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

Modelos Preditivos para Seleção de Solicitações de Compensação de Crédito Tributário

Leon Sólon da Silva

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

Orientador
Prof. Dr. João Carlos Felix Souza

Co-orientador
Prof. Dr. Rommel Novaes Carvalho

Brasília
2016

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

da Silva, Leon Sólon
dm Modelos Preditivos para Seleção de Solicitações de Compensação de Crédito Tributário / Leon Sólon da Silva; orientador João Carlos Felix Souza; co orientador Rommel Novaes Carvalho. -- Brasília, 2016.
72 p.

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

1. Mineração de Dados. 2. Modelos Preditivos. 3. Compensação de Crédito Tributário. I. Felix Souza, João Carlos, orient. II. Novaes Carvalho, Rommel, co orient. III. Título.

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

Modelos Preditivos para Seleção de Solicitações de Compensação de Crédito Tributário

Leon Sólon da Silva

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

Prof. Dr. João Carlos Felix Souza (Orientador)
EPR/UnB

Prof. Dr. Gladston Luiz da Silva
EST/UnB

Prof. Dr. Remis Balaniuk
Universidade Católica de Brasília

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

Brasília, 27 de julho de 2016

Dedicatória

Dedico esse trabalho à minha família: pais, sogros, esposa e filhos. Minha esposa por sempre ter me incentivado a buscar cada vez mais. Meus pais por serem meus exemplos de superação e por terem tornado minha vida tão diferente da que eles tiveram.

Agradecimentos

Agradeço aos meus orientador e coorientador, João Carlos Félix Souza e Rommel Novaes Carvalho por me motivarem a trabalhar em algo que se tornou minha principal paixão profissional. Ao meu ex-orientador Mauricio Ayala Rincón por ter me ensinado a não me contentar somente com o bom. Ao Auditor-Fiscal e amigo Márcio Vital Santos de Araújo, especialista em compensação de crédito, pelas enormes contribuições no trabalho. Aos meus chefes durante esse período de dois anos pelo apoio e paciência: Suely Nunes Braga, Aparecido Xavier de França e Gérson D'agord Schaan.

Resumo

Um dos principais objetivos das administrações tributárias é promover a justiça fiscal e uma forma de alcançá-la é selecionar corretamente os contribuintes para fiscalização de forma a focalizar naqueles que apresentam maior risco de não cumprir suas obrigações tributárias. Com a tendência global de redução de custos, recursos e quadro de profissionais, em contraste com o aumento de contribuintes, volume de tributos e processos a serem analisados, é primordial que as administrações tributárias trabalhem de forma mais eficiente. Nesse contexto, a gestão de riscos é uma ferramenta importante para melhorar a alocação de recursos e aumentar a efetividade na seleção dos contribuintes que realmente devem ser fiscalizados. O presente trabalho avalia o processo atual de seleção de solicitações de compensação de crédito tributário e propõe melhorias na sua gestão de riscos. Mais especificamente, propomos melhorar a escolha de solicitações de compensação de crédito tributário a serem analisadas por Auditores-Fiscais da Receita Federal do Brasil a partir da utilização de técnicas de análise e mineração de dados. Utilizando essas técnicas, criamos modelos preditivos para, com base no histórico de análise manual dos Auditores-Fiscais, tentar predizer os riscos de uma solicitação de compensação de crédito ser ou não devida. As simulações da aplicação dos modelos preditivos apresentaram resultados promissores e nos leva a crer que terá performance melhor que o processo de trabalho atual. A Secretaria da Receita Federal do Brasil abarcou o projeto, que está em fase de implementação para o uso dos modelos preditivos na seleção das solicitações de compensação de crédito de todas as Regiões Fiscais do Brasil.

Palavras-chave: mineração de dados, modelos preditivos, compensação de crédito tributário

Abstract

One of the main goals of every tax administration is safeguarding tax justice. For that matter, accurate taxpayers' auditing selection plays an important role. Current scenario of economic recession, budget cuts and tax professionals' hiring difficulty combined with growth of both population and number of enterprises presents the necessity of a more efficiently approach from tax administration in order to meet its objectives. The present work intends to show how data mining techniques usage helps better understand the profile of non compliant tax payers who claim for tax compensation (kind of tax refund). Basically, we used knowledge discovery in databases (KDD) from previous tax refund claims that were manually analyzed by tax officers and create predictive models in order to classify not yet audited claims. Moreover, we present results on the adoption of these predictive models towards selection improvement of those who claims that are more likely to be rejected in Secretariat of Federal Revenue of Brazil (RFB). The results show that this approach is an efficient way for selecting tax payers rather than not using it.

Keywords: data mining, predictive models, tax refund

Sumário

Lista de Figuras

Lista de Tabelas

Capítulo 1

Introdução

Os cidadãos de qualquer nação necessitam de alguma forma de bens e serviços providos pelo estado. Muitas teorias apontam que entidades privadas são mais eficientes que as públicas, mas podem levar a distorções e injustiças. Em [25], Giambiagi apresenta o conceito de bens públicos em contrapartida à conhecida teoria do bem estar social. A última prega que sempre haverá distorções na sociedade, em consequência dos mercados competitivos, de modo que sempre que houver benefício de um cidadão haverá uma contrapartida negativa para outro. A teoria dos bens públicos refuta esse pensamento focalizado no fornecimento de serviços públicos pelo estado. A partir dessa teoria, o fato de o estado prover um bem ou serviço a um cidadão não implica necessariamente no prejuízo de outros.

Tomando por base a teoria dos bens públicos, as políticas públicas são essenciais para o bem estar de um sociedade. Para que eles sejam bem executados, por sua vez, as nações necessitam de recursos financeiros para atender seus cidadãos por meio de políticas públicas. Um estado não prospera sem receitas provenientes de tributos pagos pelos contribuintes.

Nesse contexto se enquadram órgãos típicos de estados que realizam a coleta de tributos: as administrações tributárias. Mais que executar puramente as atividades arrecadatórias, os órgãos coletores devem atender aos objetivos de estado com eficiência, mas sempre levando em conta a justiça fiscal. No Brasil a justiça fiscal é bem representada por dois dos princípios que regem o direito tributário: da isonomia e da capacidade contributiva. O primeiro se baseia no art. 5º da Constituição Federal que prega que todos são iguais perante a lei, sem distinção de qualquer natureza [14], enquanto o segundo determina que a tributação seja adequada à capacidade de pagar de cada contribuinte. Quem tem maior capacidade contributiva paga mais, quem tem menor capacidade contributiva paga menos, e quem não tem capacidade contributiva não paga [3]. Assim, para que haja verdadeira justiça fiscal, os princípios devem ser observados para todos os processos de

trabalho das administrações tributárias brasileiras (união, estados e municípios).

Os tributos federais brasileiros são administrados pela Secretaria da Receita Federal do Brasil (RFB), que acumula as competências de administração tributária e aduaneira. A RFB é responsável por processos de trabalho que tem grande impacto no provimento de recursos ao estado. Dentre as atividades da RFB, muitos dos macroprocessos tem maior nível de maturidade e, por isso, possuem maior justiça fiscal em sua execução, pois passaram por diversas melhorias e utilizam as mais variadas técnicas e tecnologias para contemplar os objetivos do estado. Processos mais atuais, no entanto, que não sofreram muitas alterações e escrutínio de anos sob prova, e possuem mais espaço para serem alterados com o fim de melhorar a efetividade. Os processos da RFB que chegam aos noticiários e chamam mais atenção da sociedade são a arrecadação e fiscalização (tributos internos e aduana), mas outros tem características que podem ser melhoradas para garantir a justiça fiscal.

O foco dessa pesquisa é o processo de gestão do crédito tributário, primordial para se garantir, além de justiça fiscal, que a arrecadação não seja impactada por utilização dos créditos de forma indevida pelos contribuintes, conforme será apresentado em Capítulo próprio (2).

Antes do advento de técnicas estatísticas e ferramentas de tecnologia da informação as seleções de contribuintes para fiscalização e as análises de processos de compensação de crédito eram realizadas de forma aleatória e empírica na maioria dos casos. O fim da década de 1990 e início de 2000, no entanto, trouxe um aumento expressivo no número de contribuintes e de informações prestadas por eles às administrações tributárias (tax returns) e as administrações tiveram de reestruturar seus processos de trabalho.

No Brasil, o principal tributo sobre a renda de pessoas físicas passou a ser eletrônico desde 1991, com expressivo aumento de declarações deste tipo em 1996 chegando em 2014 com mais de 30 milhões de declarações [56]. Tendência semelhante é observada para as pessoas jurídicas, com um grande crescimento no número de empresas nas últimas décadas, chegando a mais de 16 milhões em 2013 [30]. Além do aumento de informações, com as restrições orçamentárias geradas pelas crises financeiras e pela racionalização dos gastos do governo, houve uma grande necessidade de tornar a seleção de contribuintes e de processos analisados mais eficiente.

Mais especificamente, como apresentado em detalhes mais adiante, a compensação de crédito tributário, instrumentalizada pelo Pedido Eletrônico de Restituição ou Ressarcimento e da Declaração de Compensação (PER/DCOMP), criado no ano de 2003, possui espaço para melhorar sua efetividade e promover uma melhor justiça fiscal. O processo atual conta com uma boa gestão de riscos semi-automatizada que garante boa parte da análise, mas, devido às restrições apresentadas anteriormente, a quantidade de solicitações

de compensação que necessitam de atuação de análise manual dos Auditores-Fiscais da Receita Federal do Brasil na análise do crédito pleiteado pelo contribuinte tem aumentado sobremaneira.

Nesse sentido, o presente trabalho analisa o macroprocesso de gestão do crédito tributário e como ele pode ser melhorado a partir da utilização de técnicas de análise e mineração de dados. Mais especificamente, são propostas alterações no processo de seleção de solicitações de compensação de crédito que são analisadas de forma manual pelos Auditores-Fiscais da Receita Federal do Brasil. Foram realizadas análises de variáveis e características dos contribuintes utilizando ferramentas estatísticas para definir quais mais impactam na decisão de indeferir (não homologar) uma solicitação de compensação de crédito tributário. Uma vez com as características determinadas, modelos preditivos foram criados usando diferentes algoritmos. Ademais, como forma de avaliar se as melhorias propostas são realmente factíveis, foram realizadas simulações de seleção baseada em modelos preditivos e comparados os resultados utilizando um indicador de performance do processo de trabalho. A seguir é apresentado em mais detalhes o problema a ser endereçado.

1.1 Definição do Problema

O processo de compensação de crédito será detalhado no Capítulo 2, mas adiantamos alguns conceitos para definir o problema em estudo. A compensação de crédito foi criada para facilitar o cumprimento de obrigações tributárias por parte do contribuinte, no sentido de se utilizar um conceito básico de contabilidade de anulação de um débito por um respectivo crédito. Nesse sentido, a RFB permite que os contribuintes utilizem um crédito que possuam junto à administração tributária com um débito que tenham junto à mesma.

Um exemplo simples seria uma empresa (contribuinte pessoa jurídica) que esteja devendo imposto sobre produtos industrializados (IPI) e possui um crédito junto a RFB de imposto de renda (IR). O contribuinte pode utilizar o crédito de IR para compensar o débito de IPI e extingui-lo junto à fazenda (1.1).

Muitas vezes para o contribuinte é mais interessante a realização da compensação em detrimento à restituição ou ressarcimento, tendo em vista que o primeiro possui um trâmite menos complicado e é liberado mais rapidamente. Diferente da restituição, a extinção do débito relativo à compensação ocorre na entrega da declaração, cabendo à RFB homologar essa compensação a posteriori. Quando se decorre um prazo de 5 (cinco) anos, a compensação é homologada (deferida/aceita) tacitamente, ou seja, mesmo que o crédito pleiteado pelo contribuinte não seja analisada por Auditores-Fiscais, se

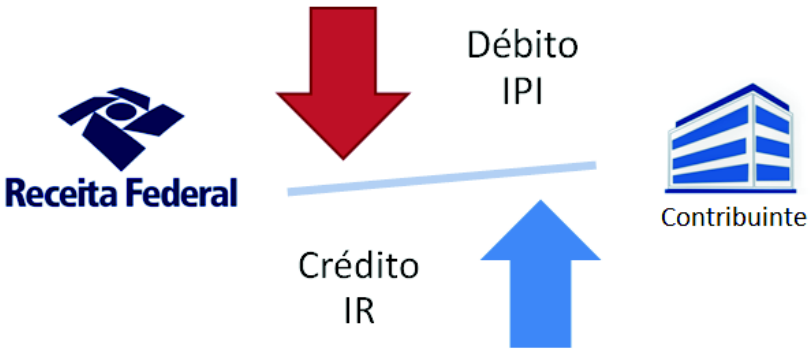

Figura 1.1: Exemplo de Compensação de Crédito

expirado o prazo, extingue-se o débito. Esse fato aumenta bastante o risco identificado para o processo de trabalho, tendo em vista que, selecionar bem quais solicitações de compensação de crédito devem ser trabalhadas, é essencial para evitar o impacto negativo na arrecadação líquida.

Dessa forma, o contribuinte deve solicitar um pedido de compensação utilizando o programa "Pedido Eletrônico de Restituição, Ressarcimento ou Reembolso e Declaração de Compensação"(PER/DCOMP), com preenchimento semelhante à Figura 1.2. Esses pedidos são recebidos e analisados pelo Sistema de Controle de Crédito (SCC), um aplicativo desenvolvido pelo Serviço Federal de Processamento de Dados (Serpro) que realiza uma série de verificações de riscos para definir o tratamento correto das solicitações. Auditores-Fiscais especialistas em compensação de crédito, reunindo grande experiência e conhecimento em fraudes, identificam indicadores de riscos a serem incluídos na análise do Sistema de Controle de Crédito.

O SCC executa rotinas que permitem agregar os pedidos e documentos relacionados a um mesmo crédito, analisar os documentos retificadores e os pedidos de cancelamento. Por fim, com base em análises e em alguns parâmetros de risco de indeferimento (não acolhimento do crédito solicitado), defere, indefere ou classifica o processo de solicitação para análise do Auditor-Fiscal. Caso o sistema identifique que o documento apresenta consistência dos dados e sendo suficientes as informações, a análise do crédito é concluída por procedimentos eletrônicos.

Entretanto, na análise eletrônica do crédito, podem ser identificadas situações de risco ou identificada necessidade de intervenção por servidor da Secretaria da Receita Federal do Brasil (RFB) em interface do sistema.

Além disso, atendendo a critérios de avaliação de risco, interesse e relevância dos documentos, o PER/DCOMP pode ser indicado para tratamento manual e decidido o aprofundamento da análise.

Com o crescimento do número de solicitações de compensação, o estoque de processos

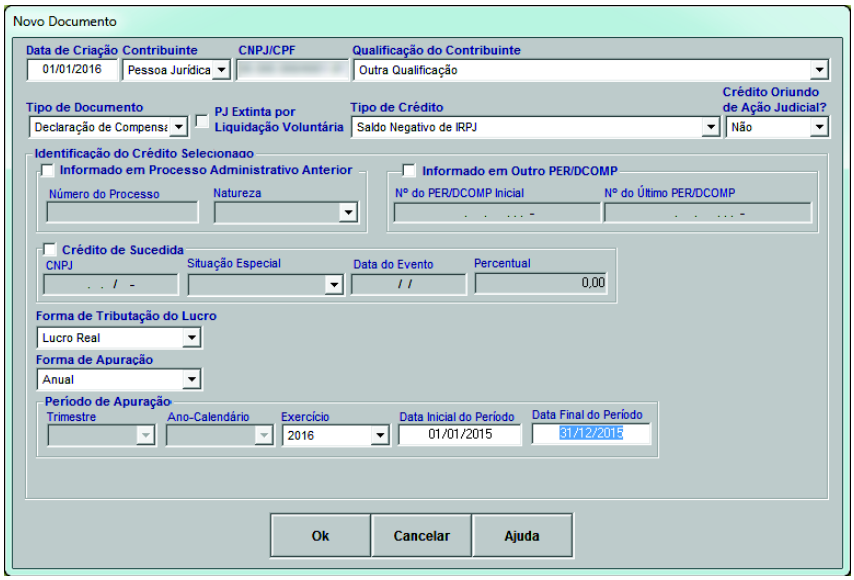

Figura 1.2: Programa PER/DCOMP

que são classificados como necessidade de intervenção do usuário e análise manual também se elevou sobremaneira. Muitas unidades não têm condições de analisar amiúde todos os processos e podem fazer escolhas que não são as mais indicadas na seleção de quais solicitações devem ser trabalhadas.O impacto das compensações na arrecadação de crédito é considerável e sua gestão é de essencial importância para que a RFB alcance seus objetivos. A importância do tema e a justificativa de sua escolha são apresentadas em mais detalhes a seguir.

1.2 Justificativa do Tema

As administrações tributárias são normalmente conhecidas por arrecadar e, por isso, o processo de arrecadação é que recebe um maior enfoque da sociedade e dos contribuintes. A compensação do crédito tributário, no entanto, resulta num grande impacto na arrecadação líquida, uma vez que os débitos que os contribuintes possuem com a RFB, ou seja, aquilo que um dia seria arrecadado, será anulado por uma compensação de um crédito que o contribuinte possui junto à administração tributária. Quando esse cancelamento (deferimento da solicitação de compensação) é realmente devido, é uma boa maneira de se relacionar com o contribuinte, mas quando isso não ocorre, há um risco de se perder arrecadação líquida anulando um débito que não deveria ser compensado.

Em períodos de ajuste fiscal e readequação da economia brasileira, selecionar bem as solicitações de compensação de crédito tributário é essencial para manter a justiça fiscal e garantir a receita que o país precisa para implementar as políticas públicas. Ademais, valores de compensação crescem muito a cada ano, conforme apresentado em 1.3, e analisar

corretamente as solicitações é vital para a boa saúde da arrecadação federal. Conforme pode se observar, os gráficos apresentam a quantidade de solicitações e o valor que elas representam desde 2003, quando a compensação eletrônica foi implementada, até 2014 [1].

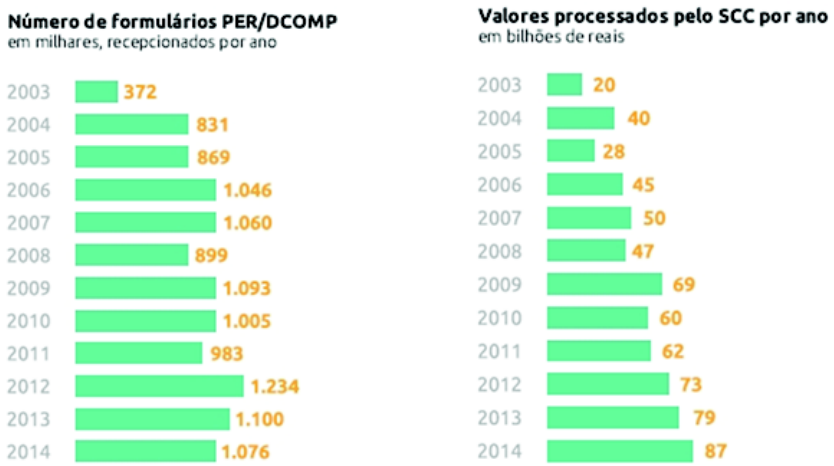

Figura 1.3: Volume Anual de Solicitações de Compensação de Crédito

Desse modo, resta claro que o tema é de grande interesse para a Secretaria da Receita Federal do Brasil, bem como para a sociedade brasileira que tem nas políticas públicas a base para que consolide uma democracia sólida e sustentável. Além do ganho salientado, as contribuições tecnológicas esperadas são apresentadas no próximo Subcapítulo.

1.3 Contribuição Tecnológica Esperada

Com o presente trabalho espera-se apresentar a aplicabilidade de modelos preditivos em problemas relacionados a compensação de crédito tributário de modo a permitir que outras administrações tributárias possam melhorar seus processos de seleção utilizando ferramentas de mineração de dados.

1.4 Estrutura do Documento

O documento está organizado como se segue:

- O Capítulo 2 apresenta a Secretaria da Receita Federal do Brasil, a administração tributária e aduaneira brasileira. O Capítulo traz ainda a explanação dos conceitos de compensação de crédito tributário e o problema da solicitação de compensações de crédito.

- O Capítulo 3 apresenta a fundamentação teórica relacionada ao problema alvo da dissertação, a metodologia de pesquisa utilizada, o modelo conceitual CRISP-DM, as descrições teóricas de algoritmos e técnicas de mineração de dados utilizados na solução proposta e os trabalhos correlatos à presente pesquisa;
- O Capítulo 4 apresenta a solução proposta, perpassando todas as fases da metodologia CRISP-DM, desde o entendimento do negócio até a fase de implantação. O Capítulo traz ainda os resultados obtidos na aplicação de técnicas de mineração de dados para construção de modelos preditivos e o impacto simulado na seleção de solicitações de compensação de crédito com base em riscos de indeferimento, bem como o plano para implementação na Secretaria da Receita Federal do Brasil (RFB);
- O Capítulo 5 conclui a dissertação e traz possíveis trabalhos futuros.

Capítulo 2

A Secretaria da Receita Federal do Brasil

As administrações tributárias e aduaneiras são órgãos de estado com características próprias. No caso brasileiro, as duas administrações se confundem num mesmo órgão de estado: a Secretaria da Receita Federal do Brasil (RFB). Este Capítulo apresenta alguns detalhes necessários para a compreensão do problema foco do trabalho.

Inicialmente é detalhado o papel da Secretaria da Receita Federal do Brasil (RFB), um detalhamento do processo de trabalho que se pretende melhorar utilizando técnicas de mineração de dados. Em seguida é apresentada uma análise de contexto para compreender onde se encaixa a administração tributária brasileira no âmbito interno e externo.

Por fim, o Capítulo traz a definição da compensação de crédito, desde sua criação até as legislações atuais que permitem a realização solicitações de compensação de crédito tributário por meio eletrônico.

2.1 Estrutura

A Secretaria da Receita Federal do Brasil (RFB) faz parte do Ministério da Fazenda e conta com 10.399 (dez mil trezentos e noventa e nove) Auditores-Fiscais [59] para realizar as atividades de administração tributária e aduana. Em muitos países, como França, Estados Unidos da América e Inglaterra as duas administrações não se confundem.

Cabe à RFB a administração de tributos internos e políticas fiscais: imposto de renda, imposto sobre produtos industrializados, contribuição social sobre o lucro líquido, etc. Como administração aduaneira a RFB deve facilitar e controlar o comércio exterior.

A missão da RFB é "Exercer a administração tributária e aduaneira com justiça fiscal e respeito ao cidadão, em benefício da sociedade."e a visão "Ser uma instituição de excelência em administração tributária e aduaneira, referência nacional e internacional.".

Os valores são "Respeito ao cidadão, integridade, lealdade com a Instituição, legalidade, profissionalismo e transparência". A missão, visão e objetivos estratégicos da RFB são apresentadas na Figura 2.1 [56].

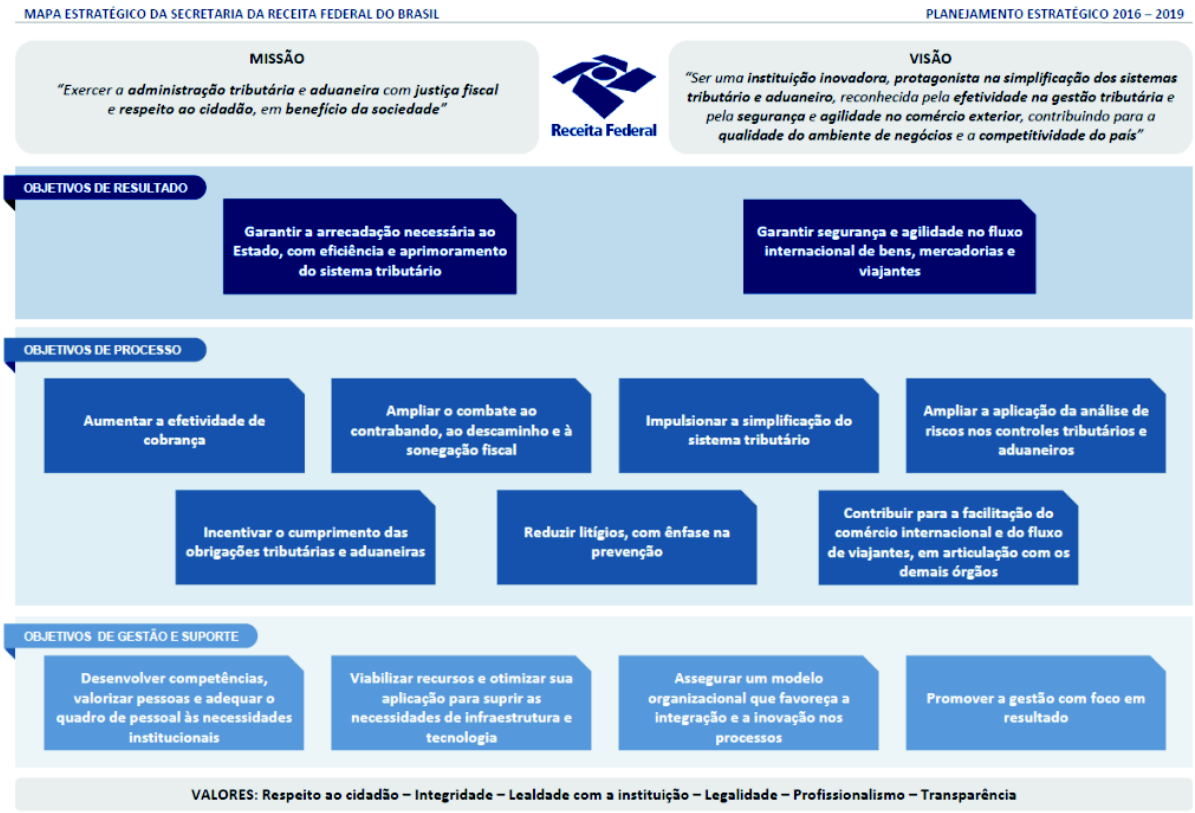

Figura 2.1: Mapa Estratégico RFB - 2016 a 2019

Os próximos Subcapítulos trazem um maior detalhamento do nível de maturidade da RFB em diferentes aspectos de gestão (contexto interno) e como o órgão se encaixa no panorama do estado brasileiro e as relações com outras organizações internacionais (contexto externo).

2.2 Contexto

Importante para se compreender o problema objeto do presente trabalho é avaliar o contexto em que ele se encontra. Nesta seção é apresentada uma breve descrição dos contextos interno e externo da Secretaria da Receita Federal do Brasil (RFB).

2.2.1 Contexto Interno

A gestão por processos teve um grande avanço nos últimos dois anos na organização. Muitos de seus processos finalísticos e de suporte foram completamente remodelados com o fim de se adequar para um modelo mais racional que permeia as unidades, em contrapartida a uma orientação por organogramas.

Um exemplo é a lotação e a capacitação de servidores, que antes era definida somente em termos de unidades da organização, mas hoje é realizada com foco em processos. Ademais, existe equipe especializada em realizar avaliações de melhorias na performance de processos. A cadeia de valor da RFB está em sua segunda versão e conta com os principais processos de trabalho modelados e disponibilizados para todos os servidores da casa, conforme Figuras 2.2 e 2.3.

Destaca-se o macroprocesso de Gestão do Crédito Tributário, mais especificamente o processo Gerir o Direito Creditório de Contribuinte que será apresentado em mais detalhes em seção própria.

A Gestão de Projetos é um dos processos de gestão com maior nível de maturidade na casa, contando com metodologia própria, fortemente baseada no PMBoK [31], desde 2008. Gestores escolhem o portifólio com base na criticidade e importância dos projetos e programas. Projetos classificados como estratégicos institucionais, nacionais e regionais.

A gestão estratégica na RFB teve seu primeiro ciclo quadri-anual finalizado em 2015. Objetivos de 2016-2019 e indicadores foram definidos com grande participação tanto de gestores de todos os níveis (estratégicos, táticos e operacionais) quanto por servidores. Há um grande esforço na comunicação institucional do mapa estratégico, bem como o acompanhamento de indicadores, que é realizado trimestralmente pela cúpula estratégica. Conforme recomendação obrigatória do Tribunal de Contas da União (TCU), a RFB confecciona e segue os Planos Diretores de Tecnologia da Informação (PDTI) alinhados aos objetivos estratégicos e suas ações, que possuem orçamento reservado de forma apartada de projetos menos relevantes.

Os objetivos do ciclo atual mais relacionados à gestão do crédito tributário, foco de análise desta pesquisa, são "Aumentar a percepção de equidade na atuação da instituição","Aumentar a efetividade dos mecanismos de garantia do crédito tributário"e "Elevar a percepção de risco e a presença fiscal". Para o ciclo dos próximos quatro anos um objetivo estratégico terá grande impacto no objeto do trabalho em questão "Ampliar a avaliação de riscos nos processos tributários e aduaneiros"onde indicadores serão criados para mensurar o quão efetivos estão sendo os processos e como a gestão de riscos impacta na melhoria dos resultados dos processos finalísticos.

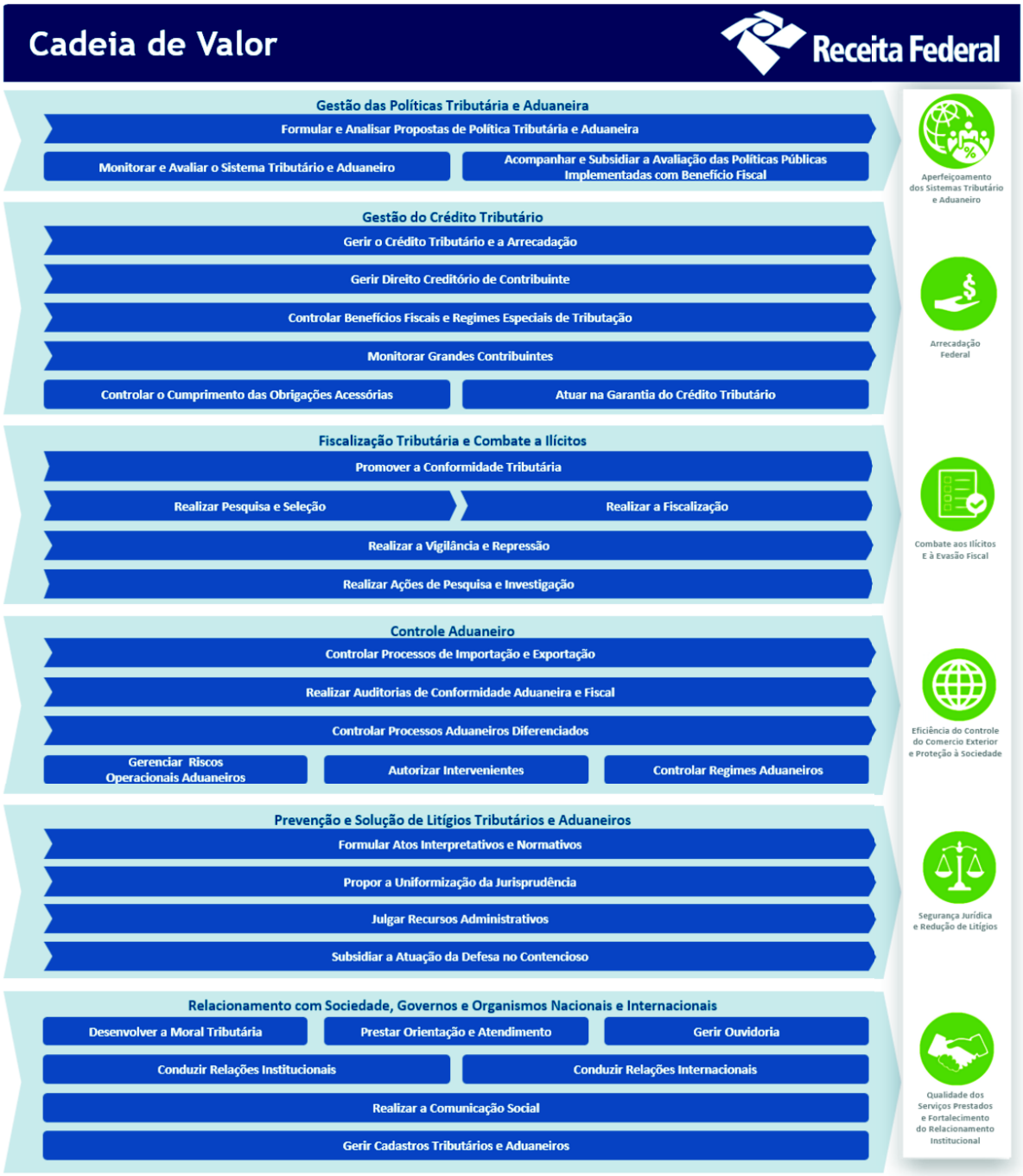

Figura 2.2: Cadeia de Valor da RFB - Processos Finalísticos

2.2.2 Contexto Externo

O entendimento do contexto externo é de extrema importância não somente para avaliar os benefícios das melhorias proporcionadas na gestão de riscos e as oportunidades a serem aproveitas, mas também para identificar quais atores serão afetados direta e indiretamente pelas alterações propostas (mudanças nos processos de trabalho a partir da avaliação de riscos). Desse modo, os próximos Subcapítulos tentam detalhar tal contexto para auxiliar

Figura 2.3: Cadeia de Valor da RFB - Processos de Suporte

nas etapas posteriores de definição de critérios e para a própria avaliação de riscos.

O trabalho de qualquer administração tributária esbarra em dificuldades culturais e sociais pelo fato de a arrecadação de tributos ser muitas vezes vista como algo pejorativo. Em frase do ex-Ministro da Fazenda, Joaquim Levy, ele afirma que "empresas não gostam de pagar impostos no Brasil"[27], mas é fato que não se resume a pessoas jurídicas. O problema se intensifica quando o país se encontra na fase atual de escândalos de corrupção, o que torna a compreensão da importância do provimento de receita para o estado mais distante da realidade. Isso prejudica o entendimento do fato de que qualquer país deve ter uma administração tributária competente e justa para prover as políticas públicas, infraestrutura, saúde, educação, segurança e outros aspectos essenciais para a vida dos cidadãos.

A legislação aplicada à Secretaria da Receita Federal é bastante ampla, contemplando desde a Constituição Federal de 1988 até acordos comerciais internacionais, passando por políticas fiscais elaboradas pelo Ministério da Fazenda. Destaca-se a Lei nº 5.172, de 25 de outubro de 1966 [13], denominado o Código Tributário Nacional, que traz as normas gerais de direito tributável aplicáveis à União, Estados e Municípios e a própria Constituição que traz muitas regras para o arcabouço jurídico tributário.

A maioria das instituições que fazem parte do governo brasileiro estão submetidos a diretrizes, políticas, auditorias e fiscalizações e a RFB não é exceção. Desse modo, além de se adequar a normas relacionadas a tributação e comércio exterior, a RFB deve atender às auditorias e recomendações dos órgãos competentes do executivo e do serviço público federal. Como parte do Ministério da Fazenda, a RFB é uma secretaria que está submetida às diretrizes e políticas fiscais do Ministério. Por ser parte da administração direta, está sob jurisdição do Ministério da Transparência, Fiscalização e Controle (antiga Controladoria-Geral da União), Tribunal de Contas da União (TCU) e pode sofrer auditorias dos dois órgãos.

Por ser uma administração aduaneira, a RFB está sujeita aos acordos internacionais

de comércio, pois muitos tem relação direta com a facilitação do comércio exterior, ou até no combate a ilícitos em conjunto com outros países. Guias de melhores práticas, manuais de processos de trabalho e outros documentos referenciais também são de grande importância para o contexto da gestão de riscos do crédito tributário. As principais fontes para tais recursos para a RFB são a Organização para a Cooperação Econômica e Desenvolvimento Econômico (OCDE) [44], a Organização Mundial de Aduanas (OMA) [45] e o Centro Interamericano de Administrações Tributárias (CIAT) [10].

Uma tendência dos últimos dois anos que pode impactar o atingimento dos objetivos e motivador para uma gestão de riscos ainda mais premente é a crise econômica que o Brasil atravessa. A RFB é demandada a obter mais recursos justamente em períodos onde é mais difícil para os contribuintes. O contexto econômico, no entanto, pode ser favorável no sentido de haver uma necessidade ainda maior de melhorar a gestão do crédito tributário e pode ser uma oportunidade para alterações no processo que não seriam avaliadas em outros momentos.

A principal parte interessada da RFB claramente é, mais que a União, o estado e os cidadãos brasileiros. Os tributos arrecadados pela RFB são a base financeira para desenvolvimento das políticas públicas e possibilitam receita para o país. Com relação aos processos aduaneiros, as empresas são importantes partes interessadas, pois são diretamente influenciadas e esperam um comércio exterior menos burocrático o possível para alavancar a comercialização de mercadorias e serviços com outros países.

2.3 Compensação de Crédito Tributário

A gestão do crédito tributário é um dos principais macroprocessos de qualquer administração tributária, pois nele estão os processos de arrecadação, cobrança administrativa e compensação de crédito. Esse macroprocesso (Figura 2.4) é responsável por mais de 1,2 trilhão de reais de arrecadação anual no Brasil [58]. Grande parte está relacionada diretamente à arrecadação espontânea, cobrança administrativa e à fiscalização. Porém, conforme apresentado na introdução (1), os valores de compensação de crédito impactam bastante na arrecadação líquida potencial e seu volume em reais é cada vez maior (87 bilhões em 2014).

Assim como as pessoas físicas, no Brasil, as empresas (pessoas jurídicas) possuem alguns direitos relativos ao crédito que possuem junto à administração tributária federal (Secretaria da Receita Federal do Brasil - RFB). As empresas podem solicitar a restituição do crédito tributário que possuem, o ressarcimento do crédito ou a compensação de débitos com créditos obtidos previamente.

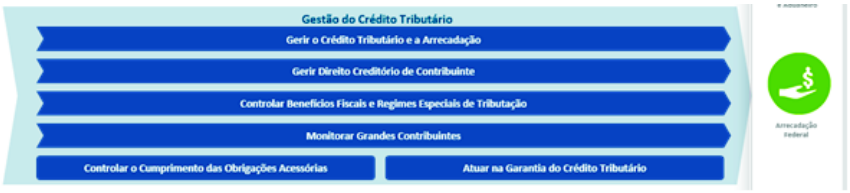

Figura 2.4: Macroprocesso Gerir o Crédito Tributário

A compensação por norma geral, está prevista no art. 156, inciso II da Lei nº 5.172, de 25 de outubro de 1966 [13], Código Tributário Nacional (CTN), e disciplinada pelo artigo 170, que determina as condições e garantias para o uso do referido direito. Destaca-se que o artigo 170 do CTN não autoriza diretamente a compensação, estando esse condicionado a lei que lhe autorize a aplicação.

Em 1991, a matéria foi tratada em legislação própria com a edição da Lei nº 8.383/91 [15], que autorizou o contribuinte ao pagamento de tributos mediante compensação com outros tributos federais, desde que da mesma espécie. Ocorreram alterações posteriores pela Lei nº. 9.430/96 [16], quando foi estendida a possibilidade de compensação com tributos de espécies distintas, e com a Lei nº. 10.637/02. A partir desses diplomas legais, houve a instrumentalização do instituto da compensação do indébito tributário.

Em relação aos pedidos de restituição, de ressarcimento e de compensação de tributos e contribuições observa-se que a RFB regulamentou, inicialmente, o mecanismo da compensação na Instrução Normativa SRF nº. 21, de 10 de março de 1997 [51], alterada pela IN SRF nº. 73, de 15 de setembro de 1997 [52], em que foi estabelecido que a compensação de tributos seria formalizada mediante pedido de compensação protocolados em processos administrativos no caso de tributos e contribuições de diferentes espécies. A compensação poderia ser feita na própria contabilidade do contribuinte desde que abrangesse débitos de mesma natureza e de períodos posteriores.

Contudo, foi editada a IN SRF nº. 210, de 30 de setembro de 2002 [53], que revogou a Instrução Normativa nº. 21/97 e dispôs que a compensação passaria obrigatoriamente a ser realizada por meio do protocolo de pedidos específicos, seja para compensação entre débitos de mesma natureza ou tributos de espécies distintas.

Posteriormente, a Instrução Normativa SRF nº. 320/2003 [54] aprovou a atual sistemática de compensação, em que, regra geral, o contribuinte ficou obrigado a efetuar a compensação, restituição ou ressarcimento de seus créditos por meio de uma declaração eletrônica denominada de PER/DCOMP.

Os contribuintes não podem se valer de todos os tipos de crédito para compensação. Além do pagamento a maior de tributos, por erro no preenchimento de (DARF), por exemplo, os principais tipos de crédito que podem ser compensados são [55]:

- Cofins embalagens (P.4, Art 51. Lei 10.833/03)
- Cofins não cumulativo Exportação
- Cofins não cumulativo Mercado interno
- Contribuição previdenciária indevida ou a maior
- IPI Residual
- IRRF de cooperativas
- IRRF de juros sobre capital próprio
- Pis/Pasep embalagens (P.4, Art 51. Lei 10.833/03)
- Pis/Pasep não cumulativo Exportação
- Pis/Pasep não cumulativo Mercado interno
- Reintegra
- Ressarcimento de IPI
- Retenção - Lei nº 9.711/98
- Salário-família/Salário-maternidade
- Saldo negativo de CSLL
- Saldo negativo de IRPJ

A atual metodologia trouxe mais facilidade, rapidez e seguranças aos contribuintes para exercerem seu direito de pleitear junto à Receita Federal do Brasil a restituição, de ressarcimento e de compensação de tributos e contribuições. Tais facilidades, no entanto, também aumentaram o risco de compensações indevidas deixarem de ser analisadas, trazendo uma maior responsabilidade na análise tempestiva das solicitações encaminhadas por meio eletrônico.

O próximo Capítulo traz a fundamentação teórica do presente trabalho, de modo a embasar os experimentos realizados e permitir uma melhor compreensão das técnicas utilizadas na criação dos modelos preditivos para melhoria da seleção de compensação de crédito tributário.

Capítulo 3

Fundamentação Teórica

3.1 Metodologia

Em um contexto de pesquisa científica e tecnológica, valem os preceitos apresentados em [26] para classificar o presente trabalho, conforme detalhado em seguida. Mais especificamente, com relação à estruturação da pesquisa, foi utilizado um processo modelo para trabalhos de mineração de dados utilizado em aplicações de diferentes áreas de conhecimento, conforme apresentado em seção própria deste Capítulo como conceituação inicial e, de forma mais detalhada, no Capítulo 3.

3.1.1 Método da Pesquisa

Método, segundo Gil em [26] é uma forma de pensar para se chegar à natureza de um determinado problema, quer seja para estudá-lo ou explicá-lo. Nas pesquisas científicas, compreender o método utilizado é de suma importância para seguir padrões de geração de conhecimento acadêmico e tecnológico.

Em [26], são apresentadas formas de classificação de qualquer pesquisa científica ou tecnológica em diferentes dimensões. As classificações se dão pela natureza da pesquisa, pela abordagem, pelos objetivos do estudo, pela estratégia utilizada e pelas técnicas de coleta de informações que embasam o trabalho.

Baseado nas classificações apresentadas em [26], o presente trabalho se enquadra:

Natureza: Gil [26] afirma que pesquisas "puras"e "aplicadas"não são mutuamente exclusivas e indica que tratá-las como se fossem inteiramente diferentes é um equívoco. Apesar dessa conclusão, Gil referencia essa classificação como sendo bastante usada na literatura, pois que foi utilizada para enquadrar o atual estudo. Tendo em vista o caráter explicitamente tecnológico, para melhoria de um processo de trabalho, a pesquisa foi classificada como de natureza aplicada

Forma de abordagem: o presente trabalho tem abordagem predominantemente quantitativa. O uso de técnicas de mineração de dados, baseados em métodos estatísticos, torna a pesquisa focalizada em interpretações numéricas dos fenômenos descritos e observados, o que nos leva a essa classificação

Objetivos: explicativa, pois se enquadra nos estudos de quais variáveis influenciam num determinado resultado. De fato, o presente trabalho está mais focalizado em conhecer o porquê de uma solicitação ser deferida e outra não. O estudo tem algumas características descritivas, no entanto, como se observa pelos conceitos detalhados em [26], tais pesquisas tem por objetivo estudar as características de um grupo: sua distribuição por idade, sexo, procedência, nível de escolaridade, estado de saúde física e mental etc. Esse conhecimento descritivo, no entanto, é um objetivo secundário na pesquisa em questão

Estratégia: a classificação de estratégia que este estudo mais se aproxima é a pesquisa *ex-post facto*. Tais pesquisas se caracterizam por avaliar a influência entre variáveis independentes noutra dependente a partir de fatos pretéritos [26]

3.1.2 Estruturação da Pesquisa

A estruturação da pesquisa, em termos mais amplos, seguiu os passos identificados na Figura 3.1, desde a formulação do problema da pesquisa até a implantação da solução concebida para responder seus questionamentos.

Mais especificamente, a estruturação deste trabalho teve como base um padrão de mercado para mineração de dados, o *Cross Industry Stardard Process for Data Mining* (CRISP-DM). O padrão estabelece etapas que devem ser seguidas para realizar qualquer trabalho de mineração de dados, de forma a não subestimar nenhuma delas para garantir que os possíveis resultados tenham validade. As próximas seções apresentam cada uma dessas etapas e como foram trabalhadas. Os detalhes de cada etapa dessa metodologia são apresentadas no próximo Subcapítulo.

3.2 Mineração de Dados

Com aplicações das mais diversas, a mineração de dados ganha espaço em todos os tipos de negócio privados ou públicos. Pode-se conceituar a mineração de dados como um conjunto de técnicas estatísticas e de programação para buscar encontrar padrões e permitir inferências a partir de bases de dados.

Alguns autores diferenciam a mineração de dados, referindo-se como o uso de técnicas, da extração de conhecimento em bases de dados (*Knowledge Discovery in Databases* ou

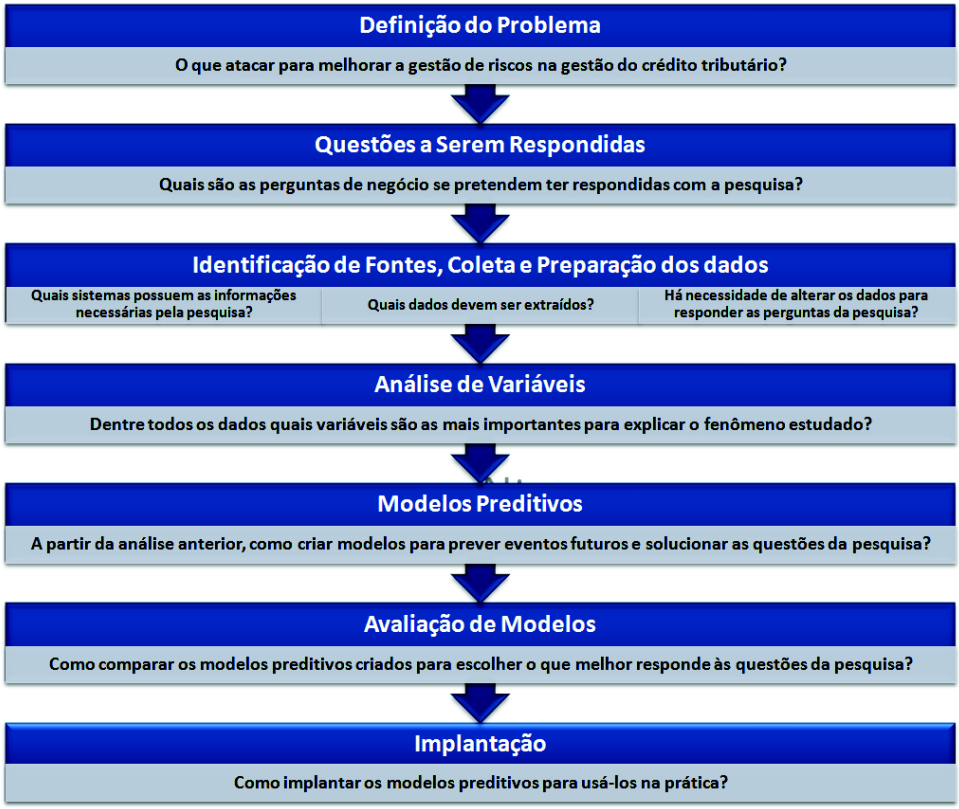

Figura 3.1: Estruturação da Pesquisa

KDD em inglês), que seria o resultado das análises. Fayyad et al [21] define KDD como o desenvolvimento de métodos e técnicas para tirar algum sentido dos dados, mas não distingue conceitualmente as duas definições, mas somente salienta o foco da análise de dados: produção de conhecimento.

Historicamente, ainda segundo [21], o termo "mineração de dados"é mais utilizado por comunidades de estatísticos, analistas de dados e sistemas de informação, enquanto KDD foi cunhado pela primeira vez no primeiro Workshop KDD em 1989 por Piatetsky-Shapiro [46] para enfatizar que o conhecimento é o grande foco de qualquer descoberta a partir de dados.

Os próximos Subcapítulos detalham um metodologia de referência para trabalhos de mineração de dados, apresentam as diferenças entre a extração de conhecimento a partir de análise supervisionada e não supervisionada.

3.2.1 CRISP-DM

Projetos de análise de dados possuem uma metodologia de referência, utilizada por diferentes áreas de conhecimento, chamada *Cross Industry Standard Process for Data Mining* (CRISP-DM) [9]. O CRISP-DM é uma metodologia que não depende da tecnologia utilizada e se apresenta como um modelo de referência para implementação de processos de análise de dados para diversas áreas de negócio.

A metodologia descreve cada fase por que um processo de mineração de dados deve passar. As fases possuem importância equivalente e não devem ser super ou subestimadas com o risco de o processo da análise de dados não atingir os objetivos de negócio. O processo também define alguns ciclos em que uma fase pode ser executada mais de uma vez.

O processo descrito pelo modelo de referência possui seis fases, que se iniciam no entendimento do negócio e terminam na implantação. As seis fases do CRISP-DM são [9]:

Entendimento do Negócio

Qualquer processo de análise de dados é projetado para responder a um ou mais questionamentos de negócio com fim de aperfeiçoar um processo de trabalho. Na fase de entendimento dos dados do CRISP-DM, essas questões são levantadas, bem como quais serão as possíveis melhorias propostas para melhorar a solução de problemas e as linhas gerais de como se dará essa solução. As melhorias apresentadas poderão ser qualitativas ou quantitativas e justificam a utilização de métodos de mineração de dados no problema objeto.

Conforme [9], a fase inicial focaliza no entendimento dos objetivos e requisitos do projeto a partir de uma perspectiva de negócio para então converter esse conhecimento numa definição de problema de mineração de dados e um plano preliminar é desenhado para alcançar os objetivos.

Entendimento dos Dados

Uma vez entendidas as questões de negócios, deve-se entender quais são as informações e dados necessários para responder aos questionamentos e atingir os objetivos identificados na fase de entendimento do negócio. Na fase de entendimento dos dados, todas as fontes de informação necessárias para realizar a análise de dados são pesquisadas e definidas. Os primeiros *insights* e principais padrões são identificados nos primeiros contatos com as informações pesquisadas nas fontes de informação identificadas. Cada questionamento de negócio deve ser mapeado com as respectivas fontes de informação (sistemas, bases de

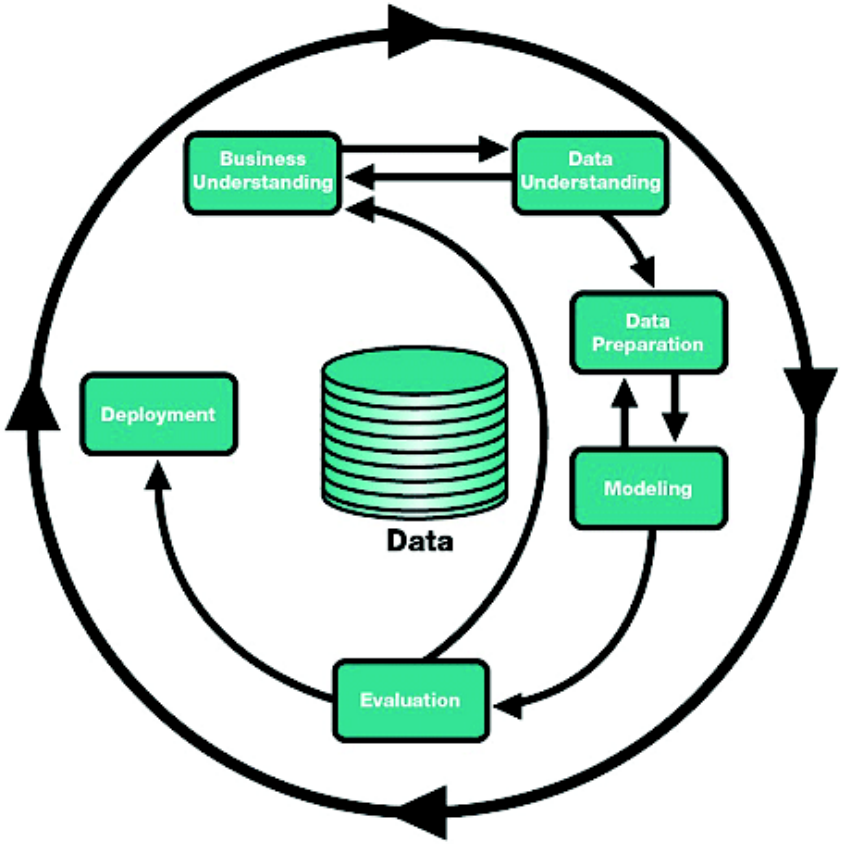

Figura 3.2: Modelo de referência CRISP-DM [9]

dados, sítios na internet, etc) como forma de parear cada objetivo a ser alcançado com as fontes e identificar o quanto antes as possíveis lacunas e falta de informações necessária.

Chapman [9] define que existe uma ligação bem próxima entre as fases entendimento do negócio e o entendimento dos dados. A formulação do problema de mineração de dados e do plano para solucioná-lo necessita de algum entendimento dos dados disponíveis.

Preparação dos Dados

A preparação dos dados cobre todas as atividades para construir o conjunto de dados final para análise a partir dos dados brutos iniciais [9]. Essa fase, conforme o modelo, pode ser realizada diversas vezes. Algumas atividades da fase são seleção de tabelas, registros e atributos, bem como a análise de consistência dos dados e a construção de novos atributos para utilização de ferramentas de modelagem [9].

Modelagem

A fase posterior à de preparação aos dados traz a seleção de várias técnicas de modelagem para aplicação nos dados preparados, bem como são realizadas calibragem de parâmetros dos diferentes métodos utilizados para encontrar o melhor resultado. Normalmente há muitos topos de técnicas para atacar um mesmo problema de mineração de dados. Algumas técnicas requerem formatos específicos. Há uma grande relação entre as fases de modelagem e de preparação dos dados, muitas vezes são encontrados problemas na preparação somente no momento da aplicação de técnicas de modelagem, o que pode ensejar numa nova execução da fase anterior [9].

Avaliação

A partir desse estágio no projeto foram criados um ou mais modelos que parecem ter uma alta qualidade, numa perspectiva de análise de dados. Antes de seguir com a implementação do modelo final, é importante realizar uma avaliação mais rígida e rever os passos executados para criação do modelo e garantir que ele alcança os objetivos de negócio [9].

A partir do gráfico que define os passos do CRISP-DM, pode-se perceber que existe a possibilidade de retornar a passos anteriores, sempre com o foco de atender aos objetivos e metas definidos no passo de entendimento do negócio. Conforme [9], ao final da fase é avaliado se interessa ao negócio a implementação dos resultados da mineração de dados e se o mesmo será utilizado ou não.

Implantação

A criação final do modelo geralmente não é o fim do projeto. Normalmente o conhecimento adquirido deve ser organizado e apresentado de uma forma que o cliente possa usá-lo [9]. Em muitos casos, a implementação se dá por pessoas diferentes da equipe que realizou as etapas anteriores e, por esse motivo, deve-se apresentar os resultados finais de forma a que seja possível a aplicação no negócio.

3.2.2 Análise Supervisionada

Conforme destacado por [28], a análise supervisionada é aquela em que o aprendizado de máquina se dá a partir de resultados previamente conhecidos. Para inferir se um paciente pode ter ou não diabetes a partir de resultados de anos de diagnósticos e reconhecer as fotos de uma pessoa a partir do aprendizado de uma imagem do rosto da pessoa, são exemplos de análises supervisionadas.

Pode-se declarar o problema de análise supervisionada como uma forma de inferir os resultados futuros a partir do aprendizado do passado. Tem-se de antemão o conhecimento

da variável dependente, aquela que se quer predizer no futuro, a partir de características das observações anteriores, as variáveis independentes ou explanatórias. Como apresentado em seção específica, a criação de modelos preditivos, foco do presente trabalho, é uma forma de análise supervisionada.

Um elemento chave para a avaliação de métodos de análise supervisionada é o critério de erro utilizado para realizar as aproximações dos modelos preditivos, ou seja, a distância entre o que foi inferido e a realidade. Os métodos mais comuns de calcular erros para métodos supervisionados são o mínimos quadrados e máxima verossimilhança.

Existem basicamente dois tipos de análise supervisionada: regressão e classificação. Na regressão a variável dependente é contínua, ou seja, não é discreta. Muitos algoritmos e técnicas podem ser utilizadas para regressão, como a Regressão Linear Simples. A classificação é utilizada quando a variável dependente é discreta, ou seja, é dividida em classes. Técnicas de classificação se estendem desde métodos estatísticos, como a Regressão Logística, até Árvores de Classificação, passando por Redes Neurais, Redes Bayesianas dentre outros.

A próxima subseção traz a definição de análises em que não se tem o conhecimento prévio: análises não supervisionadas.

3.2.3 Análise Não Supervisionada

A análise não supervisionada é uma espécie de aprendizado sem um professor [28]. Nesse tipo de análise não está a disposição, ou não são utilizadas, informações previamente conhecidas. Os algoritmos utilizados desse tipo, portanto, não permitem saber se os resultados do modelo estão corretos ou não, bem como não permitem avaliar os possíveis erros relacionados às observações trabalhadas.

O conhecimento dos especialistas do negócio de que problema a ser atacado faz parte é essencial nesse tipo de análise. Isto porque são inferidas as características que irão direcionar a análise para a criação dos modelos. Alguns tipos de técnicas não supervisionadas são a Análise de Aglomerados (*Cluster Analysis*) e Análise de Cesta de Mercado (*Market Basket Analysis*).

3.3 Modelos Preditivos

A Inferência Estatística é uma ferramenta muito utilizada para, a partir de informações e padrões obtidos com base numa amostra, estender os resultados avaliados para a respectiva população [6]. Predizer o futuro baseado nos acontecimentos do passado é a principal característica e atrativo para o uso de modelos preditivos na solução dos mais diversos

problemas, desde a previsão de erosão em terrenos [42] até a identificação se um paciente com hepatite C possui fibrose hepática [22].

Para criar modelos preditivos pode-se utilizar tanto algoritmos de regressão quanto de classificação. Alguns dos métodos e técnicas utilizadas no presente trabalho são a Regressão Logística, Árvores de Classificação e Naïve Bayes. As próximas subseções trazem os conceitos de cada algoritmo.

3.3.1 Regressão Logística

As regressões estatísticas são utilizadas para descobrir as relações de certas características (variáveis independentes) com um resultado analisado (variável dependente). Muitos problemas podem ser modelados para utilização das regressões estatísticas. Como um tipo de análise supervisionada, as regressões também podem ser utilizadas para criar modelos preditivos a partir de bases de dados com conhecimento prévio dos resultados.

A Regressão Logística é um tipo de regressão estatística onde a variável dependente é dicotômica, ou seja, aceita valores 0 (zero) ou 1 (um). A utilização de tal ferramenta facilita a modelagem de problemas de criação de modelos preditivos para classificação, diferente de outros tipos de regressão que resultam em valores quantitativos ao invés de fatores (ou classes). De fato, conforme [29], a Regressão Logística é o método estatístico mais utilizado quando variável independente é discreta.

Apesar de ser uma regressão, em termos estatísticos, em mineração de dados o algoritmo é frequentemente rotulado como sendo de classificação, ou seja, quando se deseja descrever o relacionamentos entre variáveis independentes (contínuas ou discretas) com uma variável dependente discreta.

Antes de apresentar como se constrói uma função utilizada para modelos de Regressão Logística, um modelo de regressão mais simples é detalhado. A Regressão Linear Simples é utilizada nos casos em que a variável dependente é contínua e será detalhada em seguida para que se possa compreender a necessidade de se utilizar a Regressão Logística em casos em que a linear não apresenta bons resultados. Ademais, busca-se mostrar como as regressões são estruturadas, como suas funções são definidas, como são os algoritmos para buscar os melhores coeficientes e como são avaliados os modelos criados usando regressões estatísticas.

Regressão Linear Simples

A Regressão Linear Simples responde a uma variável dependente quantitativa ao invés de discreta. A fórmula dessa regressão é simples do tipo $Y \approx \beta_0 + \beta_1 x_1 + \epsilon$, onde Y é a variável dependente, X_1 a variável independente ou explanatória, β_0 é a intercepção da

função com o eixo vertical e β_1 o coeficiente angular da reta e ϵ é o erro do modelo, ou aquilo que não é explicado pela regressão. Aplicando a Regressão Linear Simples para problemas reais, a função apresenta o quanto as variáveis independentes influenciam na variável dependente, ou seja, quanto mais β_1 for diferente de 0 (zero), mais influência a variável explanatória X_1 terá sobre Y.

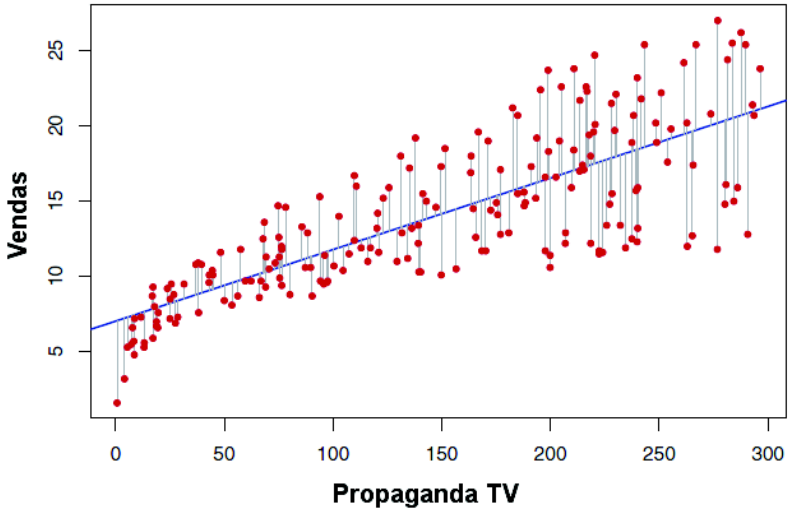

Figura 3.3: Regressão Linear Simples [33] (modificado pelo autor: tradução livre)

Um exemplo de Regressão Linear simples (com uma variável independente) pode ser vista na Figura 3.3. O exemplo apresenta uma regressão para explicar a relação entre a quantidade de propagandas veiculadas na TV de um certo produto e a quantidade de vendas do produto [33]. Visualmente há claramente uma tendência de aumento das vendas com o aumento das veiculações, conforme pode-se verificar pelos pontos no gráfico (vermelhos). Para tentar explicar como se dá a influência entre a propaganda e as vendas, uma regressão é utilizada para se aproximar o máximo possível aos pontos pré-existentes (reta azul).

A forma utilizada para garantir que essa é a melhor regressão é realizar um cálculo para avaliar qual das possíveis retas é a mais adequada para aquele conjunto de pontos (distância entre pontos e a reta representados no gráfico). Existem diversas formas de calcular as diferenças entre as previsões da regressão e as observações reais e escolher a melhor delas. O método mais utilizado para Regressões Lineares é o Método dos Mínimos Quadrados que realiza uma soma de todos os quadrados dos erros ϵ e tenta buscar os me-

lhores coeficientes (β_0 e β_1) para reduzir ao máximo essa soma [33]. Denota-se, portanto, a soma dos quadrados dos erros como:

$$SQ_{erro} = \sum_{i=1}^{\mathrm{n}} (\epsilon_i)^2 \tag{3.1}$$

Se cada ponto representando uma observação é denotado y_i e cada estimativa pela regressão $\hat{\beta}_0 + \hat{\beta}_1 x_i$, pode-se reescrever 3.1 como

$$SQ_{erro} = \sum_{i=1}^{\mathrm{n}} (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i)^2 \tag{3.2}$$

Desse modo, para encontrar os coeficientes que tenham a menor soma de quadrados de erros, a partir das equações anteriores, chega-se às definições dos coeficientes β_0 e β_1 [33].

$$\begin{gathered} \hat{\beta}_1 = \frac{\sum_{i=1}^{\mathrm{n}} (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{\mathrm{n}} (x_i - \bar{x})^2}, \\ \hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x} \end{gathered} \tag{3.3}$$

A Regressão Linear Simples é muito importante para conhecer melhor como modelar um problema, mas muitas vezes não se aplica a situações com variáveis independentes que não são contínuas. Para isso, se pode utilizar a Regressão Logística, que utilizar uma função não linear para adequar o resultado para que fique entre 0 (zero) e 1 (um): a Função *Logit*.

Função Logística

A Regressão Linear Simples tem aplicabilidade bastante variada e, ainda que seja conhecida por muitos anos, é bastante utilizada até hoje [33]. Alguns problemas, no entanto, não tem na Regressão Linear Simples uma boa solução, como é o caso onde a variável dependente é discreta e dicotômica, ou seja, quando aceita somente dois valores. De fato, conforme Figura 3.4, temos que não há função linear que resulte numa boa aproximação dos resultados dicotômicos, quase sempre resultando em erros não desprezíveis.

Para o caso exemplificado, existem funções não lineares que melhor se adequam e podem ser utilizadas em problemas com saídas discretas. Para aquelas com resultados dicotômicos é necessária uma função que aceite valores de entrada variando de $-\infty$ a $+\infty$ de menos a mais infinito e a resposta sempre contida entre 0 e 1. Uma função que atende esses requisitos é a função logística. Essa função sempre tem uma forma de "S"e consegue

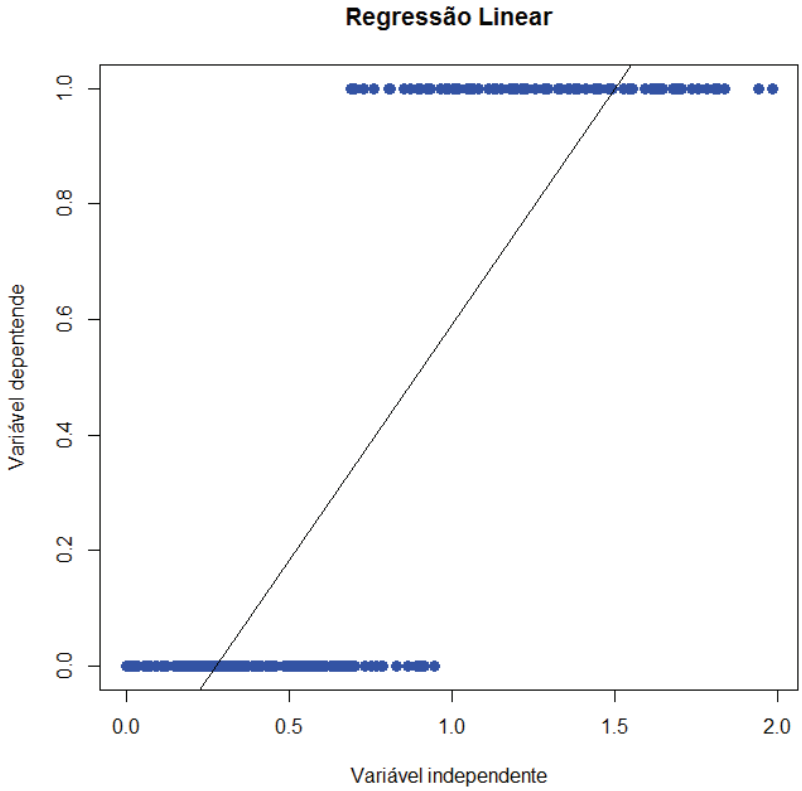

Figura 3.4: Regressão Linear Simples com Y dicotômico

obter uma predição mais sensível que as funções lineares [33]. A função logística pode ser descrita como:

$$P(X) = \frac{e^{\beta_0+\beta_1 X}}{1 + e^{\beta_0+\beta_1 X}}$$

Claramente, se $e^{\beta_0+\beta_1 X}$ for próximo de 0 (zero), $P(X)$ também o é, se $e^{\beta_0+\beta_1 X}$ tender a $+\infty$, $P(X)$ tende a 1 (um) e atende aos princípios de, para qualquer valor de X, o resultado está entre 0 e 1. Para facilitar a aplicação da Regressão Logística em função de $X, \beta_0 e \beta_1$, com alguma manipulação, podemos chegar à equação:

$$\frac{P(X)}{1 - P(X)} = e^{\beta_0+\beta_1 X}$$

A expressão à esquerda é a probabilidade de um evento ocorrer, sobre a probabilidade de ele não ocorrer. Em inglês, o resultado é chamada de *odds* e é muito utilizado em

bolsas de apostas e corridas de cavalos por ser uma forma mais simples de apresentar probabilidades [33]. Essa função pode variar de 0 a $+\infty$. Finalmente, para chegarmos a uma equação linear em X podemos aplicar o logaritmo nos dois lados da equação, obtendo:

$$logitP(X) = log(\frac{P(X)}{1 - P(X)}) = \beta_0 + \beta_1 X_1$$

A Figura 3.5 mostra o mesmo exemplo anterior com dados dicotômicos utilizando a função logística.

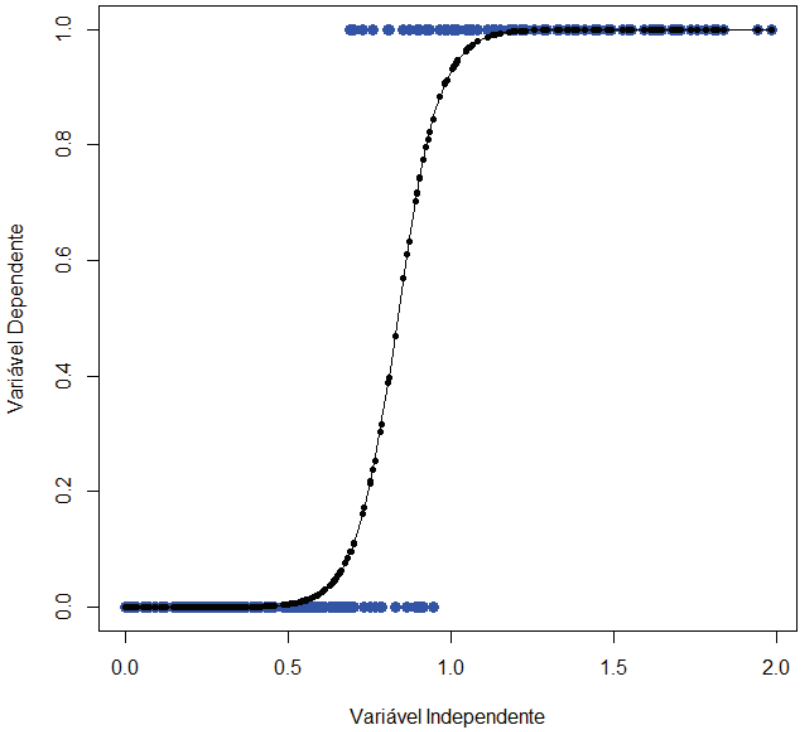

Figura 3.5: Regressão Logística

Desse modo, a Regressão Logística pode ser utilizada para conhecer as influências que as variáveis explanatórias se prestam à variável dependente e também para criação de modelos preditivos. A partir de um conjunto de coeficientes conhecidos pode-se aplicar a fórmula para predizer o valor de Y a partir dos valores de X_1, X_2....

Uma desvantagem na utilização da técnica é a estimativa dos coeficientes da regressão, mas a facilidade na aplicação em problemas que se encaixem no modelo superou os problemas no presente trabalho, conforme detalhado no Capítulo 4.

Duas diferenças são primordiais entre a Regressão Linear Simples e a logística, além da natureza das variáveis analisadas: interpretação dos coeficientes e a função utilizada para encontrar os melhores coeficientes para criação do modelo preditivo.

A interpretação dos valores se dá de um modo diferente da Regressão Linear Simples, pois a relação entre os coeficientes β_1, β_2 com a variável dependente não é linear, ou seja, se β_1 aumenta em uma unidade, Y não aumenta de forma linear. Mas, independente do valor de X, se β_1 é positivo, então o crescimento de X será associado ao incremento de $P(X)$ e vice-versa [33].

A função utilizada pela Regressão Logística para criar os modelos preditivos e encontrar os melhores coeficientes β_0 e β_1 são calculados com o Método de Máxima Verossimilhança, ao invés do método dos mínimos quadrados, como na Regressão Linear Simples. Desse modo, os coeficientes devem ser calculados para que a função de verossimilhança seja a maior possível.

Com os coeficientes calculados, conseguimos utilizar o resultado da Regressão Logística para inferir sobre novos valores de variáveis explanatórias. O Capítulo 4 apresenta como foi utilizada a Regressão Logística em dois momentos da análise, primeiramente para selecionar as variáveis independentes que mais influenciam na dependente e num segundo momento como um dos algoritmos de classificação para criação dos modelos preditivos.

3.3.2 Árvores de Classificação

As árvores de decisão são uma família de algoritmos que são utilizados para separar um conjunto de observações em subconjuntos disjuntos, a partir da análise das propriedades de cada observação. Elas podem ser usadas tanto para problemas de regressão quanto classificação. Tendo em vista a aplicação que teremos para o método, detalhado no Capítulo 4, nos atemos às Árvores de Classificação.

As Árvores de Classificação possuem aplicações similares aos métodos identificados nos subcapítulos anteriores e também endereçam problemas onde se busca conhecer a influência de variáveis explanatórias numa variável dependente e discreta. Esse método divide de forma iterativa os dados da entrada em conjuntos disjuntos, de forma binária em cada passo.

Os algoritmos para criar árvores de classificação mais conhecidos são o *Classification and Regression Tree (CART)* [5] e o C4.5 [50].

Cada nó interior da árvore corresponde a uma propriedade da obervação. Para cara valor possível de cada propriedade existe uma aresta. Cada folha representa um dos

subconjuntos, cujo caminho até a raiz traz os valores de cada propriedade para definir seus elementos [2]

Um exemplo de Árvore de Classificação pode ser visto na Figura 3.6 onde Castellon et. al [8] apresentam uma aplicação da técnica para classificar uma empresa entre noteira (emite notas fiscais falsas) e não-noteira.

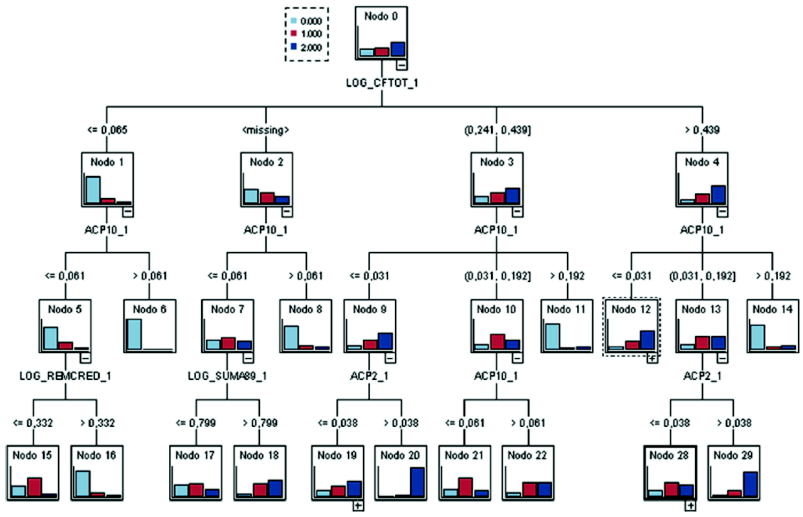

Figura 3.6: Árvore de Classificação de empresas

As Árvores de Classificação podem ser utilizadas para muitos fins, mas será utilizado para criar modelos preditivos conforme apresentado no Capítulo 4. Mais especificamente será utilizado o algoritmo de *Random Forests*, que cria diversas árvores de classificação e, a partir de alguns parâmetros de avaliação, escolhe aquelas que melhor se adequam às observações.

3.3.3 Redes Bayesianas

Conforme [38] Redes Bayesianas são modelos gráficos utilizados para buscar padrões a partir de incertezas, onde os nós representam as variáveis (discretas ou contínuas) e os arcos (ou arestas) representam conexões diretas entre as primeiras. Essas conexões são frequentemente ligações causais, ou seja, definem uma relação de causa e efeito entre variáveis. Ademais, os modelos utilizando Redes Bayesianas representam o quão significativas (fortes) são as relações entre as variáveis, permitindo conclusões probabilísticas e também sua atualização à medida em que novas informações surgem.

Redes Bayesianas são muito úteis para buscar em informações disponíveis em bases de dados as relações de causalidade entre diferentes características e também podem ser utilizadas como algoritmos de classificação. Tendo em vista essas características, esses algoritmos são utilizados para criação de modelos preditivos em diferentes áreas e

problemas desde genética [34] e prognóstico de câncer de mama [24], até a identificação de compras fracionadas [7].

Desse modo, é possível utilizar essa técnica para identificar a relação entre variáveis independentes e uma outra dependente. No exemplo utilizado na seção de Regressão Linear Simples e Logística, onde se avalia a relação entre a quantidade de propagandas veiculadas e o número de vendas, essa relação poderia ser verificada utilizando Redes Bayesianas. De fato, em Capítulo próprio é apresentado como foram utilizado os algoritmos bayesianos para criação de modelos preditivos para buscar analisar o risco de uma solicitação de compensação de crédito tributária deve ser ou não deferida.

Muitos são os algoritmos da família de Redes Bayesianas, a seguir são apresentados as técnicas Naïve Bayes e Tree-Augmented Naïve Bayes.

Naïve Bayes

Naïve Bayes é a versão mais simples de Redes Bayesianas. O algoritmo utiliza conexões fortes entre os nodos, e relaciona cada variável independente (explanatória) com a dependente, sem identificar relacionamento entre as primeiras, ou seja, as considera como sempre independentes. Apesar de sua simplicidade, muitas são as aplicações com bons resultados, além da grande performance na utilização prática [62].

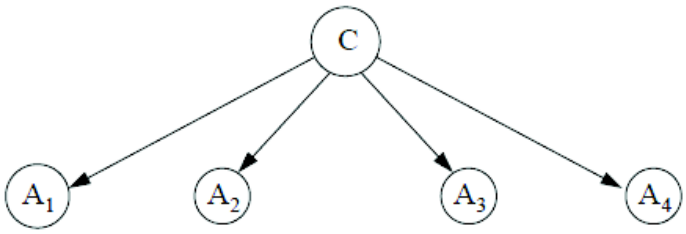

Figura 3.7: Exemplo de uma Rede Bayesiana utilizando Naïve Bayes [62]

Tree-Augmented Naïve Bayes

Tree-Augmented Naïve Bayes (TAN), como detalhado em [63], flexibiliza a premissa de que as variáveis explanatórias (independentes) são completamente independentes. A estrutura de árvore utilizada pelo algoritmo original (variável dependente como raiz e as independentes como folhas) é alterado para permitir que conexões entre as variáveis independentes sejam construídas. Apesar da estrutura ser menos rígida, somente é permitido

a uma variável independente ter relação com a dependente, por óbvio, e com outra independente. Essa flexibilização permite a representação de modelos mais complexos, o que pode levar a melhorias na performance, conforme resultados apresentados em [7].

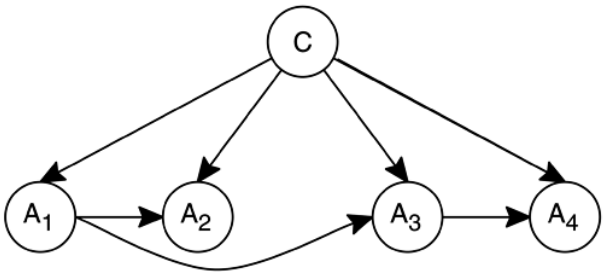

Figura 3.8: Exemplo de Rede Bayesiana utilizando *Tree-Augmented Naïve Bayes* [35]

3.3.4 Avaliação de Modelos

Conforme mostrado nas subseções anteriores existem muitas formas, técnicas e algoritmos para criar modelos preditivos. Essencial, portanto, haver uma maneira de avaliar qual o modelo mais interessante para atacar o problema objeto das predições. Nesse sentido, apresenta-se algumas medidas que podem ser utilizadas para se definir se um modelo é melhor que o outro.

Matriz Confusão

As matrizes de confusão confrontam as predições dos modelos com a classificação real dos dados. É uma forma simples de avaliar se um modelo é adequado ou não, observando na diagonal os acertos do modelo preditivo e nos outros elementos os erros (falsos positivos e negativos). A matriz, conforme Tabela 3.1, é quadrada e o número de linhas e colunas é a quantidade de classes utilizadas na predição. Conforme será apresentado, utilizaremos somente duas classes (solicitação de compensação deferida ou indeferida) e, por isso, é apresentada uma matriz com duas linhas e duas colunas.

A primeira linha apresenta a quantidade de observações que o modelo previu como sendo da classe um (positiva, quando possui somente duas classes), sendo que a primeira coluna é o que o modelo realmente acertou (verdadeiros positivos) e a segunda são as

Tabela 3.1: Conceito de Matriz de Confusão

	Referência	
	Verdadeiro Positivo (VP)	Falso Positivo (FP)
Predição	Falso Negativo (FN)	Verdadeiro Negativo (VN)

observações classificadas de forma errada (falso positivos). A segunda linha são as observações classificadas pelo modelo preditivo como sendo da classe dois (negativa, quando possui somente duas classes), sendo as quantidades da primeira coluna as previsões erradas (falsos negativos) e a segunda as previsões corretas (verdadeiros negativos).

A partir da Matriz de Confusão, pode-se analisar os resultados dos modelos preditivos a partir de alguns indicadores. As próximas subseções apresentam os conceitos da acurácia, sensibilidade, especificidade, *F-score* e da curva de características de operação do receptor.

Acurácia

A acurácia (AC) também é chamada de índice total de acertos, ou seja, o número total de predições corretas sobre o número de observações da amostra:

$$AC = (VP + VN)/(VP + FP + VN + FN) \tag{3.4}$$

Esse indicador é interessante quando não há uma grande diferença entre acertar os positivos e os negativos, ou seja, quando não importa se VP é maior que VN e vice-versa. No caso de essa diferenciação ser importante, deve-se observar outros indicadores, como a sensibilidade, especificidade e F-score.

Sensibilidade

A sensibilidade é a proporção dos verdadeiros positivos que são corretamente identificados por um modelo preditivo [64]. O total de observações realmente positivas são os verdadeiros positivos e os falsos negativos (como é falso negativo, é, na realidade, positivo) e, desse modo, pode-se calcular a sensibilidade (SE) utilizando a fórmula abaixo:

$$SE = VP/(VP + FN) \tag{3.5}$$

A sensibilidade (ou *recall*) nos mostra o quanto um modelo preditivo é bom ao tentar prever as observações positivas. Numa utilização em modelos preditivos para um teste de diagnóstico, como traz Zhu, Zeng e Wang em [64], se a sensibilidade for de 99% tem-se

que se for testado um paciente que possui a doença, há uma chance de 99% de que o paciente será identificado no teste como tendo a doença .

Em conjunto com a acurácia e a especificidade, pode trazer informações importantes na comparação de modelos preditivos.

Especificidade

A especificidade é semelhante à sensibilidade, mas calcula o índice de acerto dos modelos preditivos para observações falsas. O número total de observações falsas são a soma entre os verdadeiros falsos e os falsos positivos. Dessa forma, a especificidade (ES) é calculada do seguinte modo:

$$ES = VN/(VN + FP) \tag{3.6}$$

A mesma utilização anterior [64] nos remete ao exemplo de a especificidade de 99% como um paciente que não tem doença ser diagnosticado pelo modelo preditivo como não tendo doença.

Obviamente que a acurácia, sensibilidade e especificidade não podem ser analisadas separadamente e que, em conjunto, trazem muitas informações sobre o modelo preditivo avaliado. Ainda com essa análise conjunta podemos incorrer em perda de informações importantes do modelo preditivo. Na próxima subseção veremos uma medida que traz outras informações para se avaliar um modelo preditivo utilizando média harmônica.

F-score

A definição do *F-score*, também conhecido como *F-measure*, se baseia nas definições de sensibilidade e de precisão. A precisão (PR), que não detalhamos nas subseções anteriores, é definida por:

$$PR = VP/(VP + FP) \tag{3.7}$$

A precisão é, portanto, uma taxa de acurácia da predição de casos positivos. O F-score, por sua vez, é uma média harmônica entre a sensibilidade e a precisão e pode ser medido a partir da seguinte fórmula:

$$F - SCORE = \frac{2 * SE * PR}{SE + PR} \tag{3.8}$$

Essa medida faz com que tenhamos uma ponderação entre as predições positivas e negativas, para que não tenhamos um desbalanceamento nas análises de observações po-

sitivas e negativas. Essa medida, no entanto, não leva em consideração o número de verdadeiros negativos, conforme apontado em [48].

Os indicadores comparativos apresentados nessa seção foram utilizados nos experimentos conforme detalhamos no Capítulo 4. A próxima seção traz um levantamento de pesquisas realizadas que se assemelham ao presente trabalho.

Validação Cruzada com k Repetições

Para avaliar os resultados de modelos preditivos, precisamos dividir a base de dados utilizada para treinamento, validação e testes. Quando a quantidade de observações nas bases possuem um tamanho reduzido, muitas vezes é difícil ter quantidade suficiente para separar os dados em 3 (três). Nesses casos podemos utilizar a base de treinamento e separar uma pequena parte para também realizar a validação. Uma forma de minimizar os riscos de escolhermos uma parte do treinamento para validação que não corresponda a distribuição original, é interessante utilizar uma técnica denominada "validação cruzada"(*cross validation*) [36].

Na validação cruzada, a base de treinamento, com n elementos, é dividia em k pedaços iguais. Em cada uma das k etapas, o algoritmo de validação cruzada separa a base em duas, sendo uma do tamanho $n-(n/k)$ para realizar o treinamento e outra com tamanho n/k para realizar a validação. O algoritmo perfaz os testes X vezes, sempre escolhendo uma parte diferente para realizar a validação. Com todos os resultados, calcula-se a média dos erros (acurácia, sensibilidade, especificidade ou outro) e apresenta o resultado, normalmente com um intervalo de confiança. A Figura 3.9 exemplifica a execução do algoritmo.

Essa técnica é interessante para não viciar a validação, mesmo não havendo base de dados de grande tamanho para criar os modelos preditivos. Segundo [61] os valores mais populares para k variam de 5 (cinco) a 10 (dez), o que é "o suficiente para dar uma estimativa que é estatisticamente provável que seja precisa, a um custo 5-10 vezes maior do tempo de computação". Conforme Capítulo 4, utilizaremos o método de validação cruzada com 10 (dez) repetições. O próximo Capítulo apresenta alguns dos trabalhos correlatos mais afeitos à utilização de técnicas de mineração de dados nas administrações públicas e em problemas similares aos endereçados por este.

3.4 Trabalhos Correlatos

Muitas administrações tributárias tem utilizado ferramentas de mineração de dados para reduzir o risco do não cumprimento das obrigações tributárias (*tax compliance risk*) [43]. Apesar de ser um tema de grande interesse, percebe-se que muitos trabalhos não são apre-

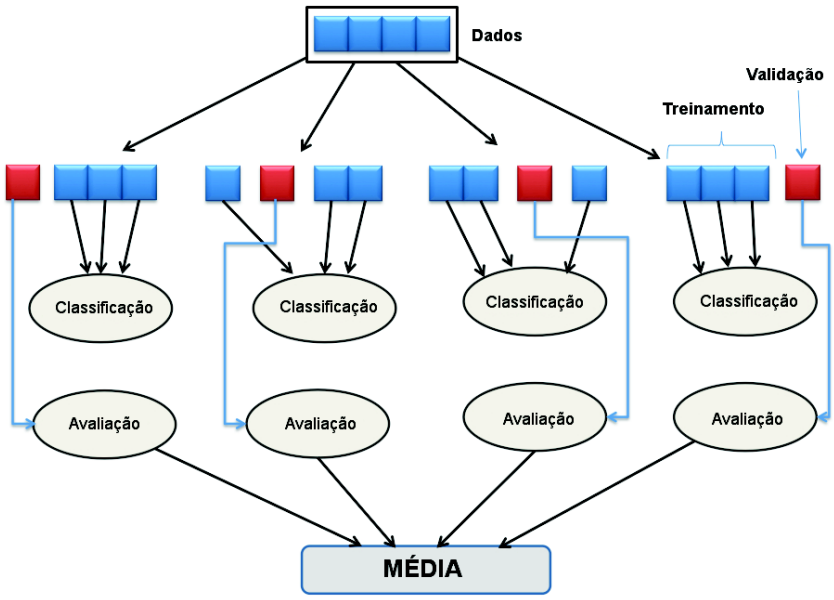

Figura 3.9: Validação Cruzada (tradução livre do autor) [19]

sentados em congressos e se restringem a publicações dos próprios órgãos governamentais. Uma razão pode ser a dificuldade em apresentar os resultados da aplicação de técnicas de mineração nos assuntos tributários, tendo em vista que, em sua maioria, são de caráter sigiloso.

Uma grande fonte de informações, compartilhamento de conhecimentos aplicados, metodologias e melhores práticas são as organizações intergovernamentais, destacando-se a Organização Mundial de Aduanas (OMA) [45] e a Organização para a Cooperação e Desenvolvimento Econômico (OCDE) [44]. Numa recente pesquisa entre os países que fazem parte da OCDE foi apresentado um quadro comparativo que traz a utilização de mineração de dados para detecção de fraudes tributárias [43].

As publicações internas das administrações tributárias também trazem muitos estudos que podem ser aplicados por outros apresentam metodologias desenvolvidas com base em análises estatísticas para a criação de sistemas baseados em regras e modelos de risco. Em sua grande maioria, as administrações se focalizam nos riscos de não cumprimento das obrigações tributárias criando indicadores que permitem ranking dos contribuintes por risco de conformidade.

Nossas pesquisas encontraram trabalhos relacionados a fraudes relativas a sonegação, elisão fiscal (*tax avoidance*) e restituição de crédito (*tax refund*), mas não se encontrou

trabalho com correlação direta ao nosso, ou seja, relacionado a compensação tributária. Isto não quer dizer que as outras administrações tributárias não estejam utilizando métodos estatísticos e de mineração de dados para definir os riscos de indeferimento das compensações se utilizando de modelos preditivos, mas não foi possível identificar estudos que endereçam o problema de forma direta.

As referências que mais se aproximam ao presente trabalho são apresentados como resultado de estudo da OCDE conforme Figura 3.10. A administração tributária dos Estados Unidos da América (*Internal Revenue Service - IRS*) utiliza mineração mineração de dados para diversos fins, de acordo com [43], dentre os quais estão a mensuração do risco de não cumprimento das obrigações tributárias pelo contribuinte, detecção de fraude fiscal e atividades criminais, detecção de fraude por parte dos contribuintes que recebem créditos fiscais e lavagem de dinheiro [12]. Assim, o IRS realiza trabalhos com mineração de dados para melhorar a seleção de processos de compensação tributária, mas a OCDE cita os trabalhos como resultado de questionários e visitas que realizam nos países consultados e não como uma referência a publicações científicas abordando o assunto.

Technique Applied	USA	Canada	Australia	UK	Bulgaria	Brazil	Peru	Chile
Neural Networks	✓	✓		✓	✓		✓	✓
Decision Tree	✓	✓	✓				✓	✓
Logistic Regression	✓		✓	✓	✓			
SOM			✓					✓
K-means			✓					✓
Support Vector Machines	✓		✓					
Visualization Techniques	✓					✓		
Bayesian Networks			✓					
K-Nearest Neighbour			✓					
Association Rules							✓	
Fuzzy Rules							✓	
Markov Chains						✓		
Time Series		✓						
Regression				✓				
Simulations	✓							

Figura 3.10: Comparativo de técnicas de mineração de dados utilizadas pelas administrações tributárias

Uma aplicação interessante de ferramentas de mineração realizada pela administração tributária do Chile é apresentada por Castellón González e Velásquez em [8]. O artigo foi elaborado por auditor-fiscal da administração tributária do Chile (*Servicio de Impuestos Internos*) e pesquisador da Universidade do Chile e tem como objetivo principal identificar de padrões de fraudes fiscais, principalmente para aqueles contribuintes que utilizam notas fiscais falsas, utilizando diferentes técnicas de mineração de dados. Os pesquisadores se utilizaram de várias técnicas para diferentes propósitos, alguns para clusterização e outros para determinação de variáveis (características dos contribuintes) relevantes para definir padrões de fraude e não fraude. Para separar o universo de contribuintes (agru-

par/clusterizar), o artigo utiliza os Mapas Auto-Organizáveis [37] e algoritmos de *Neural Gas* [40]. Para definir características de grupos de fraudes/não fraudes foram utilizadas as árvores de classificação [5]. Finalmente, para inferência de casos de fraude/não fraude, foram utilizadas Árvores de Classificação, Redes Bayesianas e Regressão Logística.

Além disso, Corvalão [11] desenvolve um modelo formal para classificação dos contribuintes a partir dos dados de movimentação mensal que são apresentados ao setor de fiscalização. A proposta busca preservar as características econômicas e regionais de cada empresa, valendo-se da análise de agrupamentos. Após esta fase foram construídos modelos probabilísticos que serão usados para relacionar os contribuintes com maiores indícios de irregularidades. Assim propõe que esta relação poderá ser utilizada para direcionar a seleção das empresas a serem auditadas.

Piccirilli [47], desenvolveu um modelo para classificar os contribuintes do Imposto Sobre Serviços de Qualquer Natureza (ISS) que apresentaram alguma irregularidade, a partir da aplicação do algoritmo de árvore de classificação. O trabalho foi desenvolvido no Município de Goiânia, abrangendo o cenário apresentado no ano de 2011.

Outra referência correlata é a tese de mestrado de Jani Martikainen [39] em que apresenta resultado de estudos em que afirma que a administração tributária da Austrália (*Australian Taxation Office - ATO*) utiliza modelos para detectar restituições (*tax refund*) de alto risco. Ainda de acordo com o autor, a ATO evitou o pagamento de restituições na ordem de US$665.000.000,00 (seiscentos e sessenta e cinco milhões de dólares americanos) entre 2010 e 2011 com base nas ferramentas de mineração de dados. Conforme será possível verificar adiante, os métodos utilizados pela administração australiana, apesar de atacarem o mesmo problema e ter o mesmo objetivo, são diferentes dos utilizados neste trabalho. A ATO utiliza modelos de restituição com base em algoritmos de descoberta de redes sociais que detecta ligação entre os indivíduos, empresas, parcerias ou declarações de tributos. Os modelos são atualizados e refinado para melhorar a detecção e aumentar o reconhecimento de novas fraudes.

Outros processos de trabalho da própria Secretaria da Receita Federal do Brasil foram melhorados consideravelmente pela aplicação de seleção baseada em modelos preditivos, como a seleção de contribuintes retidos em malha fiscal [18] e destacando-se o Sistema de Seleção por Aprendizado de Máquina (Sisam) [20]. O sistema conta com ferramentas de aprendizagem de máquina que se utiliza de algoritmos e modelos estatísticos criados especialmente para o problema de seleção de declarações de importação a partir do aprendizado do histórico de importações [32] [60] . Devido à especificidade, e daí o seu sucesso no problema a que se propõe solucionar, se torna custosa a adequação e utilização das técnicas já desenvolvidas pelo Sisam, para a melhoria do processo de seleção de solicitações de compensação de crédito ou a qualquer outro problema que não tenha a mesma

similaridade.

Capítulo 4

Solução Proposta e Resultados Obtidos

A gestão de riscos do processo de trabalho de Gestão do Crédito Tributário na Secretaria da Receita Federal do Brasil passa por diferentes subprocessos que se utilizam dos conhecimentos de Auditores-Fiscais especialistas. O Sistema de Controle de Crédito, conforme detalhado em seção anterior, internaliza regras definidas por tais servidores e as utiliza no tratamento das solicitações de compensação de crédito. Uma forma de aperfeiçoar a gestão de riscos é aproveitar melhor a experiência dos Auditores-Fiscais a partir da criação de modelos preditivos que tenham por base trabalhos já realizados. Estas são importantes fontes de informações a serem extraídas por algoritmos de aprendizagem de máquina e utilizadas na melhor seleção dos processos de compensação.

Este capítulo traz o detalhamento da solução proposta para o problema da seleção de solicitações de compensação de crédito tributário, seguindo os passos determinados pela metodologia CRISP-DM [9].

4.1 Entendimento do Negócio

O objetivo principal do trabalho na perspectiva de negócio da administração tributária brasileira é aumentar a justiça fiscal e a gestão de riscos relacionadas às análises de solicitação de compensação de crédito. Para tal, deve-se reduzir os riscos de não selecionar compensações para análise manual que são indevidas, ou seja, aquelas em que o contribuinte não tem direito ao crédito, mas ainda assim o solicita, seja por desconhecimento ou por fraude. A justiça fiscal se dá no momento em que os contribuintes que tem direito à compensação, tenham o débito homologado (deferimento) sem necessidade de análise manual, enquanto aqueles que não estão conformes e não devem ter direito ao crédito sejam analisados com rigor necessário (indeferimento).

Tendo em vista o grande impacto em termos de arrecadação líquida, conforme antecipado na justificativa do tema (1), melhorias no processo de seleção das solicitações de compensação de crédito podem reverter um grande volume de crédito tributário que seria compensado indevidamente. De fato, a melhoria de alguns pontos percentuais na performance da seleção desses processos pode elevar a arrecadação líquida em centenas de milhões de reais.

Uma grande vantagem das administrações tributárias em geral é sua grande concentração de informações sobre os contribuintes. Isto se dá pela obrigação a que estão submetidos para enviarem informações em forma de declarações (*tax returns*), e também por dados coletados em investigações, fiscalizações e diligências, onde documentos, balanços, contabilidade são reservados ao fisco para análise dos Auditores-Fiscais.

Especificamente com relação às compensações de crédito, a principal fonte de informação para atender os objetivos de negócio são as análises manuais das solicitações de compensação de crédito realizada pelos Auditores-Fiscais da RFB. A partir de uma base de análises altamente especializadas, é possível extrair o conhecimento para dar suporte a uma análise supervisionada no sentido de criar modelos preditivos que melhorem a seleção das solicitações de compensação. De forma estruturada, há interesse em responder aos questionamentos para atender os objetivos de negócio:

1. Quais são as características estatisticamente mais importantes dos contribuintes e de suas solicitações de compensação (informações da PER/DCOMP) para definir se haverá deferimento ou não do pleito?

2. Para novas solicitações de crédito, qual deve ser selecionada para análise com base nos riscos de deferimento/indeferimento (homologação/não homologação)?

Para melhorar a gestão de riscos no processo de gestão do crédito tributário, muitas outras pergunta podem ser respondidas na utilização de mineração de dados, como a detecção de agrupamentos de contribuintes que realizem determinado tipo de fraude, ou a análise dos deferimentos automáticos para buscar possíveis fraudes que passem despercebidas. Neste trabalho, no entanto, o foco se dá nos questionamentos 1 e 2 acima.

Uma característica importante do negócio que elevou a complexidade no entendimento e preparação dos dados conforme apresentado nos próximos Capítulos, é como se dá a solicitação de reconhecimento do crédito junto à administração tributária e sua(s) posterior(es) utilização em forma de compensação. Um contribuinte pode realizar quantas compensações quiser a partir de um mesmo crédito. Por exemplo, ele pode solicitar o reconhecimento de um crédito de imposto de renda (*income tax*) que pagou a mais e depois utilizar esse crédito diversas vezes para abater dívidas de imposto sobre produtos industrializados, um tipo de imposto sobre valor agregado (*value added tax*).

Desse modo, a análise das solicitações não pode se dar de forma unitária e deve-se unir todas as solicitações de um mesmo crédito em "famílias". Os deferimentos e indeferimentos analisados, portanto, serão realizados para famílias de PER/DCOMPs, acrônimo que representa as declarações de solicitação de crédito e de compensação.

Especialistas em compensação de crédito foram consultados para avaliar se o agrupamento em famílias faz sentido. A análise por família foi corroborada pelo fato de as características principais que podem influenciar na homologação (deferimento) ou não (indeferimento) de uma solicitação estarem mais relacionadas ao crédito utilizado, mais do que em cada solicitação de compensação em separado. Um exemplo de agrupamento de solicitações de compensação (DCOMP) de um mesmo crédito de R$400,00 em uma família é ilustrado pela Figura 4.1

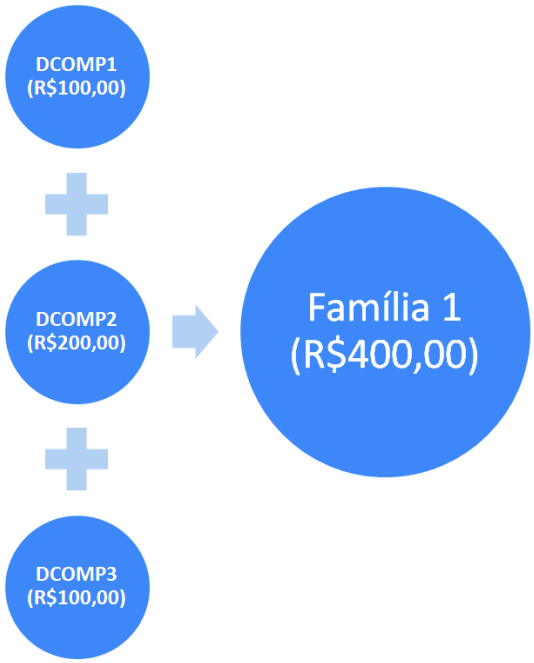

Figura 4.1: Família de Solicitações de Compensação

A partir dos conhecimentos extraídos das análises dos Auditores-Fiscais e, uma vez respondidas as perguntas, é possível identificar o que caracteriza uma solicitação deferida ou indeferida, bem como permite criar escalas de risco para classificação das solicitações de compensação. Desse modo, utilizando técnicas de mineração de dados, pretende-se atingir o objetivo final de melhorar a análise de riscos das solicitações com o fim de aumentar a arrecadação líquida.

4.2 Entendimento dos Dados

As informações relacionadas a solicitações de compensação de crédito e as variáveis e características dos contribuintes que as solicitam se encontram em diferentes bases de dados, de diferentes sistemas e arquiteturas. Um facilitador para o caso da administração tributária brasileira é a consolidação de informações em diversos data marts, cada um com uma visão do contribuinte.

Alguns indicadores foram construídos a partir de conhecimento prévio de especialistas em compensação tributárias para que não fossem usados somente dados brutos e características irrelevantes dos contribuintes analisados. Finalmente, além das informações de crédito tributário foram agregados outras informações dos contribuintes disponíveis em cadastros e sistemas de acesso a declarações de tributos diversos (tax returns).

Para este trabalho inicial foram utilizadas compensações de contribuintes sob a jurisdição da 1a Região Fiscal da Secretaria da Receita Federal do Brasil, que abrange as unidades federativas do Distrito Federal, Goiás, Mato Grosso, Mato Grosso do Sul e Tocantins. A coleta de informações foi realizada utilizando diversos softwares disponibilizados na Secretaria da Receita Federal do Brasil.

O início das extrações foi marcado por uma decisão de projeto no sentido de analisar num só conjunto todas as compensações independente do tipo de crédito e do ano em que foram trabalhadas. Desse modo, as perguntas formuladas no entendimento do negócio foram respondidas sem apartar as compensações pelos critérios de tempo e tipo do crédito originário. A quantidade de famílias por tipo de crédito pode ser visto na Tabela 4.1

A extração inicial continha cerca de 18000 famílias de compensações, que contemplava tanto trabalhos realizados manualmente por auditores fiscais quanto processos deferidos/ indeferidos automaticamente pelo Sistema de Controle de Crédito (SCC). Com os dados disponíveis partiu-se para a preparação dos mesmos conforme Subcapítulo a seguir.

4.3 Preparação dos Dados

Com a coleta inicial realizada o processo seguiu para a limpeza dos dados e seleção das compensações que importam para a análise e resposta aos questionamentos conforme entendimento do negócio. A manipulação dos dados foi realizada com o auxílio do software estatístico R (*R Project*) [49]. Alguns tratamentos básicos foram realizados para permitir a utilização dos dados para análise de variáveis e criação de modelos preditivos, dentre eles:

- Retirada de acentos
- Exclusão de linhas com poucas informações ou nulas

Tabela 4.1: Solicitações por tipo de crédito

Cofins embalagens (P.4, Art 51. Lei 10.833/03)	4
Cofins não cumulativo Exportação	524
Cofins não cumulativo Mercado interno	757
Contribuição previdenciária indevida ou a maior	189
IPI Residual	34
IRRF de cooperativas	104
IRRF de juros sobre capital próprio	39
Outros créditos	0
Pis/Pasep embalagens (P.4, Art 51. Lei 10.833/03)	4
Pis/Pasep não cumulativo Exportação	517
Pis/Pasep não cumulativo Mercado interno	750
Reintegra	1
Ressarcimento de IPI	2851
Retenção - Lei nº 9.711/98	2238
Salário-família/Salário-maternidade	1692
Saldo negativo de CSLL	1983
Saldo negativo de IRPJ	2998

- Transformação de tipos de dados

Os dados recuperados representavam observações relacionadas a compensações trabalhadas tanto manualmente quanto automaticamente pelo SCC. Tendo em vista que o objetivo é melhorar a seleção dos processos que serão trabalhados pelos Auditores-Fiscais um a um, foram removidos aquelas famílias que tiveram pedido deferido/indeferido automaticamente, restando somente as que tiveram intervenção humana.

Uma decisão importante, que levou de volta ao entendimento do negócio e dos dados foi o ponto de corte para definir se uma família de compensações foi indeferida ou não.

Tendo em vista que se realizou uma junção de todas as solicitações de compensação de um mesmo crédito, algumas dessas podem ter sido deferidas (a compensação foi aceita) quanto indeferidas. A análise pode alterar para cada limite de parâmetro deferido, o que nos leva a uma decisão de considerar acima de 70% de indeferimento como uma família indeferida e abaixo de 70% como uma família deferida. O limite foi corroborado pelos especialistas em compensação de crédito que entenderam ser mais conservador num primeiro momento, podendo ajustar caso o número de predições de solicitações a serem indeferidas fosse baixo.

Finalmente, para permitir uma análise estatística mais confiável, foram removidas as solicitações de compensações de tipos de créditos que apresentavam quantidades pouco significativas de acordo com o gráfico da Figura 4.1. Desse modo, foram removidas todas as solicitações por tipo de crédito com menos de 200 observações. Com a remoção, os

tipos de crédito que permanecem, com as respectivas quantidades podem ser vistas na Figura 4.2.

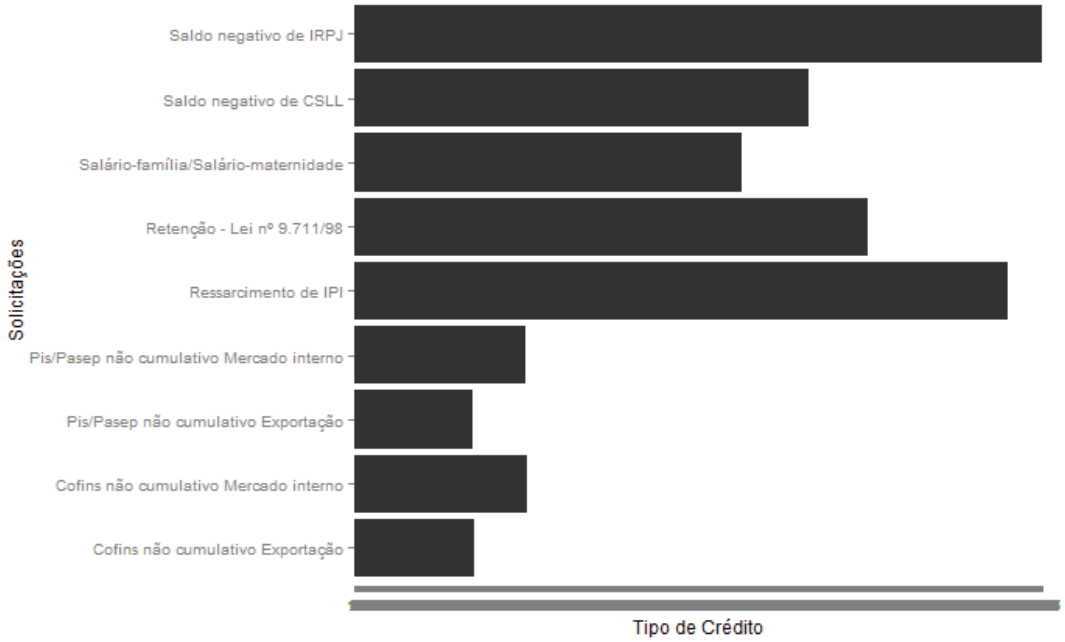

Figura 4.2: Solicitações por Tipo de Crédito

Os valores das compensações por tipo de crédito (tributo) não se dão de modo uniforme, se concentrando claramente nos saldos negativos de imposto de renda de pessoa jurídica, conforme Figura 4.3.

4.4 Modelagem

Na modelagem são iniciadas as respostas aos questionamentos elencados na fase de entendimento do negócio. Primeiramente uma avaliação de importância das características foi realizada utilizando Regressão Logística. De posse de informações das variáveis mais importantes, a criação dos modelos preditivos utilizando se iniciou utilizando métodos de Regressão Logística, Naïve Bayes e *Random Forests* (algoritmo que utiliza árvores de classificação).

1. Análise de variáveis: Com os dados preparados inicia-se a análise de variáveis (ou características do contribuinte) que são estatisticamente mais importantes para definir se uma família de compensações é indeferida ou não. Para extrair esses conhecimentos a partir da base classificada (indeferido/ deferido) utilizando a Regressão Logística. Essa técnica de regressão considera as interferências de um conjunto de

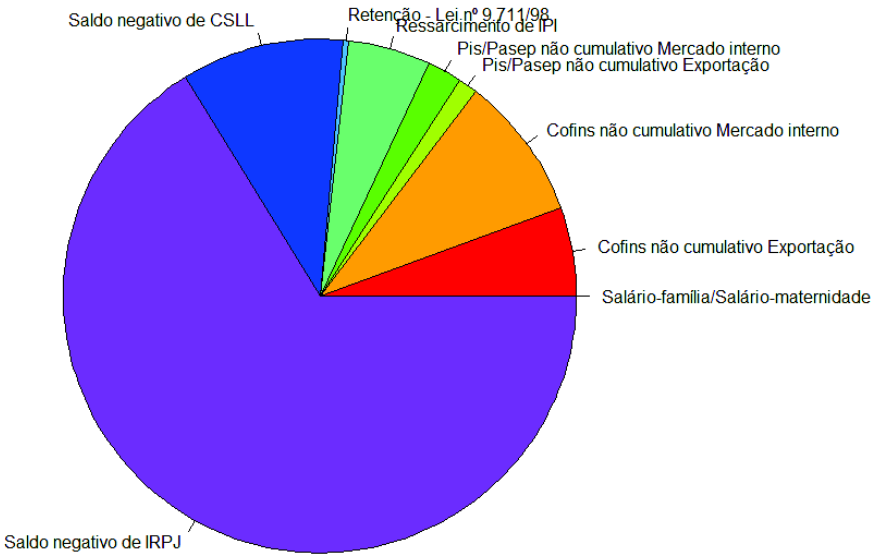

Figura 4.3: Valor das Solicitações por Tipo de Crédito

variáveis explanatórias para determinar uma variável dependente dicotômica, ou seja, que possui somente dois valores. As variáveis explanatórias no nosso modelos são todas as características do contribuinte e informações consideradas possivelmente relevantes da solicitação de compensação, enquanto a variável dependente é o indeferimento da família de compensações, que pode ser 0 (deferida) ou 1 (indeferida). A partir dessa análise, das 8 (oito) variáveis previamente selecionadas, somente 4 (quatro) delas se mostraram estatisticamente significativas para explicar o modelo. Essas características foram escolhidas e outras descartadas para iniciar a construção de modelos preditivos.

2. Modelos preditivos: A partir da análise das variáveis mais importantes, partiu-se para a confecção dos modelos preditivos com o fim de melhorar a seleção de processos de compensação de crédito tributário. As ferramentas escolhidas foram a própria Regressão Logística, Naive Bayes e *Random Forests*.

Primeiramente, foi realizada a Regressão Logística inicial para avaliar quais variáveis não tem influência no deferimento ou indeferimento de uma solicitação de compensação de crédito. Do total de variáveis e características do contribuinte, foram escolhidas 8 (oito) delas para iniciar as análise de importância e significância estatística e para construção dos modelos. Por questões de sigilo as características dos contribuintes são apresentadas de forma descaracterizada, tendo em vista que o resultado do trabalho poderá ser utilizado pela Secretaria da Receita Federal do Brasil para predição e seleção de processos de compensação.

Uma vez removidas as variáveis independentes menos importantes, para todos os modelos a base foi separada em duas, uma de treinamento e outra para testes na proporção 80 para 20%. Para garantir uma boa escolha em cada algoritmo utilizado, e para dispensar a criação de uma base de validação, a base de treinamento foi utilizada para realizar a chamada validação cruzada (*cross-validation*), que permite que uma mesma base seja usada tanto para treinamento quanto para validação. Foi utilizada uma validação cruzada de 10 configurações de base diferentes (*10-fold cross-validation*).

Todos os algoritmos de classificação para criação dos modelos preditivos utilizaram o pacote *caret* do *R Project*. Na verdade, o *caret* é um "meta-pacote"e se utiliza de outros pacotes com as implementações dos classificadores, mas facilita sobremaneira a utilização de vários métodos a partir da utilização da função *train*. Basicamente altera-se somente o parâmetro *method* para escolher o algoritmo preditor. Os métodos e pacotes utilizados foram:

- Regressão Logística: pacote *caTools* que implementa o algoritmo *Logit Boost* criado por Jerome Friedman et al. [23]
- Naïve Bayes: pacote *bnclassify* que implementa o algoritmo padrão apresentado em [41]
- Árvores de Classificação (*Random Forests*): pacote *randomForest* que implementa o algoritmo de *Random Forests* de Breiman [4]

Após o treinamento de cada algoritmo, os resultados foram confrontados com a base de testes. Em todos os casos foram geradas as matrizes de confusão e calculados a sensibilidade, especificidade, acurácia e *f-measure*. Os resultados das matrizes de confusão são apresentados nas Tabelas 4.2, 4.3 e 4.4. Os cálculos de especificidade, acurácia, sensibilidade e *f-measure* estão na Tabela 4.5.

Tabela 4.2: Matriz de confusão - Naive Bayes

		Referência	
		Deferido	Indeferido
Predição	Deferido	545	234
	Indeferido	239	541

Tabela 4.3: Matriz de confusão - Regressão Logística

		Referência	
		Deferido	Indeferido
Predição	Deferido	443	171
	Indeferido	341	604

Tabela 4.4: Matriz de confusão - *Random Forests*

		Referência	
		Deferido	Indeferido
Predição	Deferido	590	295
	Indeferido	195	479

Tabela 4.5: Resultados modelos preditivos

	Acurácia	Espec.	Sensib.	*F-measure*
Naive Bayes	0.6966	0.6981	0.6952	0.6973
Reg. Logística	0.6716	0.7794	0.5651	0.6337
Random Forests	0.6857	0.6194	0.7513	0.7062

Tendo em vista que foram realizadas diversas análises durante o primeiro ano do mestrado, fora detalhadas as análises realizadas inicialmente para que, na subseção de avaliação 4.5, sejam apresentadas outras análises que foram confeccionadas até a presente etapa das pesquisas realizadas.

4.5 Avaliação

A Regressão Logística se mostrou uma boa ferramenta para identificar as variáveis mais importantes na predição, pois foram realizados testes com todas as 8 (oito) variáveis iniciais e os resultados para os três algoritmos de modelos preditivos não passou de 0.53 de acurácia. Com as variáveis escolhidas com a regressão, o melhor modelo para acurácia apresentou 0.69 (Naive Bayes) e o melhor para *f-score* (*f-measure*) de 0.70 (*Random Forests*). Desse modo, a escolha de selecionar as variáveis antes de criar os modelos preditivos se mostrou uma prática importante na inferência do indeferimento de processos de compensação.

Os modelos preditivos tiveram uma performance bastante similar, com vantagem para o Naive Bayes com relação à Acurácia, Regressão Logística para especificidade e *Random Forests* para Sensibilidade e *F-measure*. Uma característica importante para o negócio, no entanto, é avaliar qual a percentagem de processos que deveriam ser indeferidos foram realmente preditos de forma correta, ou seja, o índice de predição de negativos é essencial para a seleção de processos de compensação de crédito tributário. Isto porque é menos custoso trabalhar um processo como predito indeferido e no fim ele ser na realidade deferido do que o contrário. Deferir processos que deveriam ser indeferidos é um erro mais importante. O *F-measure* auxilia nessa análise, mas dá o mesmo peso para acertos de ambas as classes, deferido e indeferido. Nesse quesito, o modelo que utiliza o algoritmo de *Random Forests* é o mais adequado, tendo em vista que a predição de número negativos é de 0.71, maior que os 0.63 da Regressão Logística e 0.69 do algoritmo Naïve Bayes.

Uma predição de mais de 70% de quais processos de compensação deveriam ser indeferidos é muito promissora para o problema em questão. Muitos auditores-fiscais analisam grandes números de processos manualmente e, em grande maioria, estes são deferidos, ou seja, poderiam ser homologados de forma automática. O trabalho apresentou, portanto, resultados interessantes que respondem bem às questões do entendimento do negócio.

Durante a realização desses trabalhos notou-se que algumas alterações poderiam ser feitas para melhorar a aplicabilidade dos métodos utilizados e das premissas e decisões tomadas no início do mestrado. Pretendeu-se buscar algo além do que a acurácia na predição, que não deixa de ser importante, mas não leva em consideração os valores solicitados nas compensações. Para melhorar o processo de trabalho, deve-se obter modelos preditivos que, não só acertem quais seriam as solicitações deferidas ou indeferidas, mas também que efetivamente aumentasse o valor total indeferido. Do contrário, para que realizar trabalho árduo de analisar grandes massas de dados, criar modelos, se a escolha simples baseada somente no valor for mais eficiente? Noutras palavras, se tenho 1000 (mil) processos para analisar e capacidade para trabalhar 500 (quinhentos), preciso de algo que seja melhor que escolher os processos com base no valor.

Ademais, percebeu-se que uma análise conjunta de todos as solicitações de compensação, poderiam esconder padrões interessantes que ocorrem somente para determinado tipo de crédito. Conforme apresentado, os créditos que os contribuintes tem com a RFB podem ter várias fontes, desde um pagamento a maior que deve ser devolvido, até crédito de Imposto Sobre a Renda de Pessoa Jurídica que foi aferida acima do que o contribuinte devia.

Desse modo, os próximos Subcapítulos apresentam algumas análises que foram feitas para buscar uma melhoria nos modelos preditivos, realizando análises por tipo de crédito, bem como a definição de uma análise multi-critério a partir da criação de indicador para melhor avaliar a performance da seleção de solicitação de compensação. Essa análise permite comparar o processo atual de seleção com aquele proposto pelos modelos preditivos.

4.5.1 Critério para avaliação de performance

Além dos indicadores utilizados para selecionar o melhor modelo preditivo (acurácia, sensibilidade, especificidade e f-score), precisamos verificar se a seleção de solicitações de compensação realizada a partir dos modelos é melhor que o modo de seleção atual.

Uma boa forma de se avaliar a efetividade da seleção de solicitações de compensação é calcular o valor indeferido por solicitação trabalhada manualmente. Isso traz para a análise mais um critério para escolha: o valor das compensações solicitadas. Para comparar os métodos de escolha foram calculados dos indicadores da seguinte forma:

- Calculou-se o valor de crédito indeferido por solicitação analisada
- Criou-se, com base nos processos trabalhos, um modelo preditivo que vai predizer se uma solicitação de compensação de crédito vai ser deferida ou indeferida
- Uma simulação de seleção de solicitações foi realizada de modo a determinar quais seriam e quais não seriam trabalhadas de acordo com a decisão do modelo preditivo. O modelo indicou o que seria ou não escolhido para análise dos Auditores-Fiscais
- Foi medido o valor de crédito indeferido por solicitação analisada a partir da seleção do modelo preditivo

De posse de um indicador para medir a performance do processo pode-se mensurar como se dão os resultados com o processo atual e com uma simulação do processo com propostas de melhorias. O próximo passo foi realizar a mensuração dos indicadores utilizando modelos preditivos para cada tipo de crédito e avaliando se a seleção a partir deles foi melhor que o processo atual (escolha pelo maior valor).

4.5.2 Análise por Tipo de Crédito Tributário

Para testar o indicador criado, apartamos as solicitações por tipo de crédito, pois essa é a forma com que os processos são analisados na prática nas unidades da RFB. Aproveitou-se o resultado do modelo preditivo de Regressão Logística, por ter sido o melhor dos algoritmos utilizados, mesmo que não estatisticamente pelo fato de os intervalos de confiança se mostraram sobrepostos. Para a RFB, o interessante é selecionar aquelas solicitações de crédito que tem mais chances de estarem erradas, ou seja, aquelas em que o contribuinte não tem o direito de compensar aquele crédito, ou ainda pior, nem sequer deveria ter aquele crédito junto à RFB. Desse modo, foram realizadas simulações em que alguns processos deixariam de ser trabalhados e, com base nessa estimativa (que não pode ser divulgada por sigilo funcional) comparamos a seleção de processos analisados a partir dos modelos preditivos e a seleção somente baseada no valor das compensações.

A simulação levou em conta a análise de 1084 (mil e oitenta e quatro) solicitações de créditos relativos a PIS/PASEP e Cofins relativos a mercado interno [57] que foram analisadas manualmente por Auditores-Fiscais da RFB da 1a Região Fiscal (DF, GO, TO, MT e MS). Com essa base é possível simular uma seleção de solicitações baseadas em técnicas de mineração de dados, mais especificamente a criação de modelos preditivos que dizem o risco de uma solicitação ser ou não deferida.

Para comparar efetivamente se os modelos preditivos podem ser utilizados na prática, temos que realizar o cálculo dos indicadores do processo original e das predições. Vários

Tabela 4.6: Comparativo entre processo atual e proposto

Tipo de Crédito	Processo Atual	Modelo Preditivo
PIS/Cofins Interno	R$526.351,80/processo	R$618.541,60/processo

dos modelos preditivos tiveram resultados piores que o processo atual, com base no indicador utilizado, mas outros tiveram resultados superiores, podendo ser utilizados em situações reais para validar na prática o modelo encontrado. Destaca-se na Tabela 4.6 o modelo preditivo que teve o melhor resultado dentre as análises:

A efetividade aumenta porque nem todos os 1084 processos são trabalhados, mas somente aqueles que são indicados pelo modelo como tendo alto risco de serem indeferidos (conforme apresentado anteriormente, o modelo só possui duas predições: deferido ou indeferido).

Desse modo, a implementação dos modelos preditivos para a seleção de solicitações de compensação de crédito na RFB foi sugerida aos gestores do processo de Gerir o Crédito Tributário. Com o resultado, há uma grande expectativa com relação à melhoria do processo de trabalho ora avaliado e as técnicas utilizadas pelo presente trabalho se mostraram desde já bastante interessantes, conforme apresentado no próximo subcapítulo.

4.6 Implementação

Qualquer alteração em processos que levam a impactos na arrecadação federal deve ser bastante estudada para que os novos procedimentos não resultem em queda no valor arrecadado. No caso de compensações de crédito, ao invés de arrecadar, a administração tributária anula um crédito que tem com o contribuinte, ou seja, deixa de recolher aos cofres públicos aquele valor.

De fato, é compreensível aos gestores avaliarem os riscos e vantagens de uma alteração considerável nos processos de trabalho. Nesse sentido, foram identificados os principais riscos da implementação e explicitadas algumas simulações de como seria a utilização dos modelos preditivos para integrar a carteira de indicadores de risco no momento de decidir se uma solicitação de compensação será selecionada para análise manual ou não.

Supondo que o resultado do modelos preditivos apresentados como solução proposta seja satisfatório e seja possível implantar uma nova sistemática de seleção de processos de compensação de crédito que serão analisados manualmente para implantação na Receita Federal. O modelo impõe metas de análise de processos devido ao valor solicitado de compensações, ou seja, as compensações de maior valor serão analisadas em detrimento das de menor valor.

Desse modo, é possível haver uma desconfiança dos gestores com relação ao atual modelo, pois alguns processos de menor valor podem ser escolhidos para análise em detrimento de um de maior. Considerando a situação, podemos criar uma árvore de decisão para um caso hipotético que poderia representar as decisões dos gestores na hora de selecionar quais processos serão trabalhados manualmente.

Numa seleção de processos real existem mais trabalho do que pessoas para proferir com a análise. Se imaginar-se que o estoque de processos seja de 3 e tem-se capacidade de analisar somente um deles conforme listagem abaixo:

- Processo A de solicitação de compensação de R$1.000.000
- Processo B de solicitação de compensação de R$750.000
- Processo C de solicitação de compensação de R$500.000

O modelo atual trabalharia somente o processo de maior valor, mesmo que seu risco de indeferimento (crédito não concedido ao contribuinte) seja menor. Pois se existir modelo preditivo que consiga estimar as probabilidades de indeferimento conforme se segue:

- Processo A tem 30% de chances de ser indeferido
- Processo B tem 50% de chances de ser indeferidos
- Processo C tem 75% de chances de ser indeferido

Se o processo A é escolhido, automaticamente a RFB vai deferir os outros dois processos automaticamente, ou seja, processo B e C terão crédito concedido e esse valor saíra dos cofres públicos. Desse modo, tem-se que a árvore de decisão conforme Figura 4.4.

Desse modo, tem-se que o crédito líquido para cada alternativa é a que segue:

- E(A) = (R$300.000 – R$750.000 – R$500.000) + (-R$700.000 – R$750.000 – R$500.000) = -R$2.900.000
- E(B) = (R$450.000 – R$1.000.000 – R$500.000) + (-R$300.000 – R$1.000.000 – R$500.000) = -R$2.850.000
- E(C) = (R$375.000 – R$1.000.000 – R$750.000) + (-R$125.000 – R$1.000.000 – R$750.000) = -R$3.250.000

Dessa forma, a melhor escolha é deixar de analisar os processos A e C para escolher o B, ou seja, nem sempre os processos de maior valor devem ser escolhidos em detrimento dos de menor valor.

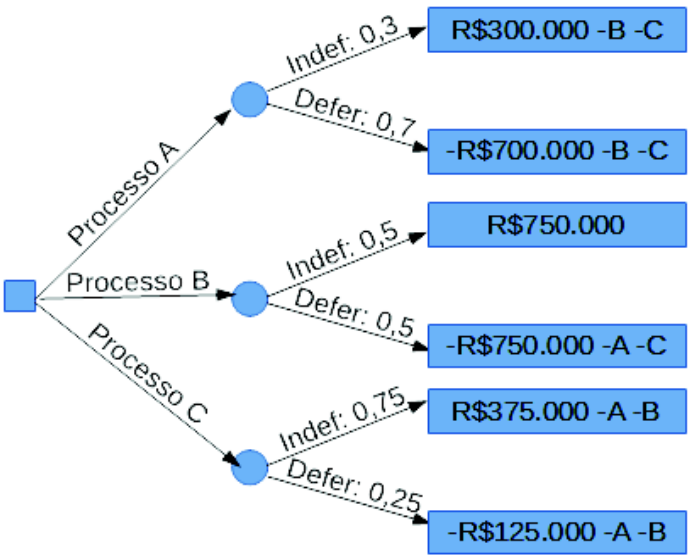

Figura 4.4: Avaliação de Riscos de Implementação - árvore de decisão

Após uma análise minuciosa dos passos realizados durante todo o processo de mineração de dados, os resultados de simulação do uso dos modelos preditivos foram apresentados para a administração da Secretaria da Receita Federal do Brasil e trabalho foi muito bem aceito. Os modelos preditivos serão utilizados pela Subsecretaria de Arrecadação e Cobrança (Suara/RFB) do órgão como um indicador de riscos de indeferimento das solicitações de compensação de crédito para todo o país. As estimativas mais conservadoras do retorno na utilização dos métodos e modelos apresentados pelo presente trabalho é da ordem de milhões de reais, conforme detalhamos no último Capítulo (5).

As estimativas poderão ser analisadas à medida em que os processos de solicitação de compensação de crédito de maior risco, e selecionados de acordo com os modelos preditivos, forem efetivamente analisados manualmente por Auditores-Fiscais. No próximo Capítulo apresenta-se como esses resultados futuros poderão influenciar em trabalhos futuros.

O Capítulo 4 apresenta a solução proposta para a criação de modelos preditivos para a seleção de solicitações de compensação de crédito tributário, detalhando a metodologia utilizada, os passos executados em cada etapa da análise de dados até se chegar aos resultados preliminares.

Este Capítulo traz a solução para o problema objeto do mestrado. Primeiramente foi apresentada a metodologia científica em que se baseou o trabalho, bem como quais foram as técnicas utilizadas para criação dos modelos preditivos. Para cada fase do processo, os passos realizados são detalhados.

O Capítulo 5 apresenta a conclusão da dissertação e apresenta os trabalhos futuros que poderão ser realizados a partir da etapa atual do projeto.

Capítulo 5

Conclusão e Trabalhos Futuros

As administrações tributárias são órgãos essenciais ao estado. O provimento de recursos por meio de coleta de tributos é primordial para qualquer sociedade ter condições de receber dos governos os produtos de políticas públicas. O recolhimento de tributos não pode, no entanto, se dar de qualquer forma, pois deve se ater a princípios básicos de justiça fiscal e capacidade contributiva. A seleção do que será auditado e o que não será é uma das principais fontes de possíveis injustiças, pois os contribuintes que cumprem com suas obrigações tributárias devem ter seu relacionamento com a administração tributária facilitada enquanto aquele que não cumpre, deve ser selecionado e fiscalizado com o rigor necessário.

De fato, de acordo com estudos e levantamentos da Organização para Colaboração do Desenvolvimento Econômico (OCDE) em [43], e a partir de colaborações de diversas administrações tributárias, apresenta uma pirâmide de classificação dos contribuintes com relação ao cumprimento de obrigações tributárias e quais devem ser as medidas adotadas para tratamento de cada um. A pirâmide pode ser vista na Figura 5.1 e mostra a importância de tratar os contribuintes da forma devida.

A compensação de crédito tributário tem um grande impacto, tanto positivo, quanto possivelmente negativo para os contribuintes brasileiros. Isto porque homologar compensações de crédito que são indevidas podem gerar grandes distorções econômicas: o contribuinte que a realiza a compensação indevida de forma intencional, tem vantagens de competitividade com relação àquele que segue a risca o cumprimento de suas obrigações. Com o voluma de compensações aumentando bastante, maior é a responsabilidade da Secretaria da Receita Federal do Brasil (RFB) em saber separar aquelas compensações corretas das incorretas.

Nesse sentido, o trabalho propôs uma melhoria na gestão de riscos para o processo de gestão do crédito tributário e, dentre todos os seus subprocessos, focalizou na seleção das solicitações de compensação de crédito tributário por encontrar nele, uma forma de

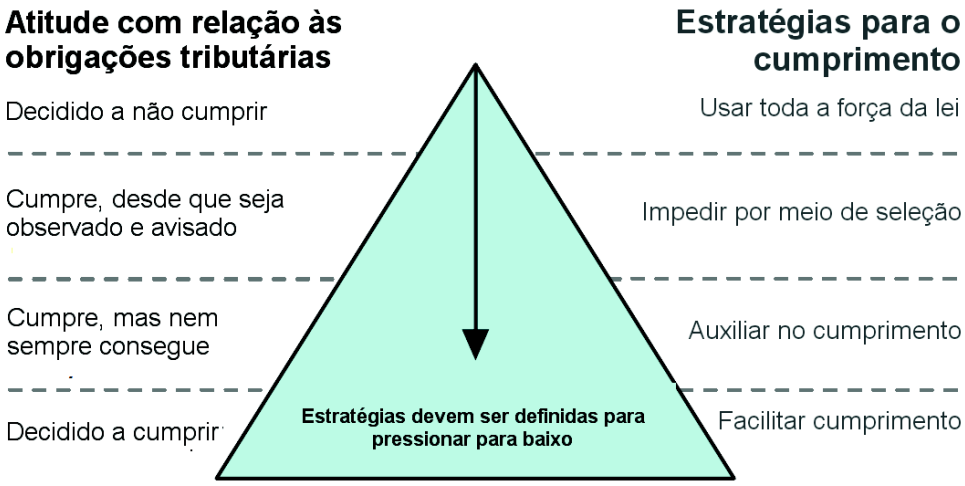

Figura 5.1: Pirâmide OCDE (livre tradução pelo autor)[43]

atingir uma maior justiça fiscal e ainda tentar garantir que a arrecadação não sofra com corrosões de compensações de valores indevidos.

O volume de compensações, instituídos por declarações eletrônicas no sistema Pedido Eletrônico de Ressarcimento, Restituição, Reembolso e Declaração de Compensação (PER/DCOMP), é muito grande para que seja possível a análise manual dos Auditores-Fiscais da RFB, o que seria o ideal para garantir que compensações indevidas sempre seriam auditadas. Diante dessa impossibilidade, e tendo em vista os grandes avanços computacionais nas últimas décadas, a utilização de técnicas de mineração de dados se torna primordial para análise de tendências e padrões com o fim de melhorar a seleção de trabalhos. Uma grande vantagem do subprocesso escolhido é a possibilidade de extrair informações de bases de compensações já trabalhadas pelos Auditores-Fiscal e, a partir de algoritmos específicos, criar modelos que possam aperfeiçoar o entendimento dos riscos de uma solicitação ainda não auditada ser devida ou não.

Os resultados das análises são promissores, conforme simulações para alguns tipos de crédito, mas indicaram que é necessário mais um ciclo do CRISP-DM, revisitando o entendimento do negócio para avaliar se o problema pode ser modelado com faixas de riscos ao invés de indicar de modo dicotômico se uma família de compensações deve ser ou não homologada. Ademais, a inclusão no modelo de penalizações para sensibilidade e levando em consideração o valor das compensações seria interessante para tornar sua aplicação mais direta. Isto porque verificou-se a necessidade de criar um indicador nesse sentido, o que mostrou uma deficiência no modelo. A partir dos resultados dos modelos preditivos, podemos inferir o quanto de arrecadação que seria mantida a partir da não homologação de compensações indevidas que seriam deixadas de ser auditadas se o método

de escolha não fosse realizado com o proposto neste trabalho.

Conforme apresentado no Capítulo 4, a melhoria na seleção das solicitações de compensação para o crédito originário de PIS/PASEP Mercado Interno foi de 17,5%. A partir de informações internas do Sistema de Controle de Crédito (SCC), temos que cerca de 20% dos valores processados pelo sistema são separados para análise manual dos Auditores-Fiscais. Em valores de 2014 o volume total foi de R$84 bilhões, ou seja, aproximadamente R$16,8 bilhões foram analisados manualmente. Se 45% desse valor fosse não homologado (proporção da amostra da 1ª Região Fiscal), uma melhoria na performance dos valores não homologados de 1 (um) ponto percentual se reverte em R$ 75 milhões de reais anuais em arrecadação.

Obviamente os modelos podem não ser tão efetivos para todos os tipos de crédito, mas o baixo gasto para implementação das análises e ferramentas utilizadas e propostas, combinadas a seu grande potencial, faz com que o custo/benefício seja bastante vantajoso para a RFB, sugerindo que pode-se utilizar dessas novas possibilidades da tecnologia e das ferramentas de análise para encontrar outras formas de selecionar o que deve ser trabalhado.

Com relação aos trabalhos futuros, a base utilizada nas análises apresentadas foi restrita a uma região fiscal, o que pode enviesar os resultados dos modelos preditivos, não só pelas características dos contribuintes da região, mas também pelo número reduzido de solicitação de compensações tributárias avaliadas. Um primeiro passo, portanto, seria a análise de uma base nacional, ou de uma grande região fiscal, que inclua a unidade federativa de São Paulo, por exemplo. O acesso à totalidade das informações é bastante complexa, no entanto, tendo em vista que a RFB possui muitas restrições contratuais e tecnológicas com seu principal prestador de serviços de TI, Serviço Federal de Processamento de Dados (Serpro) e a extração de bases se dá por meio de consultas limitadas a *datawarehouses* e sistemas online.

Outras análises que podem ser feitas são regressões logísticas em bases separadas por tipo de crédito. Desse modo variáveis que não apresentaram importância na base como um todo, pode ser mostrar estatisticamente relevante para casos específicos de diferentes tipos de crédito. Uma engenharia de atributos (*feature engineering*) também poderia ser realizada para contemplar mais indicadores de riscos a partir de informações que não foram incluídas nas análises, como movimentação financeira e dados de notas fiscais eletrônicas. Tendo em vista a facilidade de criação de modelos preditivos utilizando diferentes algoritmos, deve-se utilizar outras ferramentas de mineração de dados para tentar alcançar melhores resultados de predição.

Ademais, seria de grande interesse a comparação dos modelos preditivos com a seleção feita por critérios escolhidos por especialistas no assunto. Alguns estudos estão sendo

realizados na administração tributária brasileira para definir os principais critérios na inferência se um processo de compensação será indeferido ou não. O método utilizado para a determinação das variáveis é o análise hierárquica de processos (AHP), onde questionários são elaborados por conhecedores da técnica e respondidos por especialistas no assunto em que se quer avaliar.

Os resultados do trabalho, portanto, se mostraram de interesse à Secretaria da Receita Federal do Brasil (RFB). Após um ou dois ciclos CRISP-DM, espera-se ser possível utilizar os modelos criados para melhorar a gestão de riscos do crédito tributário, mais especificamente para melhor selecionar as solicitações de compensação de crédito tributário. Como contribuição tecnológica esperada, conforme descrito no Capítulo 1, o trabalho também se mostrou promissor, tendo seus resultados detalhados em artigo científico intitulado *"Predictive Models on Tax Refund Claims-Essays of Data Mining in Brazilian Tax Administration"* [17] submetido e aceito no 4^{th} *International Conference on Electronic Government and the Information Systems Perspective - EGOVIS 2015*, parte da 26^{th} *DEXA Conference* (Qualis/Capes B1) que ocorreu em Valência, Espanha, de 1º a 4 de setembro de 2015.

Referências

[1] Serviço federal de processamento de dados (serpro) - sítio principal. https://www.serpro.gov.br/tema/noticias-tema/de-volta-para-o-contribuinte. (Acessado em 09/12/2014). 6

[2] Decision tree learning, July 2015. Page Version ID: 673463734. 29

[3] Luciano Amaro. Direito tributário. *NÚCLEO*, 30:1, 2009. 1

[4] Leo Breiman. Random forests. *Machine learning*, 45(1):5–32, 2001. 46

[5] Leo Breiman, Jerome Friedman, Charles J Stone, e Richard A Olshen. *Classification and regression trees*. CRC press, 1984. 28, 37

[6] Brian Caffo. *Regression Models for Data Science in R*. Leanpub, February 2015. 22

[7] Rommel N Carvalho, Leonardo Sales, Henrique A Da Rocha, e Gilson Libório Mendes. Using bayesian networks to identify and prevent split purchases in brazil. In *BMAW UAI*, pages 70–78, 2014. 30, 31

[8] Pamela Castellón González e Juan D Velásquez. Characterization and detection of taxpayers with false invoices using data mining techniques. *Expert Systems with Applications*, 40(5):1427–1436, 2013. 29, 36

[9] Pete Chapman, Julian Clinton, Randy Kerber, Thomas Khabaza, Thomas Reinartz, Colin Shearer, e Rudiger Wirth. Crisp-dm 1.0 step-by-step data mining guide. 2000. x, 19, 20, 21, 39

[10] CIAT. Ciat: Inter american center of tax administrations main page. http://www.ciat.org/, 2015. (Acessado em 20/06/2015). 13

[11] Eder Daniel Corvalão et al. Classificação de contribuintes: um modelo em duas fases. 2009. 37

[12] WR Cory, K Michael Reynolds, RF DeMara, M Georgiopoulos, A Gonzalez, e R Eaglin. Tracking dirty proceeds: exploring data mining technologies as tools to investigate money laundering. *Police Practice and Research*, 4:163–178. 36

[13] Presidência da República. L5172. http://www.planalto.gov.br/ccivil_03/leis/L5172.htm, 1966. (Acessado em 07/05/2016). 12, 14

[14] Presidência da República. *Constituição (1988)*. Constituição da República Federativa do Brasil. Senado, Brasília, 1988. Disponível em: `http://www.planalto.gov.br/ccivil_03/constituicao/constituicao.htm`, acessado em 01 de outubro de 2014. 1

[15] Presidência da República. L8383. `http://www.planalto.gov.br/ccivil_03/leis/L8383.htm`, 1991. (Acessado em 07/05/2016). 14

[16] Presidência da República. L9430. `http://www.planalto.gov.br/ccivil_03/leis/L9430.htm`, 1996. (Acessado em 07/05/2016). 14

[17] Leon Sólon da Silva, Rommel Novaes Carvalho, e João Carlos Felix Souza. Predictive models on tax refund claims-essays of data mining in brazilian tax administration. In *International Conference on Electronic Government and the Information Systems Perspective*, pages 220–228. Springer, 2015. 57

[18] Leon Solon da Silva, Henrique de Carvalho Rigitano, Rommel Novaes Carvalho, e Joao Carlos Felix Souza. Bayesian networks on income tax audit selection - a case study of brazilian tax administration. In *Bayesian Modeling Application Workshop (BMAW)*, 2016. 37

[19] Joan Domenech. File:esquema-kfold.jpg - wikimedia commons. `https://commons.wikimedia.org/wiki/File:Esquema-kfold.jpg`. (Acessado em 05/04/2016). x, 35

[20] Luciano A. Digiampietri et al. Conference paper: Uses of artificial intelligence in the brazilian customs fraud detection system. 2008. 37

[21] Usama Fayyad, Gregory Piatetsky-Shapiro, e Padhraic Smyth. From data mining to knowledge discovery in databases. *AI magazine*, 17(3):37, 1996. 18

[22] Xavier et al. Forns. Identification of chronic hepatitis C patients without hepatic fibrosis by a simple predictive model. *Hepatology*, 36(4):986–992, October 2002. 23

[23] Jerome Friedman, Trevor Hastie, e Robert Tibshirani. Additive logistic regression: a statistical view of boosting. *Annals of Statistics*, 28:2000, 1998. 46

[24] Olivier Gevaert, Frank De Smet, Dirk Timmerman, Yves Moreau, e Bart De Moor. Predicting the prognosis of breast cancer by integrating clinical and microarray data with bayesian networks. *Bioinformatics*, 22(14):e184–e190, 2006. 30

[25] Fabio Giambiagi e Ana Cláudia Duarte de Além. *Finanças públicas: teoria e prática no Brasil*. Elsevier Brasil, 2008. 1

[26] Antonio Carlos Gil. Como elaborar projetos de pesquisa. *São Paulo*, 5:61, 2002. 16, 17

[27] O Globo. Reportagem - levy diz que maioria das empresas não gosta de pagar impostos no brasil. `http://oglobo.globo.com/economia/levy-diz-que-maioria-das-empresas-nao-gosta-de-pagar-impostos-no-brasil-15735691`, 2016. (Acessado em 07/05/2016). 12

[28] Trevor Hastie, Robert Tibshirani, Jerome Friedman, e James Franklin. The elements of statistical learning: data mining, inference and prediction. *The Mathematical Intelligencer*, 27(2):83–85, 2005. 21, 22

[29] David W Hosmer Jr e Stanley Lemeshow. *Applied logistic regression*. John Wiley & Sons, 2004. 23

[30] IBGE. Instituto brasileiro de geografia e estatística (ibge) - sítio principal. `http://www.ibge.gov.br`, 2014. (Acessado em 09/12/2014). 2

[31] Project Management Institute. *A Guide to the Project Management Body of Knowledge: PMBOK Guide*. PMI Standard. Project Management Institute, 2013. 10

[32] Jorge Jambeiro Filho e Jacques Wainer. Hpb: A model for handling bn nodes with high cardinality parents. *Journal of Machine Learning Research (JMLR)*, 9:2141–2170, 2008. 37

[33] Gareth James, Daniela Witten, Trevor Hastie, e Robert Tibshirani. *An introduction to statistical learning*, volume 112. Springer, 2013. x, 24, 25, 26, 27, 28

[34] Ronald Jansen, Haiyuan Yu, Dov Greenbaum, Yuval Kluger, Nevan J Krogan, Sambath Chung, Andrew Emili, Michael Snyder, Jack F Greenblatt, e Mark Gerstein. A bayesian networks approach for predicting protein-protein interactions from genomic data. *Science*, 302(5644):449–453, 2003. 30

[35] Liangxiao Jiang, Harry Zhang, e Zhihua Cai. A novel bayes model: Hidden naive bayes. *Knowledge and Data Engineering, IEEE Transactions on*, 21(10):1361–1371, 2009. 31

[36] Ron Kohavi et al. A study of cross-validation and bootstrap for accuracy estimation and model selection. In *Ijcai*, volume 14, pages 1137–1145, 1995. 34

[37] Teuvo Kohonen. The self-organizing map. *Proceedings of the IEEE*, 78(9):1464–1480, 1990. 37

[38] Kevin B Korb e Ann E Nicholson. *Bayesian artificial intelligence*. CRC press, 2010. 29

[39] Jani Martikainen et al. Data mining in tax administration-using analytics to enhance tax compliance. *Department of Information and Service Economy. Aalto University*, 2012. 37

[40] Thomas Martinetz, Klaus Schulten, et al. *A "neural-gas" network learns topologies*. University of Illinois at Urbana-Champaign, 1991. 37

[41] Marvin Minsky e Oliver G Selfridge. *Learning in random nets*. MIT Lincoln Laboratory, 1960. 46

[42] R. P. C. Morgan, D. D. V. Morgan, e H. J. Finney. A predictive model for the assessment of soil erosion risk. *Journal of Agricultural Engineering Research*, 30:245–253, 1984. 23

[43] OCDE. Tax administration 2013 - comparative information on oecd and other advanced and emerging economies. Technical Report 2308-7331, Organisation for Economic Co-operation and Development, Paris, 2013. x, 34, 35, 36, 54, 55

[44] OCDE. Organização para a cooperação e desenvolvimento economico (ocde) main page. http://www.oecd.org/, 2015. (Acessado em 08/06/2015). 13, 35

[45] OMA. Organização mundial de aduanas (oma) - sítio principal. http://www.wcoomd.org/. (Acessado em 08/06/2015). 13, 35

[46] Gregory Piatetsky-Shapiro. Knowledge discovery in real databases: A report on the ijcai-89 workshop. *AI magazine*, 11(4):68, 1990. 18

[47] Tiago Leveergger Piccirilli. Mineração de dados aplicada a classificação dos contribuintes do iss. 2013. 37

[48] David Martin Powers. Evaluation: from precision, recall and f-measure to roc, informedness, markedness and correlation. 2011. 34

[49] R Project. R: The r project for statistical computing main page. https://www.r-project.org/, 2015. (Acessado em 08/06/2015). 42

[50] J Ross Quinlan. C4. 5: Programming for machine learning. *Morgan Kauffmann*, page 38, 1993. 28

[51] RFB. In srf nº21 - 1997. http://normas.receita.fazenda.gov.br/sijut2consulta/link.action?idAto=13301&visao=original, 1997. (Acessado em 07/05/2016). 14

[52] RFB. In srf nº73 - 1997. http://normas.receita.fazenda.gov.br/sijut2consulta/link.action?idAto=14243&visao=original, 1997. (Acessado em 07/05/2016). 14

[53] RFB. In srf nº210 - 2002. http://normas.receita.fazenda.gov.br/sijut2consulta/link.action?idAto=15083&visao=original, 2002. (Acessado em 07/05/2016). 14

[54] RFB. In srf nº320 - 2003. http://normas.receita.fazenda.gov.br/sijut2consulta/link.action?idAto=15210&visao=original, 2003. (Acessado em 07/05/2016). 14

[55] RFB. In rfb nº1300 - 2012. http://normas.receita.fazenda.gov.br/sijut2consulta/link.action?visao=compilado&idAto=38972, 2012. (Acessado em 07/05/2016). 14

[56] RFB. Secretaria da receita federal do brasil (rfb) - sítio principal. http://www.receita.fazenda.gov.br, 2014. (Acessado em 09/12/2014). 2, 9

[57] RFB. Contribuição para o pis/pasep e cofins main page. http://www.receita.fazenda.gov.br/pessoajuridica/pispasepcofins/, 2015. (Acessado em 12/06/2015). 49

[58] RFB. Análise do resultado da arrecadação — secretaria da receita federal do brasil. http://idg.receita.fazenda.gov.br/dados/receitadata/arrecadacao/analise-do-resultado-da-arrecadacao/analise-do-resultado-da-arrecadacao, 2016. (Acessado em 07/05/2016). 13

[59] RFB. Quantitativo de cargos — secretaria da receita federal do brasil. http://idg.receita.fazenda.gov.br/acesso-a-informacao/servidores/quantitativo-de-cargos, 2016. (Acessado em 07/05/2016). 8

[60] Norton Trevisan Roman, Cristiano Ferreira, Luis Meira, Rodrigo Carvalho Rezende, Luciano Digiampietri, e Jorge Jambeiro Filho. Attribute-value specification in customs fraud detection: a human-aided approach. 2009. 37

[61] Stuart Jonathan Russell, Peter Norvig, John F Canny, Jitendra M Malik, e Douglas D Edwards. *Artificial intelligence: a modern approach*, volume 2. Prentice hall Upper Saddle River, 2003. 34

[62] Harry Zhang. The optimality of naive bayes. *AA*, 1(2):3, 2004. x, 30

[63] Fei Zheng e Geoffrey I Webb. Tree augmented naive bayes. In *Encyclopedia of Machine Learning*, pages 990–991. Springer, 2011. 30

[64] Wen Zhu, Nancy Zeng, Ning Wang, et al. Sensitivity, specificity, accuracy, associated confidence interval and roc analysis with practical sas® implementations. *NESUG proceedings: health care and life sciences, Baltimore, Maryland*, pages 1–9, 2010. 32, 33