investigação tributária visa combater, dentre outros crimes, a lavagem de dinheiro nas operações de comércio exterior. De forma análoga à fiscalização aduaneira, sua origem se dá a partir de denúncias ou cruzamento de dados. Porém, sua atuação pode ocorrer no momento das transações ou posteriormente, dentro do prazo prescricional penal.

Em todos esse casos, o sucesso do trabalho da RFB conta com o bom uso das suas bases de dados em conjunto com a competência e qualidade do auditor-fiscal responsável. Acredita-se que seja possível transpor, ainda que parcialmente, esse conhecimento empírico dos especialistas para um modelo de mineração de dados que aumente a produtividade e a tempestividade das análises com vistas à seleção de contribuintes para fiscalização e investigação.

5.2 Recursos disponíveis

Esta seção apresenta os recursos tecnológicos de infraestrutura e dados disponíveis na RFB para a realização do presente trabalho.

5.2.1 Infraestrutura

A RFB conta com um *Laboratório de Tecnologia Contra a Lavagem de Dinheiro* (Lab-LD) integrante da *Rede Nacional de Laboratórios contra Lavagem de Dinheiro* (Rede-LAB) do Ministério da Justiça. Além da infraestrutura de hardware, o laboratório possui diversas

ferramentas para análises estatísticas e mineração em grandes volumes de dados. As seguintes plataformas estão atualmente disponíveis para realização deste trabalho:

R[3] é uma linguagem para computação estatística que permite, por meio de pacotes externos, a agregação de outras funcionalidades como as de mineração de dados.

RStudio[4] é um software livre que é usado como ambiente de desenvolvimento integrado para R.

H2O[5] é uma máquina virtual Java otimizada para fazer processamentos distribuídos e algoritmos de aprendizado de máquina paralelas em *clusters*. *H2O* é integrável ao R a partir do pacote *R-H2O*.

Contágil é uma ferramenta de extração, manipulação e análise de dados desenvolvida pela própria RFB. Possui integração com o R e, das ferramentas apresentadas, é a única capaz de interagir diretamente com as bases de dados de produção e em *Data Warehouse* [75].

5.2.2 Dados

Desde a década de 90 o governo brasileiro trata todo o trâmite aduaneiro de forma eletrônica por meio do Sistema Integrado de Comércio Exterior Brasileiro[6] (*Siscomex*). De forma análoga são tratados em diversos sistemas própios da RFB todas informações relativas aos tributos federais. Assim, é possível afirmarmos que todos os dados necessários ao trabalho estão disponíveis em bases de dados eletrônicas.

Quanto à existência de dados rotulados para classificação (variável dependente), a RFB possui diversas frentes de trabalho que atuam no comércio exterior: fiscalização aduaneira de portos, aeroportos e fronteiras; fiscalização aduaneira dentro do território nacional; repressão aduaneira; investigação de crimes tributários (ver Seção 5.1 para maiores detalhes). Todas essas áreas possuem informações provenientes de ações do fisco federal ocorridas no passado sobre fraudes em exportações para serem utilizadas numa análise supervisionada.

[3] `https://cran.r-project.org/`
[4] `https://www.rstudio.com/`
[5] `http://www.h2o.ai/`
[6] `http://www.portalsiscomex.gov.br/`

5.3 Restrições Legais aplicáveis ao presente trabalho

A presente pesquisa realiza-se a título de mestrado profissional, com seu tema tendo sido proposto em conjunto com a RFB. Tal fato é facilitador para a condução dos trabalhos pelo próprio interesse da instituição no seu sucesso. Contudo, cabe esclarecer que em obediência à legislação em vigor[7], determinadas restrições se impõem quanto ao local de manipulação dos dados e quanto à apresentação de resultados parciais e finais.

O parágrafo 2º do artigo 2º da Portaria RFB n º 2.344 [77], de 24 de março de 2011 é explícito ao caracterizar, dentre outras condutas, que a divulgação de informações, agregadas ou não, mesmo não expondo a identificação do contribuinte, caracteriza descumprimento do dever de sigilo funcional previsto no art. 116, inciso VIII, da Lei Nº 8.112 [78], de 1990.

In verbis

> Art. 2º São protegidas por sigilo fiscal as informações sobre a situação econômica ou financeira do sujeito passivo ou de terceiros e sobre a natureza e o estado de seus negócios ou atividades, obtidas em razão do ofício para fins de arrecadação e fiscalização de tributos, inclusive aduaneiros, tais como:
>
> I - as relativas a rendas, rendimentos, patrimônio, débitos, créditos, dívidas e movimentação financeira ou patrimonial;
> II - as que revelem negócios, contratos, relacionamentos comerciais, fornecedores, clientes e volumes ou valores de compra e venda;
> III - as relativas a projetos, processos industriais, fórmulas, composição e fatores de produção.
>
> § 1º Não estão protegidas pelo sigilo fiscal as informações:
>
> I - cadastrais do sujeito passivo, assim entendidas as que permitam sua identificação e individualização, tais como nome, data de nascimento, endereço, filiação, qualificação e composição societária;
> II - cadastrais relativas à regularidade fiscal do sujeito passivo, desde que não revelem valores de débitos ou créditos;
> III - agregadas, que não identifiquem o sujeito passivo; e
> IV - previstas no § 3º do art. 198 da Lei Nº 5.172, de 1966.
>
> § 2º A divulgação das informações referidas no § 1º caracteriza descumprimento do dever de sigilo funcional previsto no art. 116, inciso VIII, da Lei Nº 8.112, de 1990.

Decorre portanto que em nenhum momento o tratamento dos dados do presente trabalho poderá ser realizado fora das dependências e equipamentos da RFB.

Decorre ainda que os documentos produzidos devem observar as regras de sigilo fiscal e funcional ainda que em prejuízo da clareza do trabalho.

[7] art. 199 da Lei n º 5.172 [76], de 25 de outubro de 1966 – Código Tributário Nacional (CTN) e Portaria RFB n º 2.344 [77], de 24 de março de 2011.

5.4 Critérios de resultado para sucesso da mineração de dados

Elencam-se abaixo critérios para aceitação dos resultados:

1. Possibilidade do modelo desenvolvido ser implementado nos sistemas da RFB;
2. Validação de bases de testes realizadas por métricas aceitas pela comunidade de mineração de dados;
3. Identificação pelo modelo dos casos já conhecidos pela RFB com acurácia e especificidade medidas em bases de avaliação (base que não participaram das fases de treinamento e testes) superiores às conseguidas atualmente.

Capítulo 6

Entendimento e Preparação dos Dados

O presente capítulo apresenta as fases de entendimento e preparação dos dados. As seções deste capítulo estão distribuídas da seguinte forma: a Seção 6.1 apresenta as bases de origem dos dados, seus atributos e descrição; a Seção 6.2 apresenta a análise exploratória realizada sobre os dados, a análise de consistência, de varância e das suas distribuições; a Seção 6.3 analisa a correlação entre os atributos; a Seção 6.4 busca identificar distorções nos dados e analisar os *outliers*; a Seção 6.5 analisa a linearidade na relação entre os atributos; a Seção 6.6 seleciona os modelos mais adequados aos dados para os testes de indução; a Seção 6.7 prepara os dados para indução do modelo, seus testes e avaliação.

6.1 Coleta de dados inicial e descrição das bases

Dentre o amplo conjunto de informações eletrônicas disponíveis na RFB, foram identificadas aquelas que se supõe serem as mais adequadas ao atingimento dos objetivos propostos. Assim, em reuniões com especialistas da RFB na área de investigação dos crimes de lavagem de dinheiro no comércio exterior e das áreas de fiscalização aduaneira e de vigilância e repressão aduaneira, levantou-se de forma empírica quais dados representativos da atividade econômica do contribuinte seriam capazes de explicar o comportamento da variabilidade nos valores exportados e das fraudes na exportação.

Foram coletados inicialmente 77 atributos, oriundos de 8 bases de dados distintas, contendo, além dos dados representativos da atividade econômica do contribuinte, informações cadastrais, sociais e características das mercadorias exportadas.

A seguir são apresentadas de forma sumária a descrição dos dados coletados. Para melhor entendimento, os dados foram agrupados em bases que refletem características afins dos atributos.

Base *Arrecadação* (BArr) Os atributos selecionados da base *Arrecadação* indicam o total de tributos federais efetivamente recolhidos pelas empresas exportadoras. Traz também informações oriundas da base de cálculo de alguns tributos que indicam os diversos valores totais de receitas das empresas.

Os dados são compostos de seis atributos numéricos e originários de três fontes:

1. Declarações diversas prestadas por contribuintes à RFB;
2. Demonstrativo de apuração de contribuições sociais (Dacon);
3. Bases do Banco Central do Brasil (BC) - quanto aos valores do efetivo recolhimento do *Documento de Arrecadação de Receitas Federais* (DARF).

Neste trabalho esses atributos encontram-se referenciados como *atributo 1* a *atributo 6*.

Base *Cadastros* (BCad) Os atributos selecionados da base *Cadastro* apresentam informações quanto à identificação da empresa exportadora, o tipo de atividade econômica realizada, sua situação cadastral atual e passada (ativa, inativa ou suspensa).

Estes dados são compostos de 14 atributos, sete deles do tipo *caracter*[1], dois atributos de data e cinco categóricos. Todos têm origem nas diversas declarações de interesse do fisco federal que são prestadas pelos contribuintes ao longo do ano. Elas refletem portanto a última informação transmitida à RFB. Neste trabalho encontram-se referenciados como *atributo 7* a *atributo 20*.

Base *Comércio Exterior* (BCE) Os atributos selecionados da base *Comércio Exterior* trazem as movimentações realizadas no comércio exterior pelas empresas exportadoras. Nesta base se encontram as informações sobre os valores e quantitativos exportados e importados em cada declaração de exportação (DE) e declaração de importação (DI), respectivamente. Demais informações relativas às características das mercadorias são encontradas na base *Notas Fiscais Eletrônicas*.

Estes dados têm origem no Sistema Integrado de Comércio Exterior Brasileiro[2] (*Siscomex*) e são compostos de quatro atributos do tipo numérico sendo referenciados neste trabalho como *atributo 21* a *atributo 24*.

Base *Contribuições, Tributos e Benefícios Fiscais* (BCTBF) Esta base contém os tributos e contribuições que são apurados e declarados pelas empresas por meio de

[1] Mantiveram-se aqui atributos do tipo *caracter* para permitir a identificação das empresas.
[2] `http://www.portalsiscomex.gov.br/`

programas específicos. Foram selecionados dessa base os valores declarados como devidos, os créditos existentes para compensação e os benefícios fiscais informados.

Estes dados são compostos de três atributos do tipo numérico e têm origem nas seguintes declarações apresentadas pelos contribuintes:

1. Declaração de Contribuições Federais (DCTF)
2. Declaração de Benefícios Fiscais (DBF)

Neste trabalho são referenciados como *atributo 28* a *atributo 30*.

Base *Empregados* (BEmp) Os atributos selecionados da Base *Empregados* refletem indiretamente, por meio dos pagamentos da Guia da Previdência Social (GPS), a mão de obra empregada em cada empresa exportadora. São compostos de três atributos do tipo numérico e são referenciados dentro deste trabalho como *atributos 28* a *atributo 30*.

Base *Movimentações Financeiras* (BMF) Os atributos selecionados da base *Movimentações Financeiras* apresentam informações sobre transações em moeda nacional, estrangeira e cartões de crédito. Compreendem as operações de débito/crédito (moeda nacional) e compra, venda e transferências (moeda estrangeira).

Os dados originam-se das Declarações de Informações sobre Movimentação Financeira - DIMOF prestadas pelos bancos, cooperativas de crédito e associações de poupança e empréstimo. Compões-se de onze atributos do tipo numérico e são referenciados como *atributo 31* a *atributo 41*.

Base *Notas Fiscais Eletrônicas* (BNFe) Os atributos selecionados da Base *Notas Fiscais Eletrônicas* indicam os documentos fiscais de trânsito de mercadorias e serviços quando da sua aquisição ou quando da saída para comercialização pelas empresas exportadoras. Trazem dados pormenorizados dos insumos usados nas indústrias exportadoras e das mercadorias adquiridas para posterior exportação. A origem desses dados é o Sistema Público de Escrituração Digital (*SPED*).

Esta base é composta de sete atributos numéricos, sete atributos categóricos, e três atributos do tipo caracter[3]. Totalizam assim dezessete atributos e são referenciados com *atributo 42* a *atributo 58*.

[3]Mantiveram-se aqui atributos do tipo caracter para permitir análises empíricas quanto ao tipo de mercadoria objeto da NFe

Base *Retenções de Impostos na Fonte* (BRIF) Os atributos selecionados da base *Retenções de Impostos na Fonte* indicam o recolhimento de tributo por parte das empresas em nome de outrem quando da ocorrência de algum pagamento. Tais dados abrangem inclusive aqueles incidentes sobre pagamentos enviados ao exterior. Esta informação é complementar às informações contidas na base *Arrecadação* e não se encontrando, portanto, coletada de forma duplicada.

Estes dados são originários da Declaração do Imposto de Renda Retido na Fonte (DIRF) e é composto de 19 atributos do tipo numérico semdo, neste trabalho, referenciados como *atributo 59* a *atributo 77*.

Variável dependente - rotulagem de atributo Não há nas bases de dados da RFB uma classificação explícita dos contribuintes quanto à suspeição de lavagem de dinheiro ou de operarem exportações de forma fictícia. É necessário que essa base seja construída a partir de várias fontes diferentes e que foram identificadas na fase de entendimento do negócio (ver Seção 5.2.2). Dessa forma, criou-se um atributo adicional, binário, contendo a classificação quanto à ocorrência de alguma irregularidade cometida pela empresa nas exportações e que pudesse caracterizar fraude na exportação. Este atributo criado possui 2.719 registros e apenas dois valores: *suspeito* e *não suspeito*. No rótulo *suspeito* estão o conjunto de ocorrências verificadas nas atividades fiscais, para o rótulo *não suspeito* foram usadas as verificações realizadas pela RFB e que não resultaram em sanções às empresas.

Os rótulos encontram-se desbalanceados na proporção de 1 rótulo *suspeito* para cada 3 *não suspeito*.

Apesar de ser um atributo *classificado*, pois recebeu a classificação quanto à suspeição, neste trabalho dá-se a ele o nome de atributo *rotulado* para evitar confusões semânticas com os dados resultantes de modelos classificadores.

6.2 Exploração e verificação da qualidade dos dados

Abaixo são apresentados os resultados obtidos a partir da exploração e verificação da qualidade dos dados.

6.2.1 Análise de consistência dos dados

Foram feitas diversas análises de consistências dos dados. Abaixo encontram-se os principais achados:

Análise de unicidade dos dados Não foram encontrados dados duplicados nas bases. Contudo, verificou-se a existência de duas empresas com nome idêntico e CNPJs diversos. De acordo com consulta diretamente na base de produção verificou-se que a informação está correta.

Identificação de *Missing Values* Com exceção dos atributos da base *cadastro* e da base *comércio exterior*, todos os demais apresentaram *Missing Values*. Os atributos 2, 3, 4, 5, 6, 27, 30, 35, 36, 37, 39, 40, 62, 64, 67, 68, 71, 72 e 73 apresentaram mais de 60% de dados faltantes:

A ausência desses dados deve-se à diferença de obrigações acessórias entre empresas. A depender do porte da empresa, de opção de tributação ou de legislação específica, determinados atributos de fato não existem, pois não ocorrem ou não há obrigação de informá-los à RFB.

6.2.2 Identificação de atributos numéricos com dados constantes ou com variação em poucos registros

A partir da análise da variação dos registros de cada atributo, foi identificado que os seguintes atributos possuíam valores constantes ou apresentava variação de valores em menos de 0,2% dos registros: atributos 5, 6, 36, 37, 71 e 72.

6.2.3 Análise de distribuições

A análise das distribuições de frequência dos dados que envolvem montantes financeiros mostrou, em todos os casos, forte concentrações assimétricas à esquerda (assimetria negativa). Tal fato corresponde ao esperado pois a frequência de empresas tende a cair à medida que os montantes financeiros que indicam suas atividade aumentam: é mais frequente empresas de pequeno porte (com pequenos montantes financeiros) que empresas muito grandes (com grandes montantes financeiros).

As Figuras 6.1, 6.2, 6.3 e 6.4 apresentam para 4 variáveis [4] o histograma correspondente. Por sua semelhança a uma distribuição log-normal [79], a mesma figura apresenta para cada variável seu correspondente normalizado pela função: $f(x) = log(y)$

[4] A análise foi realizada em todas as variáveis. Aqui constam apenas 4 de forma exemplificativa.

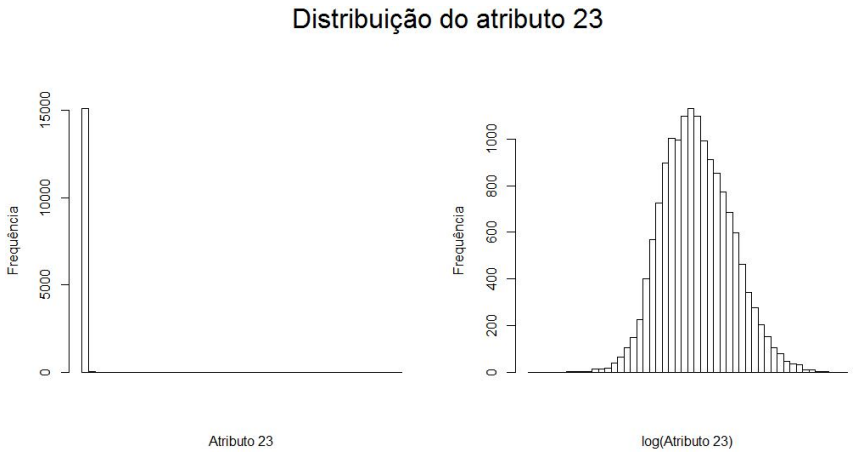

Figura 6.1: Distribuição do atributo 23.

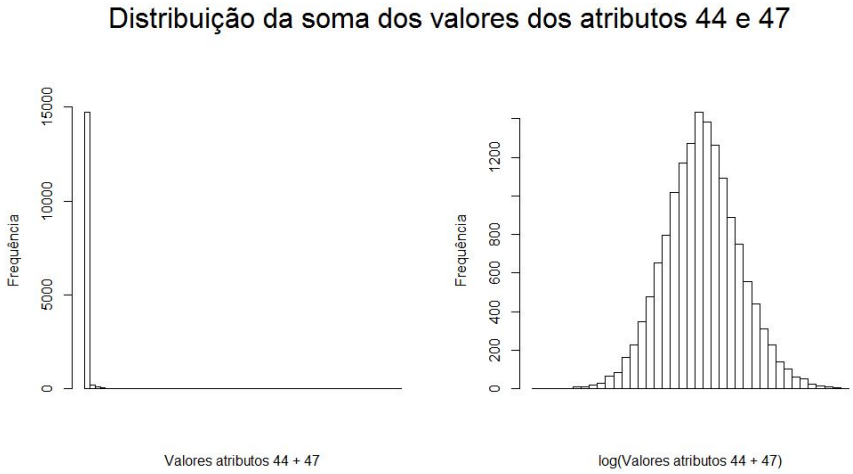

Figura 6.2: Distribuição da soma dos valores dos atributos 44 e 47.

6.3 Análise de correlação entre variáveis

Após serem retiradas da base de dados todas as variáveis que possuíam mais de 60% de *Missing Values* (Seção 6.2.1), bem como todas as variáveis com baixa variância (Seção 6.2.2), procedeu-se à analise da correlação dos demais atributos.

A Figura 6.5 mostra o quadro de cruzamento das distribuição de correlações dos atributos numéricos. Percebe-se a existência de alta correlação entre alguns atributos (elipses estreitas e azul escuras indicando correlação próxima de 1). Percebe-se também a inexistência de correlações negativas entre atributos (representações com cores tendentes ao vermelho). A análise da viabilidade para retiradas desses atributos será feita na Seção 6.7, oportunidade em que serão tratados assuntos pertinentes à preparação dos dados para indução do modelo.

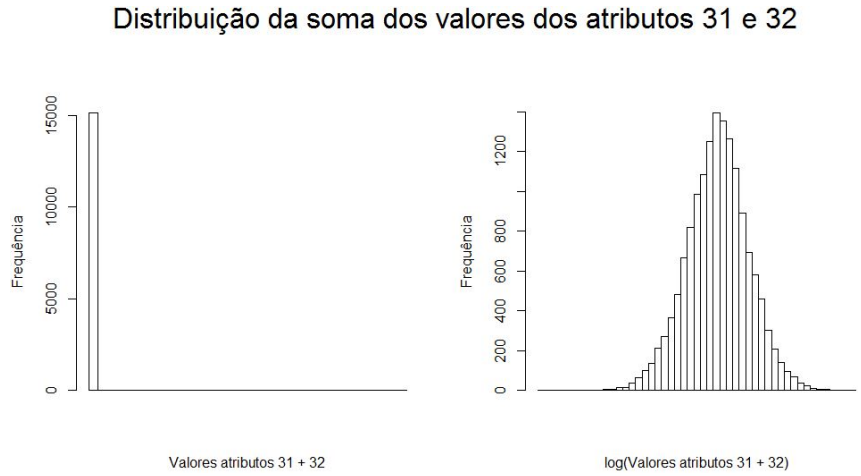

Figura 6.3: Distribuição da soma dos valores dos atributos 31 e 32.

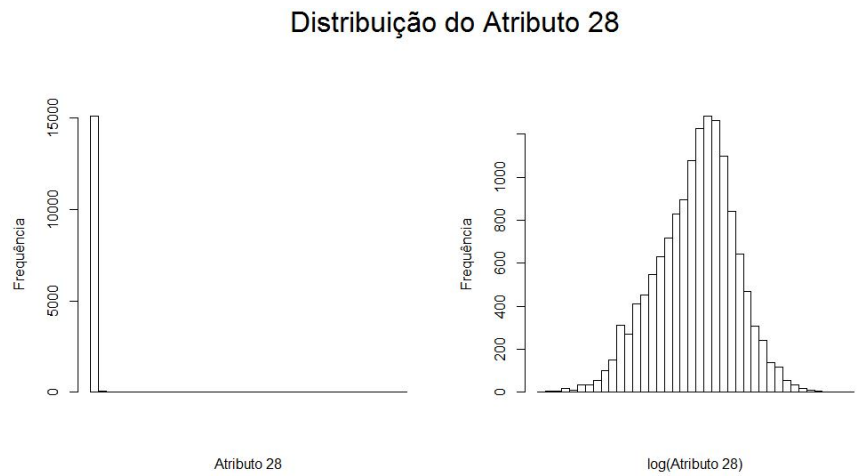

Figura 6.4: Distribuição do atributo 28.

6.4 Análise de distorções e de *outliers*

A fim de se verificar distorções nos dados, todos atributos numéricos foram divididos em 10 decis[5] e analisados na forma de *boxplots*. Para esta análise, foram descartados os atributos identificados da Seção 6.2.1 cujo quantitativo de *Missing Values* superou 60% do total de dados coletados por atributo.

Segundo Dawson [80], a análise de dados a partir de *boxplots* apresentam melhor resultado em distribuições normais ou assemelhadas. Assim, devido à distribuição log-normal dos dados, identificada na Seção 6.2.3, procedeu-se à transformação dos dados pela função: $f(x) = log(y)$.

[5] A escolha do número de decis ocorreu em função de uma melhor visualização dos dados: uma divisão menor, por exemplo em quartis, não seria capaz de bem representar as distorções; já uma divisão em quantidades maiores poderia poluir desnecessariamente o gráfico.

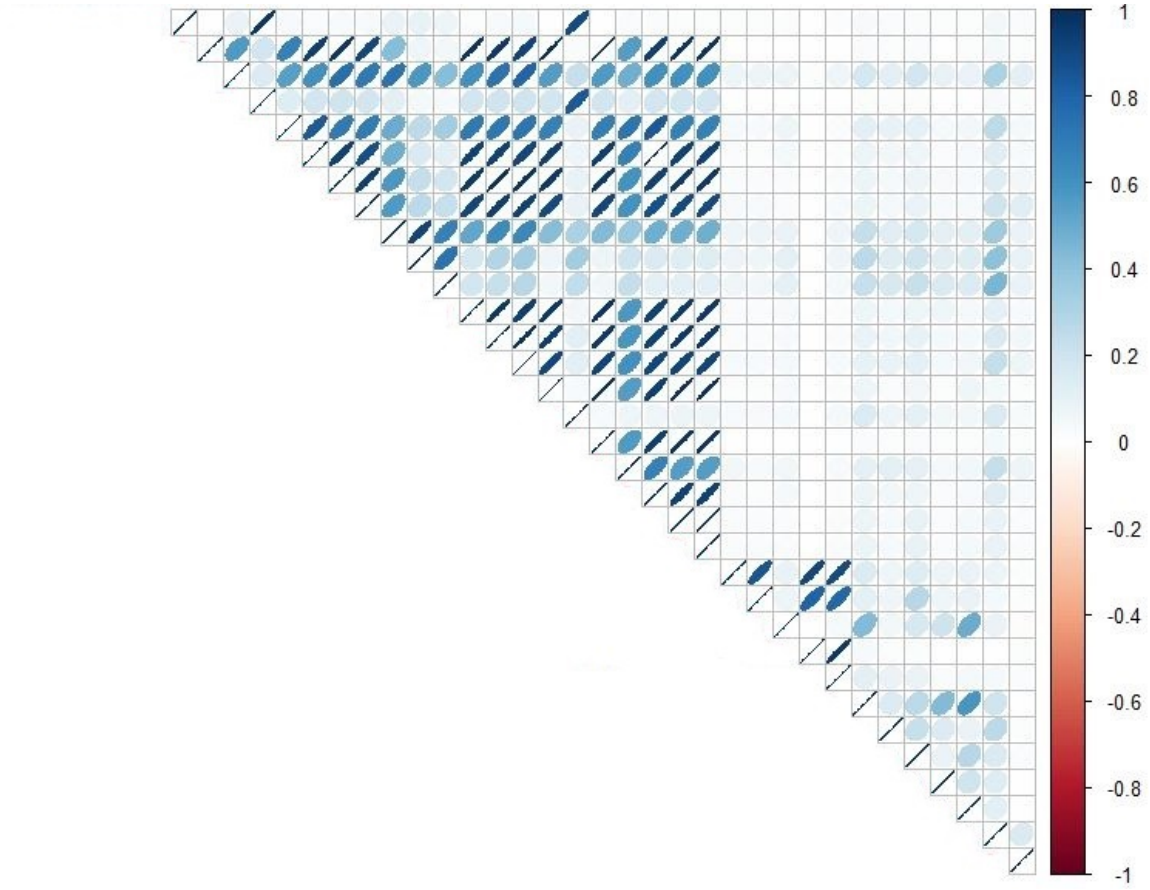

Figura 6.5: Correlação entre os atributos.

Foram gerados 32 gráficos contendo em cada um deles o *boxplot* relativo a cada atributo. Em todos os casos verificou-se a existência de valores extremamente altos. Pela análise pormenorizada desses dados conclui-se que não se tratam de erros nas bases de dados, mas sim de dados reais, pertencentes a grandes grupos empresariais que, de fato, destacam-se quanto aos valores.

Observam-se também nestes gráficos, em 21 atributos, a existência de valores extremamente pequenos. Feita a verificação direta e individualmente na base de dados em produção, percebeu-se que naqueles atributos oriundos das bases *Arrecadação* e *Contribuições, Tributos e Benefícios Fiscais* havia 23 dados claramente inconsistentes. Como esses dados apresentavam valores de exportação inexpressivos (inferiores a R$ 10.000,00), foram retirados da base. Assim, excetuando-se esses citados registros, conclui-se pela inexistência de *Missing Values*.

Observou-se ainda que nos decis centrais, entre o 2º e o 9º decis, alguns *boxplots* apresentam nítida diferença de amplitude entre os quartis centrais. Tal fato revela assimetria na distribuição log-normal desses atributos sem contudo indicar necessariamente problemas de coleta.

6.5 Análise dos relacionamentos entre atributos

A análise dos relacionamentos entre atributos pode revelar significativas anomalias entre os dados, além de esboçar, ainda que isoladamente e sem a percepção de todas as variáveis, a natureza da relação entre eles. Porém, combinados 2 a 2, teríamos para os 69 atributos (excetuam-se aqui os atributos do tipo *caracter*) deste trabalho 2.346 ($C_{69,2} = 2.346$) análises possíveis.

Assim, pela inviabilidade de se esgotar a análise, optou-se por efetuar a análise apenas contra o atributo que julgamos mais representativo de distorções. Entende-se que dado o objetivo deste trabalho, as relações e as anomalias contra este atributo são as mais relevantes. Chamaremos esse atributo de *atributo paradigma*.

Conforme já explicitado na Seção 6.2.3, usou-se a distribuição log-normal transformada em normal para todos os atributos em análise. Assim, este atributo foi dividido em 20 partes[6], cada uma correspondendo a 5 percentis.

Relacionamentos lineares

O primeiro gráfico apresentado na Figura 6.6 mostra relação linear com o *atributo paradigma* e de uma variável sem *Missing Values*. A perceptível variância desse atributo ao longo dos *boxplots* é explicável por sua não necessária correlação com o *atributo paradigma*.

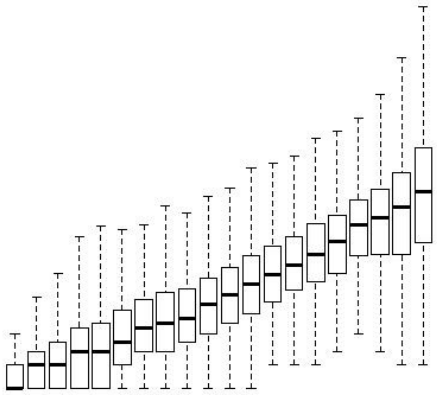

Figura 6.6: Relacionamento entre os atributo 23 e o log do atributo 21.

[6] A escolha do número de partes, tal qual se fez na Seção 6.3, se deu em função de uma melhor visualização dos dados: por tentativa e erro, julgou-se que uma divisão menor não seria capaz de bem representar as relações entre atributos; já uma divisão em quantidades maiores poderia poluir desnecessariamente o gráfico.

A Figura 6.7 mostra, como outro exemplo de relações lineares encontradas, duas variáveis que também apresentam variabilidade aparentemente linear quando confrontadas com o *atributo paradigma*. A ausência observada em alguns *boxplots* da marcação e plotagem dos quartis inferiores se deve a muitos valores iguais a zero existentes nas bases de dados, o que, para a análise em questão, não invalida a constatação da linearidade no relacionamento.

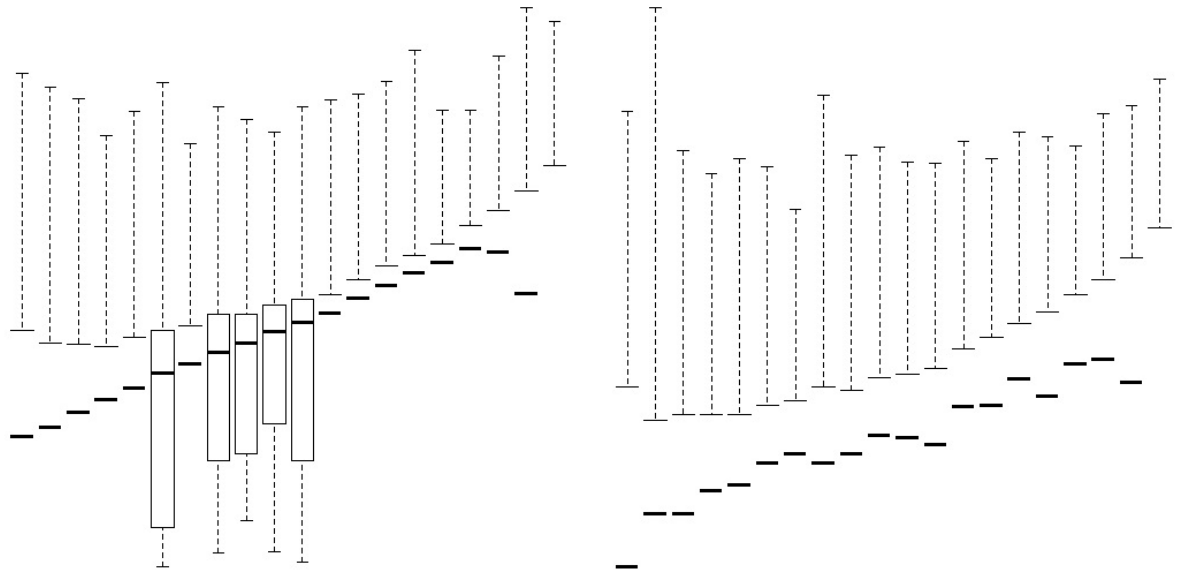

Figura 6.7: Relacionamento aparentemente linear entre atributos.

Relacionamentos não lineares

Devido ao fato de que determinados algoritmos não generalizam bem quando trabalham com dados não-lineares (é o caso por exemplo do *k-means* e dos algoritmos de regressão linear e suas variantes como o *Generalized Linear Models*) passamos à busca deste tipo de relação entre as variáveis.

Como exemplo de relacionamentos não lineares encontrados nas análises, apresentam-se na Figura 6.8 a plotagem de dois importantes atributos. A forma do gráfico sugere a existência de uma variação mais acentuada, dada pela variação da inclinação, à medida que o *atributo paradigma* aumenta.

Relacionamentos não identificáveis visualmente

As demais variáveis mostraram comportamentos erráticos ou de ordem não identificável quando comparadas com o *atributo paradigma*. A Figura 6.9 traz dois casos dessa relação.

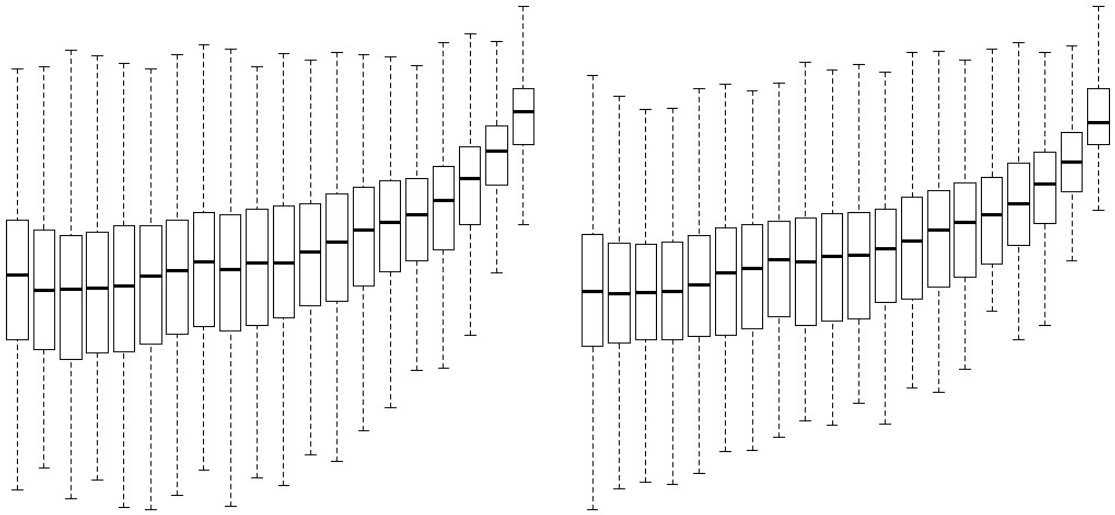

Figura 6.8: Relacionamento não-linear entre atributos.

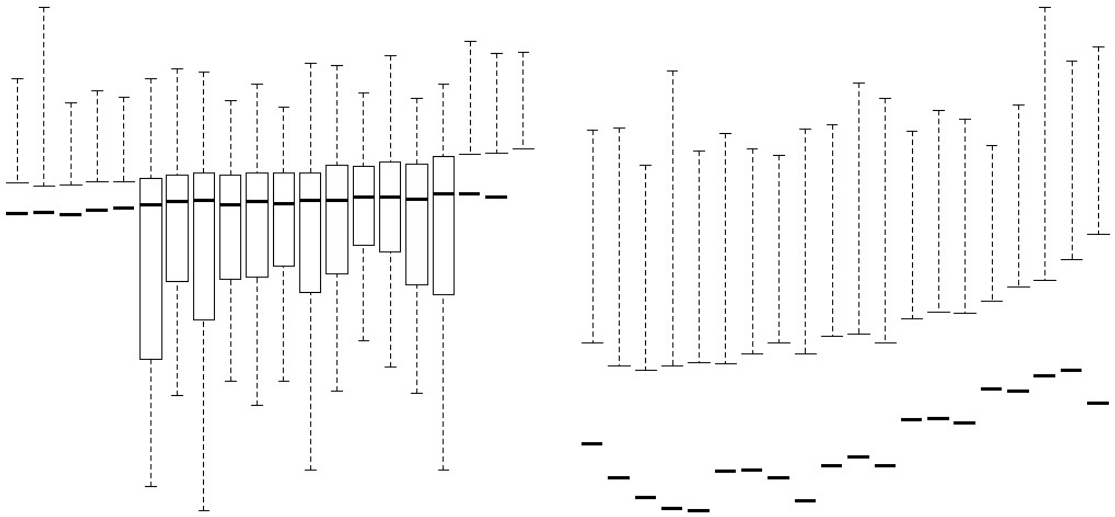

Figura 6.9: Relacionamentos entre atributos não identificáveis visualmente.

A diversidade de relacionamentos entre os atributos identificados nessa seção terão importante papel na definição das abordagens que serão utilizadas na indução do modelo realizada no Capítulo 7.

6.6 Escolha dos prováveis modelos

A partir do entendimento do negócio, especificamente na Seção 5.2.2, temos, do ponto de vista dos dados, a possibilidade de utilizarmos uma análise supervisionada partindo dos dados classificados como suspeitos de fraude à exportação. Evidentemente, sempre há a alternativa de uma análise não supervisionada que, apesar de poder se apresentar com um direcionamento diferente daquele dado pelos dados rotulados, pode descobrir formas de atuação de fraudes diversas do padrão utilizado pelos dados rotulados.

Ainda a partir do entendimento do negócio, especificamente na Seção 5.2.1, temos que a ferramenta *H2O* nos traz as seguintes alternativas de algoritmos a serem adotados:

- *Generalized Linear Models* (GLM) [81]
- *Distributed Random Forest* (DRF)
- *Gradient Boosting Machine* (GBM)
- *K-Means* [17]
- *Deep Learning*
- *Naïve Bayes* [82]

Preliminarmente, excluiremos três destes algoritmos em função dos estudos realizados nas seções anteriores: o GLM por ser um modelo que pressupõe que o valor esperado da variável resposta é uma função linear das covariáveis [81], fato que empiricamente sabe-se não verdadeiro; K-Means pela natureza não normal de vários atributos, por não serem balanceadas as contribuições de cada atributo e por não estar entre os objetivos do projeto a descoberta de clusters; Naïve Bayes pelo fato de que as variáveis são bastante dependentes umas das outras.

Assim, a indução do modelo será feita de forma supervisionada para *Distributed Random Forest* e *Gradient Boosting Machine* e de forma não supervisionada, como detector de anomalias, para *Deep Learning* na função de *Autoencoder*.

6.7 Preparação dos dados para indução dos modelos

A partir do conhecimento obtido com o entendimento dos dados, com a verificação da sua qualidade e com a seleção de algoritmos de mineração a serem testados, procedeu-se aos seguintes ajustes nas bases de dados.

Eliminação de registros As bases de dados oriundas da RFB apresentam, após agregadas por empresa, um total de 15.265 registros. Durante a busca de distorções e *outliers* realizada na Seção 6.4, foram identificados 23 registros que após análise foram eliminados da base.

Eliminação de atributos Para os três algoritmos onde serão realizados os treinamentos e testes de indução do modelo optou-se por remover os atributos que apresentaram mais de 60% de *Missing Values* dentre seus registros (atributos identificados na Seção 6.2.1) e baixa variância (atributos identificados na Seção 6.2.2).

Para a indução de modelos supervisionados, optou-se por remover também os atributos altamente correlacionados identificados na Seção 6.3 mantendo-se os atributos com menor quantidade de *Missing Values*. A fim de preservar anomalias por ventura existentes em um atributo, mas não existente no atributo correlacionado, tal decisão não foi aplicada na preparação dos dados do algoritmo *Deep Learning*.

Particionamento das Bases As bases de dados após a eliminação acima permaneceu com 15.242 registros, dos quais 2.719 possuem um atributo binário indicando a classificação quanto à suspeição da empresa. Maiores detalhes sobre a coleta na RFB destes dados de suspeição podem ser obtidas na Seção 6.1.

Para o treinamento no algoritmo de *Deep Learning Autoencoder*, não houve nenhum particionamento da base de dados, sendo portanto utilizado na sua construção todos os 15.242 registros.

Para o treinamento dos algoritmos supervisionados, procedeu-se à separação da base original em dois grupos que receberam os nomes de: *dados rotulados* e *dados não rotulados* de acordo com a existência de classificação prévia dos dados por parte da RFB.

Os *dados rotulados* foram novamente divididos, agora de forma aleatória, em outros dois grupos que receberam os nomes de: *base de treinamento* e *base de avaliação*. Coube à *base de treinamento* 75% dos registros aleatoriamente selecionados e à *base de avaliação* os 25% restantes.

Para verificar a capacidade de generalização dos modelos supervisionados foram construídos a partir da *base de treinamento* 10 subconjuntos de mesmo tamanho e mutuamente exclusivos denominados *folds* e rotulados como *k-folds* sendo k um número de 1 a 10 representativo do subconjunto.

Balanceamento dos dados A proporção entre dados rotulados como *suspeito* e *não suspeito* é de aproximadamente 1:3. Optou-se por não fazer ajustes prévios nas bases de treinamento, pois todas implementações de algoritmos supervisionados no *H2O* permitem a parametrização para balanceamento durante as etapas de treino e testes.

Capítulo 7

Indução do Modelo e Análise de Resultados

Este capítulo apresenta a indução de modelos para as técnicas selecionadas no Capítulo 6. Apresenta ainda os testes destes a partir de métricas adequadas a cada técnica empregada. O capítulo se divide inicialmente em três seções, uma para cada técnica selecionada. Por fim, numa quarta seção, apresenta-se um comparativo entre os melhores resultados obtidos.

Conforme visto na Seção 6.6, a indução ocorrerá utilizando-se duas técnicas supervisionadas e uma não supervisionada. Foram adotadas as seguintes técnicas de explanação de acordo com o tipo de supervisão:

Técnicas supervisionadas primeiramente apresentamos um quadro resumo comparativo entre os diversos parâmetros de ajustes[1] testados naquela técnica, utilizando-se a métrica *logloss* como parâmetro de seleção. A seguir, para o modelo de melhor resultado (menor *logloss*), apresenta-se uma análise mais detalhada das métricas de treinamento utilizadas. Em todos os casos, usou-se *Cross-Validation*[2] como forma de testar a maior ou menor adequação aos dados dos modelos gerados.

Técnicas não-supervisionadas apresentam-se todos os modelos gerados a partir dos dados que melhor explicaram os modelos supervisionados. Como critério de avaliação objetiva, ainda que não conclusivo, comparam-se os resultados obtidos de forma não-supervisionada com dois grupos de referência: o atributo rotulado pela RFB (ver Seção 6.1); e a classificação pelos modelos supervisionados.

[1] Além dos testes em parâmetros próprios de cada técnica, também foi testada a indução dos modelos com e sem balanceamento do atributo de classe.

[2] Ainda que o treinamento tenha ocorrido de forma balanceada, o *fold* usado para testes manteve-se na proporção dos dados originais.

7.1 *Gradient Boosting Machine* (GBM)

Essa seção apresenta os resultados obtidos pelo uso da técnica supervisionada *Gradient Boosting Machine*.

A partir dos dados preparados na fase anterior (Seção 6.7), foram gerados modelos com e sem balanceamento. Em cada caso variou-se a quantidade de árvores (40, 55 e 70 árvores) e sua profundidade máxima permitida no algoritmo (3, 6 e 10). A Tabela 7.1 mostra os valores de *logloss* em *Cross-Validation* com (10 *folds*) para escolha do modelo que melhor reduziu os erros.

Percebe-se nesta tabela que a variação no número de árvores e o balanceamento de classes tiveram um papel secundário na redução do *logloss*, sendo que a diminuição da profundidade máxima das árvores apresentou um papel nitidamente mais relevante. Assim, nas subseções que se seguem, apresentaremos uma análise mais detalhada do modelo que apresentou maior redução de erros na média dos extratos de *Cross-Validation* - o modelo *GBM_model_7*.

Tabela 7.1: Média dos valores de *logloss*

Balanceamento	Profundidade Máxima	Nº de Árvores	Identificação do Modelo	*Valor de logloss em Cross-Validation*
Verdadeiro	3	55	GBM_model_7	0.3841
Falso	3	70	GBM_model_12	0.3848
Verdadeiro	3	70	GBM_model_13	0.3855
Falso	3	55	GBM_model_6	0.3865
Verdadeiro	3	40	GBM_model_1	0.3873
Verdadeiro	6	55	GBM_model_9	0.3887
Falso	3	40	GBM_model_0	0.3898
Verdadeiro	6	40	GBM_model_3	0.3915
Falso	6	40	GBM_model_2	0.3948
Falso	6	55	GBM_model_8	0.3969
Falso	6	70	GBM_model_14	0.4090
Verdadeiro	6	70	GBM_model_15	0.4106
Falso	10	40	GBM_model_4	0.4243
Verdadeiro	10	40	GBM_model_5	0.4311
Verdadeiro	10	55	GBM_model_11	0.4466
Falso	10	55	GBM_model_10	0.4513
Falso	10	70	GBM_model_16	0.4675
Verdadeiro	10	70	GBM_model_17	0.4787

Análise de curva ROC

A Figura 7.1 apresenta a curva ROC a partir dos valores obtidos por *Cross-Validation* do modelo *GBM_model_7* em 10 *folds*. A área abaixo da curva tem cobertura ligeiramente

maior que 90% do total do gráfico. Percebe-se também que o modelo é capaz de atingir aproximadamente 40% de taxa de verdadeiros positivos (tpr) sem contudo apresentar acréscimos da taxa de falsos positivos (fpr).

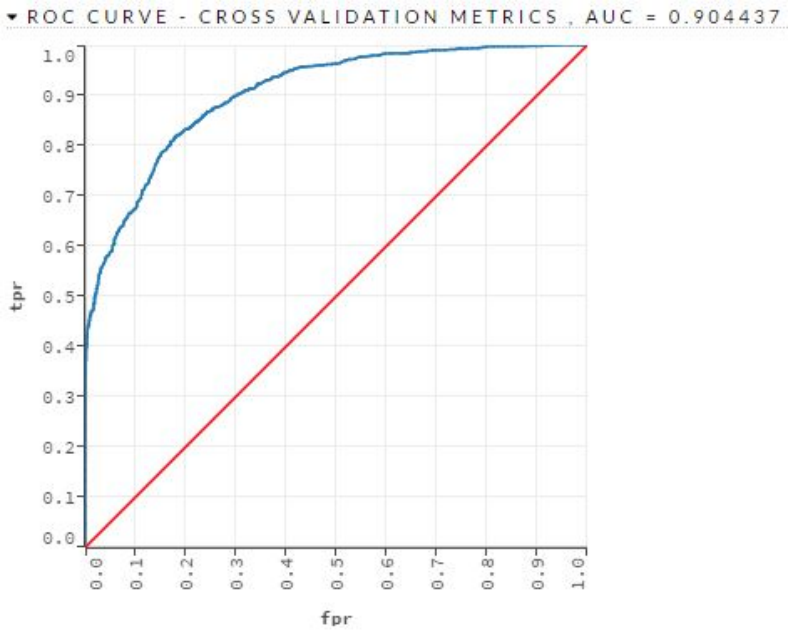

Figura 7.1: Curva ROC - *Cross-Validation* do modelo *GBM_model_7*.

A Tabela 7.2 apresenta os *threshold* da curva ROC em função das seguintes cinco métricas: *f1 measure*, acurácia e *matthews correlation coefficient* (mcc) absoluto.

Tabela 7.2: Valores de *threshold* e métricas correspondentes. Modelo *GBM_model_7*

threshold	0,4776	0,4877	0,5328
f1 measure	***0,8423***	0,8412	0,836
acurácia	0,8402	***0,8402***	0,8398
precisão	0,8333	0,838	0,8589
recall	0,8515	0,8445	0,8142
especificidade	0,8287	0,8358	0,8655
mcc absoluto	0,6805	0,6803	***0,6806***
verdadeiros negativos(%)	0,8287	0,8358	0,8655
falsos negativos(%)	0,1485	0,1555	0,1858
falsos positivos(%)	0,1713	0,1642	0,1345
verdadeiros positivos(%)	0,8515	0,8445	0,8142

Entre os *threshold* 0,4776 e 0,5328 é possível encontrar os valores máximos das métricas evidenciadas na Tabela 7.2. O mcc de 68% em toda essa faixa mostra que o modelo é superior ao acaso e possui forte correlação entre a predição e os dados analisados. A

acurácia de 84,0% para o *threshold* em 0,4877 tem 84,4% de especificidade: esses valores são superiores aos valores obtidos atualmente pela RFB. A métrica *f1 measure*, métrica menos suscetível a distorções causadas por desbalanceamento, apresenta um valor máximo de 84,2% no *threshold* 0,4776, esse valor é próximo ao apresentado nas métricas *acurácia* e *especificidade* para este mesmo *threshold*.

Apesar dos bons resultados aqui apresentados, para uma avaliação definitiva do modelo será necessária a sua avaliação com bases de dados que não participaram das fases de teste.

Sumário das métricas de *Cross-Validation*

A Tabela 7.3 apresenta, de forma pormenorizada para cada extrato, as principais métricas calculadas em cada um dos dez *folds* do *Cross-Validation* realizado no treino do modelo.

Tabela 7.3: Métricas de *Cross-Validation* em cada *fold*. Modelo *GBM_model_7*.

	cv_1_valid	cv_2_valid	cv_3_valid	cv_4_valid	cv_5_valid
acurácia	0,8309859	0,8290909	0,8103448	0,8426966	0,810219
auc	0,90505415	0,91832894	0,90100664	0,90949297	0,8990261
f1 *measure*	0,8321678	0,83154124	0,8358209	0,8141593	0,80451125
logloss	0,38251752	0,35180092	0,399944	0,37501177	0,39449012
mcc	0,67391026	0,6660656	0,6283373	0,6920764	0,6222782
mse	0,123996876	0,11433302	0,13248594	0,12261349	0,12998162
precisão	0,7677419	0,7785235	0,7692308	0,9108911	0,7753623
recall	0,90839696	0,8923077	0,9150327	0,736	0,8359375
rmse	0,3521319	0,33813167	0,3639862	0,3501621	0,36052963
especificidade	0,7647059	0,7724138	0,69343066	0,9366197	0,7876712

	cv_6_valid	cv_7_valid	cv_8_valid	cv_9_valid	cv_10_valid
acurácia	0,82129276	0,8487395	0,8470149	0,78688526	0,8392157
auc	0,8829589	0,9224095	0,9143193	0,8889868	0,9049157
f1 *measure*	0,80497926	0,83928573	0,8509091	0,8115942	0,84410644
logloss	0,42048335	0,34712043	0,37045687	0,408433	0,38299102
mcc	0,640091	0,69751424	0,70640475	0,59044075	0,6787611
mse	0,13735178	0,11012612	0,11740025	0,13675356	0,1237972
precisão	0,80833334	0,8173913	0,7852349	0,7329843	0,82835823
recall	0,8016529	0,86238533	0,9285714	0,90909094	0,8604651
rmse	0,37061003	0,33185256	0,3426372	0,36980206	0,35184827
especificidade	0,8380282	0,8372093	0,7746479	0,66225165	0,8174603

Dado o caráter de normalidade da distribuição das médias pelo teorema do limite central, observamos que todos os *folds* aleatóriamente formados geraram métricas dentro do intervalo de dois desvios padrões, dentro de um intervalo de confiança de 95%. Não há portanto necessidade da geração de nova amostragem para formação de novos *folds*.

Tabela 7.4: Sumário das métricas de *Cross-Validation*

	média	desvio padrão
acurácia	0,8266	0,0132
auc	0,9046	0,0083
f1 *measure*	0,8269	0,0113
logloss	0,3833	0,0157
mcc	0,6596	0,0253
mse	0,1249	0,0062
precisão	0,7974	0,0325
recall	0,8650	0,0404
rmse	0,3532	0,0088
especificidade	0,7884	0,0518

Análise de sensibilidade dos atributos

A Tabela 7.5 apresenta os atributos que mais influenciaram o modelo. São apresentados os atributos cujo percentual de importância foi superior a 1%.

Tabela 7.5: Análise de sensibilidade dos atributos. Modelo *GBM_model_7*.

Atributo	Importância Relativa	Importância escalar	Percentual
44	1411.2264	1.0	0.6700
75	343.4708	0.2434	0.1631
26	94.4672	0.0669	0.0448
47	67.9543	0.0482	0.0323
31	44.0234	0.0312	0.0209
61	24.6602	0.0175	0.0117

A baixa quantidade de atributos mostra um modelo bem mais simples que o percebido na fase de entendimento do negócio (a coleta inicial dos dados mostrou que a *RFB* utiliza-se de 77 atributos em seu modelo atual) e fortemente baseado no *atributo 44*. Porém, é preciso ressaltar que o modelo que resultará desta pesquisa é um modelo auxiliar: não se pretende (conforme visto no Capítulo 3) a substituição integral dos servidores que hoje atuam neste domínio na RFB pelo algoritmo aqui desenvolvido. O reforço das evidências do crime de LD a partir de um conjunto amplo de atributos é essencial para o prosseguimento das investigações e cumprimento de pressupostos processuais legais.

7.2 *Distributed Random Forest* (DRF)

Essa seção apresenta os resultados obtidos pelo uso da técnica supervisionada *Distributed Random Forest*.

A partir dos dados preparados na fase anterior (Seção 6.7), foram gerados modelos com e sem balanceamento. Em cada caso variou-se a quantidade de árvores (20, 50 e 80 árvores) e sua profundidade máxima permitida no algoritmo (3, 20 e 50). A Tabela 7.6 mostra os valores de *logloss* por *Cross-Validation* (10 *folds*) para escolha do modelo que melhor reduziu os erros.

Percebe-se nesta tabela que o algoritmo se beneficia do aumento do número de árvores, mas não responde bem se estas tiverem um limite de profundidade muito raso. O balanceamento de classes não interferiu na redução do *logloss*. Assim, nas subseções que se seguem, apresentaremos uma análise mais detalhada do modelo que apresentou maior redução de erros na média dos extratos de *Cross-Validation* - o modelo *DRF_model_17*.

Tabela 7.6: Média dos valores de logloss

Balanceamento	Profundidade Máxima	Nº de Árvores	Identificação do Modelo	*Valor de logloss em Cross-Validation*
Falso	50	80	DRF_model_17	0.4137
Verdadeiro	20	80	DRF_model_12	0.4170
Falso	50	50	DRF_model_5	0.4174
Verdadeiro	50	80	DRF_model_16	0.4195
Falso	20	80	DRF_model_13	0.4196
Falso	20	50	DRF_model_1	0.4220
Verdadeiro	20	50	DRF_model_0	0.4225
Verdadeiro	50	50	DRF_model_4	0.4251
Verdadeiro	3	80	DRF_model_14	0.4444
Falso	3	50	DRF_model_3	0.4451
Verdadeiro	50	20	DRF_model_10	0.4463
Falso	3	20	DRF_model_9	0.4479
Falso	3	80	DRF_model_15	0.4480
Verdadeiro	20	20	DRF_model_6	0.4486
Verdadeiro	3	50	DRF_model_2	0.4518
Falso	50	20	DRF_model_11	0.4534
Verdadeiro	3	20	DRF_model_8	0.4535
Falso	20	20	DRF_model_7	0.4743

Análise de curva ROC

A Figura 7.2 apresenta a curva ROC a partir dos valores obtidos por *Cross-Validation* do modelo *DRF_model_17* em 10 *folds*. A área abaixo da curva tem cobertura de aproximadamente 89% do total do gráfico. Percebe-se também que o modelo é capaz de atingir aproximadamente 25% de taxa de verdadeiros positivos (tpr) sem contudo apresentar aumento na taxa de falsos positivos (fpr).

A Tabela 7.7 apresenta os *threshold* da curva ROC em função das seguintes cinco métricas: *f1 measure*, acurácia e *matthews correlation coefficient* (mcc) absoluto.

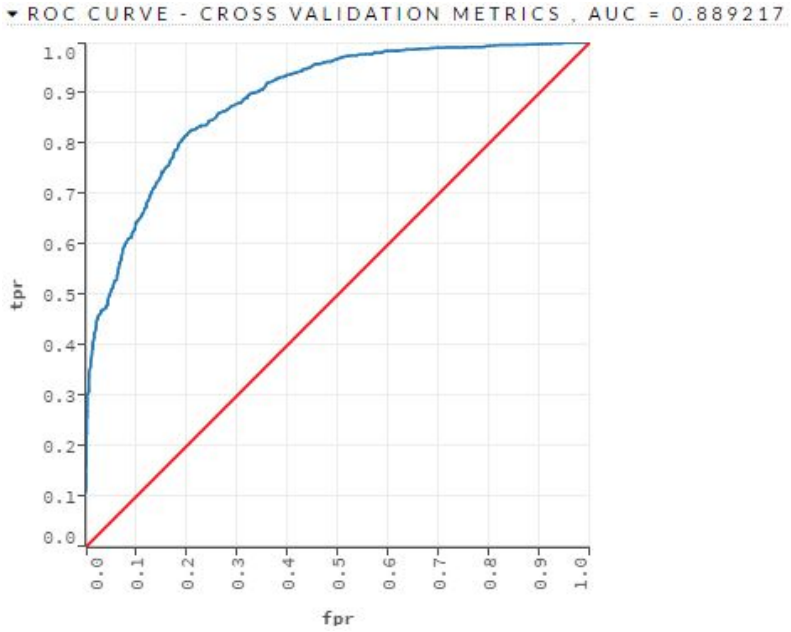

Figura 7.2: Curva ROC - *Cross-Validation* do modelo *DRF_model_17*.

Tabela 7.7: Valores de *threshold* e métricas correspondentes. Modelo *DRF_model_17*.

threshold	0,4497	0,4721	0,4721
f1 measure	***0,8048***	0,8032	0,8032
acurácia	0,808	***0,8088***	0,8088
precisão	0,7865	0,7942	0,7942
recall	0,8239	0,8124	0,8124
especificidade	0,7933	0,8054	0,8054
mcc absoluto	0,6168	0,6174	***0,6174***
verdadeiros negativos(%)	0,7933	0,8054	0,8054
falsos negativos(%)	0,1761	0,1876	0,1876
falsos positivos(%)	0,2067	0,1946	0,1946
verdadeiros positivos(%)	0,8239	0,8124	0,8124

Entre os *threshold* 0,4497 e 0,4721 é possível encontrar os valores máximos das métricas evidenciadas na Tabela 7.7. O mcc de 62% em toda essa faixa, ainda que menor que o apresentado pelo modelo GBM, mostra-se superior ao acaso e possui forte correlação entre a predição e os dados analisados. A acurácia de 80,9% para o *threshold* em 0,4721 tem 80,5% de especificidade: esses valores são substancialmente superiores aos valores obtidos atualmente pela RFB. A métrica *f1 measure*, métrica menos suscetível a distorções causadas por desbalanceamento, apresenta um valor máximo de 80,5% no *threshold* 0,4497, esse valor é próximo ao apresentado nas métricas *acurácia* e *especificidade* para este mesmo *threshold*.

Apesar dos bons resultados aqui apresentados, para uma avaliação definitiva do modelo será necessária a sua avaliação com bases de dados que não participaram das fases de teste.

Sumário das métricas de *Cross-Validation*

A Tabela 7.8 apresenta, de forma pormenorizada para cada extrato, as principais métricas calculadas em cada um dos dez *folds* do *Cross-Validation* realizado no treino do modelo.

Tabela 7.8: Métricas do *Cross-Validation* em cada *fold*. Modelo *DRF_model_17*.

	cv_1_valid	cv_2_valid	cv_3_valid	cv_4_valid	cv_5_valid
acurácia	0,807971	0,8490566	0,82711864	0,7983539	0,83219177
auc	0,89891994	0,90634286	0,9069201	0,9026555	0,9061795
f1 *measure*	0,82033896	0,84962404	0,8118081	0,82807016	0,8292683
logloss	0,39598224	0,38607734	0,3828824	0,38510218	0,3836612
mcc	0,64072394	0,70432204	0,6524139	0,63340086	0,6652426
mse	0,12918423	0,12230067	0,12256015	0,12778515	0,12463981
precisão	0,73333335	0,8014184	0,79710144	0,72392637	0,8095238
recall	0,9307692	0,904	0,8270677	0,9672131	0,85
rmse	0,3594221	0,34971514	0,3500859	0,35747048	0,35304365
especificidade	0,69863015	0,8	0,8271605	0,6280992	0,81578946

	cv_6_valid	cv_7_valid	cv_8_valid	cv_9_valid	cv_10_valid
acurácia	0,7844523	0,8174905	0,7892857	0,8069498	0,7756654
auc	0,8823338	0,88583	0,85246694	0,89579105	0,8468902
f1 *measure*	0,8038585	0,8125	0,7944251	0,8015873	0,7790262
logloss	0,4304992	0,42222032	0,4744024	0,39465404	0,48163927
mcc	0,5893763	0,6351009	0,579942	0,61371374	0,5593687
mse	0,14067689	0,13728583	0,15600584	0,1322239	0,16068353
precisão	0,72254336	0,82539684	0,7702703	0,79527557	0,72727275
recall	0,9057971	0,8	0,8201439	0,808	0,83870965
rmse	0,37506917	0,37052104	0,39497575	0,36362603	0,4008535
especificidade	0,6689655	0,83458644	0,75886524	0,80597013	0,7194245

Dado o caráter de normalidade da distribuição das médias pelo teorema do limite central, observamos que todos os *folds* aleatóriamente formados geraram métricas dentro do intervalo de dois desvios padrões, dentro de um intervalo de confiança de 95%. Não há portanto necessidade da geração de nova amostragem para formação de novos *folds*.

Tabela 7.9: Sumário das métricas de *Cross-Validation*. Modelo *DRF_model_17*.

	média	desvio padrão
acurácia	0,8089	0,0154
auc	0,8884	0,0148
f1 *measure*	0,8131	0,0134
logloss	0,4137	0,0253
mcc	0,6274	0,0289
mse	0,1353	0,0091
precisão	0,7706	0,0270
recall	0,8652	0,0386
rmse	0,3675	0,0121
especificidade	0,7557	0,0489

Análise de sensibilidade dos atributos

A Tabela 7.10 apresenta os atributos que mais influenciaram o modelo. São apresentados os atributos cujo percentual de importância foi superior a 1%.

Apesar do modelo DRF apresentar uma quantidade de atributos maior que o modelo GBM para explicar 99% da variabilidade dos dados, ainda é baixa a quantidade de atributos usados por ele quando comparados aos usados pela RFB identificados na fase de entendimento do negócio (a coleta inicial dos dados mostrou que a *RFB* utiliza-se de 77 atributos em seu modelo atual). De forma análoga ao modelo GBM, o *atributo 44* é o mais importante (ainda que aqui apresente uma importância significativamente menor). Aqui também é preciso ressaltar que o modelo que resultará desta pesquisa é um modelo auxiliar: não se pretende (conforme visto no Capítulo 3) a substituição integral dos servidores que hoje atuam neste domínio na RFB pelo algoritmo, pois, além do algoritmo apresentar falsos positivos que precisam ser manualmente eliminados, o reforço das evidências do crime de LD a partir de um conjunto amplo de atributos é essencial para o prosseguimento das investigações e cumprimento de pressupostos processuais legais.

7.3 *Deep Learning Autoencoder* (DLA)

Esta seção apresenta as anomalias encontradas pelo uso da técnica não supervisionada *Deep Learning Autoencoder* (DLA).

Serão buscadas anomalias em dois diferentes arranjos dos dados:

Arranjo 1 Dados formados pelos atributos oriundos da análise de sensibilidade realizada pela técnica GBM no modelo *GBM_model_7*, discriminados neste capítulo na Seção 7.1.

Tabela 7.10: Análise de sensibilidade dos atributos. Modelo *DRF_model_17*.

Atributo	Importância Relativa	Importância escalar	Percentual
44	4450.9927	1.0	0.1215
32	3274.3716	0.7356	0.0894
26	2757.0867	0.6194	0.0752
75	2724.8052	0.6122	0.0744
47	2614.6648	0.5874	0.0713
60	2323.1382	0.5219	0.0634
31	1797.7269	0.4039	0.0491
22	1725.1727	0.3876	0.0471
74	1252.5055	0.2814	0.0342
70	1245.8973	0.2799	0.0340
2	1223.3253	0.2748	0.0334
28	1150.0692	0.2584	0.0314
59	1133.2535	0.2546	0.0309
25	896.7281	0.2015	0.0245
29	839.4394	0.1886	0.0229
69	754.1373	0.1694	0.0206
63	667.1016	0.1499	0.0182
61	632.4728	0.1421	0.0173
46	583.1119	0.1310	0.0159
42	571.3697	0.1284	0.0156
38	522.4429	0.1174	0.0143
66	512.0089	0.1150	0.0140
43	494.8244	0.1112	0.0135
75	470.0620	0.1056	0.0128
45	365.9323	0.0822	0.0100

Arranjo 2 Dados formados pelos atributos oriundos da análise de sensibilidade realizada pela técnica DRF no modelo *DRF_model_17*, discriminados neste capítulo na Seção 7.2.

A Tabela 7.11 apresenta os principais parâmetros do *Autoencoder* utilizados nos modelos dos dois arranjos. Optou-se por não variar os parâmetros na busca de um melhor ajuste, pois o objetivo aqui não é o de melhorar a eficiência da rede neural para uma melhor cópia, e sim o de criar uma cópia imperfeita da entrada na saída conforme visto na Seção 2.5.

Para a definição das camadas internas foi adotado o ajuste proposto por Hinton et al. [57] com a diminuição da quantidade de neurônios pela metade até a camada central na fase de *encode* e seu espelhamento para as camadas de *decode*.

Tabela 7.11: Parâmetros utilizados nos modelos gerados por DLA

Parâmetro	Valores para o *Arranjo 1*	Valores para o *Arranjo 2*
activation[3]	Rectifier	Rectifier
hidden[4]	3, 2, 3	12, 6, 3, 6, 12
epochs[5]	10	10
autoencoder[6]	true	true
reproducible[7]	true	true

7.3.1 O erro de reconstrução nos modelos DLA

A mensuração do erro na rede neural se deu pela medida do *mean squared error* (MSE) entre os valores correspondentes aos neurônios da saída e os valores dos neurônios da entrada.

Os gráficos apresentados nas Figuras 7.3 e 7.4 mostram o log do erro de reconstrução para o *Arranjo 1* e *Arranjo 2*. Em ambos observa-se nitidamente que a sua porção esquerda possui concavidade negativa com a curva de erro praticamente estável na maior parte do gráfico (entorno de 2^{-14} para o primeiro gráfico, e 2^{-13} para o segundo gráfico). Na porção extrema direita de âmbos, percebe-se que a concavidade altera-se para positiva de forma brusca com o valor do *log* dos erros crescendo aparentemente de forma exponencial. Essa porção indica as anomalias encontradas.

A captura do ponto onde se deu a inflexão pode ser feita de forma análoga ao cálculo da derivada segunda de uma função. Para os gráficos em questão esses pontos foram 14151 e 13530, para respectivamente primeiro e segundo gráficos.

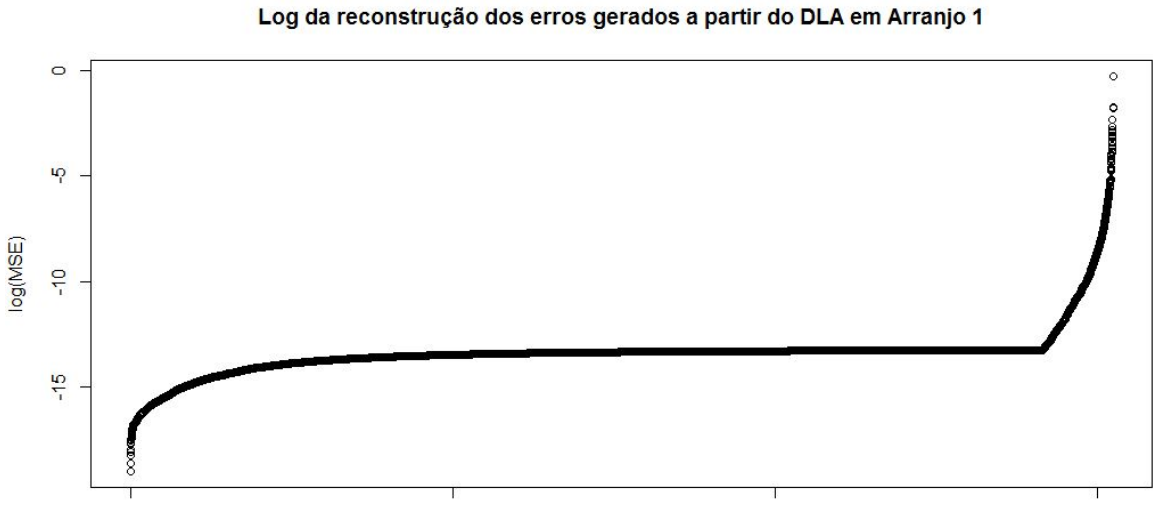

Figura 7.3: Log do erro de reconstrução pela função MSE - Arranjo 1.

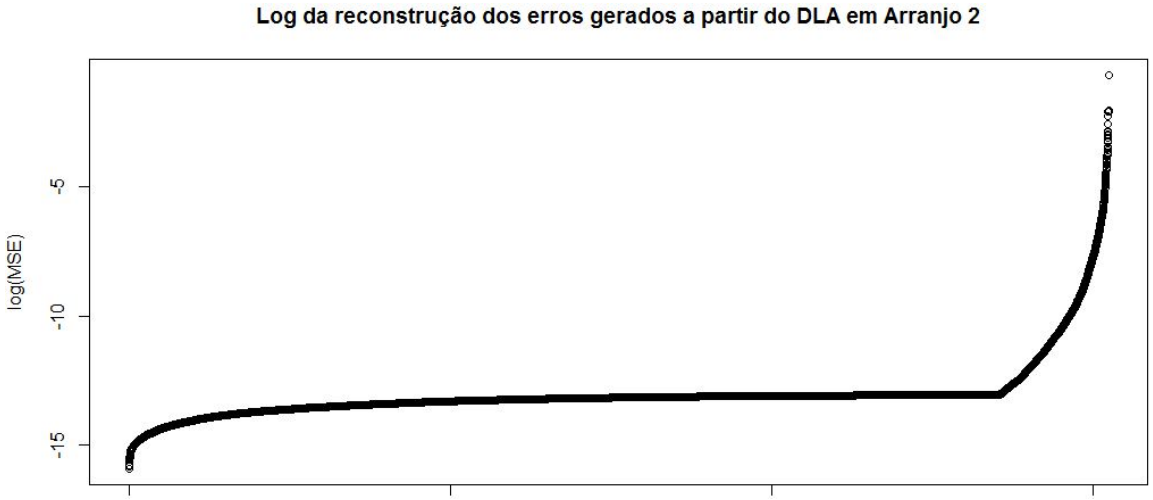

Figura 7.4: Log do erro de reconstrução pela função MSE - Arranjo 2.

7.3.2 Análise dos modelos DLA

Para análise dos modelos não supervisionados faremos dois tipos de confrontação com as anomalias detectadas. Uma em relação aos dados rotulados pela RFB e outra em relação aos dados rotulados nos classificadores supervisionados expostos nas Seções 7.1 e 7.2.

DLA e dados rotulados pela RFB A Figura 7.5 apresenta dois gráficos relativos ao erro de reconstrução após a modelagem por DLA usando os arranjos de atributos 1 e 2 assim definidos no início desta seção. Essa plotagem encontra-se na cor preta e será denominada de *plotagem principal*.

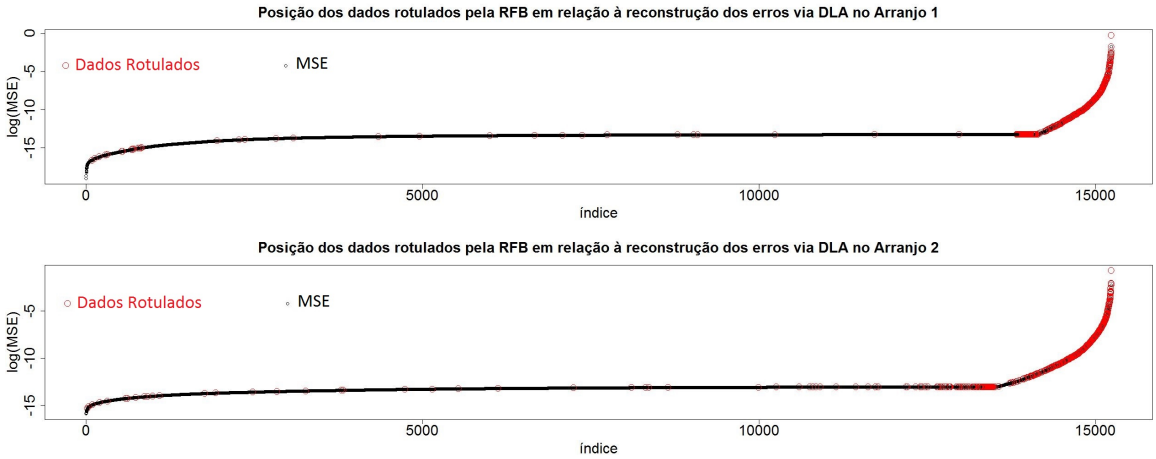

Figura 7.5: Distribuição dos dados rotulados como *suspeito* sobre o erro de reconstrução.

Sobreposto à *plotagem principal*, encontram-se plotadas em vermelho as informações oriundas da RFB que dão conta dos casos manualmente classificados como exportações fictícias e consequentemente suspeitos de lavagem de dinheiro.

Em ambos gráficos observa-se uma concentração da plotagem oriunda da RFB sobre a parcela de concavidade positiva da *plotagem principal*. Porém, para os atributos do *Arranjo 2*, observa-se uma dispersão maior de dados rotulados ao longo da porção da *plotagem principal* com concavidade negativa. Observa-se ainda que em ambos os gráficos há uma concentração de dados rotulados na parte próxima à inflexão das concavidades.

Assim, se for considerado apenas os dados rotulados como critério de avaliação, a detecção de anomalias no *Arranjo 1* tem a capacidade de representar melhor os dados conhecidos atualmente pois os mantêm menos dispersos.

DLA e dados rotulados pelos classificadores supervisionados Sobreposto à *plotagem principal* na Figura 7.6, encontram-se plotadas em vermelho as informações oriundas dos resultados obtidos nos classificares GBM e DRF.

De forma semelhante à observada na Figura 7.5, percebemos em ambos gráficos uma concentração de dados rotulados na convexidade positiva da *plotagem principal*. Porém, na região de convexidade negativa, há um espalhamento dos dados rotulados, principalmente naqueles plotados sobre o erro de reconstrução oriundo do *Arranjo 2*.

Assim, tendo como critério de avaliação a dispersão dos dados classificados, pode-se dizer que a detecção de anomalias no *Arranjo 1* tem a capacidade de representar melhor os dados classificados por GBM pois os mantêm menos dispersos.

Avaliação pela simplicidade do modelo Outro ponto à favor da detecção não supervisionada realizada sobre o *Arranjo 1* é a sua maior simplicidade.

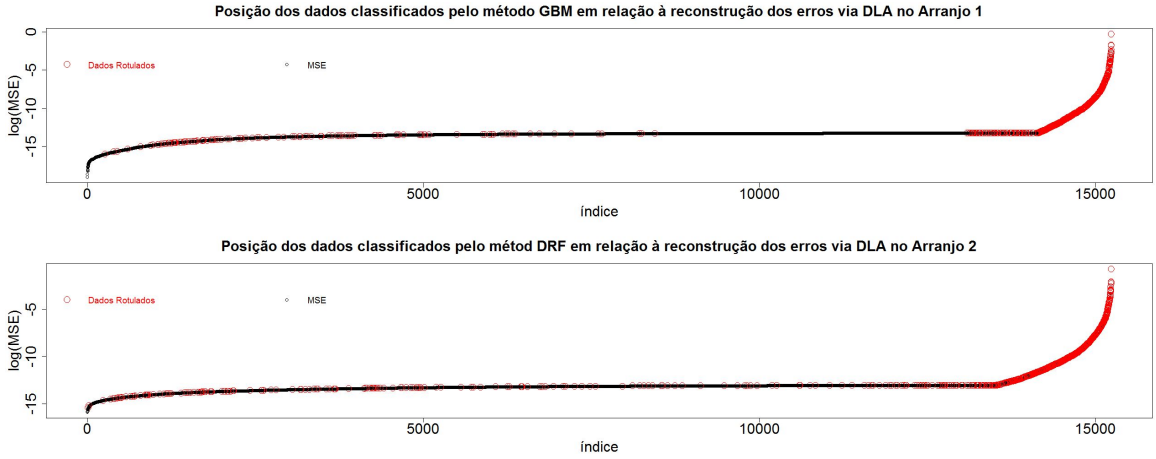

Figura 7.6: Distribuição dos dados classificados como *suspeito* de forma supervisionada sobre a plotagem do erro de reconstrução.

- O *Arranjo 1* consegue explicar melhor a partir de menos atributos: são utilizados 6 atributos no *Arranjo 1* e 25 atributos no *Arranjo 2*.
- Observando-se a Tabela 7.11, percebe-se também que a parte oculta da rede neural que tratou o *Arranjo 1* é menor em número de camadas e neurônios. Isso resulta em uma rede neural mais simples.

7.4 Seleção do Modelo

A escolha dos modelos será feita a partir dos testes realizados nas seções anteriores. Inciaremos com a escolha entre os modelos supervisionados a partir de sua métrica e, posteriormente com a escolha entre este e o modelo não supervisionado.

7.4.1 Comparação entre as métricas dos modelos GBM e DRF

A Tabela 7.12 apresenta as métricas obtidas pelos modelos GBM (Seção 7.1) e DRF (Seção 7.2). Em todas elas foi testada a hipótese (H_0) das médias serem iguais em ambos modelos. A coluna *p-value* apresenta a probabilidade de H_0 ser verdadeiro. Observa-se que, à exceção da métrica *recall*, todas as demais apresentam diferença estatisticamente significante entre as médias, sendo então possível afirmar que o método GBM tem um melhor desempenho em todas as outras métricas que o método DRF.

Tabela 7.12: Métricas obtidas pelos modelos GBM e DRF

	GBM		DRF		
	Média	Desvio Padrão	Média	Desvio Padrão	*p-value*
acurácia	0,8266	0,0132	0,8089	0,0154	<2.2e-16
auc	0,9046	0,0083	0,8884	0,0148	<2.2e-16
f1 measure	0,8269	0,0113	0,8131	0,0134	<2.2e-16
logloss	0,3833	0,0157	0,4137	0,0253	<2.2e-16
mcc	0,6599	0,0253	0,6274	0,0289	<2.2e-16
mse	0,1249	0,0062	0,1353	0,0091	<2.2e-16
precisão	0,7974	0,0325	0,7706	0,0270	<2.2e-16
recall	0,8650	0,0404	0,8652	0,0386	0,0909
rmse	0,3532	0,0088	0,3675	0,0121	<2.2e-16
especificidade	0,7884	0,0518	0,7557	0,0489	<2.2e-16

7.4.2 Comparação entre os resultados dos modelos GBM e DLA

Parte dessa comparação encontra-se realizada na Seção 7.3.2, quando se comparou os modelos DLA gerados em dois diferentes arranjos de atributos.

A Figura 7.7 mostra uma ampliação da área onde foram detectadas anomalias no *Arranjo 1*, sobreposta pelos dados classificados no modelo GBM. Visualmente observa-se uma cobertura quase total dessa área pelos resultados do algoritmo GBM. Ao se observarem os dados pormenorizadamente, percebe-se que essa cobertura é realmente ampla: de 1092 dados com anomalias detectadas, 1107 foram classificados pelo GBM, ou seja, aproximadamente 92,9%.

O total de dados classificados como verdadeiro pelo modelo GBM foi de 1286, dado que 1107 estão na área de detecção de anomalias, 179 estão dispersos nas área de concavidade negativa.

7.4.3 Modelo escolhido

O modelo baseado na técnica GBM é o modelo escolhido, pois, além de apresentar melhores métricas em comparação com o modelo baseado em DRF, cobre praticamente toda a área onde se encontram as maiores anomalias detectadas pelo modelo DLA. Não obstante a escolha do modelo GBM, registra-se que o modelo DLA obteve resultados quase tão bons quanto os obtidos pelos modelos supervisionados. Isso indica que em outros cenários, na inexistência de dados rotulados, o DLA é uma alternativa bastante promissora.

Uma opção de escolha de modelo (a depender de resultados em trabalhos futuros) seria a conjugação dos modelos GBM e DLA. Conforme apresentado na Figura 7.7, há pontos não identificados pelo GBM na concavidade positiva da curva de reconstrução de erro do DLA. Uma análise futura pela RFB das empresas representadas por esses pontos

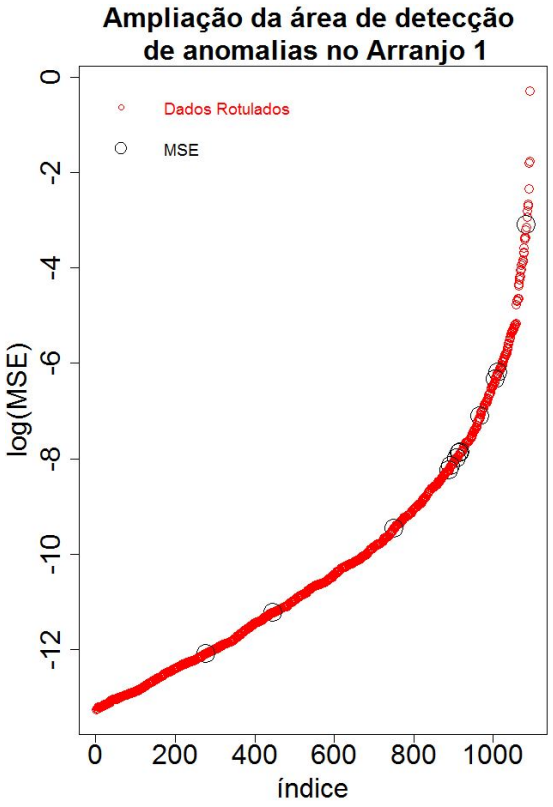

Figura 7.7: Área de anomalias detectadas sobreposta pelos dados rotulados a partir do GBM.

pode indicar padrões de comportamento que escaparam à representação dada pelos dados rotulados como *suspeito*.

Capítulo 8

Validação do Modelo e Índice de Prioridades

O presente capítulo apresenta a validação do modelo proposto no Capítulo 7 e, a partir desta validação, propõe um índice que indica uma ordem de prioridade para atuação da RFB nos casos classificados como suspeitos.

A avaliação será feita de duas formas:

1. avaliação das métricas a partir da aplicação do modelo escolhido (GBM) a uma base de dados classificada pela RFB não utilizada para a construção do modelo;
2. avaliação empírica dos resultados da aplicação do modelo escolhido (GBM) a uma base de dados não classificada pela RFB.

8.1 Avaliação por métricas

A partir do modelo selecionado e ajustado (Seção 7.4), foi realizada a predição da *base de avaliação* (ver Seção 6.7) com o respectivo cálculo das métricas frente aos dados rotulados pela RFB. As subseções abaixo apresentam os resultados.

8.1.1 Análise de Curva ROC

A Figura 8.1 apresenta a curva ROC a partir dos valores obtidos pela aplicação do modelo 'GBM_model_7' à base de avaliação.. A área abaixo da curva tem cobertura pouco inferior a 90% do total do gráfico.

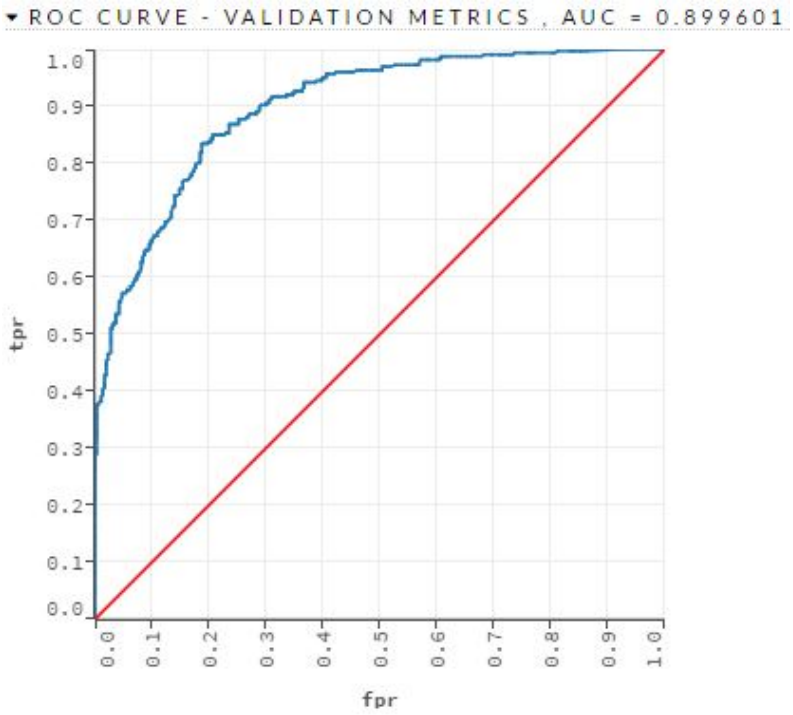

Figura 8.1: Curva ROC gerada na avaliação do modelo GBM.

A Tabela 8.1 apresenta os *threshold* da curva ROC em função das seguintes cinco métricas: *f1 measure*, acurácia, precisão, especificidade e *matthews correlation coefficient* (mcc) absoluto. A proximidade dos valores obtidos nessas métricas com os valores obtidos no sumário das métricas de Cross-Validation apresentados na Tabela 7.2 indicam que o modelo GBM escolhido não apresenta super ajuste, sendo capaz de manter uma performance próxima à obtida na fase de treinamento e testes mesmo quando exposto a dados novos.

Tabela 8.1: Valores de *threshold* e métricas correspondentes para o modelo GBM.

threshold	0.4460	0.4803	0.9952	0.9952	0.4803
f1 *measure*	***0.8189***	0.8186	0.0061	0.0061	0.8186
acurácia	0.8202	***0.8231***	0.5234	0.5234	0.8231
precisão	0.7898	0.8029	***1.0***	1.0	0.8029
recall	0.8502	0.8349	0.0031	0.0031	0.8349
especificidade	0.7927	0.8123	1.0	***1.0***	0.8123
mcc absoluto	0.6425	0.6466	0.0400	0.0400	***0.6466***
tnr	0.7927	0.8123	1.0	1.0	0.8123
fnr	0.1498	0.1651	0.9969	0.9969	0.1651
fpr	0.2073	0.1877	0	0	0.1877
tpr	0.8502	0.8349	0.0031	0.0031	0.8349

8.1.2 Gráfico de ganhos e alavancagem cumulativas

A Figura 8.2 mostra o gráfico de ganhos e alavancagem cumulativas obtido pela ordenação decrescente da probabilidade de uma empresa ser classificada como suspeita na base de avaliação do algoritmo GBM. No eixo X encontram-se os percentis dessa ordenação. No eixo Y a taxa de verdadeiros positivos atingida para aquele percentil. Percebe-se que ao atingir o primeiro decil, a linha preta (ganhos) indica que há cerca de 20% dos verdadeiros positivos capturados, ou seja, 10% das empresas com maior probabilidade retornam 20% dos verdadeiros positivos. Essa relação 20:10 é mostrada na linha verde (alavancagem) que, para o primeiro decil, encontra-se bem próxima de 2.0. Acompanhando-se essa linha (verde), temos que até o terceiro decil é possível manter a taxa de aproximadamente 20:10 visto no primeiro decil, ou seja, uma análise de 30% dos casos é capaz de abarcar 60% dos verdadeiros positivos descoberto pelo modelo.

Assim, a partir do exposto, conclui-se que na hipótese da RFB não efetuar a fiscalização de todas as empresas classificadas pelo GBM como *suspeito*, a observância da ordenação decrescente da probabilidade dada pelo algoritmo é relevante, pois traz alavancagem de até 2 vezes.

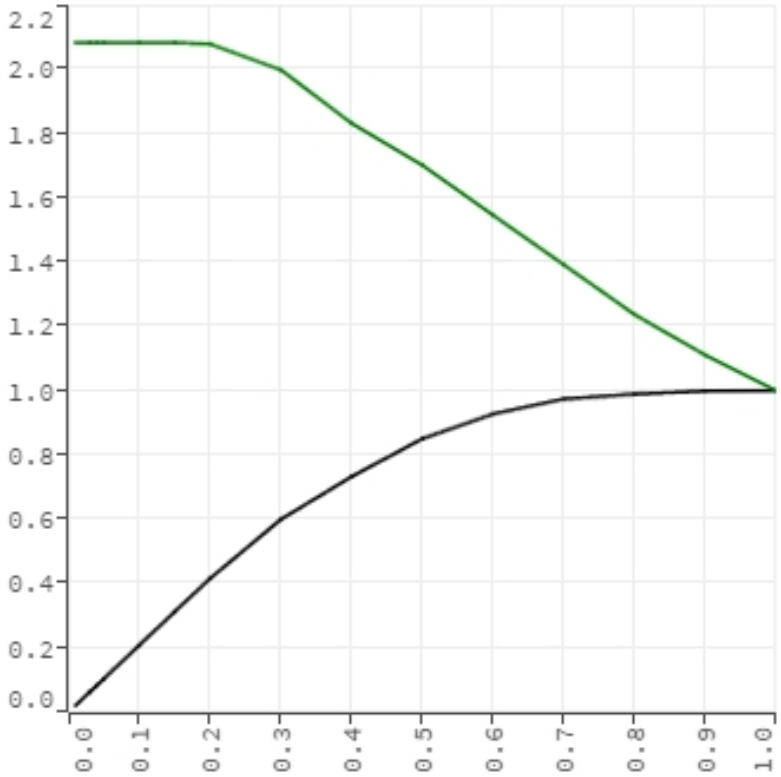

Figura 8.2: Gráfico de *Gain/Lift* da classificação por GBM da base de testes.

8.2 Avaliação empírica

Ainda que usemos como parâmetro de avaliação as métricas calculadas sobre uma parcela dos dados rotulados pela RFB, esta somente será definitiva quanto à sua eficiência prática quando no futuro tivermos ações fiscalizatórias e investigativas conclusivas em pelo menos uma amostra estatisticamente significativa dos dados não rotulados. Porém, a depender do tamanho da empresa e do volume exportado, uma operação de fiscalização ou de investigação pode demorar entre meses e anos. Ainda que fossem apenas autuações simples do fisco federal, os tempos mínimos são longos devido a toda sorte de prazos previstos na legislação brasileira: prazos de intimação, de ciência, para fornecimento de documentação, de diligências, de notificações, etc.. Soma-se ainda o fato de que não se esperaria do fisco que este se diligenciasse a confirmar os casos classificados como *não suspeitos*, investindo recursos em ações que se supõe com baixa probabilidade de sucesso.

Assim, pelo exposto, não obstante os achados desse trabalho terem sido encaminhados à área responsável para inclusão na programação de trabalhos de 2017, faz-se a opção por realizar uma validação empírica, mais sumária, e que envolve tanto as empresas cujos dados foram classificados como *suspeito*, quanto aquelas classificados como *não suspeito*. A partir de uma amostragem aleatória, sem reposição e estatisticamente significativa realizada sobre os dados não rotulados pela RFB e classificados pelo modelo, fez-se uma análise caso a caso da admissibilidade da sua suspeição a partir de pressupostos empíricos de irregularidade.

As subseções a seguir apresentam o processo de amostragem e a avaliação.

8.2.1 Determinação da quantidade de empresas a serem amostradas

Trata-se aqui de identificar uma amostra de n empresas dentre a população de empresas que efetuaram exportações nos anos de 2014 e 2015, não rotuladas pela RFB e classificadas pelo algoritmo GBM.

Como a classificação feita pelo GBM é uma classificação em apenas duas categorias (*suspeito* e *não suspeito*) estamos diante de uma distribuição de *Bernoulli*. Como não queremos que uma empresa seja selecionada mais de uma vez, esta amostragem será feita sem reposição.

A variância σ^2 de uma distribuição de *Bernoulli* [83] é dada por:

$$\sigma^2 = p(1-p) \tag{8.1}$$

sendo p o valor da proporção entre o quantitativo de empresas classificadas como suspeitas e o total de empresas. Considerando que a classificação como suspeito pelo GBM em sucesso foi de 1286 em 15242 empresas (N), temos que p é igual a 0.08437213 e, portanto a variância é de 0.07325347.

Desta forma, para o cálculo de n usando a técnica de amostragem aleatória sem reposição para porporções proposta por Bolfarine e Bussab [83] adaptada de [84] temos a seguinte equação:

$$n_0 = \frac{Z^2.\sigma^2}{E^2} \tag{8.2}$$

Para um intervalo de confiança de 95% bicaudal, temos pela tabela da distribuição normal padrão acumulada que $Z = 1,96$. Dada uma margem de erro de 5% temos que $n_0 = 112,5$. Dado que a relação de razão entre n_o e N é inferior a 0,05, não há necessidade de se fazer o ajuste para população finita ficando portanto $n = n_0$. Arredondando-se o valor de n_0, temos que a amostra será de 113 empresas.

Utilizando-se a função 'sample_n' com semente igual a '1' do pacote *dplyr*[1] do aplicativo 'R' (ver código no Apêndice A), procedeu-se à seleção das 113 empresas. O quantitativo de empresas amostradas subdividido por sua classificação feita pelo modelo GBM está apresentado na Tabela 8.2.

Tabela 8.2: Sumário dos quantitativos da classificação GBM nos dados

Suspeito	Não Suspeito
14	99

8.2.2 Análise de pressupostos em relação à classificação feita pelo GBM

Foi realizada uma análise das empresas classificadas e sorteadas conforme seção anterior. Empiricamente foi possível deduzir algumas relações entre atributos reveladoras de suspeição da empresa quanto às suas atividades comerciais. Ressalta-se que essas relações não são necessariamente correspondentes às identificadas pelos algoritmos testados no Capítulo 7.

Assim, foram identificadas seis situações que abrangeram os 14 casos amostrados classificados como *suspeito*. Algumas dessas situações também se mostraram presentes em casos amostrados classificados como *não suspeito* e, neste caso, foram buscadas as possíveis justificativas.

[1] `https://cran.r-project.org/web/packages/dplyr/index.html`

Concluiu-se que a classificação feita pelo GBM era satisfatória sendo possível em todos os casos amostrados encontrar explicações empíricas para a classificação dada.

Devido ao caráter estratégico que essas informações possuem, e considerando o exposto na Seção 5.3, optou-se por não apresenta-las pormenorizadamente neste documento. Seu conteúdo foi apresentado em documento próprio da RFB com classificação *reservado*.

8.3 Proposta de índice de prioridade para atuação da RFB

A proposta que aqui se apresenta leva em conta os seguintes fatores:

- impacto financeiro da operação do fisco;
- probabilidade da classificação *suspeito* nos dados da empresa;
- setor econômico. .

Impacto financeiro da operação do fisco Não há dúvidas de que quanto maior o volume de dinheiro lavado, maior o prejuízo à sociedade. Assim, é razoável supor que as empresas que exportam um maior volume financeiro tenham um peso maior no índice que aquelas cujo valor exportado seja menor. Uma forma de expressar esse peso seria a divisão do atributo representativo do valor exportado em dez partes (10 decis). Destarte empresas pertencentes ao primeiro decil receberiam peso igual a 1, empresas pertencentes ao segundo decil receberiam peso igual a 2 e assim sucessivamente até que as empresas pertencentes ao maior decil, aquelas portanto pertencentes ao décimo decil recebem o peso igual a 10. Para efeito da Equação 8.3 este fator será chamado de *If*.

Probabilidade da classificação *suspeito* nos dados da empresa Um problema facilmente perceptível na aplicação do fator anterior (Impacto financeiro da operação do fisco) é que podemos ter empresas com *If* igual a 10, porém com baixa probabilidade de ser um *verdadeiro positivo* na classificação como *suspeito*. Além disso, a Seção 8.1.2 mostrou a alavancagem que se obtém de verdadeiros positivos ao se priorizar o tratamento das empresas suspeitas de fraudarem as exportações a partir dessa probabilidade fornecida pelo modelo GBM. Assim, é desejável a associação ao fator *If* de um peso capaz de relativizar as empresas de maior valor exportado em função da probabilidade de sucesso da correta classificação pelo modelo GBM. Para efeito da Equação 8.3 este fator será chamado de *P(s)*.

O algoritmo GBM, ao classificar os dados de uma empresa como *suspeito*, calcula a probabilidade[2] de sucesso dessa classificação. Propõe-se que esse valor da probabilidade da empresa ser classificada como *suspeito* seja utilizado como um peso associado ao fator *If*, multiplicando-o: $If.P(s)$. Dessa forma obteremos um número variando entre 0 e 10 associado a cada empresa.

Setor econômico Estudos internos da Coordenação de Pesquisa e Investigação da RFB mostram que há um aumento na arrecadação espontânea das empresas de um dado setor econômico após uma operação de investigação incidir sobre outras empresas deste mesmo setor. Além disso, é razoável supor que determinados setores econômicos tenderão a apresentar o fator *If* mais alto que outros devido a fatores ligados à demanda internacional de certos produtos e também pelo valor maior que habitualmente determinados produtos possuem.

Os atributos ligados ao setor econômico são categóricos e portanto não compõe a Equação 8.3. Assim, propõe-se que o cálculo do índice de prioridade para atuação da RFB seja aplicado separadamente para cada setor econômico, ficando a cargo da RFB sua aplicação de forma a contemplar todos setores.

Proposta de índice

Por se tratar de um índice de prioridades, valores mais baixos serão atribuídos a empresas que se acredita serem mais prioritárias. Para tanto, o índice será formado pelo inverso da multiplicação dos fatores *If* e *P(s)* e calculado separadamente para cada grupo de empresas com o mesmo setor econômico.

O índice é dado pela fórmula:

$$\text{Índice} = \frac{1}{If.P(s)} \tag{8.3}$$

Depreende-se que este índice poderá variar entre $0,1 \leq \text{Índice} < \infty$. O valor de 0,1 seria dado à uma grande empresa exportadora, pertencente ao decil 10 ($If = 10$) e com probabilidade de verdadeiro positivo igual a 1 ($P(s) = 1$). No outro extremo, teríamos uma pequena empresa exportadora, pertencente ao decil 1 ($If = 1$) e com probabilidade de verdadeiro positivo muito próximo de 0 ($P(s) \approx 1$).

Pela fórmula apresentada (e pela proposição de sua aplicação), temos uma distribuição das ações do fisco em todas atividades econômicas detectadas, não apenas nos grandes volumes exportados, mas também naqueles mais flagrantes que são os detentores das maiores probabilidades de fraude.

[2]O algoritmo GBM obedece aos 3 axiomas da probabilidade propostos por [85].

Capítulo 9

Conclusões e Trabalhos Futuros

Este capítulo apresenta as conclusões e os resultados obtidos no presente trabalho. Apresenta ainda, na última seção, os trabalhos futuros que se vislumbram a partir deste estudo.

9.1 Conclusões

A vinculação entre exportações fictícias e o crime de lavagem de dinheiro é, antes de tudo, uma vinculação formal, um pressuposto legal da legislação brasileira. Assim, neste trabalho, buscou-se a identificação das exportações fictícias como forma suficiente para a suspeição do cometimento do crime de lavagem de dinheiro. Os dados que deram origem ao atributo de classe na mineração de dados para indicar as exportações fictícias encontram-se distribuídos em várias áreas da RFB responsáveis por atuar no comércio exterior. Por não ser trivial a coleta e a integração desses dados, a montagem deste atributo constituiu a tarefa mais trabalhosa no presente trabalho. Tal tarefa pode, no futuro, ser minimizada com um tratamento prévio dessa informação via *Data Warehouse*.

A coleta inicial de dados apresentada na Seção 6.1 e realizada junto a terceiros especialistas se mostrou suficiente, porém, percebe-se, foi além do necessário. O modelo escolhido, *Gradiente Boosting Machine* (GBM), foi capaz de identificar as exportações fictícias rotuladas pela RFB com aproximadamente 80% de acurácia, de precisão e de especificidade, e utilizando-se de apenas 6 dos 77 atributos levantados.

Os dados disponibilizados pelo *Data Werehouse* foram utilizados sem tratamento adicional, ainda que tenham sido encontrados uma quantidade relativamente pequena de dados inconsistentes: 23 em 15265 registros.

Muitos atributos foram eliminados da análise devido à baixa variância e à existência de alto percentual (maior que 60%) de *Missing Values*. É preciso ressaltar que não se tratou de erro na transferência e registro de dados, mas sim de imposições da legislação brasileira que distingue as informações prestadas, quando da declaração das empresas, de

acordo com suas características mercantis, sociais e econômicas. Apesar da eliminação desses atributos, permaneceu preservada a totalidade dos registros extraídos para análise, representando a totalidade das empresas que efetuaram exportações no período abrangido por esse trabalho (2014 e parte de 2015), independentemente do seu porte.

A alta cardinalidade dos atributos relacionados à NCM e país de destino da mercadoria, apontada na fase de entendimento do negócio como uma possível dificuldade, não se mostrou importante neste trabalho por dois motivos: optou-se pelo uso de dados agregados por empresa em todo o período analisado (2014 e parte de 2015), o que reduziu o volume de dados; e a não relevância desses atributos na análise de sensibilidade nos modelos estudados.

As relações não lineares e o caráter log-normal dos dados, ambos evidenciados na fase de entendimento dos dados, constituem importantes observações quanto ao tipo de tratamento a ser dado na análise de dados da RFB. Não apenas as áreas ligadas às exportações, mas várias áreas de seleção de contribuintes para fiscalização de tributos externos podem se beneficiar desses achados.

A técnica de *Deep Learnig Autoencoder* foi utilizada no início deste trabalho quando não haviam sido coletados dados suficientes para uma classificação supervisionada. Ainda que esta não tenha sido a técnica presente no modelo escolhido, quando comparada ao *Gradiente Boosting Machine*, ela foi capaz de identificar cerca de 92,9% dos mesmos achados. Assim, sua capacidade de lidar com dados não lineares e sem supervisão pode ter aplicação em outras áreas da *Rede-LAB* e da RFB tais como a identificação de ganhos patrimoniais a descoberto e identificação de declarações indevidas de despesas médicas.

A avaliação a partir de métricas do modelo escolhido, frente a dados rotulados não apresentados na fase de treino e testes (Seção 8.1), mostrou-se capaz de atingir quase 90% da área da curva ROC. Com medidas de f1 *measure* e acuracia de 0,819 e 0,823 respectivamente. O mcc superou 0.5 com resultado de 0.647 para o *threshold* de acurácia máxima. O gráfico de ganhos e alavancagem cumulativas (Figura 8.2) mostrou que é possível atingir 60% dos verdadeiros positivos do modelo atuando em 30% dos casos.

A avaliação empírica do modelo (Seção 8.2) evidenciou várias relações importantes entre os atributos, várias delas aparentemente determinantes para a suspeição das empresas. Da mesma forma, evidenciou-se determinados atributos (atributos 31 e 61) para os quais, a princípio, não se percebe relação direta com a classificação da exportação fictícia.

9.2 Resultados obtidos

Abaixo estão listados os resultados imediatos alcançados com o presente trabalho.

- Desenvolvimento de um modelo preditivo para identificação de exportadores suspeitos de operarem lavagem de dinheiro no comércio exterior a partir de exportações fictícias com métricas próximas a 80% de acurácia.
- Proposição de um índice que indica uma ordem de prioridade para a investigação de casos de exportações fictícias pela RFB.
- Identificação dos atributos mais relevantes para explicar a exportação fictícia.
- Disseminação da metodologia utilizada neste trabalho, seus resultados e prioridades identificadas, a todos Escritórios de Pesquisa e Investigação da RFB, de forma a que possam ser utilizados em suas análises.
- Inclusão dos achados classificados como *suspeito* na programação de fiscalização aduaneira do ano de 2017.
- Proposição de integração, via *Data Warehouse*, dos dados que deram origem ao atributo de classe[1] (dados rotulados).
- Identificação de 23 inconsistências na base do *Data Warehouse*.
- Produção do artigo acadêmico [2] intitulado *Deep Learning Anomaly Detection as Support Fraud Investigation in Brazilian Exports and Anti-Money Laundering* aceito e apresentado no 15º *IEEE International Conference on Machine Learning and Applications*.

9.3 Trabalhos futuros

A partir do presente trabalho vislumbram-se diversos caminhos a serem percorridos no futuro. Oportunamente, a RFB juntamente com o Serviço Federal de Processamento de Dados (SERPRO) encontra-se em processo licitatório para aquisição e implantação de uma ferramenta corporativa de *Business Intelligence*. Assim, a implementação do modelo aqui desenvolvido na nova plataforma possibilitará o seu uso corporativo e com características gerenciais.

Um aperfeiçoamento do presente trabalho na busca de um classificador de suspeição de lavagem de dinheiro, via exportação fictícia, com foco em transações e rotas de mercadorias seria de muita utilidade nas atividades de repressão e controle de fronteiras com provável aumento de especificidade em relação aos métodos atuais.

[1] Maiores informações quanto à dispersão dos dados que deram origem ao atributo de classe podem ser obtidas na Seção 5.2.2

[2] Cópia do artigo encontra-se no Apêncice C deste trabalho

A disseminação dentro da RFB dos bons resultados alcançados com o *Deep Learning Autoencoder* poderá encontrar diversas aplicações nas seleções de contribuintes para fiscalização, não apenas na área de comércio exterior, mas também nas fiscalizações de tributos internos e investigação. Assim, propõe-se a criação de um módulo específico do sistema Contágil para *Deep Learning Autoencoder* de forma a facilitar a utilização da técnica por aqueles que não estejam habituados a usar ferramentas de mineração de dados como o *H2O*.

O combate à lavagem de dinheiro não é um trabalho apenas da RFB, vários outros órgãos estão envolvidos. Embora os *Laboratórios de Tecnologia contra Lavagem de Dinheiro* (Lab-LD) localizados fora da RFB não tenham acessos às bases da RFB para uso do modelo desenvolvido, a troca de informações sobre as técnicas utilizadas neste trabalho, além de disseminar os conhecimentos adquiridos, trarão críticas construtivas e a possibilidade de ajuste no presente trabalho.

Referências

[1] Senator, T. E., H. G. Goldberg, J. Wooton, M. A. Cottini, A. F. Umar Khan, C. D. Klinger, W. M. Llamas, M. P. Marrone e R. W. H. Wong: *The financial crimes enforcement network AI system (FAIS) : identifying potential money laundering from reports of large cash transactions*. The AI magazine, 16(4):21–39, 1995, ISSN 0738-4602. `http://cat.inist.fr/?aModele=afficheN&cpsidt=2985240`, acesso em 2016-04-25TZ. 6

[2] Goldberg, Henry G. e Ted E. Senator: *Restructuring Databases for Knowledge Discovery by Consolidation and Link Formation*. Em *Proceedings of the First International Conference on Knowledge Discovery and Data Mining*, páginas 136–141. AAAI Press, 1995. 6

[3] Larik, Asma S. e Sajjad Haider: *Clustering based Anomalous Transaction Reporting*. Procedia Computer Science, 3:606–610, 2011, ISSN 18770509. 6

[4] Kenaya, Riyadh e Ka C. Cheok: *Euclidean ART Neural Networks*. 2008. 6

[5] Khan, Nida S., Asma S. Larik, Quratulain Rajput e Sajjad Haider: *A Bayesian Approach for Suspicious Financial Activity Reporting*. International Journal of Computers and Applications, 35(4):181–187, janeiro 2013, ISSN 1206-212X. 6, 7

[6] Friedman, Nir, Dan Geiger e Moises Goldszmidt: *Bayesian Network Classifiers*. Machine Learning, 29(2-3):131–163, novembro 1997, ISSN 0885-6125, 1573-0565. `http://link.springer.com/article/10.1023/A:1007465528199`, acesso em 2016-12-21TZ. 6, 7

[7] Raza, Saleha e Sajjad Haider: *Suspicious activity reporting using dynamic bayesian networks*. Procedia Computer Science, 3:987–991, 2011, ISSN 1877-0509. 6

[8] Murphy, Kevin Patrick: *Dynamic Bayesian Networks: Representation, Inference and Learning*. 2002. 6

[9] Rajput, Quratulain, Nida Sadaf Khan, Asma Larik e Sajjad Haider: *Ontology Based Expert-System for Suspicious Transactions Detection*. Computer and Information Science, 7(1), janeiro 2014, ISSN 1913-8997, 1913-8989. `http://www.ccsenet.org/journal/index.php/cis/article/view/30883`, acesso em 2016-04-25TZ. 7

[10] Horrocks, Ian, Peter Patel-Schneider, Harold Boley, Said Tabet, Benjamin Grosof e Mike Dean: *{SWRL}: A Semantic Web Rule Language Combining {OWL} and {RuleML}*. maio 2004. `http://www.w3.org/Submission/SWRL/`, acesso em 2016-12-21TZ. 7

[11] Rohit, Kamlesh e Patel Dharmesh: *Review on Detection of Suspicious Transaction in Anti-Money Laundering Using Data Mining Framework*. International Journal for Innovative Research in Science and Technology, 1(8):129–133, fevereiro 2015. 7

[12] Wang, Xingqi e Guang Dong: *Research on Money Laundering Detection Based on Improved Minimum Spanning Tree Clustering and Its Application*. 2009, ISBN 978-0-7695-3888-4. https://www.infona.pl//resource/bwmeta1.element.ieee-art-000005362309, acesso em 2016-12-21TZ. 7

[13] Zhou, Gengui e Mitsuo Gen: *Genetic algorithm approach on multi-criteria minimum spanning tree problem*. European Journal of Operational Research, 114(1):141–152, abril 1999, ISSN 0377-2217. http://www.sciencedirect.com/science/article/pii/S0377221798000162, acesso em 2016-12-21TZ. 7

[14] Hawkins, D. M.: *Multivariate outlier detection*. Em *Identification of Outliers*, Monographs on Applied Probability and Statistics, páginas 104–114. Springer Netherlands, 1980, ISBN 978-94-015-3996-8 978-94-015-3994-4. http://link.springer.com/chapter/10.1007/978-94-015-3994-4_8, acesso em 2016-12-21TZ, DOI: 10.1007/978-94-015-3994-4_8. 7, 8

[15] Ng, Andrew Y., Michael I. Jordan e Yair Weiss: *On Spectral Clustering: Analysis and an algorithm*. Em *Advances in Neural Information Processing Systems*, páginas 849–856. MIT Press, 2001. 7

[16] Khac, N. A. Le e M. T. Kechadi: *Application of Data Mining for Anti-money Laundering Detection: A Case Study*. Em *2010 IEEE International Conference on Data Mining Workshops*, páginas 577–584, dezembro 2010. 7

[17] Jain, Anil K.: *Data clustering: 50 years beyond K-means*. Pattern Recognition Letters, 31(8):651–666, junho 2010, ISSN 0167-8655. http://www.sciencedirect.com/science/article/pii/S0167865509002323, acesso em 2016-12-21TZ. 7, 45

[18] Haykin, Simon: *Redes Neurais - Principios E Prática*. Bookman, Porto Alegre, 2 edition edição, 2003, ISBN 978-85-7307-718-6. 7

[19] Lin, S. e B. W. Kernighan: *An Effective Heuristic Algorithm for the Traveling-Salesman Problem*. Operations Research, 21(2):498–516, abril 1973, ISSN 0030-364X. http://pubsonline.informs.org/doi/abs/10.1287/opre.21.2.498, acesso em 2016-12-21TZ. 7

[20] Keyan, L. e Y. Tingting: *An Improved Support-Vector Network Model for Anti-Money Laundering*. Em *2011 Fifth International Conference on Management of e-Commerce and e-Government*, páginas 193–196, novembro 2011. 7

[21] Cortes, Corinna e Vladimir Vapnik: *Support-vector networks*. Machine Learning, 20(3):273–297, setembro 1995, ISSN 0885-6125, 1573-0565. http://link.springer.com/article/10.1007/BF00994018, acesso em 2016-12-21TZ. 7

[22] Liu, R., X. l Qian, S. Mao e S. z Zhu: *Research on anti-money laundering based on core decision tree algorithm*. Em *2011 Chinese Control and Decision Conference (CCDC)*, páginas 4322–4325, maio 2011. 7

[23] Zhang, Tian, Raghu Ramakrishnan e Miron Livny: *BIRCH: An Efficient Data Clustering Method for Very Large Databases.* Em *Proceedings of the 1996 ACM SIGMOD International Conference on Management of Data*, SIGMOD '96, páginas 103–114, New York, NY, USA, 1996. ACM, ISBN 978-0-89791-794-0. http://doi.acm.org/10.1145/233269.233324, acesso em 2016-12-21TZ. 7

[24] Safavian, S. Rasoul e David Landgrebe: *A survey of decision tree classifier methodology.* Relatório Técnico, setembro 1990. https://ntrs.nasa.gov/search.jsp?R=19910014493, acesso em 2016-12-21TZ. 7, 12

[25] Umadevi, P. e E. Divya: *Money laundering detection using TFA system.* Em *International Conference on Software Engineering and Mobile Application Modelling and Development (ICSEMA 2012)*, páginas 1–8, dezembro 2012. 7

[26] Han, Jiawei, Hong Cheng, Dong Xin e Xifeng Yan: *Frequent pattern mining: current status and future directions.* Data Mining and Knowledge Discovery, 15(1):55–86, agosto 2007, ISSN 1384-5810, 1573-756X. http://link.springer.com/article/10.1007/s10618-006-0059-1, acesso em 2016-12-21TZ. 7

[27] Cao, Dang Khoa e Phuc Do: *Applying Data Mining in Money Laundering Detection for the Vietnamese Banking Industry.* Em Pan, Jeng Shyang, Shyi Ming Chen e Ngoc Thanh Nguyen (editores): *Intelligent Information and Database Systems*, Lecture Notes in Computer Science, páginas 207–216. Springer Berlin Heidelberg, março 2012, ISBN 978-3-642-28489-2 978-3-642-28490-8. http://link.springer.com/chapter/10.1007/978-3-642-28490-8_22, acesso em 2016-12-21TZ, DOI: 10.1007/978-3-642-28490-8_22. 7

[28] Yang, Yiling, Xudong Guan e Jinyuan You: *CLOPE: A Fast and Effective Clustering Algorithm for Transactional Data.* Em *Proceedings of the Eighth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '02, páginas 682–687, New York, NY, USA, 2002. ACM, ISBN 978-1-58113-567-1. http://doi.acm.org/10.1145/775047.775149, acesso em 2016-12-21TZ. 7

[29] Sharma, Anuj e Prabin Kumar Panigrahi: *A Review of Financial Accounting Fraud Detection based on Data Mining Techniques.* International Journal of Computer Applications, 39(1):37–47, fevereiro 2012, ISSN 09758887. arXiv: 1309.3944. 7

[30] Nerlove, Marc: *Univariate and multivariate log-linear and logistic models.* Rand Corp, 1973. 7

[31] Filho, Jorge Jambeiro: *Tratamento Bayesiano de Interações entre atributos de Alta Cardinalidade.* Tese de Doutoramento, Unicamp, setembro 2007. http://www.bibliotecadigital.unicamp.br/document/?code=vtls000426153&print=y, acesso em 2016-06-04TZ. 8

[32] Filho, Jorge Jambeiro e Jacques Wainer: *Using a Hierarchical Bayesian Model to Handle High Cardinality Attributes with Relevant Interactions in a Classification Problem.* Em *Proceedings of the 20th International Joint Conference on Artifical*

Intelligence, IJCAI'07, páginas 2504–2509, San Francisco, CA, USA, 2007. Morgan Kaufmann Publishers Inc. http://dl.acm.org/citation.cfm?id=1625275.1625679, acesso em 2016-03-11TZ. 8

[33] Chandola, Varun, Arindam Banerjee e Vipin Kumar: *Anomaly Detection: A Survey*. ACM Comput. Surv., 41(3):15:1–15:58, julho 2009, ISSN 0360-0300. http://doi.acm.org/10.1145/1541880.1541882, acesso em 2017-01-03TZ. 8

[34] Hodge, Victoria J. e Jim Austin: *A Survey of Outlier Detection Methodologies*. Artificial Intelligence Review, 22(2):85–126, outubro 2004, ISSN 0269-2821, 1573-7462. http://link.springer.com/article/10.1007/s10462-004-4304-y, acesso em 2017-01-03TZ. 8

[35] Agyemang, Malik, Ken Barker e Rada Alhajj: *A comprehensive survey of numeric and symbolic outlier mining techniques*. Intelligent Data Analysis, 10(6):521–538, janeiro 2006, ISSN 1088-467X. http://content.iospress.com/articles/intelligent-data-analysis/ida00266, acesso em 2017-01-03TZ. 8

[36] Markou, Markos e Sameer Singh: *Novelty detection: a review—part 2:: neural network based approaches*. Signal Processing, 83(12):2499–2521, dezembro 2003, ISSN 0165-1684. http://www.sciencedirect.com/science/article/pii/S0165168403002032, acesso em 2017-01-03TZ. 8

[37] Markou, Markos e Sameer Singh: *Novelty detection: a review—part 1: statistical approaches*. Signal Processing, 83(12):2481–2497, dezembro 2003, ISSN 0165-1684. http://www.sciencedirect.com/science/article/pii/S0165168403002020, acesso em 2017-01-03TZ. 8

[38] Patcha, Animesh e Jung Min Park: *An overview of anomaly detection techniques: Existing solutions and latest technological trends*. Computer Networks, 51(12):3448–3470, agosto 2007, ISSN 1389-1286. http://www.sciencedirect.com/science/article/pii/S138912860700062X, acesso em 2017-01-03TZ. 8

[39] Rousseeuw, Peter J. e Annick M. Leroy: *Robust Regression and Outlier Detection*. John Wiley & Sons, fevereiro 2005, ISBN 978-0-471-72537-4. 8

[40] Barnett, Vic e Toby Lewis: *Outliers in Statistical Data*. Wiley, Chichester ; New York, 3 edition edição, abril 1994, ISBN 978-0-471-93094-5. 8

[41] Beckman, R. J. e R. D. Cook: *Outlier ... ... s*. Technometrics, 25(2):119–149, maio 1983, ISSN 0040-1706. http://dx.doi.org/10.1080/00401706.1983.10487840, acesso em 2017-01-03TZ. 8

[42] Bakar, Z. A., R. Mohemad, A. Ahmad e M. M. Deris: *A Comparative Study for Outlier Detection Techniques in Data Mining*. Em *2006 IEEE Conference on Cybernetics and Intelligent Systems*, páginas 1–6, junho 2006. 8

[43] Goodfellow, Ian, Yoshua Bengio e Aaron Courville: *Deep Learning*. The MIT Press, Cambridge, MA, novembro 2016, ISBN 978-0-262-03561-3. 9, 14, 15

[44] LeCun, Yann, Yoshua Bengio e Geoffrey Hinton: *Deep learning*. Nature, 521(7553):436–444, maio 2015, ISSN 0028-0836, 1476-4687. `http://www.nature.com/doifinder/10.1038/nature14539`, acesso em 2016-06-03TZ. 9

[45] Nisbet, Robert, Gary Miner e John Elder IV: *Handbook of Statistical Analysis and Data Mining Applications*. Academic Press, maio 2009, ISBN 978-0-08-091203-5. Google-Books-ID: U5np34a5fmQC. 9

[46] Chapman, Pete, Julian Clinton, Randy Kerber, Thomas Khabaza, Thomas Reinartz, Colin Shearer e Rudiger Wirth: *CRISP-DM 1.0 Step-by-step data mining guide*. IBM, agosto 2000. 10

[47] Friedman, Jerome H.: *Greedy Function Approximation: A Gradient Boosting Machine*. The Annals of Statistics, 29(5):1189–1232, 2001, ISSN 0090-5364. 12

[48] Friedman, Jerome H.: *Stochastic gradient boosting*. Computational Statistics & Data Analysis, 38(4):367–378, fevereiro 2002, ISSN 0167-9473. 12

[49] Kuhn, Max e Kjell Johnson: *Applied Predictive Modeling*. Springer, New York, 2013 edition edição, setembro 2013, ISBN 978-1-4614-6848-6. 12, 13

[50] Friedman, Jerome, Trevor Hastie e Robert Tibshirani: *Additive logistic regression: a statistical view of boosting (With discussion and a rejoinder by the authors)*. The Annals of Statistics, 28(2):337–407, abril 2000, ISSN 0090-5364, 2168-8966. 12

[51] Ridgeway, Greg: *Generalized Boosted Models: A Guide to the GBM Package*. ResearchGate, 1:1–12, janeiro 2005. 12

[52] Breiman, Leo: *Bagging predictors*. Machine Learning, 24(2):123–140, agosto 1996, ISSN 0885-6125, 1573-0565. 13

[53] Rossini, A. J., Luke Tierney e Na Li: *Simple Parallel Statistical Computing in R*. Journal of Computational and Graphical Statistics, 16(2):399–420, junho 2007, ISSN 1061-8600. 13

[54] Breiman, Leo: *Random Forests*. Machine Learning, 45(1):5–32, outubro 2001, ISSN 0885-6125, 1573-0565. 13

[55] Strobl, Carolin, Anne Laure Boulesteix, Achim Zeileis e Torsten Hothorn: *Bias in random forest variable importance measures: Illustrations, sources and a solution*. BMC Bioinformatics, 8:25, 2007, ISSN 1471-2105. 14

[56] Hecht-Nielsen, R.: *Theory of the backpropagation neural network*. Em *International 1989 Joint Conference on Neural Networks*, páginas 593–605 vol.1, 1989. 14

[57] Hinton, G. E. e R. R. Salakhutdinov: *Reducing the Dimensionality of Data with Neural Networks*. Science, 313(5786):504–507, julho 2006, ISSN 0036-8075, 1095-9203. 15, 57

[58] Powers, David Martin: *Evaluation: from Precision, Recall and F-measure to ROC, Informedness, Markedness and Correlation*. dezembro 2011, ISSN 2229-3981. http://dspace.flinders.edu.au/xmlui/handle/2328/27165, acesso em 2016-12-22TZ. 16

[59] Matthews, B. W.: *Comparison of the predicted and observed secondary structure of T4 phage lysozyme*. Biochimica et Biophysica Acta (BBA) - Protein Structure, 405(2):442–451, outubro 1975, ISSN 0005-2795. 19

[60] Grupo de Egmont: *100 Casos de Lavagem de Dinheiro*. COAF, 2001. 20, 24

[61] Nações Unidas, UNODC: *UNODC marca Dia Nacional de Prevenção à Lavagem de Dinheiro*, outubro 2013. https://www.unodc.org/lpo-brazil/pt/frontpage/2013/10/29-unodc-marca-dia-nacional-de-prevencao-a-lavagem-de-dinheiro.html, acesso em 2016-05-02TZ, Acesso em: 02/05/2016. 20

[62] Brasil: *Lei 9613, de 03 de março de 1998*, 1998. http://www.planalto.gov.br/ccivil_03/LEIS/L9613.htm, acesso em 2016-03-28TZ. 21

[63] Amaral, Leandro Freitas: *Lavagem de Dinheiro*, maio 2015. http://www.coaf.fazenda.gov.br/backup/pld-ft/sobre-a-lavagem-de-dinheiro, acesso em 2016-03-28TZ, Acesso em: 28/03/2016. 21

[64] Receita Federal do Brasil: *Portaria RFB nº 671, de 07 de fevereiro de 2014*, 2014. 22

[65] Duarte, Sinval: *Receita Federal de São Paulo ganha laboratório contra lavagem de dinheiro*, abril 2014. https://www.justica.gov.br/noticias/receita-federal-de-sao-paulo-ganha-laboratorio-contra-lavagem-de-dinheiro, acesso em 2016-05-02TZ, Acesso em: 01/05/2016. 22

[66] Conselho de Controle de Atividades Financeiras: *Fases da Lavagem de Dinheiro*, junho 2014. http://www.coaf.fazenda.gov.br/links-externos/fases-da-lavagem-de-dinheiro, acesso em 2016-04-04TZ, Acesso em: 04/04/2016. 22

[67] Conselho de Controle de Atividades Financeiras: *Casos e Casos - I Coletânea de Casos Brasileiros de Lavagem de Dinheiro*. COAF, 2011. 24

[68] He, Ping: *A typological study on money laundering*. Journal of Money Laundering Control, 13(1):15–32, janeiro 2010, ISSN 1368-5201. 24

[69] Madinger, John: *Money Laundering: A Guide for Criminal Investigators, Third Edition*. CRC Press, dezembro 2011, ISBN 978-1-4398-6912-3. 24

[70] Greene, Olivia: *Trade-Based Money Laundering*, julho 2015. https://www.dhgllp.com/Portals/4/ResourceMedia/publications/Risk-Advisory-Trade-Based-Money-Laundering.pdf, acesso em 2016-05-02TZ, Acesso em: 01/05/2016. 24

[71] Brasil: *Lei nº 9.613, de 3 de março de 1998*, 1988. http://www.planalto.gov.br/ccivil_03/leis/L9613.htm, acesso em 2016-12-21TZ. 24

[72] Brasil: *Decreto-Lei nº 3.689, de 3 de outubro de 1941*, 1941. http://www.planalto.gov.br/ccivil_03/decreto-lei/Del3689.htm, acesso em 2016-12-21TZ. 24

[73] Brasil: *Decreto-Lei nº 37, de 18 de novembro de 1966*, 1966. http://www.planalto.gov.br/ccivil_03/decreto-lei/Del0037.htm, acesso em 2016-12-26TZ. 29

[74] Ministério da Fazenda: *Portaria MF nº 203, de 14 de maio de 2012*, 2012. http://normas.receita.fazenda.gov.br/sijut2consulta/link.action?visao=anotado&idAto=37965, acesso em 2016-12-26TZ. 29

[75] Chaudhuri, Surajit e Umeshwar Dayal: *An Overview of Data Warehousing and OLAP Technology*. SIGMOD Rec., 26(1):65–74, março 1997, ISSN 0163-5808. http://doi.acm.org/10.1145/248603.248616, acesso em 2016-12-26TZ. 31

[76] Brasil: *Lei nº 5.172, de 25 de outubro de 1966*, 1966. https://www.planalto.gov.br/ccivil_03/leis/L5172.htm, acesso em 2016-12-26TZ. 32

[77] Receita Federal do Brasil: *Portaria RFB nº 2344, de 24 de março de 2011*, 2011. http://sijut2.receita.fazenda.gov.br/sijut2consulta/imprimir.action?visao=original&idAto=30552, acesso em 2016-12-26TZ. 32

[78] Brasil: *Lei nº 8.112, de 11 de dezembro de 1990*, 1990. https://www.planalto.gov.br/ccivil_03/leis/L8112cons.htm, acesso em 2016-12-26TZ. 32

[79] Limpert, Eckhard, Werner A. Stahel e Markus Abbt: *Log-normal Distributions across the Sciences: Keys and Clues On the charms of statistics, and how mechanical models resembling gambling machines offer a link to a handy way to characterize log-normal distributions, which can provide deeper insight into variability and probability—normal or log-normal: That is the question*. BioScience, 51(5):341–352, maio 2001, ISSN 0006-3568, 1525-3244. http://bioscience.oxfordjournals.org/content/51/5/341, acesso em 2016-12-21TZ. 38

[80] Dawson, Robert: *How Significant Is a Boxplot Outlier?* Journal of Statistics Education, 19(2), janeiro 2011, ISSN 1069-1898. 40

[81] Nelder, J. A. e R. J. Baker: *Generalized Linear Models*. Em *Encyclopedia of Statistical Sciences*. John Wiley & Sons, Inc., 2004, ISBN 978-0-471-66719-3. http://onlinelibrary.wiley.com/doi/10.1002/0471667196.ess0866.pub2/abstract, acesso em 2016-12-21TZ. 45

[82] Witten, Ian H. e Eibe Frank: *Data Mining: Practical Machine Learning Tools and Techniques, Second Edition*. Morgan Kaufmann, julho 2005, ISBN 978-0-08-047702-2. 45

[83] Bolfarine, Heleno e Wilton de Oliveira Bussab: *Elementos de amostragem*. Edgard Blücher, 2005, ISBN 978-85-212-0367-4. Google-Books-ID: a_fqPwAACAAJ. 66, 67

[84] Cochran, William G.: *Sampling Techniques, 3rd Edition*. John Wiley & Sons, New York, 3rd edition edição, janeiro 1977, ISBN 978-0-471-16240-7. 67

[85] Kolmogorov, A. N.: *Foundations of the Theory of Probability*. Chelsea Pub Co, S.l., 2 edition edição, junho 1960, ISBN 978-0-8284-0023-7. 69

Apêndice A

Código em linguagem R

```
## Carga das bibliotecas utilizadas

library(dplyr)
library('corrplot') #package corrplot
library("arules", lib.loc="~/R/win-library/3.3")
library("Hmisc", lib.loc="~/R/win-library/3.3")
library(h2o)

#########################
### Tratamentos iniciais

## Leitura do arquivo consolidado em csv
Tabelao <- read.csv("D:/Analise.Dados.PPCA/Dissertacao/Arq.Brutos/Tabelao4_Emp.csv")
Tabelao <- Tabelao[,-c(3, 79:82)]
Tabelao [redacted] <- as.character(Tabelao [redacted]

## Gera tabela com empresas que [redacted]
TabelaoM0 <- Tabelao %>% filter [redacted] > 0)

#######################################
### Entendimento e preparação dos dados


## Verifica a variância dos atributos numéricos
Varia <- as.logical(summarise all(TabelaoM0, function(col)
ifelse(is.numeric(col),var(col)==0,F)))
colnames(TabelaoM0[Varia])

## Verifica a qte de dados faltantes nos atributos
zeros <- as.logical(summarise all(TabelaoM0, function(col)
ifelse(is.numeric(col),sum(col==0)/151.74 >60 ,F)))
colnames(TabelaoM0[zeros]) # Mostra quais colunas têm mais de 60% de 0

## Plota histogramas
par(mfrow=c(1,2), oma = c(0, 0, 2, 0))

    # [redacted]
    hist(TabelaoM0$[redacted],nclass = 50, main = NULL,  xlab = [redacted] ,
    ylab = "Frequência")
    hist(log(TabelaoM0 [redacted]),nclass = 50, main = NULL, xlab = "log [redacted]
    [redacted])", ylab = "Frequência")
    mtext("Distribuição dos [redacted] , outer = TRUE, , cex = 1.5)

    # [redacted]
    hist(TabelaoM0 [redacted] nclass = 50, main = NULL, xlab =
    [redacted] , ylab = "Frequência")
    hist(log(TabelaoM0$[redacted] ,nclass = 50, main = NULL, xlab =
    "log [redacted] ", ylab = "Frequência")
    mtext("[redacted] outer = TRUE, , cex = 1.5)

    # [redacted]
    hist(TabelaoM0 [redacted],nclass = 50, main = NULL, xlab = [redacted] ylab =
    "Frequência")
    hist(log(TabelaoM0 [redacted]),nclass = 50, main = NULL, xlab = "log [redacted] ylab
    = "Frequência")
    mtext( [redacted] , outer = TRUE, , cex = 1.5)

    # [redacted]
    hist(TabelaoM0 [redacted],nclass = 50, main = NULL, xlab = [redacted]", ylab =
    "Frequência")
    hist(log(TabelaoM0 [redacted]),nclass = 50, main = NULL, xlab = "log [redacted])",
    ylab = "Frequência")
    mtext( [redacted]", outer = TRUE, , cex =
    1.5)

## Análisa a correlação entre variáveis

par(mfrow=c(1,1))
```

```
Tab Limpo <- TabelaoM0[!zeros]
corrplot(cor(Tab_Limpo[, sapply(Tab_Limpo, is.numeric)]),
         method = "ellipse",
         type = "upper",
         title = "Correlação entre os atributos",
         tl.cex = .7,
         hclust.method = "centroid")# tl.pos = "ld") #plot matrix


## Busca de outliers

par(mfrow=c(8,5), mar = c(.7,0,.8,0), cex=.4)

for(i in 1:ncol(Tab Limpo)-1){
  if(is.numeric(Tab_Limpo[,i])){
    boxplot(log(Tab Limpo[,i]) ~ cut2(Tab Limpo[,i], g=10, levels.mean=TRUE),
    frame.plot=FALSE, axes=FALSE, main= colnames(Tab Limpo[i]))
  }
}

## Análise de relacionamento entre [redacted] e os demais atributos

par(mfrow=c(8,5), mar = c(.7,0,.8,0), cex=.4)

for(i in 1:ncol(Tab Limpo)){
  if(is.numeric(Tab_Limpo[,i]) & i != 4){
    boxplot(log(Tab Limpo[,i]) ~ cut2(Tab Limpo[,4], g=20, levels.mean=TRUE),
    frame.plot=FALSE, axes=FALSE, main= colnames(Tab_Limpo[i]), outline = F)
  }
}

    # Exportação dos gráficos em JPG
    par(mfrow=c(1,1), mar = c(.7,0,.8,0), cex=.4)

    for(i in 1:ncol(Tab_Limpo)){
      if(is.numeric(Tab Limpo[,i]) & i != 4){
        name <- paste("graf_",colnames(Tab_Limpo[i]), "_", i, ".jpeg", sep = "")
        jpeg(name)
        boxplot(log(Tab_Limpo[,i]) ~ cut2(Tab_Limpo[,4], g=20, levels.mean=TRUE),
        frame.plot=FALSE, axes=FALSE, main= colnames(Tab Limpo[i]), outline = F)
        dev.off()
      }
    }

    # Gera gráfico isolado [redacted]

    par(mfrow=c(1,2), mar = c(.7,0,.8,0), cex=.8)
    for(i in 8:9){
      if(is.numeric(Tab Limpo[,i]) & i != 4){
        boxplot(log(Tab_Limpo[,i]) ~ cut2(Tab_Limpo[,4], g=20, levels.mean=TRUE),
        frame.plot=FALSE, axes=FALSE, main= colnames(Tab Limpo[i]), outline = F)
      }
    }

    # Gera gráfico isolado [redacted]

    par(mfrow=c(1,2), mar = c(.7,0,.8,0), cex=.8)
    for(i in c(17,18)){
      if(is.numeric(Tab_Limpo[,i]) & i != 4){
        boxplot(log(Tab Limpo[,i]) ~ cut2(Tab Limpo[,4], g=20, levels.mean=TRUE),
        frame.plot=FALSE, axes=FALSE, main= colnames(Tab_Limpo[i]), outline = F)
      }
    }

    # Gera gráfico isolado dos valores [redacted]

    par(mfrow=c(1,2), mar = c(.7,0,.8,0), cex=.8)
    for(i in c(25,3)){
```

```
      if(is.numeric(Tab Limpo[,i]) & i != 4){
        boxplot(log(Tab_Limpo[,i]) ~ cut2(Tab_Limpo[,4], g=20, levels.mean=TRUE),
        frame.plot=FALSE, axes=FALSE, main= colnames(Tab Limpo[i]), outline = F)
      }
    }

############################################################
### Indução dos modelos, Treinamento, Testes e avaliação

# Essa fase foi realizada no H2O - Ver apêndice B
```

Apêndice B

Código em *H2O*

```
'Particionamento do dataframe em 25% e 75%
splitFrame "g", [0.75], ["g_0.750","g_0.250"], 836674

'Código para geração modelo GBM com variações em número de árvores, profundidade e
balanceamento de classes e cross-validation para 10 folds

buildModel 'gbm', {"model_id":"GBM",
"training frame":"g 0.750",
"validation_frame":"g_0.250",
"nfolds":"10",
"response_column":"classe",
"ignored columns":[ [redacted] , [redacted] ],
"ignore_const_cols":true,
"min rows":10,
"nbins":20,
"seed":-1,
"learn rate":0.1,
"sample_rate":1,
"col sample rate":1,
"score_each_iteration":false,
"score tree interval":0,
"fold_assignment":"AUTO",
"nbins top level":1024,
"nbins_cats":1024,
"r2 stopping":1.7976931348623157e+308,
"stopping_rounds":0,
"stopping metric":"AUTO",
"stopping_tolerance":0.001,
"max_runtime_secs":0,
"learn rate annealing":1,
"distribution":"AUTO",
"huber alpha":0.9,
"checkpoint":"",
"col sample rate per tree":1,
"min_split_improvement":0.00001,
"histogram type":"AUTO",
"categorical_encoding":"AUTO",
"keep cross validation predictions":false,
"keep_cross_validation_fold_assignment":false,
"build tree one node":false,
"sample_rate_per_class":[],
"col_sample_rate_change_per_level":1,
"max abs leafnode pred":1.7976931348623157e+308,
"pred_noise_bandwidth":0,
"grid id":"grid3 GBM",
"hyper_parameters":{"ntrees":["40","55","70"],
"max depth":["3","6","10"],
"balance_classes":[false,true]},
"search criteria":{"strategy":"Cartesian"}}


'Código para geração modelo DRF com variações em número de árvores, profundidade e
balanceamento de classes e cross-validation para 10 folds

buildModel 'drf',  {"model_id":"drf",
"training frame":"g 0.750",
"validation_frame":"g_0.250",
"nfolds":"10",
"response_column":"classe",
"ignored columns":[ [redacted] , [redacted] ],
"ignore_const_cols":true,
"min rows":1,
"nbins":20,
"seed":-1,
"mtries":-1,
"sample rate":0.6320000290870667,
"score each iteration":false,
"score_tree_interval":0,
```

```
"fold assignment":"AUTO",
"nbins_top_level":1024,
"nbins cats":1024,
"r2_stopping":1.7976931348623157e+308,
"stopping_rounds":0,
"stopping metric":"AUTO",
"stopping_tolerance":0.001,
"max runtime secs":0,
"checkpoint":"",
"col sample rate per tree":1,
"min_split_improvement":0.00001,
"histogram type":"AUTO",
"categorical_encoding":"AUTO",
"keep cross validation predictions":false,
"keep_cross_validation_fold_assignment":false,
"build tree one node":false,
"sample rate per class":[],
"binomial_double_trees":false,
"col sample rate change per level":1,
"grid_id":"grid-drf",
"hyper parameters":{"ntrees":["50","20","80"],
"max_depth":["20","3","50"],
"balance classes":[true,false]},
"search_criteria":{"strategy":"Cartesian"}}


'Código para geração modelo DLA com a base "Arranjo 1"

buildModel 'deeplearning', {"model_id":"deep_arr1",
"training frame":"arr1",
"nfolds":0,
"ignore const cols":true,
"activation":"Rectifier",
"hidden":[3,2,3],
"epochs":"10",
"variable importances":false,
"score_each_iteration":false,
"max hit ratio k":0,
"checkpoint":"",
"standardize":true,
"train_samples_per_iteration":-2,
"adaptive_rate":true,
"input dropout ratio":0,
"l1":0,
"l2":0,
"loss":"Automatic",
"distribution":"AUTO",
"quantile_alpha":0.5,
"huber alpha":0.9,
"score_interval":5,
"score training samples":10000,
"score_validation_samples":0,
"score duty cycle":0.1,
"stopping_rounds":5,
"stopping_metric":"AUTO",
"stopping tolerance":0,
"max_runtime_secs":0,
"autoencoder":true,
"categorical_encoding":"AUTO",
"pretrained autoencoder":"",
"overwrite_with_best_model":true,
"target ratio comm to comp":0.05,
"seed":2405,
"rho":0.99,
"epsilon":1e-8,
"nesterov accelerated gradient":true,
"max w2":"Infinity",
"initial_weight_distribution":"UniformAdaptive",
```

```
"classification stop":0,
"regression_stop":0.000001,
"score validation sampling":"Uniform",
"diagnostics":true,
"fast_mode":true,
"force load balance":true,
"single_node_mode":false,
"shuffle training data":false,
"missing_values_handling":"MeanImputation",
"quiet mode":false,
"sparse":false,
"col major":false,
"average_activation":0,
"sparsity beta":0,
"max_categorical_features":2147483647,
"reproducible":true,
"export weights and biases":false,
"mini_batch_size":1,
"elastic averaging":false}


'Código para geração modelo DLA com a base "Arranjo 2"

buildModel 'deeplearning', {"model_id":"deep_arr2",
"training frame":"arr2",
"nfolds":0,
"ignore const cols":true,
"activation":"Rectifier",
"hidden":[12,6,3,6,12],
"epochs":"10",
"variable_importances":false,
"score each iteration":false,
"max_hit_ratio_k":0,
"checkpoint":"",
"standardize":true,
"train samples per iteration":-2,
"adaptive_rate":true,
"input dropout ratio":0,
"l1":0,
"l2":0,
"loss":"Automatic",
"distribution":"AUTO",
"quantile alpha":0.5,
"huber_alpha":0.9,
"score interval":5,
"score_training_samples":10000,
"score validation samples":0,
"score_duty_cycle":0.1,
"stopping rounds":5,
"stopping_metric":"AUTO",
"stopping tolerance":0,
"max_runtime_secs":0,
"autoencoder":true,
"categorical_encoding":"AUTO",
"pretrained_autoencoder":"",
"overwrite with best model":true,
"target_ratio_comm_to_comp":0.05,
"seed":2405,
"rho":0.99,
"epsilon":1e-8,
"nesterov_accelerated_gradient":true,
"max w2":"Infinity",
"initial_weight_distribution":"UniformAdaptive",
"classification stop":0,
"regression_stop":0.000001,
"score validation sampling":"Uniform",
"diagnostics":true,
"fast_mode":true,
```

```
"force load balance":true,
"single_node_mode":false,
"shuffle training data":false,
"missing_values_handling":"MeanImputation",
"quiet_mode":false,
"sparse":false,
"col_major":false,
"average activation":0,
"sparsity_beta":0,
"max categorical features":2147483647,
"reproducible":true,
"export weights and biases":false,
"mini_batch_size":1,
"elastic_averaging":false}
```

Apêndice C

Artigo aceito para publicação no 15º IEEE International Conference on Machine Learning and Applications (IEEE ICMLA'16)

Deep Learning Anomaly Detection as Support Fraud Investigation in Brazilian Exports and Anti-Money Laundering

Ebberth L Paula*, Rommel N. Carvalho†‡, Marcelo Ladeira† and Thiago Marzagão†‡
*Coordination of Research and Investigation (COPEI)
Secretariat of Federal Revenue of Brazil (RFB), Brasilia, DF, Brazil
Email: ebberth.paula@receita.fazenda.gov.br
†Department of Computer Science (CIC)
University of Brasilia (UnB), Brasilia, DF, Brazil
Email: {mladeira,rommelnc}@unb.br
‡Department of Research and Strategic Information (DIE)
Ministry of Transparency, Monitoring and Control (MTFC), Brasilia, DF, Brazil
Email: {rommel.carvalho,thiago.marzagao}@cgu.gov.br

***Abstract*—Normally exports of goods and products are transactions encouraged by the governments of countries. Typically these incentives are promoted by tax exemptions or lower tax collections. However, exports fraud may occur with objectives not related to tax evasion, for example money laundering. This article presents the results obtained in implementing the unsupervised Deep Learning model to classify Brazilian exporters regarding the possibility of committing fraud in exports. Assuming that the vast majority of exporters have explanatory features of their export volume which interrelate in a standard way, we used the AutoEncoder to detect anomalous situations with regards to the data pattern. The databases used in this work come from exports of goods and products that occurred in Brazil in 2014, provided by the Secretariat of Federal Revenue of Brazil. From attributes that characterize export companies, the model was able to detect anomalies in at least twenty exporters.**

1. Introduction

Several authors ([1], [2], [3], and [4]) indicate money laundering cases with the use of foreign trade, thus taking advantage of the difficulties of the countries to exchange information massively, to operate the 'clean' money. The US Immigration and Customs Enforcement [5] define Trade-Based Money Laundering as "an alternative remittance system that allows illegal organizations the opportunity to earn, move and store proceeds disguised as legitimate trade. Value can be moved through this process by false-invoicing, over-invoicing and under-invoicing commodities that are imported or exported around the world". In Brazil, the law is explicit as to the application of money laundering to those who import or export goods that do not correspond to their true value [6].

According to the Egmont Group[1], "Money laundering is the process by which the criminal transforms resources from illegal activities in assets with an apparently legal source. This practice generally involves multiple transactions, to hide the source of financial assets and allow them to be used without compromising the criminals. The concealment is thus the basis for all washing operations involving money from a criminal history".

Brazilian exports are directed annually to nearly 200 countries. Thousands of invoices with tax suspension on goods destined for export are issued daily. About 50,000 legal entities directly or indirectly operated in shipping goods and merchandise abroad annually. The *Mercosur Common Nomenclature* (NCM)[2], used for the tax classification of goods, distinguishes between 9,600 types of goods and merchandise, each subject to specific legislation. Most of the variables are nonlinearly correlated and temporally dependent. It is difficult for humans to distinguish the normal state from the abnormal state only by looking at the raw data. For this reason, training a machine to learn the normal state and displaying the reconstruction error as the anomaly score is valuable.

This paper presents results of applying unsupervised deep learning AutoEncoder in databases of foreign trade of the Secretariat of Federal Revenue of Brazil with the objective of identifying exporting corporations whose explanatory variables of their export operations in 2014 show signs of divergence (anomalies) compared to regular patterns found.

This article is structured as follows: Section 2 presents

1. International group created to promote worldwide the treatment of suspected communications related to money laundering. http://www.egmontgroup.org/

2. The Mercosur Common Nomenclature (MCN) was adopted by the countries that integrate the Argentina, Brazil and Uruguay Block to foster international trade growth, make the creation and comparison of statistics easier, in addition to elaborating freight tariffs and providing other relevant information to international trade. http://bit.ly/29wHa1T

the work related to money laundering, fraud and error detection on imports. It also presents the state-of-the-art data mining techniques for high-cardinality attributes with non-linear relationships. Section 3 presents the Cross Industry Standard Process for Data Mining (CRISP-DM) methodology used in this study. Section 4 addresses the current scenario in Brazil for fraud detection in exports and combating money laundering. Section 5 presents the understanding of the data and their preparation for modeling. Section 6 addresses the modeling process and evaluation of the model. Finally, Section 7 presents the conclusion and future work.

2. Related Works

In this section we present some of the most relevant works related to the application of data mining techniques in the field of combating money laundering and fraud.

Applications developed for the financial system represent most of the articles that use data mining techniques for money laundering detection, even when searching for papers from more than ten years ago. To the best of our knowledge, there are no applications involving trade-based money laundering detection. Nevertheless, there are works that use artificial intelligence for this purpose via *Financial Crimes Enforcement Network* in 1995 [7] and 1998 [8]. Unfortunatelly, these articles do not specify the databases used.

Larik and Haider [9] approach the problem of dirty money entering the financial system with a hybrid approach for detecting anomalies in financial transactions. This approach employs unsupervised clusters to meet normal standards of behavior for clients in conjunction with the use of statistical techniques to identify the diversion of a particular transaction of the corresponding expected behavior in their group. A variant of the Euclidean Adaptive Resonance Theory (EART) is suggested to group clients into different clusters. The perspective of the authors, unlike what is discussed in this paper, is a financial institution with a focus on transactions.

Khan et al. [10] present a Bayesian network approach (BN) to analyze transactions of customers of a financial institution in order to detect suspicious patterns. Based on transaction history, the proposed approach assigns a baseline from which the transaction becomes suspect. The problem with this approach when transposed to this work domain is the absence of a relevant historical period.

Raza and Haider [11] join the two approaches mentioned above to create what they called *Suspicious Activity Reporting using Dynamic Bayesian Network* (SARDBN), a combination of clustering with *dynamic Bayesian network* (DBN) to identify anomalies in sequences of transactions. The authors created an index called *Anomaly Index Rank and using Entropy* (AIRE), which measures the degree of abnormality in an operation and compares it with a predefined threshold value to mark the transaction as normal or suspicious. This index is similar to the baseline proposed by Khan et al. [10]. However, this division into two phases appear to suffer less of the problems outlined in the previous section, because the clustering first evaluates all the customers and the AIRE evaluates transactions of a given client individually.

Rajput et al. [12] address the problem by proposing ontologies and rules written in *Semantic Web Rule Language* (SWRL). Such an approach, according to the authors, require less computation and allows the reuse of the knowledge base in similar areas.

In the money laundering domain, Sharma and Panigrahi [13] show that the technical data mining and logistic models, neural networks, Bayesian networks, and decision trees have been extensively applied to provide solutions to the problems of fraud detection and classification. From the study of forty-five articles on fraud in the financial system, the authors present four groups of approaches in mining commonly used data. Table 1 presents a summary of the survey.

TABLE 1. APPROACHES TO FRAUD DETECTION IN THE FINANCE DOMAIN

Method	% of papers
Regression models	40%
Neural Network	31%
Fuzzy Logic	16%
Genetic algorithms and specialist systems	13%

Finally, Jambeiro and Wainer [14], [15], when examining the use of Bayesian methods in a practical interest of pattern classification problem for the Secretariat of Federal Revenue of Brazil from a similar basis (databases of imports-trade and NCM) to the one proposed in this work (databases of exports-trade and NCM), showed empirically that more advanced Bayesian strategies for the treatment of high cardinality of attributes (pre-processing for cardinality reduction and substitution of conditional probability tables, Bayesian networks, default tables, decision trees and decision graphs) although they bring specific benefits, do not result in overall performance gain in our target domain. Their work then turned to propose a new Bayesian classification method, named Hierarchical Pattern Bayes (HPB). "The HPB runtime is exponential in the number of attributes, but is independent of its cardinality. Thus, in areas where the attributes are few, but have high cardinality, it is much faster" than traditional algorithms.

2.1. State-of-the-art

In this work we chose to use Deep Neural networks AutoEncoders. This tool, in addition to dealing with the problems faced by Jambeiro [14], [15], allows unsupervised (AutoEncoder) and semi-supervised detection of anomalies. When compared to most related works of the financial system, it has the advantage of performing nonlinear generalizations.

Deep Learning has emerged as one capable of reaching the state-of-the-art algorithm for various domains: Szegedy et al. [16] propose a deep convolutional neural network architecture that achieves the new state-of-the-art for classification and detection in the ImageNet Large-Scale Visual

Recognition Challenge 2014. Jaiswal et al. [17] achieve state-of-the-art performance on the FERA-2015 Challenge dataset recognizing spontaneous facial expressions. Liang et al. [18] achieve state-of-the-art of Atari Games using shallow reinforcement learning with a recently introduced Deep Q-Networks (DQN) algorithm - a combination of deep neural networks and reinforcement learning.

The future of deep learning is unsupervised learning, but it has been overshadowed by the successes of purely supervised learning [19]. Semi-supervised learning follows the same path. Problems with or without a small subset of the observations having a corresponding class label are of "immense practical interest in a wide range of applications where unlabeled data is abundant, but obtaining class labels is expensive or impossible to obtain for the entire data set" [20].

3. Methodology

This study used as reference model the Cross Industry Standard Process for Data Mining (CRISP-DM) [21], since it is a well-known data mining reference model. The CRISP-DM methodology is flexible and allows the creation of a model that fits the specific needs of projects. It is observed that the execution sequence of the phases is not rigid and depends on the results achieved in each phase (see figure 1).

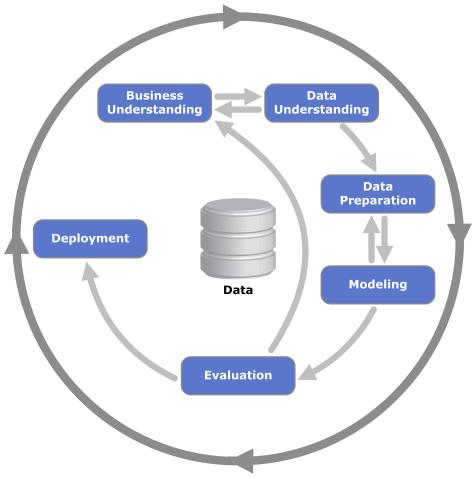

Figure 1. Phases of the CRISP-DM Process Model

The life cycle of the mining project on this methodology consists of six phases:

1) *Business understanding* This initial phase focuses on understanding the goals and project requirements from a business perspective, then converting that knowledge into a definition of the data mining problem and a preliminary plan designed to achieve the objectives.
2) *Data understanding* The data understanding phase starts with the initial data collection and continues with activities that allow the familiarization with the data, the identification of data quality problems, the discovery of the first insights into the data and/or detection of interesting subsets to form hypotheses about the unknown information. Sections 1 and 4 of this paper summarize the results of this step of the methodology.
3) *Data preparation* The data preparation phase concentrates all activities necessary to the construction of the final data set to be used in the modeling phase. Data preparation tasks are typically performed several times and not in any prescribed order. The tasks include selecting, cleaning, constructing, integrating and formatting data for modeling purposes.
4) *Modeling* At this stage, several modelling techniques are chosen and applied and its parameters are adjusted to the optimum values. Usually, there are many different techniques for the same data mining problem. Some techniques have specific requirements regarding the form of the data. Thus, it is often necessary to go back to previous phases to perform adjustments.
5) *Evaluation* In this phase, it is important to evaluate and review the steps performed to create the final model (or models), before final deployment, to make sure it achieves the business objectives. It is important to try to determine if there are any important business issues not yet considered. At the end of this phase, it is important to decide if the results are satisfactory and whether the final model should be used or not.
6) *Deployment* The knowledge obtained with the models generated must be applied in the Organization and this knowledge must be disseminated and presented to users in a way that they can use it.

4. Scenario

In Brazil, in 2012, the Law 9613/98 (amended by Law 12,683/12) [6], brought important advances in preventing and combating money laundering with the extinction of the exhaustive list of predicate criminal offenses. Now any criminal offense is considered a precedent to money laundering. This law establishes a framework to combat money laundering and related crimes in which the Secretariat of Federal Revenue of Brazil plays an important role in fiscal intelligence.

In this context, the Secretariat of Federal Revenue of Brazil is responsible, among other related duties, to "plan, coordinate and implement the tax intelligence activities in the fight against laundering and concealment of assets, rights and values" [22]. The cases that may relate to money laundering crimes are selected for investigation from various mechanisms such as complaints, audits, lawsuits, cross-checking, among others.

It is intended that the presented data mining techniques will join the currently existing mechanisms for selection

of suspected exports frauds. Besides detecting anomalies related to fraud and money laudering, the analysis of complaints against companies can also benefit from models generated by these techniques. The predictive variables for the company in question may be submitted to the model for evaluation of their suspicion.

5. Data Understanding and Data Preparation

It was identified eighty attributes that proved sufficient to characterize fraudulent exports based on the experience of the author and empirical studies conducted. These attributes are distributed in ten different dimensions:

1) *Registration* Registration data that allow unequivocal identification of the exporter and its license to operate in foreign trade.
2) *Foreign Trade* Export volumes and values, commercial classification of goods and products, origin and destination of goods and products.
3) *Tax Collection* Amounts charged and paid in fees and taxes in the years in which the exporting company conducted export activities.
4) *Financial Transactions* Transacted values in Brazilian financial institutions, consolidated per year, bank accounts (debit and credit), credit cards, and foreign exchange transactions (purchase, sale, and transfer).
5) *Tax Withheld at Source* Amounts related to taxes that companies are required to hold upon payment for services.
6) *Employees* Amounts collected by the exporting companies in the form of social security of its employees.
7) *Electronic Invoices* Information about electronic invoices emitted when the company purchases goods and products for commercialization and industrialization and about electronic invoices emitted when the company sells goods and manufactured products for export.
8) *Supplementary Obligations* Information regarding compliance with the obligation to deliver different types of declarations to the tax authorities.
9) *Inspection Operations* Information regarding tax and customs inspections already conducted in exporting companies.
10) *Others* Information concerning surveillance operations already carried out in the exporting companies.

One of the proposed models to be evaluated in the next phase of the Crisp-DM was Deep Learning AutoEncoder (see Section 6). For detection of anomalies in this model it is necessary that the "predictive attributes" reflect the phenomenon on which anomalies are sought. Thus, these eighty attributes went through two changes: 1) using Gradient Boosted Machines (GBM), we identified eighteen attributes able to explain 80% of the variability of the volumes exported by the companies; 2) for the unsupervised learning model to effectively detect anomalies related to exports, these eighteen attributes were then relativized from the formula shown below in eighteen indices, which were then used to learn the unsupervised model.

The relativization of predictive attributes is responsible for creating indexes that effectively reflect the participation of the attributes in the phenomenon in which anomalies are sought: the amount exported. For example, given the exploratory attribute *financial transactions*, the relativization transforms this attribute in *amount exported by financial transactions unit*.

Thus, the formula below indicates that given i explanatory attributes x, $Index_{x_i}$ indicates the Amount of exports for the record of a company for each unit of $ExplanatoryAttribute_{x_i}$.

$$Index_{x_i} = \frac{ExportAmount_{registry}}{ExplanatoryAttribute_{x_i}}$$

6. Modeling and Evaluation

For Data modeling we used *0xdatas H_2O software*[3] connected to *R* by *H_2O R package* [23].

H_2O is a Java Virtual Machine that is optimized for doing "in memory" processing of distributed, parallel machine learning algorithms on clusters. In this research we used just one node with 3 CPUs and 6 GB of memory allocated to H_2O.

6.1. Comparing Models

H_2O offers an array of machine learning algorithms. Deep Learning AutoEncoder [24] (encoding stage in Figure 2) and linear principal component analysis (PCA) [25] are the options available for reducing dimensionality.

To detect anomalies, Deep Learning AutoEncoder can handle this task through its decode stage (see details in Section 6.2) and, likewise, the results of dimensionality reduction obtained using the PCA method can be decoded in a deep network using only the decode stage. Thus, differences will be observed only in the coding phase. We investigated the performance differences for dimensionality reduction between the two models proposed.

In both models the same dimensionality reductions were applied. All records corresponding to companies operating directly or indirectly in exports in the Brazilian market in 2014 were processed. The processing time using AutoEncoder was substantially lower, about 20 times faster.

These results are supported by Sakurada and Yairi [26]: PCA is computationally more expensive than AutoEncoder because it "basically requires to hold all the training samples". These authors demonstrates that AutoEncoders detect subtle anomalies which PCA fails to and they "can detect anomalies even with relatively high latent dimensions while linear PCA can not".

3. A Open Source Software for data analysis, Apache 2.0 licensed, available in http://www.h2o.ai/

Another point in favor of AutoEncoders is its non-linear generalizability due to the presence of non-linear functions in both the encoder and in the decoder [24].

6.2. AutoEncoder

AutoEncoder proved to be the most appropriate method for anomaly detection task. It was much faster: PCA requires more computation power than AutoEncoder. PCA basically requires to hold all the training samples, which is also computationally expensive. AutoEncoders can detect subtle anomalies which linear PCA fails to detect and can avoid complex computation that PCA requires without degrading the quality of detecting performance.

According to Goodfellow et al. [24] AutoEncoders are neural networks that are trained to make copies of their entries in their outputs. Internally, they have a hidden layer h which is a *code* used to describe the input. These networks can be seen as consisting of two parts: An encoding function $h = f(x)$ and a decoding function $r = g(h)$ that produces the reconstruction. This architecture is presented in Figure 2. However AutoEncoders should not learn to copy perfectly, otherwise they would be useless. Restrictions in the inner layers (hidden layers) network allow such copying is only an approximation. This ultimately forces the AutoEncoder network to prioritize the most important aspects to make the copy. Thus, most often it learns the most useful properties of the data.

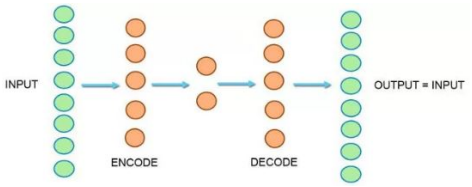

Figure 2. Layers in a AutoEncoder network

Anomaly detection using dimensionality reduction is based on the assumption that the data has variables correlated with each other and that can be embedded into a lower dimensional subspace in which normal samples and anomalous samples appear significantly different [25].

In this work, we use 18 neurons (predictive attributes) as input layer and the same 18 neurons (predictive attributes) as the output layer. The goal here is that the network learn to copy input data to the output.

As hidden layers we used one with 6 neurons, one with 3 neurons and one with 6 neurons. So, the middle layer is a 3-dimensional representation of an 18-dimensional input. The objective here was to force the network to gradually reduce the dimensionality of the input data into a format in 3 dimensions. This prevents the learning to perfectly copy the entry, as the network will have to deal with a learning process in a few dimensions. The choice of the hidden layer size with 6-3-6 was made after various tests and graphical analysis of the middle layer. It is possible (and probable) that other combinations of hidden layers would reach similar results.

Figure 3 shows a graphical representation of the middle layer. We separate by color the twenty most anomalous records, i.e. twenty records in which the network had more difficulty to create a copy. These records will be those that we consider more likely to be suspected of fraud. This graphical view also allows us to realize that the middle layer was able to create a linear separation of records, focusing on the right part of the graphics the vast majority of records (corresponding to the records where there is a pattern of behavior) and in the left, more dispersed, anomalous records considered suspects.

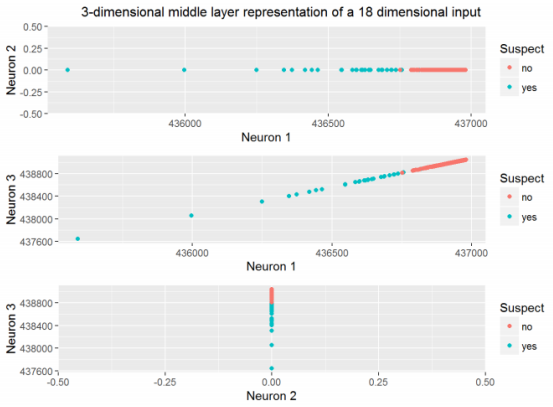

Figure 3. 18 dimensional input in 3-dimensional representation of middle layer

The adjustment of the amount of epochs[4] was done by trial and error. A very small number could greatly decrease the network sensitivity. A large number tends to overfitting. The epochs were adjusted to 50 and the activation function used was ReLU ("Rectifier" in H_2O).

All other parameters were left at default values (per-weight adaptive learning rate, no L1/L2 regularization, no Dropout). Attempted settings of these parameters, despite having effects on the ability to learn to copy the data and thus influence the value of errors when comparing the input and output of the network, did not change the order of found anomalous records. Thus, we opted for the simplest model, namely the maintenance of defaults parameters.

6.2.1. Performance Analysis. We proceeded tests to verify the performance gains using different amounts of processors. These tests are intended to serve as reference of computational power needed for future works which involve the same database, but with greater granularity.

We conducted performance tests (on one cluster) varying the amount of processors to anomaly detection task with AutoEncoder. Four tests were conducted in a *Linux Ubuntu 16.04 LTS*: 1, 2, 3 and 4 allocated processors and 12GB of ram memory. Was used a *Intel Core i5-3317U CPU @ 1.70GHz* $\times 4$. Table 2 shows the results obtained.

4. Number of epochs represents "how many times the dataset should be iterated (streamed)" [23]

TABLE 2. PERFORMANCE TESTS - VARYING THE AMOUNT OF PROCESSORS TO ANOMALY DETECTION TASK WITH AUTOENCODER

Number of allocated processors	Performance in milliseconds
1	54785
2	50512
3	49211
4	49001

6.3. Evaluation

Once the model was trained we used the mean squared error (MSE) as a measure of how distant our predictions were from the real data. MSE measures the average of the squares of the errors, that is, the difference between the estimator and what is estimated. In this case, consider x_i the value of n neurons in the input layer and $\widehat{x_i}$ the value of n neurons in the output layer. The MSE value for each record containing n attributes of an exporting company is given by the formula below:

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(x_i - \widehat{x}_i)^2$$

The higher the MSE value, the more anomalous, in relation to the pattern found in the data, a particular record is.

The MSE values are placed in ascending order and the distribution of the 170 highest values shown in Figure 4 indicate a clear change in behavior around the 20 last records.

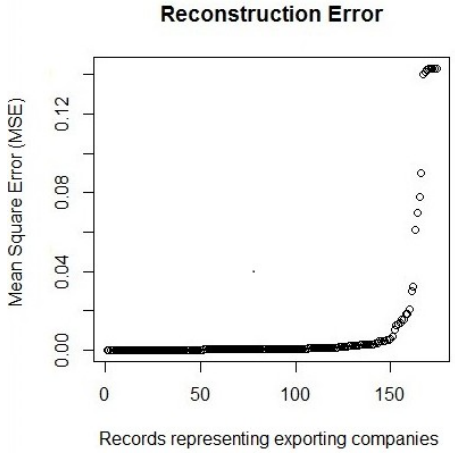

Figure 4. The one hundred and seventy largest MSE values.

In order to carry out the evaluation of the records relating to major anomalies found, the attributes of the fifty companies that presented the highest MSE were presented to third party experts in exports fraud. Preliminarily, they considered the system as efficient, since it identified some fraud cases already known by the experts. The remaining cases will be evaluated for a conclusive opinion on the effectiveness of the model.

7. Conclusion and Future Works

This paper presented an unsupervised model for detecting fraud suspects in exports. Using the *0xdatas H_2O software* connected to *R* by *H_2O R package*, the performance of two-dimensionality reduction models were evaluated under the same conditions. The tests showed a performance to reduce dimensionalities about 20 times faster using Deep Learning AutoEncoder compared with PCA. The choice of AutoEncoder algorithm is supported by previous studies that indicate the detection of anomalies is more accurate and have a better power nonlinear generalization. *0xdatas H_2O software* provides other methods of analysis unsupervised but with linear linear approach. These methods can be tested in the future for comparison with this work.

The greatest difficulty in the use of unsupervised techniques is the evaluation of the results against the business objectives to be achieved. The evaluation of third party experts is subjective and therefore can be devoid of factors perceived by the data mining algorithm. In this work, the selection of suspected cases of fraudulent exports through unsupervised Deep Learning proved to be preliminarily promising, but a more thorough assessment should be made by experts. The in-depth investigation of cases identified is not trivial and takes time. Their conclusions will be disclosed in due course. Depending on the results, adjustments in the number of hidden layers and the number of neurons may prove necessary and lead to better results. Similarly, the decrease in the number of epochs may reduce a possible overfitting that has allowed even records with indications of fraud to have a low value of MSE.

Acknowledgments

The authors would like to thank the tax auditors Leon Solon da Silva, Marcelo Renato Lingerfelt and Nildomar Jose Medeiros for their help and support in making this work possible.

References

[1] Grupo de Egmont, *100 Casos de Lavagem de Dinheiro*. COAF, 2001. [Online]. Available: http://www.coaf.fazenda.gov.br/menu/pld-ft/publicacoes/100_Casos.pdf

[2] Conselho de Controle de Atividades Financeiras, *Casos e Casos - I Coletnea de Casos Brasileiros de Lavagem de Dinheiro*. COAF, 2011. [Online]. Available: www.coaf.fazenda.gov.br

[3] P. He, "A typological study on money laundering," *Journal of Money Laundering Control*, vol. 13, no. 1, pp. 15–32, Jan. 2010. [Online]. Available: http://www-emeraldinsight-com.ez54.periodicos.capes.gov.br/doi/full/10.1108/13685201011010182

[4] J. Madinger, *Money Laundering: A Guide for Criminal Investigators, Third Edition*. CRC Press, Dec. 2011.

[5] O. Greene, "Trade-Based Money Laundering," Jul. 2015, acesso em: 01/05/2016. [Online]. Available: https://www.dhgllp.com/Portals/4/ResourceMedia/publications/Risk-Advisory-Trade-Based-Money-Laundering.pdf

[6] Brasil, "Lei 9613, de 03 de maro de 1998." [Online]. Available: http://www.planalto.gov.br/ccivil_03/LEIS/L9613.htm

[7] T. E. Senator, H. G. Goldberg, J. Wooton, M. A. Cottini, A. F. Umar Khan, C. D. Klinger, W. M. Llamas, M. P. Marrone, and R. W. H. Wong, "The financial crimes enforcement network AI system (FAIS) : identifying potential money laundering from reports of large cash transactions," *The AI magazine*, vol. 16, no. 4, pp. 21–39, 1995. [Online]. Available: http://cat.inist.fr/?aModele=afficheN&cpsidt=2985240

[8] H. G. Goldberg and T. E. Senator, "Restructuring Databases for Knowledge Discovery by Consolidation and Link Formation," in *Proceedings of the First International Conference on Knowledge Discovery and Data Mining*. AAAI Press, 1995, pp. 136–141.

[9] A. S. Larik and S. Haider, "Clustering based Anomalous Transaction Reporting," *Procedia Computer Science*, vol. 3, pp. 606–610, 2011. [Online]. Available: http://linkinghub.elsevier.com/retrieve/pii/S187705091000476X

[10] N. S. Khan, A. S. Larik, Q. Rajput, and S. Haider, "A Bayesian Approach for Suspicious Financial Activity Reporting," *International Journal of Computers and Applications*, vol. 35, no. 4, pp. 181–187, Jan. 2013. [Online]. Available: http://www.tandfonline.com/doi/abs/10.2316/Journal.202.2013.4.202-3864

[11] S. Raza and S. Haider, "Suspicious activity reporting using dynamic bayesian networks," *Procedia Computer Science*, vol. 3, pp. 987–991, 2011. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S1877050910005375

[12] Q. Rajput, N. S. Khan, A. Larik, and S. Haider, "Ontology Based Expert-System for Suspicious Transactions Detection," *Computer and Information Science*, vol. 7, no. 1, Jan. 2014. [Online]. Available: http://www.ccsenet.org/journal/index.php/cis/article/view/30883

[13] A. Sharma and P. K. Panigrahi, "A Review of Financial Accounting Fraud Detection based on Data Mining Techniques," *International Journal of Computer Applications*, vol. 39, no. 1, pp. 37–47, Feb. 2012, arXiv: 1309.3944. [Online]. Available: http://arxiv.org/abs/1309.3944

[14] J. J. Filho, "Tratamento Bayesiano de Interaes entre atributos de Alta Cardinalidade," Ph.D. dissertation, Unicamp, Sep. 2007. [Online]. Available: http://www.bibliotecadigital.unicamp.br/document/?code=vtls000426153&print=y

[15] J. J. Filho and J. Wainer, "Using a Hierarchical Bayesian Model to Handle High Cardinality Attributes with Relevant Interactions in a Classification Problem," in *Proceedings of the 20th International Joint Conference on Artifical Intelligence*, ser. IJCAI'07. San Francisco, CA, USA: Morgan Kaufmann Publishers Inc., 2007, pp. 2504–2509. [Online]. Available: http://dl.acm.org/citation.cfm?id=1625275.1625679

[16] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich, "Going Deeper With Convolutions," 2015, pp. 1–9. [Online]. Available: http://www.cv-foundation.org/openaccess/content_cvpr_2015/html/Szegedy_Going_Deeper_With_2015_CVPR_paper.html

[17] S. Jaiswal and M. F. Valstar, "Deep learning the dynamic appearance and shape of facial action units," Lake Placid, USA, 2016. [Online]. Available: http://eprints.nottingham.ac.uk/31301/

[18] Y. Liang, M. C. Machado, E. Talvitie, and M. Bowling, "State of the Art Control of Atari Games Using Shallow Reinforcement Learning," *arXiv:1512.01563 [cs]*, Dec. 2015, arXiv: 1512.01563. [Online]. Available: http://arxiv.org/abs/1512.01563

[19] Y. LeCun, Y. Bengio, and G. Hinton, "Deep learning," *Nature*, vol. 521, no. 7553, pp. 436–444, May 2015. [Online]. Available: http://www.nature.com/doifinder/10.1038/nature14539

[20] D. P. Kingma, S. Mohamed, D. Jimenez Rezende, and M. Welling, "Semi-supervised Learning with Deep Generative Models," in *Advances in Neural Information Processing Systems 27*, Z. Ghahramani, M. Welling, C. Cortes, N. D. Lawrence, and K. Q. Weinberger, Eds. Curran Associates, Inc., 2014, pp. 3581–3589. [Online]. Available: http://papers.nips.cc/paper/5352-semi-supervised-learning-with-deep-generative-models.pdf

[21] P. Chapman, J. Clinton, R. Kerber, T. Khabaza, T. Reinartz, C. Shearer, and R. Wirth, *CRISP-DM 1.0 Step-by-step data mining guide*. IBM, Aug. 2000. [Online]. Available: ftp://ftp.software.ibm.com/software/analytics/spss/support/Modeler/Documentation/14/UserManual/CRISP-DM.pdf

[22] Receita Federal do Brasil, "Portaria RFB n 671, de 07 de fevereiro de 2014."

[23] S. Aiello, T. K. a. P. Maj, and w. c. f. t. H. a. team, "h2o: R Interface for H2o," Jun. 2016. [Online]. Available: https://cran.r-project.org/web/packages/h2o/index.html

[24] I. Goodfellow, Y. Bengio, and A. Courville, "Deep Learning," 2016, book in preparation for MIT Press. [Online]. Available: http://www.deeplearningbook.org

[25] V. Chandola, A. Banerjee, and V. Kumar, "Anomaly Detection: A Survey," *ACM Comput. Surv.*, vol. 41, no. 3, pp. 15:1–15:58, Jul. 2009. [Online]. Available: http://doi.acm.org/10.1145/1541880.1541882

[26] M. Sakurada and T. Yairi, "Anomaly Detection Using Autoencoders with Nonlinear Dimensionality Reduction," in *Proceedings of the MLSDA 2014 2Nd Workshop on Machine Learning for Sensory Data Analysis*, ser. MLSDA'14. New York, NY, USA: ACM, 2014, pp. 4:4–4:11. [Online]. Available: http://doi.acm.org/10.1145/2689746.2689747