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

Uso de modelos preditivos na gestão de riscos da Fiscalização Tributária

Fabíola C. Venturini

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

Orientador
Prof. Dr. Ricardo Chaim

Brasília
2020

Venturini, Fabiola Cristina
Vu Uso de modelos preditivos na gestão de riscos da Fiscalização Tributária / Fabiola Cristina Venturini; orientador Ricardo Matos Chaim. -- Brasília, .
p.

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

1. gestão de riscos. 2. indícios de irregularidades fiscais. 3. mineração de dados. 4. modelos preditivos. 5. redes neurais. I. Matos Chaim, Ricardo, orient. II. Título.

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

Uso de modelos preditivos na gestão de riscos da Fiscalização Tributária

Fabíola C. Venturini

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

Prof. Dr. Ricardo Chaim (Orientador)
CIC/UnB

Prof. Dr. Rosalvo Ermes Streit Prof. Dr. João Carlos Felix Souza

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

Brasília, 08 de outubro de 2020

Dedicatória

Dedico este trabalho a minha família, a meu esposo Jabes, por ser meu companheiro e sempre me incentivar, aos meus filhos Ricardo e Rodrigo, pela inspiração e luz que trazem para minha vida, a meus pais Dionizio (*in memorian*) e Jaci, por serem exemplos de perseverança, coragem, dedicação e honestidade e por todo amor que deles recebi e com eles aprendi a compartilhar.

Agradecimentos

Agradeço a todos cuja participação e conhecimentos contribuíram para realização deste trabalho:

Ao meu orientador, Ricardo Matos Chaim, pelo seu entusiasmo contagiante e por ter compartilhado comigo seus conhecimentos e me mostrado o caminho a seguir.

Aos meus colegas Auditores Fiscais da Receita do DF que me ajudaram a compreender melhor nossos processos de trabalho, pelas informações e pelo apoio que me foram dados.

Agradeço ainda aos professores do Programa de Pós-Graduação em Computação Aplicada da Universidade de Brasília (PPCA-UnB) pelo comprometimento, pela seriedade e dedicação ao Mestrado Profissional e a nós, alunos.

O presente trabalho foi realizado com apoio da Coordenação de Aperfeiçoamento de Pessoal de Nível Superior - Brasil (CAPES), por meio do Acesso ao Portal de Periódicos.

Resumo

Visando verificar de que forma a gestão de riscos da administração tributária do DF pode ser aprimorada por meio do uso de modelos preditivos de mineração de dados, o estudo partiu da caracterização do processo de gestão de riscos dos tributos indiretos pelo Fisco Distrital à luz das boas práticas de gestão de risco aplicáveis à Administração Pública local, envolveu a identificação de modelos preditivos aplicáveis ao processo de seleção de contribuintes a fiscalizar, bem como sua criação e avaliação, que levou em conta o resultado das ações fiscais já realizadas. Nesta pesquisa, exploratória e aplicada, a revisão teórica foi orientada por meio de uma pesquisa bibliométrica a qual resultou na escolha de modelos preditivos baseados em regressão logística e em redes neurais; reuniões foram realizadas para obter informações junto aos auditores responsáveis pela seleção das empresas a fiscalizar e um questionário foi aplicado para colher opiniões dos auditores responsáveis pelas auditorias; os dados armazenados em sistemas corporativos foram estudados e extraídos para obtenção das informações relativas às variáveis de interesse identificadas, as quais foram utilizadas na criação e treinamento dos modelos. Foram criados modelos preditivos capazes de mapear conjuntos de empresas que correspondem a aproximadamente metade das empresas auditadas e mais de 80% do crédito constituído (89% no caso da rede neural modelo). Desta forma, foi possível concluir que a utilização de modelos preditivos tem o potencial de otimizar a aplicação dos recursos disponíveis e maximizar os resultados alcançados.

Palavras-chave: Gestão de risco, indícios de irregularidades fiscais, avaliação de risco, mineração de dados, modelos preditivos, regressão logística, redes neurais

Abstract

With the objective of verifying how the use of predictive data mining models can improve the risk management of the tax administration of the Federal District, this study started from the characterization of the risk management process used by the District Tax Administration, followed by the identification of the models applicable, of its implementation and finally of its evaluation, which considered the result of the tax actions already carried out.

In this exploratory and applied research, the theoretical review was guided by a bibliometric research that resulted in the choice of predictive models based on logistic regression and neural networks; meetings were held to obtain information from the auditors responsible for selecting the companies to be inspected and a questionnaire was applied to collect the opinions of the auditors responsible for the audits; the data stored in corporate systems were studied and extracted to obtain the information related to the identified variables of interest, which were used in the creation and training of the models.

Predictive models were created capable of mapping sets of companies that correspond to approximately half of the audited companies and to more than 80 % of the constituted credit (89 % in the case of the neural network model). Thus, it was possible to conclude that the use of predictive models has the potential to optimize the application of available resources and maximize the results achieved.

Keywords: Risk management, evidence of tax irregularities, risk assessment , data mining, predictive models, logistic regression, neural networks

Sumário

Lista de Figuras

Lista de Tabelas

Lista de Abreviaturas e Siglas

AEC Auditoria Especial Concentrada.

CGDF Controladoria-Geral do Distrito Federal.

CGU Controladoria-Geral da União.

CNAE Classificação Nacional de Atividades Econômicas.

CNN Convolutional Neural Network ou ConvNet.

CTN Código Tributário Nacional.

DMSP Declaração Mensal de Serviços Prestados.

ENAP Escola Nacional de Administração Pública.

GEPRO Gerência de Programação Fiscal.

GIM Guia Informativa Mensal do ICMS.

ICMS Imposto sobre Operações relativas à Circulação de Mercadorias e Prestação de Serviços de Transporte Interestadual e Intermunicipal e de Comunicação.

ISS Imposto sobre Serviços de Qualquer Natureza.

KDD Knowlege Database Discovery.

LFE Livro Fiscal Eletrônico.

MP Ministério do Planejamento, Orçamento e Gestão.

NFCE Nota Fiscal ao Consumidor Eletrônica.

NFE Nota Fiscal Eletrônica.

OCDE Organização para a Cooperação e Desenvolvimento Econômico.

SEEC Secretaria de Estado de Economia do DF.

SPED Sistema Público de Escrituração Digital.

TCU Tribunal de Contas da União.

Capítulo 1

Introdução

A presente pesquisa permitiu verificar que modelos preditivos baseados em aprendizado de máquina podem ser utilizados para aperfeiçoar a etapa de seleção dos indícios de irregularidades a fiscalizar, contribuindo desta forma para o aprimoramento do processo de gestão de riscos da Fiscalização Tributária do Distrito Federal.

Inicialmente os processos de tratamento das informações fiscais, recebidas em meio magnético pelo Fisco Distrital, foram comparados a boas práticas de gestão de riscos. Tal análise resultou na identificação da necessidade de aprimorar o processo de comparação dos indícios de fraudes fiscais provenientes de fontes distintas, como forma de melhorar a seleção dos alvos a fiscalizar.

A opção por utilizar modelos logísticos e de redes neurais, baseados em aprendizado de máquina, nesta tarefa foi impulsionada tanto pelo volume e complexidade dos dados envolvidos quanto pelo fato de que a metodologia tem sido utilizada para fins semelhantes, com trabalhos de Choi e Lee [1] sobre tratamento de dados financeiros, de González e Velásquez [2] e de Babu e Vasavi [3] sobre fraudes fiscais, de Hajek e Henriques [4] sobre classificação, e de Li et al. [5] sobre classificação de risco de crédito.

Os potenciais benefícios envolvidos na melhoria do processo de gerenciamento dos indícios de irregularidade, que motivaram a pesquisa, podem ser assim resumidos:

- Seleção de alvos com maior potencial de autuação;
- Melhor aproveitamento dos recursos disponíveis;
- Diminuição da dependência de intervenção humana no processo.

De um modo geral, ao longo do tempo, a busca pelo aprimoramento dos processos relacionados ao tratamento das informações recebidas pelo Fisco Distrital tem se mostrado promissora no sentido de otimizar o emprego dos recursos humanos e materiais disponíveis para o combate às fraudes fiscais, melhorando os resultados obtidos. Trata-se, portanto, de um processo contínuo, contexto no qual a presente pesquisa se insere.

Para realização do trabalho, foram utilizados conhecimentos relacionados à Ciência da Computação, à Estatísticas, à Gestão de Riscos e às áreas Tributária e Financeira.

1.1 Detalhamento do problema e justificativa do tema

1.1.1 A relação tributária

A relação entre o Estado e os cidadãos vem se modificando no curso da história humana, sendo que alguns aspectos podem ser rastreados a tempos remotos, tais como:

- A obrigação de pagar tributos;

> Nas sociedades pré-monetárias os tributos eram devidos na forma de trabalho. Um exemplo deste tipo de sociedade é o Império Inca.[...] Em sociedades sem moedas os agricultores tinham que entregar parte de sua colheita para o Governo [6, p. 7].

- A formalização de estruturas governamentais para sua cobrança;

> Durante a época dos faraós os arrecadadores de impostos eram chamados de escribas[...] Na época do Império Romano, os agricultores tributários eram encarregados de arrecadar fundos para o Governo [...] [6, p. 8].

- A percepção de que a carga tributária deve ser compatível com a realidade vigente, acerca da qual os primeiros registros provêm da Suméria, do século XXIII a.C.:

> Quando Urukagina chega ao poder na cidade-estado de Lagash, a situação da população é de total subjugo econômico tanto pelos funcionários do Palácio como pelos do Templo, a ponto de ter a população escravizada. Pois os impostos e contribuições, somados aos juros de dívidas, e ao abuso dos poderosos sobre os fracos, têm a população economicamente asfixiada. [...] realiza uma reforma social, que ficará lembrada por ser uma das primeiras reformas sociais e a primeira reforma tributária para a qual existem dados disponíveis [7, p. 7].

Por outro lado, a concepção da finalidade do tributo e das obrigações do Estado sofreram grandes alterações com o passar do tempo. Diversos autores apontam o século XIX como marco importante destas transformações:

Segundo Luiz Emygdio Rosa Júnior (Manual de Direito Financeiro Direito Tributário, 20 ed., Editora Renovar em 2007, RJ), conforme citado por Machado e Balthazar [8, p. 228], "Já a partir dos fins do século XIX, porém, aconteceu o alargamento das ações do Estado, que deixou de ser um mero expectador da atividade econômica". Sobre os ideais do iluminismo, Ricardo Lobo Torres [9, p. 149] concluiu que no Brasil fortaleceu-se a relação entre tributo e liberdade, a preocupação com a livre iniciativa e o contrato social, mas que "não chegamos, todavia, a explorar em toda a sua plenitude o liberalismo, eis que conservamos os traços da escolástica, mesclada com o ecletismo e o empirismo do iluminismo ítalo-germânico, que nos influenciou".

De acordo com Almir e Kommer [6, p. 13], em razão das ideias iluministas foram incorporadas à relação tributária conceitos de redistribuição de recursos, promoção da concorrência leal, valorização de normas claras e direito a apelações, e o posterior reconhecimento de que o Estado pode prover a todos os cidadãos serviços que não poderiam ser oferecidos pelo mercado privado, segundo eles:

> A legitimidade da tributação requer uma política justa e prática do gasto governamental. Serviços bons, providos por um Governo de maneira efetiva e eficiente, são uma condição prévia para estabelecer um clima no qual os contribuintes cumpram voluntariamente com suas obrigações [6, p. 14].

O alargamento das ações do Estado, decorrente da prestação de serviço a todos os cidadãos, pode ser exemplificado no Brasil pelos percentuais da arrecadação vinculados aos gastos com educação e saúde constantes da Constituição Federal de 1988 [10].

Entretanto, a importância da arrecadação tributária para a coletividade não garante o pagamento dos impostos devidos, a inadimplência fiscal é um fenômeno complexo relacionado a fatores econômicos, éticos e culturais tais como o aumento contínuo da carga tributária, a complexidade do sistema, a ineficiência dos gastos públicos e a percepção de impunidade dos contribuintes, conforme evidenciam Siqueira e Ramos [11].

Cabe à Administração Tributária a atribuição de buscar o adimplemento da parcela que não foi recolhida de forma espontânea. Entretanto, existem limites para esta atuação. Ricardo Lôbo Torres (Curso de direito Financeiro e Tributário, RJ, 1993, p. 186), conforme citado por Coêlho [12], apresenta um conceito da relação jurídica tributária, no qual evidencia sua complexidade e elementos importantes para os fins da presente pesquisa, como a possibilidade do Fisco exigir a apresentação de informações e a importância do

tratamento igualitário aos contribuintes:

> A relação jurídica tributária é complexa, pois abrange um conjunto de direitos e deveres do Fisco e do contribuinte. A Fazenda Pública tem o direito de exigir do contribuinte o pagamento de tributo e a prática de atos necessários a sua fiscalização e determinação; mas tem o dever de proteger a confiança nela depositada pelo contribuinte. O sujeito passivo, por seu turno, tem o dever de pagar o tributo e de cumprir os encargos formais necessários à apuração de débito; mas tem o direito ao tratamento igualitário por parte da Administração Pública e ao sigilo com relação aos atos praticados. [12, p. 35].

Este trabalho não aborda os motivos da inadimplência, embora relevantes no contexto nacional, visto que há mais de duas décadas ocorrerem esforços para realização de uma reforma tributária no Brasil, motivada, entre outros fatores, pelo aumento da carga tributária, conforme apontam Orair e Gobetti [13, p. 2].

Tampouco é objeto do presente estudo a falta de recolhimento dos valores que foram corretamente declarados e dos valores decorrentes de condutas relacionadas à elisão fiscal, assim definida por Siqueira e Ramos [11, p. 557]:

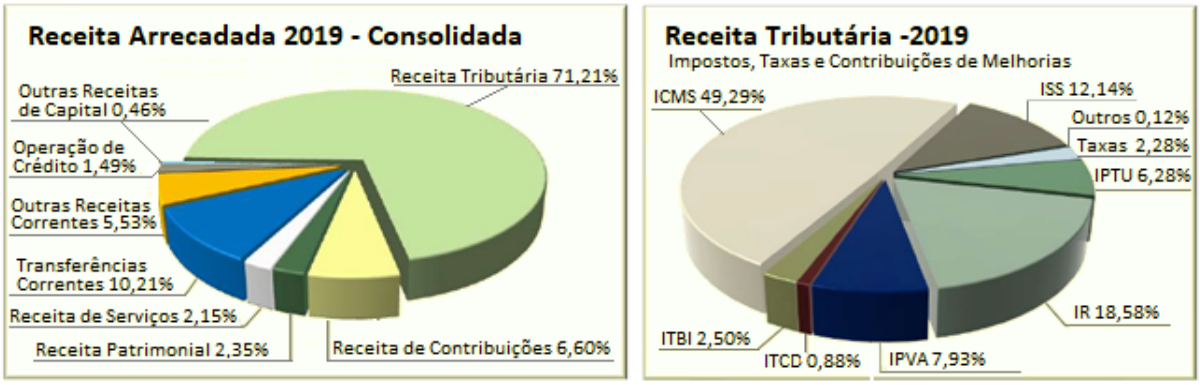
> Um outro conceito relacionado à perda de arrecadação é o da elisão fiscal (ou economia de imposto), pela qual os indivíduos reduzem seu próprio imposto de uma maneira que não era desejada pelos legisladores, mas que não foi expressamente prevista e proibida pela lei. A elisão é realizada tipicamente por meio de transações estruturadas de forma a minimizar a responsabilidade tributária. De um ponto de vista legal, a sonegação difere da elisão por ser ilegal, e daí sujeita à punição (ao menos na teoria).

A pesquisa considerou a carga tributária vigente e as condutas contrárias a legislação que implicam em declaração do imposto devido em valor menor do que o correto, situações estas que geram uma lacuna entre a arrecadação devida e a realizada, e se constituem em fonte de incerteza e, portanto, um risco para as contas públicas.

Neste trabalho são tratados especificamente os aspectos operacionais do processo adotado no combate às mencionadas condutas, quando realizadas por contribuintes do Imposto sobre Operações relativas à Circulação de Mercadorias e Prestação de Serviços de Transporte Interestadual e Intermunicipal e de Comunicação (ICMS) e do Imposto sobre Serviços de Qualquer Natureza (ISS), para mapear os pontos nos quais as técnicas de mineração de dados ainda não são utilizadas e avaliar a possibilidade de inseri-las, como forma de aprimorar o processo de gestão de riscos.

1.1.2 A arrecadação tributária no DF

A receita tributária responde pela maior parte da receita arrecadada no Distrito Federal, tendo atingido em 2019 o percentual de 72,21% do total arrecadado, sendo que o ICMS e o ISS respondem por mais da metade deste montante. Tendo totalizado, em 2019, 10,18 bilhões de Reais, ou 61,43% das Receitas Tributárias do Distrito Federal [14], dos quais 49,29% correspondem ao ICMS e 12,14% ao ISS, conforme mostrado na Figura 1.1.



Fonte: Adaptado do Relatório Balanço de 2019 [14]

Figura 1.1: Receitas tributárias: Arrecadação de 2018

Nas situações em que o valor devido de ICMS ou de ISS declarado pelo contribuinte é inferior ao correto, apenas a autoridade administrativa competente pode efetuar o lançamento do crédito tributário complementar, de modo a possibilitar sua cobrança administrativa e/ou judicial, conforme estabelecido no Código Tributário Nacional (CTN).

> Art. 142. Compete privativamente à autoridade administrativa constituir o crédito tributário pelo lançamento, assim entendido o procedimento administrativo tendente a verificar a ocorrência do fato gerador da obrigação correspondente, determinar a matéria tributável, calcular o montante do tributo devido, identificar o sujeito passivo e, sendo caso, propor a aplicação da penalidade cabível [15].

No Distrito Federal, a autoridade competente é definida pela Lei 4.717/2011:

> Art. 4º Compete ao Auditor-Fiscal da Receita do Distrito Federal: I – em caráter privativo: a) exercer as funções de lançamento, fiscalização, arrecadação e administração dos tributos de competência do Distrito Federal;[...] [16].

1.1.3 A fiscalização tributária no DF

A Secretaria de Estado de Economia do DF (SEEC) é a instituição responsável pela administração tributária no DF. Seu Regimento Interno [17] define as atividades de cada setor, levando em conta a missão, a visão e os valores da instituição [18]. O acompanhamento da arrecadação do ICMS e do ISS cabe à Fiscalização Tributária e as atividades relacionadas à identificação, quantificação, classificação dos indícios, bem como o planejamento da fiscalização, escolha das ações a executar e dos contribuintes a fiscalizar, desde a geração das ordens de serviço até a avaliação dos resultados das ações, atualmente está a cargo da Gerência de Programação Fiscal (GEPRO).

Para viabilizar a fiscalização, a legislação estabeleceu que o contribuinte deve cumprir diversas obrigações acessórias, algumas das quais implicam na apresentação de informações em meio magnético, principal insumo para detecção de irregularidades na atualidade.

> Art. 113. A obrigação tributária é principal ou acessória. [...] § 2º A obrigação acessória decorre da legislação tributária e tem por objeto as prestações, positivas ou negativas, nela previstas no interesse da arrecadação ou da fiscalização dos tributos.[...] [15].

Silva e Cerqueira [19, p. 19] relatam iniciativas adotadas pelas administrações tributárias para coibir a sonegação e aumentar a rapidez e a precisão da atuação do Fisco.

> [...] o maior avanço neste sentido foi a criação do Sistema Público de Escrituração Digital (SPED), através do Decreto Federal 6.022, de 22.01.2007 [...] constituindo-se em mais um avanço na informatização da relação entre o fisco e os contribuintes [...] O SPED consiste na modernização da sistemática do cumprimento de obrigações acessórias, transmitidas pelos contribuintes às administrações tributárias e aos órgãos fiscalizadores, utilizando-se da certificação digital para fins de assinatura dos documentos eletrônicos, garantindo assim a validade jurídica dos mesmos apenas na sua forma digital [19, p. 20].

Com tais iniciativas do Fisco, o volume dos dados apresentados pelos contribuintes tem crescido ao longo dos anos, merecendo destaque as seguintes alterações:

- Na década de 1990, os contribuintes do ICMS e do ISS do DF entregavam um arquivo por tributo ao mês, a Guia Informativa Mensal do ICMS (GIM) e a Declaração Mensal de Serviços Prestados (DMSP), de acordo com os regulamentos dos respectivos tributos então vigentes [20] e [21], com dados totalizados por categoria (compras, vendas, serviços prestados, etc.).

- A partir de 2006, com a publicação do Decreto Nº 26.529/2006 [22] que instituiu o Livro Fiscal Eletrônico (LFE), se tornou obrigatória no DF a escrituração digital dos livros fiscais do ICMS e ISS;
- A partir de 2007, com a publicação do Ajuste SINIEF 07/05, foi instituída a obrigação de emissão da Nota Fiscal Eletrônica (NFE) nas operações realizadas entre contribuintes do ICMS de alguns segmentos. Com o tempo, a obrigatoriedade foi ampliada até abarcar todas as vendas entre contribuintes do ICMS. Por meio da NFE, os dados são informados ao Fisco operação a operação, permitindo, no momento da venda, saber entre outras coisas: quem compra o que, de quem, por quanto.
- A partir de 2008, com a Lei Complementar nº 772/2008 do DF [23], passou a ser obrigatória a entrega de dados referente operações de ICMS e ISS realizadas com pagamento por meio de cartões de débito e de crédito.
- A partir de 2014, com a Portaria nº 234/2014 [24], foi instituída a obrigatoriedade escalonada da emissão da Nota Fiscal ao Consumidor Eletrônica (NFCE) no DF.

A maneira pela qual o Fisco Distrital trata as informações foi mudando junto com a forma de entrega dos dados. De um modo geral, a busca pelo aprimoramento dos processos relacionados ao tratamento das informações recebidas pelo Fisco Distrital tem se mostrado promissora no sentido de otimizar o emprego dos recursos humanos e materiais disponíveis para o combate às fraudes fiscais, melhorando os resultados obtidos. Trata-se, portanto, de um processo contínuo, contexto no qual a presente pesquisa se insere.

Atualmente, os dados em meio magnético são recebidos e armazenados em bases de dados controladas pelo Fisco, sendo submetidos à algoritmos para tratamento dos dados. A mineração de dados está, portanto, inserida em diversas etapas do processo de gerenciamento dos riscos de inadimplência e fraudes fiscais, e seu uso tem se intensificado à medida que as obrigações acessórias são alteradas, de modo que seu aprimoramento é importante para o processo como um todo.

Entretanto, a melhoria do processo de gestão de riscos da fiscalização tributária não se reduz a uma questão de melhorar os algoritmos de mineração de dados existentes ou de criar novos algoritmos. Trata-se de um processo dinâmico que abarca ações como mapear novas situações que envolvam falta de recolhimento de tributos, aperfeiçoar a legislação para obter eventuais dados que se façam necessários, além da constante adequação de procedimentos, de estratégias de alocação dos recursos disponíveis e de abordagem de cobrança, além do acompanhamento das mudanças de comportamento dos contribuintes.

Neste contexto, a presente pesquisa objetivou mapear novas situações em que o tratamento massivo dos dados pudesse propiciar a melhoria do processo de gestão de riscos,

identificar quais tipos de modelos poderiam ser usados na tarefa, criar os modelos e verificar se seus resultados confirmavam a possibilidade inicialmente identificada. Em face do amplo espectro de ações possíveis para aprimorar o processo em questão, a etapa inicial da pesquisa envolveu a análise do tratamento das informações recebidas em meio digital à luz de boas práticas adotadas no gerenciamento de riscos, tendo sido identificada a possibilidade utilizar os dados de conclusão de auditorias já realizadas para prever o resultado de novas ações fiscais e assim otimizar o desempenho da fiscalização

1.1.4 A relevância do tema e a questão de pesquisa

A pesquisa decorre da necessidade constante de aprimorar gestão de riscos no combate à inadimplência do ICMS e do ISS, sua relevância está associada aos seguintes fatos:

- Os tributos tratados (ICMS e ISS) respondem por mais de 60% da arrecadação tributária do DF, sendo relevantes para o orçamento e para a atuação do GDF;
- Apenas a fiscalização tributária pode efetuar o lançamento dos tributos não declarados, cabendo à SEEC organizar e aprimorar os processos envolvidos;
- Os dados apresentados ao Fisco geram uma grande quantidade de indícios de irregularidade a verificar, cuja priorização da averiguação tende a maximizar os resultados;
- A gama de atividades envolvidas no aprimoramento do combate à inadimplência fiscal requer uma abordagem de gerenciamento de riscos e o volume de dados tratados indica a necessidade de uso de ferramentas de mineração de dados.

Este estudo considera a hipótese de que o uso de modelos preditivos de mineração de dados, aliados a recursos de aprendizado de máquina podem aprimorar o processo de gestão de riscos da administração tributária em face da possibilidade de descoberta de padrões não detectáveis na análise humana.

Assim, visando gerar subsídios para o contínuo aprimoramento da cultura organizacional da administração tributária do DF na busca de uma gestão de riscos mais eficiente, o presente trabalho tem por objetivo responder a seguinte questão de pesquisa:

A utilização de modelos preditivos de mineração de dados para análise de indícios de irregularidades e fraudes fiscais a luz do histórico de resultado das auditorias já realizadas pode aprimorar o processo de seleção das empresas a fiscalizar e consequentemente o processo de gestão de riscos da administração tributária do DF?

1.2 Organização da dissertação

As próximas seções apresentam: o objetivo geral e os específicos; o referencial teórico; a metodologia; a análise dos dados e as conclusões, bem como reflexões sobre trabalhos futuros, as referências bibliográficas utilizadas e os anexos.

Capítulo 2

Objetivos

2.1 Objetivo geral

Verificar como o uso de modelos preditivos de mineração de dados pode aprimorar o processo de gestão de riscos da administração tributária do DF.

2.2 Objetivos específicos

Para atingir tal objetivo, foram estabelecidos os seguintes objetivos específicos:

- Caracterizar o processo de gestão de riscos aplicado aos tributos indiretos pela fiscalização tributária do DF, à luz das boas práticas de gestão de risco aplicáveis à Administração Pública Distrital;
- Identificar modelos preditivos aplicáveis à base de dados selecionada e aos objetivos da pesquisa;
- Criar os modelos preditivos para tratamento dos dados selecionados e avaliar seu desempenho;

Capítulo 3

Referencial Teórico

A fundamentação teórica utilizada na pesquisa está organizada em quatro seções:

- Gestão de risco no setor público
- Mineração de dados
- Modelos computacionais
- Trabalhos relacionados

A primeira seção apresenta especificidades da aplicação do gerenciamento de risco no âmbito da administração pública, contexto em que se insere o presente trabalho.

As demais apresentam aspectos ligados às técnicas e ferramentas utilizadas em contextos que se assemelham ao da presente pesquisa e que permitem organizar, analisar, tratar e interpretar um grande volume de dados.

3.1 Gestão de riscos no setor público

Conforme exposto no Referencial Básico de Gestão de Riscos [25], o gerenciamento de riscos, que acompanha a história humana, tem se estruturado ao longo do tempo, e nos ambientes corporativos recebeu atenção crescente ao longo das últimas décadas:

> Em termos históricos a gestão de riscos pode ser rastreada à época em que os primeiros chefes de clãs decidiram fortificar muralhas, realizar alianças com outras tribos e estocar provisões para o futuro [...] No período recente, atribui-se a Frank Knight a publicação, em 1921, de obra (Risk, Uncertainty and Profit) que se tornou referência por estabelecer conceitos, definir princípios e introduzir alguma sistematização ao tema [...] [25, p. 12]

> Cinquenta anos depois, em 1975, a revista Fortune publicou o artigo The Risk Management Revolution, um dos primeiros documentos a tratar o tema sob o enfoque corporativo e a atribuir à alta administração a responsabilidade por instituir políticas, supervisionar e coordenar as várias funções de riscos existentes em uma organização (FRASER; SIMKINS, 2010) [...] No início dos anos 90, as bases para o que conhecemos como gestão de risco foram estabelecidas, mediante a publicação de três documentos que se tornaram referência mundial no tema: o COSO I, o Cadbury e a AS/NZS 4360:1995 [25, p. 12].

O tema vem ganhando importância no âmbito da administração pública, segundo Hill e Dinsdale (*A foundation for developing risk Management learning strategies in the Public Service*, *Canadian Centre for Management Development, Learning Leaders*, 2001), publicado pela Escola Nacional de Administração Pública (ENAP) [26] em 2003:

> No setor público, uma preocupação central na gestão de riscos é o dever de cuidar do bem público – os riscos sempre devem ser gerenciados mantendo-se, em primeiro plano, o interesse público. Nesse contexto, a decisão sobre como a distribuição dos benefícios e das perdas potenciais deve ser equacionada é aspecto importante da gestão de riscos [26, p. 17] [...] Obviamente, é importante possuir uma sólida compreensão dos diferentes aspectos da gestão de riscos: o processo geral da tomada de decisões que envolvam riscos, a importância do diálogo e da comunicação de riscos, alguns conceitos fundamentais, como o princípio preventivo, e o papel da ciência e dos especialistas no processo decisório. Esses conhecimentos são vitais para todos os servidores [26].

A preocupação com a gestão de riscos na administração pública brasileira tem se materializado com a edição de normas, a realização de cursos e a elaboração de documentos relacionados a sua implantação, tanto pela União quanto por diversas unidades federadas, sendo que para os fins da presente pesquisa destacam-se:

Em 2015, a Controladoria-Geral do Distrito Federal (CGDF) criou um projeto para implantar a gestão de riscos no poder executivo local em decorrência de recomendação da Organização para a Cooperação e Desenvolvimento Econômico (OCDE) no sentido de "integrar a gestão de riscos como elemento-chave da responsabilidade gerencial, de modo a promover a integridade e prevenir a improbidade, os desvios e a corrupção" [27], segundo apontam Araújo Júnior e Pinho Filho [28, p. 8], para eles a CGDF tinha em vista que:

> [...] gerenciar riscos é uma ação relevante para as organizações de qualquer natureza, uma vez que a incerteza mina a concentração dos gestores, pois dificulta o planejamento, bem como o aperfeiçoamento de operações e processos [28, p.9].

Em abril de 2016, o Decreto 37.302/2016 estabeleceu os modelos de boas práticas para gestão de riscos e controle interno da administração pública do Distrito Federal.

> Art. 2º Devem ser utilizados como instrumentos de boas práticas técnicas e gerenciais os seguintes modelos: I - ISO 31000:2009 - Gestão de Riscos; II - ISO 19011:2011 - Diretrizes para Auditoria de Sistemas de Gestão; e III - Controle Interno - Estrutura Integrada - 2013 do Comitê de Organizações Patrocinadoras da Comissão Treadway (COSO) [29].

Em maio de 2016, o Ministério do Planejamento, Orçamento e Gestão (MP) e a Controladoria-Geral da União (CGU) publicaram a Instrução Normativa Conjunta MP/CGU nº 01/2016 [30], sobre a sistematização de práticas relacionadas à governança, à gestão de riscos e aos controles internos no âmbito do Poder Executivo Federal.

Em 2018, foi publicado o Referencial Básico de Gestão de Riscos [25], pelo Tribunal de Contas da União (TCU), que trata desde a definição de Risco, como o termo que designa a existência de incertezas capazes de ameaçar o alcance de um objetivo desejado, até a determinação da maturidade global em gestão de riscos da organização, apresentando as diretrizes para seu aperfeiçoamento no âmbito da União, bem como a legislação aplicável.

Tais publicações tendem a consolidar a cultura de gestão de riscos na Administração Pública e, segundo enfatizou o Ministro Raimundo Carreiro, Presidente do Tribunal quando da publicação do Referencial Básico de Gestão de Riscos, investir na cultura de gerenciamento de riscos tende a contribuir para o bom desempenho das organizações:

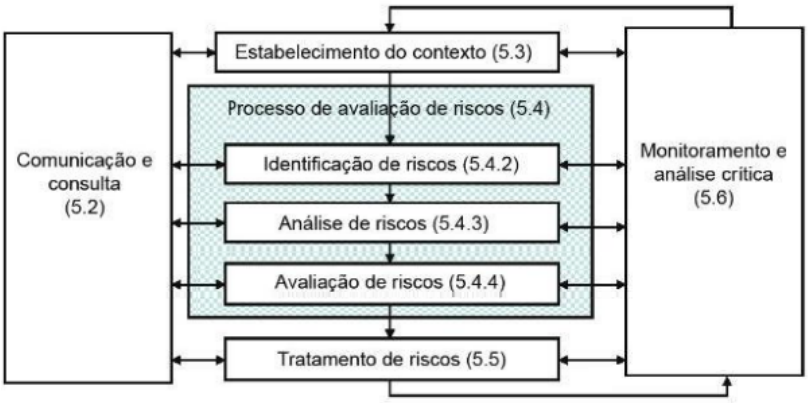
> A sociedade anseia por uma administração pública ágil e eficiente, capaz de implementar políticas e programas de governo que entreguem o melhor valor para a população. Todavia, não raras vezes essas expectativas são frustradas e, ao se analisarem as causas por trás das dificuldades da administração pública em corresponder a esses anseios, depara-se não apenas com restrições orçamentárias e deficiências de diferentes naturezas, mas principalmente com a baixa capacidade para lidar com riscos. Diante desse cenário, a gestão e o controle da aplicação dos recursos públicos com base em risco têm sido recomendações recorrentes deste Tribunal, conquanto reconheça o fato de ser um desafio para a gestão das organizações públicas determinar o quanto de risco aceitar na busca do melhor valor para os cidadãos. Apesar de não ser nova a discussão sobre a necessidade de gerenciar riscos no setor público, isso ainda é um paradigma a ser atingido. Persiste a necessidade não apenas de estruturas e processos, mas também de uma cultura de gerenciamento de riscos, a fim de contribuir para que a organização obtenha resultados com desempenho otimizado [25].

A seguir é apresentada uma visão geral dos modelos aplicáveis à Administração Pública

Distrital, definidos pelo Decreto 37.302/2016 [29], sendo abordados alguns aspectos que merecem destaque, em face do escopo do presente trabalho. Entretanto, o detalhamento das metodologias estudadas não consta do presente documento e pode ser consultado nas respectivas documentações.

3.1.1 ABNT NBR ISO 31000

A ISO 31000 [31], publicada em 2009, fornece princípios e diretrizes para a gestão de riscos, que auxiliam na adoção sistemática e confiável de processos de gestão de risco. A norma apresenta o processo de gestão de risco como uma lista de atividades coordenadas, e fornece a representação esquemática que pode ser vista na Figura 3.1.



Fonte: NBR ISO 31000

Figura 3.1: Processo de gestão de riscos - NBR ISO 31000

A avaliação de riscos, atividade na qual o presente trabalho se concentra, é tratada como uma etapa capaz de gerar conhecimentos que podem auxiliar na tomada de decisões, sendo que a comparação do nível de risco permite individualizar e priorizar o tratamento:

> A finalidade da avaliação de riscos é auxiliar na tomada de decisões com base nos resultados da análise de riscos, sobre quais riscos necessitam de tratamento e a prioridade para a implementação do tratamento. A avaliação de riscos envolve comparar o nível de risco encontrado durante o processo de análise com os critérios de risco estabelecidos quando o contexto foi considerado. Com base nesta comparação, a necessidade do tratamento pode ser considerada. Convém que as decisões levem em conta o contexto mais amplo do risco e considerem a tolerância aos riscos assumida por partes que não a própria organização que se beneficia do risco. Convém que as decisões sejam tomadas de acordo com os requisitos legais, regulatórios e outros requisitos [31, p. 18].

3.1.2 O Framework COSO

O *Committee of Sponsoring Organizations of the Treadway Commission – COSO* publicou em 1992 o *Guide Internal Control - integrated framework*, conhecido como COSO-IC [32], e em 2004, o *Enterprise Risk Management - integrated framework*, conhecido como *COSO-ERM* [33].

Segundo Araújo Júnior e Pinho Filho [28, p. 13], o COSO-IC é reconhecido por propiciar o desenvolvimento, a condução e avaliação da eficácia do controle interno.

Ao tratar do COSO-ERM, o Referencial Básico de Gestão de Riscos do TCU afirma:

> O modelo é apresentado na forma de matriz tridimensional (cubo), demonstrando uma visão integrada dos componentes que os gestores precisam adotar para gerenciar os riscos de modo eficaz, no contexto dos objetivos e da estrutura de cada organização [25, p. 15].

Este caráter multidimensional da gestão de riscos está presente desde a primeira publicação do COSO-IC, representado por um cubo cujas três faces visíveis mostram: os tipos de objetivos; os níveis da estrutura organizacional e os componentes básicos para o controle interno. A atualização do COSO-IC publicada em 2013 manteve a representação em forma de cubo. O COSO-ERM, de 2004, incorporou o COSO-IC e, entre outros acréscimos, indica que o apetite ao risco deve estar alinhado aos objetivos estratégicos:

> A presente estrutura de gerenciamento de riscos corporativos, embora não tenha por meta substituir a estrutura de controles internos das organizações, incorpora a estrutura de controle interno em seu conteúdo [34, p. 5][...] O gerenciamento de riscos corporativos requer: Alinhar o apetite a risco e a estratégia – A administração considera em primeiro lugar o apetite a risco, ao avaliar as opções estratégicas e fixar objetivos compatíveis com a estratégia escolhida, bem como desenvolver mecanismos para administrar os riscos implícitos [34, p. 20].

Dentre os diversos pontos das metodologias mencionadas no Decreto 37.302/2016 aplicáveis à gestão dos indícios de irregularidades identificados pelo Fisco Distrital, destaca –se a necessidade de conhecer do contexto interno para eficaz gestão do risco.

Acerca da importância de conhecer o contexto da organização, Hill e Dinsdale (*A foundation for developing risk Management learning strategies in the Public Service, Canadian Centre for Management Development, Learning Leaders*, 2001), conforme citado em [26], afirmam:

> Para que todos os riscos sejam levados em conta no processo de tomada de decisões, é essencial incorporar muitas perspectivas diferentes e incorporar a avaliação de riscos não como processo de unidade, mas da organização como um todo. Essa é a finalidade da gestão integrada de riscos: "processo contínuo, proativo e sistemático de compreensão, gerenciamento e comunicação de riscos a partir da perspectiva da organização como um todo. Sua meta é permitir a tomada de decisões estratégicas que contribuam para a realização dos objetivos corporativos gerais da organização" (TBS, 2001) [26, p. 20].

Neste contexto, é importante observar que as preocupações das partes interessadas devem ser levadas em consideração no processo de gestão de riscos, conforme aponta o Referencial Básico de Gestão de Riscos [25, p. 24], ao tratar da ISO 31000:

> Um dos primeiros passos da atividade de estabelecimento do contexto é identificar os fatores do ambiente, interno e externo, no qual a organização persegue seus objetivos. Não menos importante é a identificação das partes interessadas, bem como a identificação e a apreciação das suas necessidades, expectativas legítimas e preocupações, pois essas partes interessadas devem ser incluídas em cada etapa ou ciclo do processo de gestão de riscos, por meio do processo de comunicação [...] [25, p. 24].

Outra questão relevante para a presente pesquisa diz respeito ao estabelecimento do apetite ao risco e da capacidade de aceitar e tolerar determinados níveis de riscos.

> Ao definir os critérios de risco, convém que os fatores a serem considerados incluam os seguintes aspectos: [...] o nível em que o risco se torna aceitável ou tolerável[...] [31, p. 17]
> Os critérios estabelecidos para priorização de riscos levam em conta, por exemplo, a significância ou os níveis e tipos de risco, os limites de apetite a risco, as tolerâncias a risco ou variações aceitáveis no desempenho, os níveis recomendados de atenção [...] [25, p. 111]).

Os aspectos aqui apontados foram objeto do estudo de caso que consta do capítulo 5.

3.2 Mineração de dados

A mineração de dados permitir explorar uma grande quantidade de dados procurando por padrões, como regras de associação ou sequências de tempo, e viabiliza detectar a existência de relações sistemáticas, implícitas, previamente desconhecidas e potencialmente úteis, servindo tanto para compreender os dados quanto para efetuar previsões, conforme apontam Witten et al. [35, p. 3], que assim dispõe sobre sua importância:

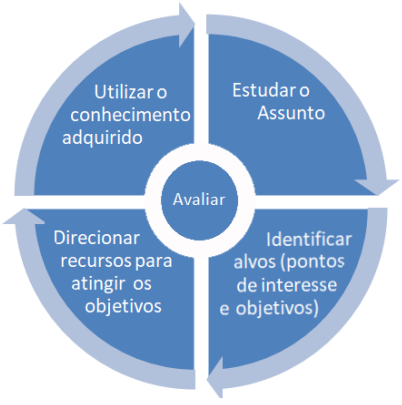
> À medida que o mundo cresce em complexidade, nos sobrecarregando com os dados que gera, a mineração de dados se torna nossa única esperança de elucidar padrões ocultos. Dados analisados de forma inteligente são um valioso recurso, podem levar a novos conhecimentos e, em ambientes comerciais, a vantagens competitivas. [35, p. 4, tradução nossa].

Para Fayyad et. al [36, p. 82], a mineração de dados corresponde ao uso de algoritmos específicos para extrair padrões de dados e se refere a uma etapa específica do processo geral de descoberta de conhecimento útil a partir de dados, o Knowlege Database Discovery (KDD), sendo que etapas do processo KDD como preparação, seleção e limpeza de dados, são essenciais para a descoberta de informações úteis.

> A aplicação cega de métodos de mineração de dados (corretamente criticados como "*data dredging*", dragagem de dados, na literatura estatística) pode ser uma atividade perigosa que leva facilmente à descoberta de padrões sem sentido [36, p. 82, tradução nossa].

Existem outras metodologias que tratam a criação e implementação de modelos de mineração dados como uma etapa do processo de descoberta de conhecimento a partir dos dados como a *SEMMA - Sample, Explore, Modify, Model e Assess by SAS Institute* [37, p. 144] ou o *CRISP-DM - Cross Industry Standard Process for Data Mining* da IBM [38, p. 23]. Cada metodologia possui suas peculiaridades, algumas tratam de etapas posteriores à mineração, como o monitoramento da solução implementada, outras não.

Entretanto, existem etapas em comum, como a organização das fases dos processos e a necessidade de conhecer o contexto e os dados. É possível observar ainda uma convergência entre metodologias de mineração de dados e frameworks de gestão de risco. A Figura 3.2 apresenta as semelhanças identificadas.



Fonte: A autora (2019)

Figura 3.2: Convergências: gestão de risco e metodologias de mineração de dados

3.2.1 Análise e preparação dos dados

Witten et al. [35] apontam que os dados reais são imperfeitos: algumas partes podem estar distorcidas, outras ausentes, desta forma é necessário conhecer os dados disponíveis antes de iniciar a implementação de qualquer modelo de detecção de padrões:

> Nada substitui o conhecimento de seus dados. Ferramentas simples que mostram histogramas da distribuição de valores de atributos nominais e gráficos dos valores de atributos numéricos (talvez classificados ou simplesmente representados graficamente em relação ao número de instância), são muito úteis [...] O tempo analisando seus dados é sempre bem gasto [35, p. 60, tradução nossa].

Das etapas do processo de mineração apresentadas por Mishra [39, p. 9] quatro estão relacionadas à obtenção e tratamento dos dados, o que evidencia a importância de conhecer atributos e sua relevância para o tema a ser tratado e podem ser assim resumidas:

- Extração de dados;
- Verificação da integridade dos dados, com supressão de redundâncias e de dados irrelevantes para o estudo;
- Combinação de informações de fontes distintas, baseada no relacionamento de atributos comuns às fontes em questão;
- Transformação dos dados e inclusão de atributos quando necessário.

De acordo com Bussab e Morettin [40, p. 1], a estatística permite ao pesquisador a análise e a compreensão dos dados relevantes para a realidade específica, objeto do estudo. O conjunto de técnicas estatísticas abrange a coleta, organização e interpretação de dados experimentais, bem como a extrapolação possibilitando inferências e predições. Neste contexto, segundo Montgomery e Runger [41, p. 128] a estatística descritiva aborda os aspectos relacionados a organização e resumo dos dados de maneira a facilitar sua interpretação e análise subsequente, enquanto a estatística inferencial está relacionada à análise dos dados e sua interpretação, visando a identificação de modelos plausíveis [40, p. 1], que permitam produzir afirmações gerais sobre dada característica de uma realidade específica a partir de dados parciais.

No capítulo 5 é detalhado o estudo de caso realizado e a correspondente análise dos dados realizada. As ferramentas utilizadas na análise estão detalhadas no capítulo 4.

Uma vez conhecidos os dados e seus contextos, torna-se necessário selecionar a forma como estes dados devem ser tratados. A próxima seção trata desta seleção.

3.2.2 Aprendizado de máquina

A mineração de dados utiliza técnicas de aprendizado de máquina e de modelagem analítica e estatística, que envolvem uma ampla gama de modelos de reconhecimento de padrões, classificação e predição, como expõe Mishra [39, p. 11]. Witten et al. [35, p. 8] entendem que o conceito de aprendizado envolve as ideias de aquisição de conhecimentos e a capacidade de utilizá-lo:

> Muitas técnicas de aprendizagem procuram por descrições estruturais do que é aprendido - descrições que podem se tornar bastante complexas e são normalmente expressas como conjuntos de regras [...] A experiência mostra que em muitas aplicações de aprendizado de máquina para mineração de dados, as estruturas de conhecimento explícito que são adquiridas, as descrições estruturais, são pelo menos tão importantes quanto a capacidade de bom desempenho em novos exemplos [...] [35, p. 8, tradução nossa] .

As estratégias de aquisição de conhecimento computacional se dividem em três classes básicas, citadas por autores como Prieto et al. [42] e LeCun et al. [43]:

- Aprendizado supervisionado: objetiva aprender um mapeamento de entradas e saídas a partir de pares rotulados de entrada e saída, de acordo com Jordan e Rumelhart (*Supervised learning with a distal teacher*, 1992), conforme citado em [42, p.244];
- Aprendizado não supervisionado: não existem saídas previamente informadas para cada entrada e o objetivo é identificar "estruturas latentes" para obter melhores estimadores de densidade de probabilidade conjunta, segundo Ghahramani (*Unsupervised learning*, *SpringerVerlag*, 2004), conforme citado em [42, p. 244]; e
- Aprendizado por reforço: segundo Sutton e Barto (*Reinforcement Learning. An Introduction*, *MITPress*, 1998), conforme citado em [42, p. 244], o mapeamento de entrada-saída é realizado através da interação continuada de um sistema de aprendizagem com seu ambiente. Segundo Mnih V. et al. (*Human-level control through deep reinforcement learning*, *Nature*, 2015), citado em [43, p. 442], resultados impressionantes têm sido obtidos quando se trata de aprender a jogar diferentes videogames .

A combinações destas classes é citada por autores como Abiodun et al. [44, p.10], sendo que o aprendizado supervisionado é o mais utilizado, segundo LeCun et al. [43].

No presente estudo foi utilizado o aprendizado supervisionado, tendo como dados rotulados e previamente conhecidos os resultados obtidos em fiscalizações anteriores e como variáveis independentes características relacionadas aos contribuintes e aos indícios.

3.2.3 Escolha do modelo

Para Fayyad et. al. [36, p. 85], os objetivos de previsão e descrição são alcançados por meio de seis métodos primários: Classificação, regressão, agrupamento, resumo, modelagem de dependências e detecção de mudanças de desvios, assim descritos:

> **Classificação**: aprender uma função que mapeia (classifica) um item de dados em uma das várias classes predefinidas. **Regressão**: identificar uma função que mapeia um item de dados para uma variável de predição de valor real e a descoberta de relações funcionais entre as variáveis. **Agrupamento**: identificar um conjunto finito de categorias ou *clusters* para descrever os dados. Intimamente relacionado ao agrupamento está o método de estimativa de densidade de probabilidade, que consiste em técnicas para estimar a partir de dados a função de densidade de probabilidade multivariada conjunta de todas as variáveis / campos no banco de dados. **Resumo**: encontrar uma descrição compacta para um subconjunto de dados, por exemplo, a derivação de regras de resumo ou associação e o uso de técnicas de visualização multivariadas. **Modelagem de Dependência**: encontrar um modelo que descreve dependências significativas entre variáveis (por exemplo, aprendizagem de redes de crença). **Detecção de Mudança e Desvio**: descobrir as mudanças mais significativas nos dados de valores medidos anteriormente ou normativos [36, p. 85, tradução nossa].

Sobre a escolha do modelo de mineração, Fayyad et al. [36, p. 86]. afirmam que a escolha do algoritmo para resolução de um problema específico é quase uma arte, dada a variedade de algoritmos existentes e considerando que não existe um algoritmo que resolva todos os problemas, uma vez que "cada técnica normalmente atende a alguns problemas melhor do que outros" [36, p. 86, tradução nossa].

Ao tratar de modelagem de dados, Donoho [45, p. 22] afirma que na prática são usadas ferramentas e pontos de vista das duas culturas de modelagem de Leo Breiman (2001): A cultura da modelagem gerativa, que permite fazer inferências, e a cultura da modelagem preditiva, que prioriza a previsão.

Mishra [39] apresenta vários cenários que permitem entender qual tipo de modelo pode ser aplicado em cada situação, merecendo destaque para os fins da presente pesquisa, os métodos de aprendizado supervisionado baseados em regressão [39, p. 111] e os métodos baseados em redes neurais [39, p. 221], importantes para criar modelos preditivos que permitem estimar o valor futuro de variáveis de interesse.

No presente estudo foram adotados modelos com características inferenciais e preditivas de regressão logística e de redes neurais, detalhados nas seções 3.4 e 3.5, tendo em vista seu uso em trabalhos relacionados a fraudes financeiras e análise de risco de crédito.

3.2.4 Treinamento do modelo

De acordo com Witten et al. [35, p. 149], o treinamento do modelo visa aprimorar o provável desempenho futuro com novos dados, e pode ser realizado utilizando técnicas de aprendizado de máquina e bases distintas para treinamento, validação e testes.

> Para prever o desempenho de um classificador em novos dados, precisamos avaliar sua taxa de erro em um conjunto de dados que não desempenhou nenhum papel na formação do classificador. Este conjunto de dados independente é chamado de conjunto de teste. Assumimos que tanto os dados de treinamento quanto os dados de teste são amostras representativas do problema subjacente [35, p. 149].

Prieto et al. [42, p. 243] enfatizam que a divisão da base de dados em três porções nem sempre é a solução adotada, em razão do risco de dividir os dados de forma desfavorável, especialmente quando não se tem muitas amostras, e apontam que a técnica de validação cruzada (*Cross-Validation*) é usada para evitar a criação de bases de validação e garantir um bom ajuste do modelo, o conjunto de treinamento é subdividido em conjuntos mutuamente exclusivos que são utilizados pelos algoritmos de aprendizagem para treinamento e validação do modelo, de forma sucessiva e alternada.

Para obtenção das bases distintas de treinamento e testes é comum dividir o conjunto de dados disponível, Tkáč e Verner [46, p. 791], ao tratarem do uso de redes neurais para pontuação de risco de crédito, mencionam o uso das proporções 50:50 e 70:30 para treinamento e testes. A proporção 70:30 foi citada por González a e Velásquez [2, p. 1433], ao tratarem de detecção de fraudes.

A divisão dos dados costuma ser feita de forma aleatória, para que os novos conjuntos de dados tenham características semelhantes ao conjunto original, minimizando o risco da divisão interferir no ajuste do modelo. Essas etapas sucessivas de treinamento objetivam generalizar o modelo, para que apresente resultados igualmente aceitáveis para novos conjuntos de dados, conforme enfatizam LeCun et al. [43, p. 437].

Conforme exposto por Fayyad et al.[36, p. 87], o treinamento pode modelar não apenas os padrões gerais nos dados, mas também algum ruído específico para o conjunto de dados de treino, resultando em um desempenho ruim do modelo quando submetido a outra base de dados, este problema é conhecido como sobre ajuste ou *overfitting*, o problema costuma ocorrer quando se utiliza um conjunto limitado de dados para treinamento do modelo. As soluções possíveis incluem validação cruzada (*cross-validation*), regularização, e outras estratégias estatísticas.

Guo et al. [47, p. 44] afirmam que a limitação de dados de treinamento pode limitar o tamanho e a capacidade de aprendizado dos modelos, e apresentam duas soluções comu-

mente usadas para obter mais dados de treinamento: generalizar mais dados utilizando esquemas de aumento de dados, como dimensionamento, rotação e corte ou coletar mais dados de treinamento com algoritmos de aprendizado fracos.

Na presente pesquisa foi utilizado o esquema de partição da base de dados em 3/4 para treinamento e 1/4 para testes, e a técnica de validação cruzada (*Cross-Validation*).

3.2.5 Avaliação dos resultados

O resultado da submissão dos dados preparados ao modelo escolhido deve ser avaliado para que se possa prever o desempenho do modelo quando submetido a novos dados, Witten et. al [35, p. 149] reforçam a importância da avaliação da taxa de erros ser feita utilizando a base de dados de teste, assim entendido o conjunto de dados que não desempenhou nenhum papel na formação do classificador.

Conforme exposto por Prati et al. [48, p. 1], para avaliação de algoritmos de aprendizado supervisionado de classificação que utilizam uma base de dados composta por exemplos de casos onde o campo com a classificação verdadeira, binária, e previamente conhecida, corresponde à variável dependente e os demais campos possuem características do caso em questão e correspondem às variáveis independentes ou explicativas, caso da presente pesquisa, é possível utilizar a tabulação cruzada entre a classificação verdadeira e a prevista pelo modelo treinado, conhecida como Matriz de Confusão ou Tabela de Contingências.

A tabela 3.2, conforme apresentada por Witten et al. [35, p. 164], corresponde à matriz de confusão para uma situação em que existem apenas dois resultados possíveis (positivo e negativo; empreste ou não empreste; com a mancha de óleo ou sem a mancha):

Tabela 3.1: Matriz de confusão

		Dados reais	
		Positivo	Negativo
Dados Previstos	Positivo	TP	FP
	Negativo	FN	TN

Conforme exposto por diversos autores, entre eles Witten et. al. tipos [35, p. 164]) e Prati et al. [48, p. 1], a Matriz de Confusão classifica os resultados em quatro tipos:

- TP do inglês *True Positive* ou verdadeiro positivo: a classe real é positiva e a classe prevista pelo modelo também;
- TN do inglês *True Negative* ou Verdadeiro Negativo: a classe real é negativa e a classe prevista pelo modelo também;

- FP do inglês *False Positive* ou falso Positivo: a classe real é negativa e a classe prevista pelo modelo é positiva, situação conhecida como erro do tipo 1;
- FN do inglês *False Negative* ou Falso Negativo: quando a classe real é positiva e a classe prevista pelo modelo é negativa, esta situação também conhecida como erro do tipo 2;

As classificações corretas correspondem aos verdadeiros positivos (TP) e aos verdadeiros negativos (TN). As medidas de aferição da qualidade do modelo levam em conta tanto os acertos por classe de resposta, quanto os acertos globais, conforme exposto por Witten et al [35, p. 164 a 177], são sensibilidade, especificidade, acurácia e F1-Score, sendo que:

Tabela 3.2: Métricas de avaliação

Métrica	**Descrição**	**Fórmula**
Acurácia	indica a taxa geral de sucesso das classificações do modelo	$\frac{\text{TP+TN}}{TP+FN+TN+FP}$
Sensibilidade (*Recall*)	fração de instâncias relevantes que são recuperadas com sucesso	$\frac{\text{TP}}{TP+FN}$
Especificidade (*precision*)	fração de instâncias recuperadas que são relevantes	$\frac{\text{TP}}{TP+FP}$
F1-score	Média harmônica entre as sensibilidade e especificidade	$\frac{2(\text{sensibilidade* especificidade})}{sensibilidade+especificidade}$

A especificidade indica o quanto os resultados são corretos, a sensibilidade indica o quanto são completos e o F1 maior indica um modelo melhor.

3.3 Modelos computacionais

A presente seção apresenta alguns aspectos relevantes relacionado aos dois modelos computacionais utilizados na busca de conhecimento a partir dos dados disponíveis realizada na presente pesquisa:

- Análise de regressão e
- Redes neurais

3.3.1 Análise de regressão

Ao tratar da análise de regressão, técnica de estatística que permite inferir e modelar a relação de uma variável dependente de resposta com variáveis explicativas, Montgomery e Runger [41, p. 265] afirmam que "a regressão é largamente utilizada e frequentemente

mal empregada", enfatizam que é necessário cuidado na escolha das variáveis explicativas para evitar o desenvolvimento de relações estatísticas entre variáveis que não estejam completamente relacionadas em um sentido de causa e efeito, e apontam o planejamento de experimentos como maneira de determinar as relações de causa e efeito e assim evitar o problema de implementação.

Existem vários modelos de regressão, com características próprias, Stulp e Sigaud [49] apresentam diversos algoritmos de regressões, enquanto Montgomery e Runger [41] apresentam aplicações envolvendo as regressões lineares simples (usadas quando há apenas um regressor), lineares múltiplas e a regressão logística, usada quando a variável resposta é categórica, ou seja, a variável apresenta como possíveis realizações uma qualidade ou atributo, como "sucesso" e "falha", e não mais uma mensuração.

Regressão logística

Ao tratar da regressão logística, Mishra [39, p. 129] aponta que existem diversos cenários da vida onde a variável de interesse é de natureza categórica a exemplo de comprar um produto ou não, aprovar um cartão de crédito ou não, merecendo destaque, em face da presente pesquisa, os seguintes aspectos:

> A regressão logística não apenas prevê uma classe de variável dependente, mas também prevê a probabilidade de um caso pertencer a um nível na variável dependente. As variáveis independentes não precisam ser normalmente distribuídas e não precisam ter variâncias iguais. [...] Todas as variáveis independentes podem ser contínuas, categóricas ou nominais [...] Se algumas variáveis independentes são categóricas, a conversão binária (criando variáveis "dummy", para cada categoria) é necessária [39, p. 129, tradução nossa].

Função logística

Montgomery e Runger [41, p. 288] avaliam a suposição de que o modelo para tratar a variável de resposta Y_i binária teria uma forma linear; concluem que os resultados sugerem a inadequação do uso do modelo linear para o tratamento em questão; apontam que existe evidência empírica considerável de que a função deve ser não-linear; e indicam que para o caso em questão, geralmente é empregada a função de resposta logit, também conhecida como função logística, cuja expressão é:

$$E(Y) = \frac{1}{1 + e^{-(\beta_0 + \beta_1 x_1 + \cdots + \beta_k x_k)}}$$

para um conjunto de variável preditiva como um vetor X= $(x_1, x_2, ..., x_k)$. Posteriormente, os autores indicam que a razão entre as probabilidades dos dois eventos (razão de chances ou *Odd ratio*), quando $E(Y)$ está relacionado a x pela função logit é:

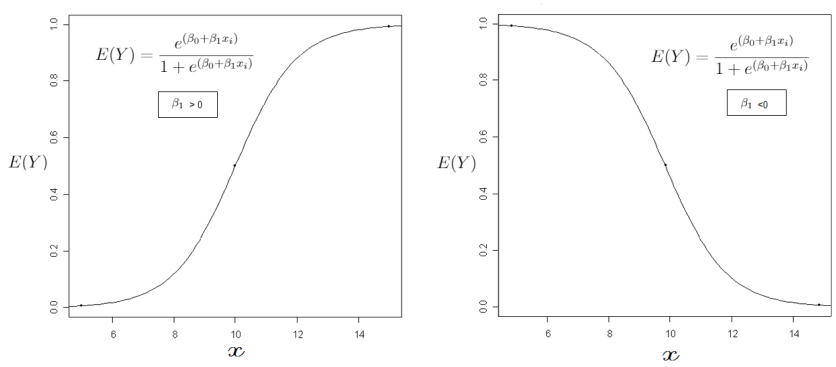
$$\frac{E(Y)}{1 - E(Y)} = e^{(\beta_0 + \beta_1 x_1 + \cdots + \beta_k x_k)}$$

e que o logaritmo natural da razão de chances é uma função linear da variável regressora, onde a inclinação β_i é a variação no logaritmo das chances resultantes do aumento de uma unidade x_i, ou seja, a razão de chances varia de e^{β_i} quando x_i aumenta uma unidade.

No mesmo sentido, Mishra [39, p. 129] relata que modelo da regressão logística corresponde ao log das chances do resultado e apresenta a equação do modelo:

$$Logit(Y) = ln\Big(\frac{E(Y)}{1 - E(Y)}\Big) = ln(e^{\beta_0 + \beta_1 x_1 + ... + \beta_k x_k}) = \beta_0 + \beta_1 x_1 + ... + \beta_k x_k$$

A Figura 3.3 apresenta duas curvas logísticas com uma variável explicativa, sendo que na primeira curva β_1 é positivo e na segunda β_1 é negativo



Fonte: Adaptado de Montgomery e Runger 2012)

Figura 3.3: Gráficos da função de respostas logísticas

Na regressão logística, a log-verossimilhança do modelo é usada para medir a adequação do ajuste conforme exposto por Montgomery e Runger [41, p. 288] e por Witten et al. [35, p. 126], sendo que os valores previstos para a variável dependente sempre estarão entre 0 e 1. As probabilidades acima de 0,50 são classificadas como evento ou "sucesso", por convenção.

A regressão logística binomial ou binária foi utilizada no presente estudo, para criação do modelo de mineração de dados, em face da necessidade de identificar, a partir do histórico do resultado das auditorias, qual contribuinte terá maior chance de ser autuado, necessidade que se amolda às características do modelo em questão, por tratar-se de

processo de aprendizagem de relações entre entradas e saídas a partir de dados de exemplo a fim de efetuar previsões da categoria de saída para novas entradas.

Importante observar que o volume de dados disponíveis para análise pode dificultar ou inviabilizar o seu tratamento manual, tornando necessário identificar as ferramentas computacionais que permitem a implementação do modelo, na presente pesquisa o modelo foi implementado com utilização do Software R, acerca do qual maiores detalhes constam no capítulo relativo à metodologia, seção ferramentas.

3.3.2 Redes neurais artificiais

Redes neurais artificiais são modelos computacionais inspirados no cérebro humano, trata-se de um sistema de processamento de informações onde as entradas são analisadas por unidades de processamento ligadas umas às outras, como neurônios no cérebro humano para transmitir informações, conforme apontado por diversos autores, como Mishra [39, p. 222] e Guo et al. [47, p. 43].

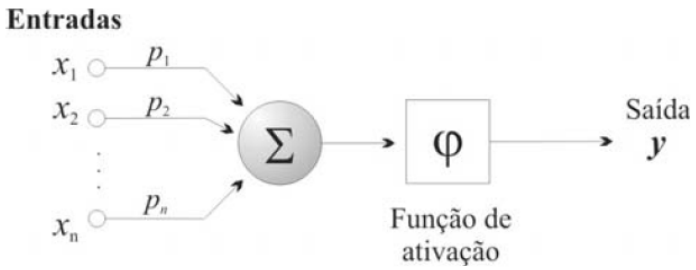
Segundo Huang et al. [50, p. 115], as redes neurais possuem a capacidade de detectar as relações subjacentes dentro de um conjunto de dados, reconhecer padrões, efetuar classificação, avaliação, modelagem, predição e controle de padrões. Enquanto Tkáč e Verner [46, p. 789, tradução nossa] afirmam que "as redes neurais apresentam níveis de eficiência, robustez e adaptabilidade que as tornam uma ferramenta valiosa para classificação, suporte a decisões, análise financeira ou *credit scoring*".

Mishra [39, p. 222] aponta que a habilidade de aprender a partir dos dados de entrada das redes neurais decorre de um processo iterativo de ajustes aplicado aos seus pesos relacionados às entradas de cada neurônio, etapa chamada de treinamento.

No treinamento das redes neurais, o algoritmo calcula um vetor de gradiente que indica se o aumento de cada peso aumentaria ou diminuiria o erro, a média de todos os exemplos de treinamento é calculada e o vetor gradiente negativo indica a direção em que o erro diminui, aproximando-o de um mínimo, onde o erro de saída é baixo, conforme exposto por LeCun et al. [43, p. 436].

Neurônios artificiais

O neurônio artificial recebe informações (dados de entradas) pelas suas conexões, executa operações sobre os dados recebidos, utilizando em funções próprias de ativação os pesos calculados nas etapas de treinamento e envia um resultado para a próxima camada, conforme apontado por diversos autores como Mishra [39, p. 129] e Li et al. [5, p. 1151]. Ferneda [51, p. 26] apresenta a representação esquemática do neurônio artificial reproduzida na Figura 3.4.

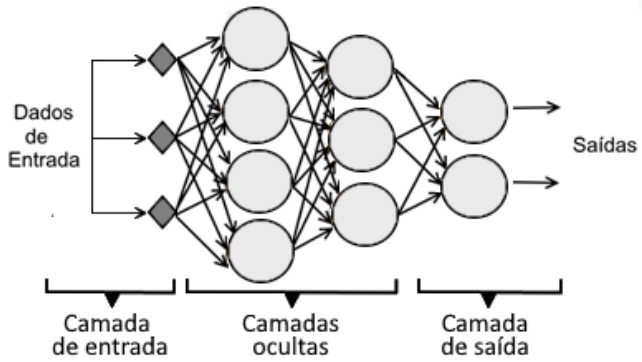


Fonte: Ferneda(2006)

Figura 3.4: Modelo de neurônio artificial

Arquitetura de redes neurais artificiais

Os neurônios são dispostos em camadas, cada camada pode ter diversos neurônios e cada rede diversas camadas, entretanto existem limitações em termos práticos relacionadas ao aumento do tamanho das redes, tanto devido à complexidade quanto ao tempo de processamento, conforme apontam diversos autores tais como Ferneda [51, p. 26], Mishra [39, p. 222] e Witten et al. [35, p. 239]. A Figura 3.5 apresenta um exemplo de rede com 3 entradas, duas camadas e duas saídas.



Fonte: A autora (2019)

Figura 3.5: Modelo de rede neural artificial

A informações entre as camadas pode fluir através da rede em uma única direção, da camada de entrada à camada de saída, isto ocorre nas redes *feed-forward* multicamadas, conforme exposto por LeCun et al. [43, p. 438] e Abiodun et al. [44, p. 7]. Existem casos em que a informação é armazenada durante o processamento e utilizada na determinação do resultado da camada de saída, este procedimento é conhecido como retro propagação (*backpropagation*), de modo que o gradiente dos pesos de um módulo pode ser calculado a partir do gradiente em relação à saída daquele módulo, fazendo com que a retro propagação seja aplicada através de todos os módulos, conforme apontam LeCun et al. [43, p. 438] e Mishra [39, p. 223].

Aspectos importantes quanto ao uso de redes neurais

Segundo LeCun et al. [43, p. 438], desde os primórdios do reconhecimento de padrão, com trabalhos publicados em 1957, o objetivo dos pesquisadores foi substituir os recursos de engenharia manual por redes multicamadas, mas a solução não foi amplamente entendida até meados dos anos 1980, a tecnologia de redes não supervisionadas e de retro propagação foi praticamente abandonada no final da década de 1990, em especial porque era comum pensar que a descida de gradiente de erro ficaria presa em mínimos locais, condição em que nenhuma alteração reduziria o erro médio, e somente em 2006 as aplicações foram retomadas.

Atualmente a tecnologia é amplamente utilizada em diversas áreas como processamento imagens, medicina, negócios, política. Exemplos de aplicações são apresentados por Huang et al. [50],Tkáč e Verner [46], LeCun et al. [43] e Abiodun et al. [44].

Apesar das inúmeras vantagens e do uso disseminado, trabalhar com redes envolve alguns aspectos que merecem destaque:

- Compreensão teórica: em que pese os métodos de aprendizado profundos terem apresentado resultados promissores, a teoria subjacente não é bem compreendida, existem dúvidas acerca de quais arquiteturas teria melhor desempenho, quantas camadas ou quantos nós por camada são apropriados para uma determinada tarefa. A arquitetura é historicamente determinada em uma base de dados [47, p. 44].
- Complexidade do tempo: quanto maior a rede maior os recursos computacionais exigidos, as primeiras CNN, por exemplo, não eram consideradas viáveis para aplicações em tempo real [47, p. 44]. A complexidade do tempo ajuda a entender os impactos de fatores como profundidade, número de filtros, tamanhos de filtros etc.
- Sobre ajuste ou *overfitting*: o problema relacionado ao treinamento de modelos de mineração de dados em geral, exposto na seção 3.2.4, pode ocorrer para redes neurais, quando o modelo estatístico se ajustar muito bem ao conjunto de dados de treinamento, mas se mostrar ineficaz para prever novos resultados.
- Convergência para mínimos locais: em alguns casos a rede não consegue um ajuste eficiente dos pesos porque os pesos convergem para mínimos locais. Entretanto, LeCun et al. apontam que:

> Na prática, mínimos locais ruins raramente são um problema com redes grandes. Independentemente das condições iniciais, o sistema quase sempre chega a soluções de qualidade muito semelhante. Resultados teóricos e empíricos recentes sugerem fortemente que os mínimos locais não são um problema sério em geral [43, p. 438].

Como forma de lidar com as incertezas associadas à utilização de redes neurais diversos estudos têm adotado o uso combinado com outros recursos, conforme exposto por Huang et al. [50, p. 131], Guo et al. [47, p. 44] e Tkáč e Verner[46, p. 794].

No presente trabalho, para lidar com tais incertezas, foi adotado o uso em paralelo de redes neurais e de regressão logística. A escolha dos métodos foi influenciada pelos trabalhos de:

- González e Velásquez [2, p. 1429] apresentam uma tabela contendo técnicas de mineração de dados usadas pelas administrações fiscais para detectar fraudes fiscais, onde é possível observar que o uso combinado de diversas técnicas, dentre elas a Regressão Logística e as redes Neurais. e
- Hajek e Henriques [4, p. 144], que trataram de fraudes financeiras e desenvolveram filtros de fraudes financeiras utilizando diversos modelos de mineração de dados, entre eles regressão logística e redes neurais.

3.4 Trabalhos relacionados

Dentre os textos estudados, foram relevantes para a execução da presente pesquisa, os seguintes trabalhos:

Hajek e Henriques [4] apresentam um estudo comparativo dos métodos de aprendizado de máquina para verificar se um sistema aprimorado de detecção de fraudes financeiras poderia ser desenvolvido combinando recursos específicos derivados de informações financeiras e comentários gerenciais em relatórios corporativos anuais. O artigo apresenta os métodos de classificação utilizados: regressão logística, classificadores bayesianos, máquinas de vetores de suporte, árvores de decisão, redes neurais e classificadores de conjuntos.

Li et al. [5] resumem diferentes aplicações de tecnologias de inteligência artificial em vários domínios da administração de empresas, incluindo finanças, indústria, varejo e gerenciamento de negócios. Eles utilizam Aprendizado de Máquina como uma combinação de matemática, finanças e ciência da computação e mostram um exemplo de uma regressão logística bem treinada que pode separar dois tipos de clientes (bons ou ruins).

González e Velásquez [2] mostram que é possível caracterizar e detectar possíveis usuários de faturas falsas em um determinado ano, dependendo das informações sobre o pagamento de impostos, o desempenho histórico e as características, o estudo aborda o uso de diferentes tipos de técnicas de mineração de dados.

Babu e Vasavi [3] aplicam um algoritmo analítico preditivo, utilizando processo gaussiano, no conjunto de dados de imposto de renda para identificar fraudes em valores fiscais.

Silva [52] apresenta o uso de regressão logística aplicada à administração fiscal.

Sousa [53] apresenta aplicações da classificação de padrões utilizando redes neurais artificiais, no contexto da administração pública no controle externo.

O trabalho de WU et al. [54] aborda o uso de técnicas de mineração para melhorar o desempenho da detecção de evasão fiscal.

Prieto et al. [42] apresentam uma visão abrangente de modelagem, simulação e implementação de redes neurais. O artigo aborda a ideia de entender melhor o sistema de gestão para tentar construir sistemas de processamento de informação inspirados em funções biológicas naturais e aborda os conceitos ligados a redes neurais, ressaltando o aspecto interdisciplinar dos sistemas neurais artificiais.

Abiodun et al. [44] realizaram um levantamento de aplicações de redes neurais no cenário do mundo real. O estudo apresenta desafios de aplicação de redes neurais artificiais, contribuições, comparar desempenhos e métodos de crítica e abrange aplicações em várias disciplinas que incluem computação, ciência, engenharia, medicina, meio ambiente, agricultura, mineração, tecnologia, clima, negócios, artes e nanotecnologia. Os autores ressaltam a importância do uso de modelos híbridos.

Choi e Lee [1] realizaram pesquisa e analisaram os métodos de aprendizado de máquina e aprendizado profundo usados em estudos de fraude financeira entre 2016 e 2017. A pesquisa compara 13 métodos de mineração distintos utilizados para detecção de fraudes financeiras em dados reais da Coréia.

Bittencourt [55] criou uma proposta de monitoramento de contribuintes baseado na análise de *outliers*, que compreende as três etapas, com uso da metodologia DEA (*Data Envelopment Analysis*), análise de séries temporais para identificação de indícios de irregularidade para avaliação e passível de inclusão em auditoria fiscal.

Embora nenhum dos artigos tenha exatamente o mesmo escopo da presente pesquisa, eles contribuem para escolher a metodologia e as técnicas de mineração de dados adotadas.

Capítulo 4

Metodologia

A presente seção apresenta o tipo de pesquisa realizada, bem como os procedimentos e ferramentas adotados em sua execução.

4.1 Tipo de pesquisa

Prodanov e Freitas [56, p. 48] afirmam que a pesquisa científica objetiva conhecer e explicar os fenômenos, obtendo respostas a questões relevantes em seu contexto, com uso dos conhecimentos acumulados e diferentes métodos e técnicas. Os autores indicam possibilidade de classificar as pesquisas quanto à natureza, aos objetivos, à forma de abordagem e aos procedimentos [56, p. 72], e enfatizam que na prática existe uma mescla de características, onde se acentuam as características de um ou outro tipo [56, p. 50].

Em se tratando de classificação quanto à natureza, a pesquisa pode ser básica (pura) ou aplicada, sendo que a pesquisa aplicada tem como característica fundamental o interesse na utilização e nas consequências práticas dos conhecimentos, conforme apontam Gil [57, p. 27] e Prodanov e Freitas [56, p. 51].

Neste contexto, este trabalho se enquadra na categoria de pesquisa aplicada, pois seu interesse fundamental é gerar conhecimentos para aplicação prática, dirigidos especificamente ao aprimoramento da gestão de riscos da Administração Tributária do DF.

No que tange à classificação quanto aos objetivos, as pesquisas podem ser classificadas como exploratórias, descritiva ou explicativas.

Nas pesquisas exploratória, segundo Prodanov e Freitas [56, p. 51], a finalidade é obter mais informações acerca do assunto investigado para facilitar a delimitação do tema, fixar objetivos ou descobrir um novo enfoque sobre o assunto. Sobre o tema, Gil [57] afirma:

> As pesquisas exploratórias têm como principal finalidade desenvolver, esclarecer e modificar conceitos e ideias, tendo em vista a formulação de problemas mais precisos ou hipóteses pesquisáveis para estudos posteriores [...]. Habitualmente envolvem levantamento bibliográfico e documental, entrevistas não padronizadas e estudos de caso [...]. [57, p. 27]

Ao tratar de pesquisas descritivas, Gil [57] afirma:

> As pesquisas deste tipo têm como objetivo primordial a descrição das características de determinada população ou fenômeno ou o estabelecimento de relações entre variáveis[...] têm por objetivo estudar as características de um grupo: sua distribuição por idade, sexo, procedência, nível de escolaridade, nível de renda, estado de saúde física e mental etc. [...] Também são pesquisas descritivas aquelas que visam descobrir a existência de associações entre variáveis[57, p. 28].

O presente trabalho apresenta características de pesquisa exploratória, visto que objetivou a obtenção de mais informações sobre a aderência do processo de gestão de risco do Fisco Distrital aos modelos de boas práticas aplicáveis ao DF em razão da legislação vigente e sobre os indícios de fraudes identificados, mas também apresenta características de pesquisa descritiva, visto que envolve a caracterização do processo de risco com a descrição de características dos indícios de fraudes existentes e dos contribuintes a eles relacionados, além de ter buscado a identificação de relação entre tais características e o resultado das ações fiscais, a fim de definir em que etapa o uso de modelos preditivos poderia ser avaliado como ferramenta de aprimoramento do processo.

Para os critérios de classificação relacionados à abordagem, as pesquisas podem ser classificadas em quantitativas e qualitativas. Segundo Prodanov e Freitas [56], a pesquisa quantitativa:

> considera que tudo pode ser quantificável, o que significa traduzir em números opiniões e informações para classificá-las e analisá-las. Requer o uso de recursos e de técnicas estatísticas (percentagem, média, moda, mediana, [...], análise de regressão etc.). [56, p. 69]

Assim, quanto aos objetivos, este estudo apresenta as características de pesquisa quantitativa com representação de atributos qualitativos por meio de variáveis fictícias (dummies).

Tais classificações se refletem no trabalho executado e nos procedimentos metodológicos adotados em sua realização.

Quanto aos procedimentos, segundo Gil [58, p. 44], a classificação da pesquisa com base nos procedimentos técnicos utilizados decorre da necessidade de traçar um modelo conceitual e operativo da pesquisa, para analisar os fatos do ponto de vista empírico, confrontando a teoria com a realidade. Sobre o delineamento da pesquisa, o autor afirma que:

> O delineamento refere-se ao planejamento da pesquisa em sua dimensão mais ampla, que envolve tanto a diagramação quanto a previsão de análise e interpretação de coleta de dados. Entre outros aspectos, o delineamento considera o ambiente em que são coletados os dados e as formas de controle das variáveis envolvidas [...] o delineamento expressa em linhas gerais o desenvolvimento da pesquisa, com ênfase nos procedimentos técnicos de coleta e análise de dados[...] [58, p. 44].

Ao tratar da classificação da pesquisa com base nos procedimentos, Gil [58, p. 44] afirma que o estudo de caso envolve o estudo profundo e exaustivo de um ou poucos objetos, de maneira que permita seu amplo e detalhado conhecimento e aponta diferentes propósitos de sua utilização.

> a) explorar situações da vida real cujos limites não estão claramente definidos; b) descrever a situação do contexto em que está sendo feita determinada investigação; e c) explicar as variáveis causais de determinado fenômeno em situações muito complexas que não possibilitam a utilização de levantamentos e experimentos [58, p. 58].

Neste contexto, o presente trabalho adotou como procedimento principal o estudo de caso, pois buscou a compreensão do processo de seleção de contribuintes da fiscalização tributária do Distrito Federal, sob a ótica das boas práticas da gestão de riscos, e envolveu a análise de dados utilizados na avaliação dos riscos das ações fiscais desenvolvidas nos últimos anos, como as características dos contribuintes fiscalizados, dos indícios de fraudes motivadores das ações fiscais realizadas e de seus resultados, para verificar se o modelos preditivos seriam capazes de gerar conhecimento novo, apto a aprimorar a seleção de alvos a fiscalizar.

4.2 Obtenção de informações e dados analisados

O estudo envolveu a análise de informações obtidas tanto em fontes bibliográficas quanto junto a auditores envolvidos no processo de fiscalização, na etapa de seleção dos contribuintes e na etapa de execução da ação fiscal. Foram analisados ainda dados de sis-

temas corporativos da SEEC e a legislação aplicável. Quanto a obtenção das informações e dados coletados, merecem destaque os seguintes aspectos:

4.2.1 Análise bibliométrica

No início do estudo foi identificada a necessidade de reunir os conhecimentos das áreas de gestão de riscos, estatística, mineração de dados, redes neurais e classificação de indícios de fraude. Como forma de identificar publicações relacionadas aos temas gestão de riscos e redes neurais, foi efetuada uma análise nas plataformas *Web of Science* e *Scopus*.

A análise bibliométrica relativa ao tema gestão de risco retornou 1.395 publicações na plataforma *Web of Science* e 1.774 na *Scopus*. A seleção por área de interesse (Negócios, Gestão e Contabilidade) e ano da publicação (2010 a 2018) diminuiu a quantidade de publicações e as obras compatíveis com a pesquisa foram selecionadas para leitura.

A Figura 4.1 mostra as palavras-chaves das 1.744 publicações inicialmente retornadas pela plataforma *Scopus* e das 62 publicações restantes depois do processo de seleção.

Fonte: A autora (2019) - utilizando o software *VOSviewer*

Figura 4.1: Palavras-chave antes e depois da seleção *Scopus*

De modo análogo, das 1.890 publicações inicialmente retornadas pela plataforma Scopus para o tema de redes neurais, 71 foram selecionadas para leitura.

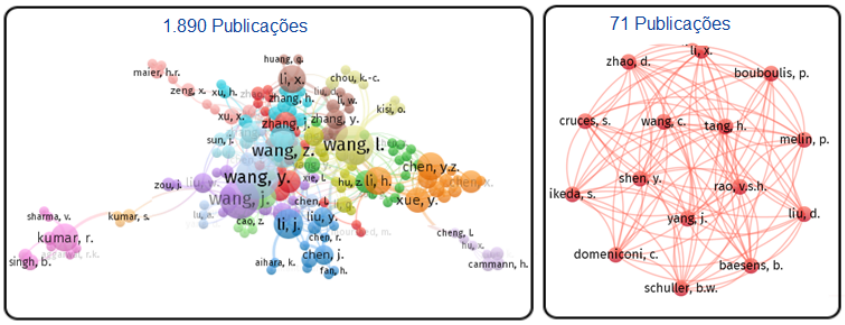
A Figura 4.2. Apresenta o mapa bibliométrico considerando os autores das publicações inicialmente localizadas e após o processo de seleção.

O Anexo I apresenta gráficos que detalham as etapas da análise bibliométrica.

4.2.2 Reuniões e aplicação de questionário

´

As informações relacionadas ao processo de seleção dos contribuintes a fiscalizar, detalhadas na avaliação do estudo de caso, foram obtidas junto aos auditores responsáveis pela atividade, por meio de reuniões de trabalho.



Fonte: A autora (2019) - utilizando o software *VOSviewer*

Figura 4.2: Processo de seleção de publicações: mapa bibliométrico

Para avaliar a possibilidade de estabelecer padrões que permitissem a comparação das ações fiscais de modo a aprimorar os critérios de seleção das empresas a fiscalizar, foi elaborado um questionário de respostas escalonadas visando colher a opinião dos auditores acerca de 20 tipos de ação fiscal quanto às seguintes características: o grau de dificuldade, o potencial de ampliação do escopo e o tempo médio dispendido.

As respostas foram definidas em graus: muito baixo, baixo, médio, alto e muito alto, associados aos números 1, 3, 5, 7 e 9. Os auditores foram orientados a avaliar apenas os quesitos para os quais conseguissem definir uma opinião em face dos critérios propostos.

O questionário, cujo modelo consta no Anexo II, foi entregue a todos os auditores que executam as ações fiscais e os resultados são apresentados na Seção 5.1.1, página 51.

4.2.3 Dados armazenados em sistemas corporativos

No estudo foram analisados dados armazenados nos sistemas corporativos da SEEC: CFI (Cadastro Fiscal), PFI (Programação Fiscal) e DAF (Desenvolvimento da Ação Fiscal). Também foram analisados os dados de arrecadação tributária publicados anualmente nos relatórios de Balanço Geral, conforme detalhado no Anexo III.

A extração das informações ocorreu por meio aplicativos que acessam ambientes que espelham a produção, criados para evitar o acesso direto aos bancos de dados dos sistemas transacionais.

Os mencionados aplicativos operam por meio de senhas pessoais, mesmo para acesso aos dados espelhados, cópias fiéis dos dados originais, tendo em vista a natureza sigilosa das informações tratadas, sendo que o conjunto dos dados analisados é composto por:

- Informações prestadas pelo contribuinte ou por terceiros, em atendimento a obrigações legais, entregues diretamente ao Fisco Distrital, ou recepcionadas pela Receita Federal do Brasil e posteriormente repassadas a SEEC, relacionadas a dados cadastrais e informações econômico-financeiras (faturamento; pagamentos; etc.).

- Informações registradas nos sistemas pelo Fisco, relativas ao histórico das fiscalizações, dos indícios fiscalizados e dos autos de infração lavrados.

As informações utilizadas foram tratadas antes de serem utilizadas na geração dos modelos, para criação das variáveis necessárias.

4.3 Variáveis utilizadas

As variáveis utilizadas nos modelos foram selecionadas a partir de opiniões colhidas junto aos auditores responsáveis pela seleção de contribuintes a fiscalizar. Todas as variáveis correspondem a dados extraídos das bases de dados da SEEC, sendo que as variáveis categóricas foram convertidas em binárias, com criação de uma variável *dummy* para cada categoria, em face da necessidade apontada no referencial teórico, na seção 3.3.1, relativa à preparação dos dados para criação do modelo logístico.

Dados cadastrais: atividade econômica (Sistema CFI)

Os contribuintes do ICMS e do ISS ao se inscreverem no Cadastro Fiscal informam a Classificação Nacional de Atividades Econômicas (CNAE) das atividades econômicas principal e secundárias que pretendem exercer. O CNAE da atividade principal foi utilizado para criação de variáveis que permitem identificar o ramo de atividade de cada contribuinte. A Tabela 4.1 apresenta as variáveis criadas e o conteúdo correspondente.

Tabela 4.1: Variáveis *dummy* para atividade econômica

Atividade Econômica Principal	**Variável *Dummy***				
	VCme	VCmu	VInd	VTran	VServ
Comércio	1	0	0	0	0
Comunicação	0	1	0	0	0
Indústria	0	0	1	0	0
Transporte	0	0	0	1	0
Serviços	0	0	0	0	1

Importante observar que todos os contribuintes analisados se enquadram em pelo menos uma das categorias acima descritas. Foi criada ainda uma variável para identificar especificamente a atividade de comércio atacadista: VAtacado, que recebe 0 para não atacadistas e 1 para atacadistas.

Dados cadastrais: localização das empresas (Sistema CFI)

Com o objetivo de verificar se existe relação entre a localização das empresas e o resultado das ações fiscais, e considerando que o Cadastro Fiscal do DF não é georreferenciado,

foi utilizada a informação da agência de atendimento da receita a qual a empresa está vinculada no Cadastro Fiscal do DF. A Figura 4.3 apresenta as áreas usualmente atendidas em cada uma da agência da receita.

Fonte: A autora (2020)

Figura 4.3: Regiões vinculadas às agências de atendimento da Receita do DF

Os contribuintes costumam estar vinculados à agência de atendimento da Receita do DF mais próxima ao endereço da empresa. Para as empresas vinculadas à Agência Empresarial, em razão do porte ou de regime de apuração diferenciado, foi realizada uma pesquisa do endereço da empresa e este dado foi usado no lugar do endereço da agência. A Tabela 4.2 apresenta as variáveis criadas para identificação da localização da empresa.

Tabela 4.2: Variáveis *dummy* para localização da empresa

Local da Empresa	**Variável *Dummy***									
	VlBra	VlBrz	VCei	VlGam	VlBan	VlPla	VlSob	VlTag	VlSia	VlOut
Brasília	1	0	0	0	0	0	0	0	0	0
Brazlândia	0	1	0	0	0	0	0	0	0	0
Ceilândia	0	0	1	0	0	0	0	0	0	0
Gama	0	0	0	1	0	0	0	0	0	0
Bandeirante	0	0	0	0	1	0	0	0	0	0
Planaltina	0	0	0	0	0	1	0	0	0	0
Sobradinho	0	0	0	0	0	0	1	0	0	0
Taguatinga	0	0	0	0	0	0	0	1	0	0
Sia	0	0	0	0	0	0	0	0	1	0
Outras UFs	0	0	0	0	0	0	0	0	0	1

Existe a possibilidade de empresas estarem vinculadas a unidades de atendimento diferentes das unidades padrão. Entretanto, tais exceções não impedem que as agências de atendimento sejam utilizadas como um indicador de localização da empresa para efeitos do presente estudo.

Dados cadastrais: situação cadastral das empresas (Sistema CFI)

Enquanto os dados de localização da empresa e atividade econômica costumam ser informados pelo contribuinte, a situação cadastral da empresa pode ser informada tanto pelo contribuinte quanto pelo Fisco. A situação "Ativo" indica que o contribuinte está exercendo as atividades normalmente e existem diversas situações que indicam inatividade temporária ou permanente do contribuinte. A variável *Dummy* VAtivo recebe um quando o contribuinte está com situação "Ativo" no cadastro e zero nos demais casos.

Dados das ações fiscais: tipo de ação (Sistema PFI)

As ações fiscais podem ser motivadas pela necessidade de verificar indícios e irregularidade que motivam a realização de verificações fiscais que podem se originar de denúncias, internas ou externas, ou de levantados mediante cruzamento das diversas informações recebidas pelo Fisco. Para cada ação fiscal realizada é emitida uma ordem de serviço que está associada a um único tipo de ação fiscal. Estes tipos podem ser assim agrupados:

- Monitoramento: envolve o acompanhamento das operações de um contribuinte por um período específico, visando verificar o regular cumprimento das obrigações tributárias e incentivar seu adimplemento voluntário.
- Auditoria: Auditoria sem escopo definido, pode decorrer de denúncia ou da necessidade de verificação de indícios obtidos por meio de dois ou mais algoritmos de mineração de dados distintos;
- Auditoria Especial Concentrada (AEC): são auditorias com escopo previamente determinado, visando a análise de um único tipo de indício de irregularidade, geralmente identificado por meio de um algoritmo de mineração de dados específico. É comum criar um tipo de AEC para cada algoritmo de mineração, a fim de facilitar a avaliação dos resultados de cada algoritmo separadamente;
- Diligência Fiscal: inclui ações fiscais realizadas com o objetivo de verificar o cumprimento de obrigações principal ou acessórias em caráter pontual e específico, podem ser decorrentes de denúncias ou algoritmos de mineração de dados;
- Outros: Envolve ações fiscais como perícias e depoimentos à justiça ou à polícia, decorrentes de ações fiscais previamente realizadas.

A tabela 4.3 apresenta os cinco grupos de tipos de ação fiscal e as quatro variáveis criadas para identificá-los:

Tabela 4.3: Variáveis *dummy* para Tipos de Ação Fiscal

Tipo de Ação Fiscal Realizada	**Variável *Dummy***			
	VAud	VAec	VMon	VDil
Auditoria	1	0	0	0
Auditoria Especial Concentrada	0	1	0	0
Monitoramento	0	0	1	0
Diligências	0	0	0	1
Outros	0	0	0	0

Dados das ações fiscais: indícios de irregularidade (Sistema PFI)

As principais fontes utilizadas na mineração de indícios de irregularidade são:

1. EFD: Escrituração fiscal digital no LFE ou no SPED [59];
2. DAS: Declarações do Simples Nacional [60];
3. Cartão: Informações de cartões de crédito/débito [23];
4. NFE: Notas Fiscais Eletrônicas [59];
5. CFI: Dados do Cadastro Fiscal do Distrito Federal [59];
6. RPC: Pagamentos realizados pelos contribuintes [59];
7. ECF: Dados dos Cupons Fiscais Emitidos [59].

Os detalhes da criação dos algoritmos de mineração de dados não são apresentados neste trabalho por questões estratégicas e porque não ser necessário para os objetivos da pesquisa. A Tabela 4.4 apresenta as fontes de dados aos principais algoritmos de mineração de dados sob análise no presente estudo.

Tabela 4.4: Fontes de dados dos indícios selecionados

Tipo de Ação relacionadas a algoritmos de mineração	*Fonte de Dados Utilizada*						
	EFD	DAS	Cartão	NFE	CFI	RPC	ECF
AEC Ali Fixa e S.Card	*X*	*X*	*X*	*X*	*X*	*X*	
AEC Auto Imune	*X*	*X*		*X*	*X*	*X*	
AEC XK	*X*	*X*		*X*	*X*		*X*
AEC Cartão	*X*		*X*	*X*	*X*	*X*	
AEC Ali Fora, Deleta, Outliers e Missing	*X*			*X*	*X*	*X*	
AEC Antecipado, Crédito podre, Costumers, Cerberus, Subprime Imobilizado e Walking Dead	*X*			*X*	*X*		

A Tabela 4.5 apresenta as variáveis criadas para identificar o algoritmo de mineração ou a situação específica que gerou os indícios das Auditorias Especiais Concentradas. A linha "Outros" engloba todas as ações que não tiveram uma fonte específica mapeada.

Tabela 4.5: Variáveis *dummy* para tipos de algoritmo de mineração de dados

Tipo de AEC	VAfi	VAli	VAnt	VAut	VCar	VCer	VCre	VCst	VDel	VImo	VLos	VMis	Vecf	VSub	VWal	VReg	VOut
AliFixa	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
AliFora	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Antecipado	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0
AutoImune	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0
Cartão	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0
Cerberus	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0
CréditoPodre	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0
costumer	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0
Deleta	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0
Imobilizado	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0
LosOtros	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0
Missing	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0
Subprime	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0
WalkingDead	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0
ECF	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0
Regime	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0
Outros	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1

Não foram objeto de estudo as bases de dados e algoritmos de mineração relacionados ao Monitoramento e à Fiscalização de Mercadorias em Trânsito, tendo as características operacionais dos setores adotarem dinâmica própria e diferenciada por situação para seleção dos contribuintes a fiscalizar.

Dados dos autos de infração (Sistema DAF)

O sistema informatizado de controle das ações fiscais permite identificar os autos de infração relacionados a cada ação fiscal concluídas.

Foi criada a variável Vautuado para identificar que assume o valor zero quando o contribuinte não foi autuado e 1 quando o contribuinte foi autuado.

Foi criado ainda um grupo de variáveis para identificar a faixa de valor da autuação do contribuinte. A análise estatística que resultou na escolha dos valores de cada faixa consta detalhada no capítulo dedicado à análise dos resultados.

A Tabela 4.6 apresenta este grupo de variáveis.

Tabela 4.6: Variáveis *dummy* para faixa de valor autuado

Faixa de valor		Variável *Dummy*				
Mín	Máx	VN1	VN2	VN3	VAutuadoQQ	VAutuadoN3
0,01	1.000,00	1	0	0	Autuado	Não Autuado
1.000,01	1.000.000,00	0	1	0	Autuado	Não Autuado
1.000.000,01		0	0	1	Autuado	Autuado
0,00	0,00	0	0	0	Não Autuado	Não Autuado

4.4 Ferramentas utilizadas

4.4.1 Extração e preparação dos dados

Para extração dos dados das bases foram utilizadas as seguintes ferramentas:

- ORACLE Database 11g, Enterprise Edition, version 11.2.0.1.0.
- Microsoft Access: com auxílio da interface de conexão ODBC (Open Database Connectivity) para ORACLE.

Importante observar que o Microsoft Access foi utilizado para a criação de todas as variáveis *dummy*, substituição da identificação dos contribuintes por uma chave artificial, em forma numeração sequencial, e para criação do arquivo de extensão .csv, que serviu de base para as análises realizadas no Software R.

4.4.2 Software R

No conjunto de ferramentas capazes de realizar a implementação dos modelos tratados nesta pesquisa, conforme exposto por Donoho [45, p. 14], o Software R se destaca por apresentar um grande número de usuários e possuir vasto material de pesquisa na rede mundial de computadores, tendo sido criado nos anos 90, por Gentleman e Ihaka, o projeto de código aberto se espalhou rapidamente e hoje é um dos ambientes dominantes de programação quantitativa usado na estatística acadêmica, o autor afirma ainda que:

> [...] o sistema R transformou a prática de análise de dados criando uma linguagem padrão que diferentes analistas podem usar para comunicar e compartilhar algoritmos e fluxos de trabalho [45, p. 27]

Foram usadas as versões R-3.4.2 e R-4.0.2, for Windows 32/64 bits do Software R, disponível em https://cloud.r-project.org, foi utilizado em conjunto com o RStudio, um ambiente de desenvolvimento integrado para programação R, conforme exposto por Wickham e Grolemund [61], disponível em http://www.rstudio.com/download.

Estatística descritiva no R

Os gráficos apresentados na avaliação dos resultados do estudo de caso foram implementados com funções disponíveis no pacote *tidyverse*. A função summary foi usada para obtenção da média, a mediana, o 1º e o 3º quartis, conforme exposto por Wickham e Grolemund [61, p. 25]. Enquanto a moda pode foi obtida por meio da criação de uma função que retorna as observações com o maior número de ocorrências.

Histogramas foram gerados com a função *qplot*, mas Wickham e Grolemund [61, p. 87] mostram como gerá-los com a função *ggplot*. Para Boxplots, foi usada a função *boxplot*.

Mineração de dados no R

Para implementação dos modelos de mineração baseados em regressão logística, foi utilizado o pacote *caret* (abreviação de Classification And REgression Training), cujo detalhamento consta na documentação do pacote [62], tendo em vista que este pacote é amplamente utilizado para criação de modelos de regressão logística.

Para implementação dos modelos baseados em redes neurais, foi utilizado o pacote *h2o*. Trata-se de uma plataforma na memória para aprendizado de máquina escalável e distribuído, amplamente utilizado para implementação de redes neurais, sendo que o detalhamento das funções está disponível na documentação do pacote [63].

Estes pacotes facilitam a criação de modelos preditivos por meio funções que simplificam os processos de:

- Divisão de dados;
- Pré-processamento dos dados;
- Seleção de recursos;
- Ajuste do modelo usando reamostragem;
- Estimativa de importância de variáveis;

Regressão logística no R

As principais funções do R utilizadas na regressão logística foram:

- Função *createDataPartition*: para criação das bases de treinamento e teste, permite separar a base de dados em partições de teste e de treinamento de forma aleatória. O parâmetro *p* permite definir a percentagem de dados da base de treinamento.
- Função *glm*, sigla de *Generalized Linear Models*: para criação do modelo de regressão logística, com uso do parâmetro *Family* igual a *"binomial"(link='logit')*.

- Funções *train* e *trainControl*: para treinamento do modelo, a segunda permitiu configurar a estratégia de treinamento com uso da técnica de validação cruzada.
- Função *predict*: para criação de um novo objeto com uma coluna que apresenta os valores preditos para a base de dados especificada.
- Funções *summary* e *confusionMatrix*: para avaliação do modelo. A primeira apresenta os coeficientes ajustados para a base de dados, bem como a significância estatística *p-valor* relacionada a cada variável explicativa do modelo, e a segunda gera a matriz de confusão e permite calcular a acurácia (*Acciuracy*), sensibilidade (*recall*) e especificidade (*precision*) do modelo.

Redes neurais no R

As principais funções utilizadas para implementação da rede neural foram:

- Função *sample.split*: para separação dos dados de teste e treinamento;
- Função *h2o.deeplearning*: para a criação e o treinamento do modelo de rede neural artificial multicamada de *feed-forward*. O comando nthreads = -1 inicia o H2O utilizando todas as CPUs disponíveis.
- Função *h2o.predict*: para usar a rede neural treinada de modo a obter valores preditos para a base de dados especificada, no caso a base de testes. Depois da previsão foi gerada a matriz de confusão para fins de avaliação do modelo, neste caso foi usada a função *table*, passando como parâmetro a coluna que contém a variável dependente na base de teste e o resultado das previsões.

4.5 Limitações e restrições do estudo

A presente pesquisa não visa à criação ou aprimoramento de algoritmos utilizados para detecção de indícios de irregularidades fiscais, trata da análise do processo de gestão de riscos existente e da averiguação da possibilidade de aplicação de recursos de mineração de dados em etapas nas quais atualmente tais recursos não são utilizados.

Desta forma, o uso de modelos preditivos para identificar novos indícios de irregularidade não foi investigado e pode ser objeto de trabalhos futuros.

Limitação dos dados tratados em razão da inscrição no cadastro fiscal

Empresas não inscritas no cadastro fiscal, cujo valor da variável Dummy VAtivo é igual a zero, estão sujeitas à fiscalização. Entretanto, os sistemas da fiscalização não

armazenam de forma estruturada sua atividade econômica principal e sua localização. Deste modo no treinamento dos modelos tratados no presente trabalho foram utilizadas apenas as ações fiscais realizadas em empresas inscritas no cadastro fiscal do DF, cujo valor da variável Dummy VAtivo é um.

Limitação dos dados tratados em razão do tipo de fiscalização

Como as ações fiscais de diligência são utilizadas tanto para obter informações que permitam atender requisições externas quanto para garantir uma percepção da presença do Fisco junto aos contribuintes, o valor do crédito constituído e a origem do indício não são preponderantes na avaliação da oportunidade e conveniência de sua realização.

Desta forma, os dados relativos ao tipo de ação fiscal diligências não foram utilizados para treinamento dos modelos.

Limitação dos dados tratados em razão do período

Nos exercícios de 2001 a 2006 a escrituração fiscal não era apresentada em meio digital. As informações das administradoras de cartão de crédito passaram a ser recebida a partir de 2008. Em 2009, O Protocolo ICMS 42/2009 estabeleceu o cronograma de massificação do uso da Nota Fiscal Eletrônica. A partir de 2009, com essas mudanças na recepção dos dados, a metodologia de identificação dos indícios adquiriu os contornos atuais.

Assim, apenas os resultados das ações realizadas a partir de 2009 foram utilizados para treinamento dos modelos preditivos.

Limitação em razão do sigilo fiscal

Importante destacar que a pesquisa envolveu análise e tratamento de dados relacionados a características de empresas, às ações fiscais executadas e seus resultados, tudo armazenado em bancos de dados institucionais, sendo que, em razão do sigilo fiscal, em nenhum momento do estudo são apresentados dados que permitam a identificação de qualquer contribuinte.

Capítulo 5

Análise dos Dados e Resultados

O presente capítulo apresenta:

- Caracterização do processo de gestão de riscos;
- Identificação dos modelos preditivos selecionados; e
- Criação dos modelos preditivos.

5.1 Caracterização do processo de gestão de riscos

A caracterização do processo permitiu identificar a etapa na qual o uso de modelos preditivos seria testado como forma de aprimorar a gestão de risco e as informações potencialmente úteis para criação de tais modelos.

5.1.1 Estudo do processo de gestão de riscos

Os três tópicos iniciais apresentam informações organizadas em termos das atividades descritas nas seções 5.3, 5.4 e 5.5 da ISO 31000, no tópico final constam observações acerca da compatibilidade dos procedimentos identificados com as boas práticas aplicáveis à Administração Pública Distrital, conforme apontado no referencial teórico.

Estabelecimento do contexto

Quanto ao contexto externo, a fiscalização tributária do DF tem seu trabalho atrelado a leis, provenientes do poder judiciário, está sujeita a órgãos de controle externo (como o Tribunal de Contas do DF e a Corregedoria Fazendária), atende a demandas de diversos órgãos, como a Polícia Civil do DF e o Ministério Público do DF. Tem como objetivo verificar o comportamento dos contribuintes no que tange ao cumprimento das obrigações

tributárias, devendo para tanto observar a legislação vigente bem como as decisões judiciais de casos específicos e de repercussão geral, visando propiciar a correta arrecadação dos tributos, de forma a garantir recursos para que o Governo do DF tenha condições de cumprir o orçamento público e atender às necessidades da sociedade em geral.

No que tange ao contexto interno, o processo de investigação de indícios de irregularidade e seleção de contribuintes a fiscalizar, objeto do presente estudo, é conduzido pela Gerência de Programação Fiscal, que está subordinada à Coordenação de Sistemas Tributários e à Subsecretaria da Receita do DF.

A interpretação da legislação tributária é a base para gerar e manter scripts de mineração de dados capazes de efetuar cruzamentos entre os dados armazenados nos bancos institucionais e identificar indícios de irregularidade, que são utilizados para selecionar os contribuintes a serem fiscalizados pelas gerências executivas (de auditoria, monitoramento, fiscalização em trânsito e malha fiscal), também são observados os objetivos estratégicos, e os recursos disponíveis.

A Figura 5.1 apresenta os contextos internos da Gerência de Programação fiscal responsável pela seleção de contribuintes do ICMS e do ISS a fiscalizar.

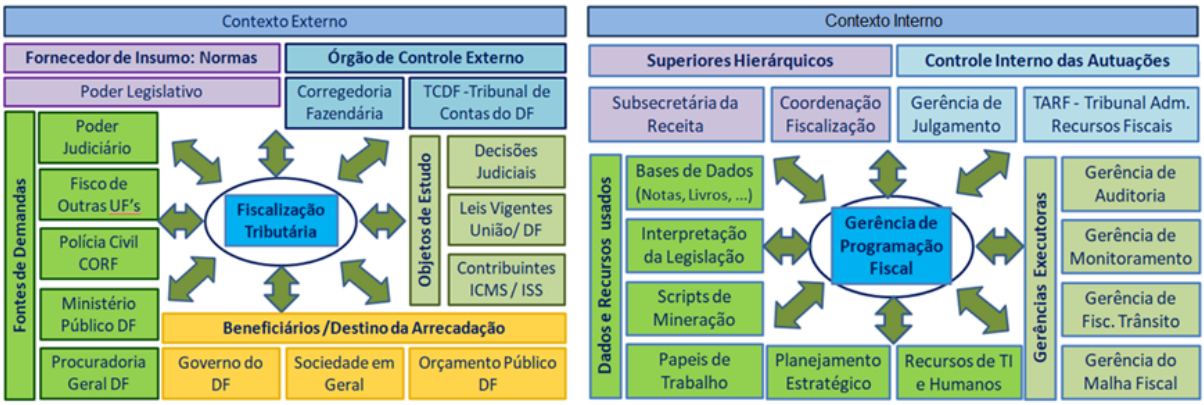


Figura 5.1: Contexto interno e externo da Gerência de Programação Fiscal

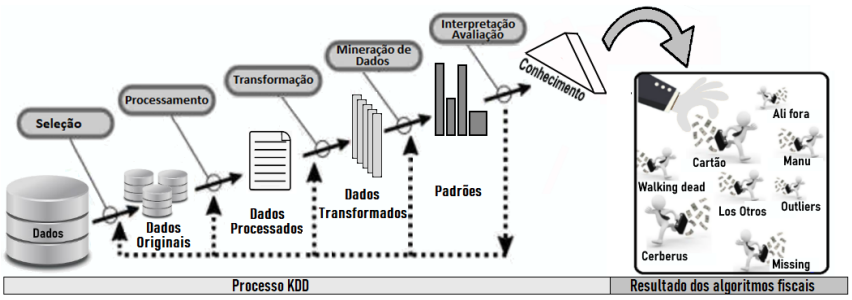
Processo de avaliação de Riscos

Os dados recebidos em meio magnético são submetidos a diversos algoritmos de mineração de dados, que efetuam cruzamentos específicos com informações das diferentes fontes e apresenta uma relação com os indícios de irregularidades identificados. Por questões estratégicas, o detalhamento dos algoritmos não consta do presente trabalho e por questões de sigilo fiscal nenhuma informação individualizada dos contribuintes auditados é exposta. Apenas dados globais são apresentados.

As listas de indícios geradas pelos algoritmos são avaliadas separadamente, e são mapeados os possíveis procedimentos fiscais nas empresas, de acordo com sua posição na

lista e com a natureza da diferença identificada. Os valores listados correspondem a situações distintas, tais como receitas omitidas cujo valor do imposto dependerá da atividade econômica da empresa ou dos produtos/serviços envolvidos; receitas não tributadas superfaturadas; imposto calculado incorretamente e créditos suspeitos. De modo que, cada tipo de indício implica em procedimentos de verificação distintos e diferentes valores a constituir, sendo que não existe um algoritmo que unifique as listas em uma "fila única".

A Figura 5.2 esquematiza o processo realizado por cada um dos algoritmos de seleção de indícios, sob à ótica do Processo de Knowlege Database Discovery (KDD).



Fonte Processo KDD: Fayyad (1996)

Figura 5.2: Processo de mineração de dados (KDD) aplicado às informações fiscais

A fiscalização dos indícios pode resultar em autuação ou encerramento da ação fiscal sem débitos. A Tabela 5.1 apresenta uma comparação anual entre a receita de origem tributária de ICMS e ISS arrecadada, extraída do Balanço Geral dos exercícios de 2006 a 2019 ([64], [65], [66], [67] e [14]) e o valor total do crédito constituído por meio de autos de infração desde 2002, oriundo do Sistema DAF.

Tabela 5.1: ICMS e ISS: receita arrecadada x autuações (R$ mil)

Ano	*Receita ICMS ISS*	*Crédito Autuado*	*%*	*Ano*	*Receita ICMS ISS*	*Crédito Autuado*	*%*
2002	2.149.173	283.480	13%	2011	6.171.451	1.175.278	19%
2003	2.601.833	738.122	28%	2012	6.821.347	534.151	8%
2004	3.085.159	343.190	11%	2013	7.502.109	1.953.460	26%
2005	3.500.512	417.642	12%	2014	8.228.595	2.528.881	31%
2006	3.939.691	488.488	12%	2015	8.281.246	1.505.878	18%
2007	4.143.667	777.953	19%	2016	9.226.484	1.002.480	11%
2008	4.730.927	922.279	19%	2017	9.550.407	1.750.887	18%
2009	4.892.566	922.166	19%	2018	10.041.904	1.140.136	11%
2010	5.543.231	1.089.976	20%	2019	10.189.433	1.821.393	18%
				Total	*110.599.735*	*19.395.840*	*18%*

É possível observar oscilações no percentual de crédito constituído em relação à arrecadação do exercício, sendo que em 17 anos a média corresponde a 18%.

A partir de 2007, toda escrituração fiscal passou a ser informada em arquivo magnético. Entretanto, a utilização de projetos de fiscalização voltados à apuração de indícios previamente obtidos por meio de algoritmos de mineração de dados se intensificou a partir de 2008, especialmente em razão do recebimento das informações apresentadas por administradoras de cartão de crédito e da implantação gradual da nota fiscal eletrônica.

A Tabela 5.2 mostra que não é uniforme o percentual de autuação das empresas incluídas em auditoria em razão dos diferentes algoritmos de mineração de dados.

Tabela 5.2: Quantidade de autuações por tipo de ação fiscal

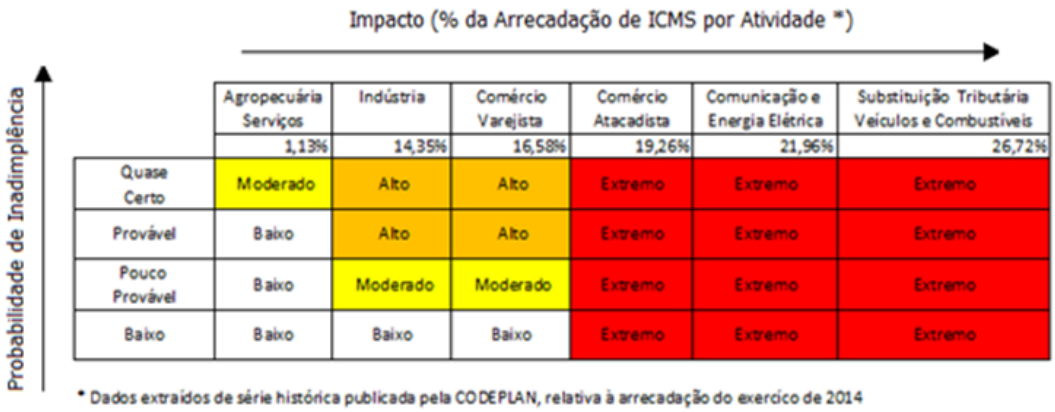
Tipo de Fiscalização	*Ações Realizadas*	*Autos Lavrados*	*% Autuado*	*Valor Autuado*
Auditoria	3.770	2.262	60	10.651.846.890,79
AEC - Missing	537	462	86	3.150.386.832,76
AEC - Cartão	2.011	1.643	81	1.526.193.959,03
Monitoramento	6.311	474	7	1.497.428.889,55
AEC - Ex-REA	19	19	100	222.765.473,82
AEC - Antecipado	207	183	88	200.281.563,89
AEC - Deleta	49	39	79	147.899.412,10
AEC - Cerberus	66	33	50	129.681.602,78
AEC - Conclusão Fiscal	4	4	100	109.236.446,14
AEC - Subprime	61	50	81	97.388.071,32
Baixa de Inscrição	9.908	1.195	12	92.623.548,05
Outros algoritmos/indícios	140.028	12.054	8	1.570.116.918,14
Total	*162.971*	*18.418*	*11*	*19.395.849.608,37*

A análise em termos de valor do indício de fraude é fundamental para a seleção de empresas a auditar. Entretanto, a existência do indício não garante que a irregularidade exista. Como exemplo, a segunda linha da Tabela 5.2 permite notar que 75 contribuintes fiscalizados em razão de indícios gerados pelos algoritmos AEC-Missing (537 - 462) não foram autuados, apesar de 86% dos contribuintes fiscalizados em decorrência deste algoritmo terem sido autuados.

Desta forma, a eficiência na seleção das empresas a auditar envolve a análise de outros fatores, cuja valoração ainda não é feita de forma automatizada. O desafio está em otimizar a utilização da capacidade operacional da fiscalização tributária de modo a investigar o conjunto mais relevante dentre os indícios disponíveis, visto que aprimorar a seleção de alvos a fiscalizar a partir desses dados é fundamental para o bom desempenho da fiscalização.

Tratamento dos riscos

O Mapa de Probabilidade e Consequência mostrado na Figura 5.3 ilustra a forma como é realizado o tratamento de riscos no âmbito da fiscalização tributária e como influenciou a estrutura organizacional da instituição.



Fonte: A autora (2018)

Figura 5.3: Matriz probabilidade consequência

As situações mapeadas de risco extremo são tratadas com monitoramento constante das empresas pela Gerências de Monitoramentos do ICMS e do ISS (GEMAE e GMISS), o risco alto é tratado predominantemente com a realização de auditoria, pela Gerência de Auditorias do ICMS e do ISS (GEAUT e GFISS).

Existe o monitoramento do cumprimento das obrigações acessórias, realizado pela Gerência do "Malha Fiscal", que emite alertas de indícios de irregularidades aos contribuintes para regularização, sendo que as irregularidades não sanadas e que possam implicar em falta de pagamento de tributos podem ser auditados, mas a fiscalização efetiva depende dos indícios disponíveis, ou seja, dos riscos mapeados.

O transporte de mercadorias é fiscalizado pela Gerência de Fiscalização de Mercadorias em trânsito (GEFMT), que utiliza métodos tais como escalas de plantão, escolha dos alvos com critérios mistos de seleção de alvos, adotando tanto indícios obtidos com mineração de dados e quanto a aleatoriedade na escolha dos alvos.

As ações fiscais são realizadas mediante a emissão de Ordem de Serviço, que inclui a empresa-alvo, o auditor responsável e o indício de irregularidade a ser investigado.

Os trabalhos são desenvolvidos com a apoio de sistemas corporativos mencionados na metodologia, PFI e DAF, que tratam da seleção de alvos e da execução da ação, bem como do Sistema Ação Fiscal em Estabelecimentos - AFE, que controla os prazos de execução.

No caso de lavratura de Auto de infração, os dados são armazenados no Sistema DAF, que foi implantado em 2005 e apresenta dados parciais a partir de 2001 e completos a partir de 2002, migrados do sistema anterior de controle das ações fiscais.

Avaliação de aderência às boas práticas aplicáveis ao DF

A caracterização do processo de gestão de risco indicou, resumidamente, que a seleção dos contribuintes a auditar passa por processos de identificação, análise e avaliação dos indícios de irregularidade (riscos). A execução periódica dos algoritmos que detectam os indícios e avaliam o risco permite monitorar o comportamento dos contribuintes ao longo do tempo. O tratamento dos riscos ocorre de acordo com a situação mapeada. Os contribuintes submetidos à fiscalização são incluídos em ordens de serviço, nas quais ficam formalizadas as responsabilidades pela execução e pelo controle do andamento da ação fiscal e de seus prazos e a estrutura organizacional leva em consideração o potencial de impacto do descumprimento das obrigações tributárias na arrecadação.

A ISO 31000 prevê quatro abordagens no tratamento dos riscos: tolerar, tratar, transferir ou terminar, cuja adoção deve considerar o potencial lesivo e o custo para mitigação do risco. Desta forma a separação das ações fiscais nos tipos Auditoria, Monitoramento, AEC e Diligência adere ao conceito de custo-benefício no tratamento dos riscos.

As boas práticas do COSO-ERM enfatizam a importância do conceito do apetite ao risco, e seu estabelecimento na organização. A estrutura organizacional da fiscalização do DF adere ao conceito de apetite ao risco, pois as atividades e setores econômicos onde a ocorrência de irregularidades possui maior potencial de impacto sobre a arrecadação foram mapeadas e sua fiscalização atribuídas a setores específicos, que ficam responsáveis pelo acompanhamento das situações que possam impactar negativamente o recolhimento dos tributos devidos. Entretanto, nas reuniões realizadas com as equipes de seleção de alvos e de execução das ações fiscais foi identificada a percepção de que a comunicação externa e divulgação dos resultados não reflete o trabalho realizado, o que prejudicaria a elevação da sensação de risco por parte dos contribuintes em geral.

As boas práticas do COSO-IC identificadas dizem respeito à definição e alinhamento dos objetivos operacional (criação de projeto de fiscalização), de comunicação interna (relatórios de acompanhamento) e de conformidade (nível de autuações alcançado), os componentes mais fortemente observados estão relacionados às atividades de monitoramento, controle e de avaliação de risco e monitoramento dos projetos. Entretanto, as informações e comunicações são restritas em face do direito ao sigilo fiscal das empresas. A separação das atividades na estrutura organizacional também é observada.

Assim, o processo de seleção de contribuintes a fiscalizar é compatível com as boas práticas de gestão de risco mapeado pela ISO 31000 e pelos frameworks COSO/IC e

COSO-ERM, possuindo fases dedicadas à identificação, avaliação e classificação dos riscos; ao controle das medidas de tratamento (auditorias, monitoramento, diligências e outras ações) e a avaliação dos resultados obtidos, permitindo a busca pelo aprimoramento no processo de gestão de risco.

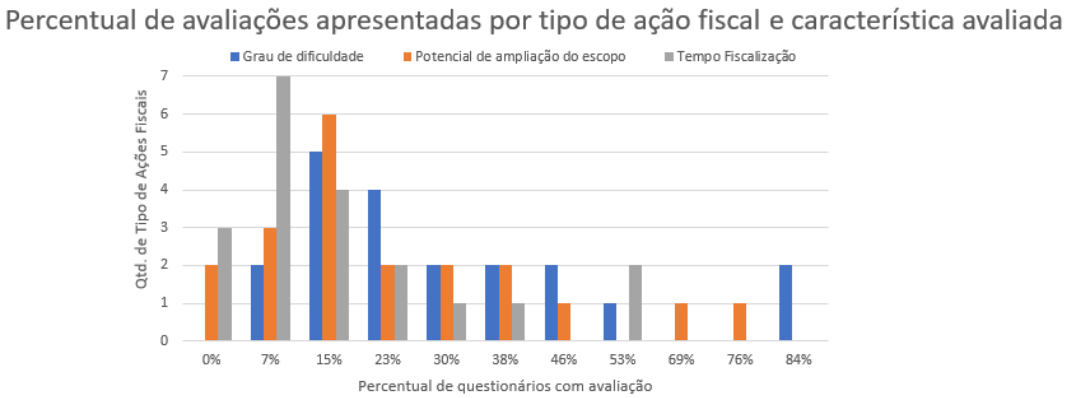
Resultado do questionário aplicado

O questionário foi respondido por 68% dos auditores envolvidos na execução das auditorias. A Tabela 5.3 apresenta o resultado das avaliações, sendo que o conceito corresponde ao valor da média das pontuações apresentadas.

Tabela 5.3: Resultado da aplicação do questionário

Tipo de ação fiscal	Grau de Dificuldade	Ampliação de Escopo	Tempo de fiscalização
AEC: Transp. Saldo Credor	Médio	-	-
AEC: Cartório	Médio	Baixo	Médio
AEC: WalkingDead	Médio	Baixo	Alto
AEC: Cartão; Missing; Manu; Ali-fora	Médio	Médio	Médio
AEC: Deleta	Médio	Médio	Alto
AEC: Customers	Alto	-	-
AEC-Subprime	Alto	Médio	-
Perícias	Alto	Médio	Muito baixo
AEC: Cartão-S.Nac.; Cred. Impr.; ExREA	Alto	Médio	Alto
AEC: Crédito podre; Imobilizado; Pro-DF	Alto	Alto	Alto
AEC: Cerberus; Reg. Especiais	Alto	Alto	Muito alto
Auditoria	Muito alto	Alto	Alto

A Figura 5.4 apresenta uma análise quantitativa dos questionários respondidos.



Fonte: A autora (2019)

Figura 5.4: Avaliação das respostas ao questionário aplicado

A análise quantitativa permite observar que dois tipos de ação fiscal não foram avaliados quanto ao potencial de ampliação de escopo; três tipos não foram avaliados quanto ao tempo de fiscalização e apenas 2 tipos de ação fiscal tiveram o grau de dificuldade avaliado por mais do que 50% dos auditores.

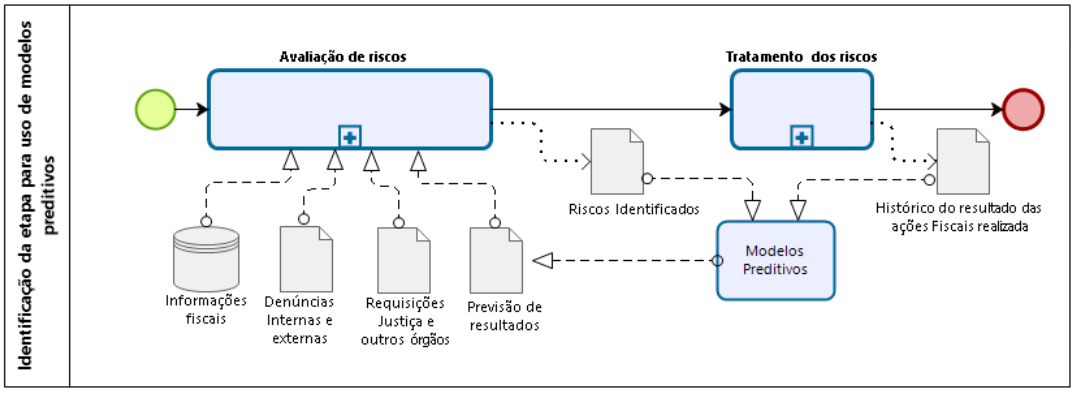
Em face da lacuna de representatividade dos parâmetros obtidos por meio do questionário, para a maioria dos tipos de ação fiscal avaliados, os dados não foram incorporados aos modelos elaborados, embora tenham servido para entender a complexidade do tema.

5.1.2 Seleção da etapa para uso de modelos preditivos

O estudo do processo de gestão de riscos sugere que a seleção dos contribuintes a fiscalizar pode ser aprimorada tendo em vista a variação identificada no percentual anual de autuações de ICMS e ISS em relação à arrecadação dos referidos tributos, mostrada na Tabela 5.1, e a diferença entre o número de empresas auditadas e o número de empresas autuadas, de cada algoritmo de seleção, mostrada na Tabela 5.2.

Neste contexto, a falta de informatização da comparação entre indícios provenientes dos diferentes algoritmos foi identificada como uma fronteira a ser investigada na busca pela otimização do uso dos recursos disponíveis, dada a possibilidade de existência de padrões não perceptíveis no tratamento manual, mas que poderiam ser eventualmente detectados com o uso de ferramentas de mineração de dados e de aprendizado de máquina.

O desafio consiste em construir modelos computacionais que permitam a comparação destes indícios e que sejam capazes de diminuir tanto a variação anual do percentual de autuação em relação ao total da receita arrecadada, quanto o percentual de empresas auditadas e não autuadas, usando para tanto o resultado das ações fiscais já realizadas, de modo que os modelos possam fazer predições para futuras ações fiscais a partir dos resultados históricos, conforme esquematizado na Figura 5.5



Fonte: A autora (2020)

Figura 5.5: Etapa na qual o uso de modelos preditivos será avaliado

5.1.3 Identificação de informações potencialmente úteis

Conforme exposto na metodologia, foi identificado junto aos auditores responsáveis pela seleção dos alvos da fiscalização quais seriam as informações potencialmente úteis para criação dos modelos preditivos.

Análises preliminares e critérios de seleção dos dados

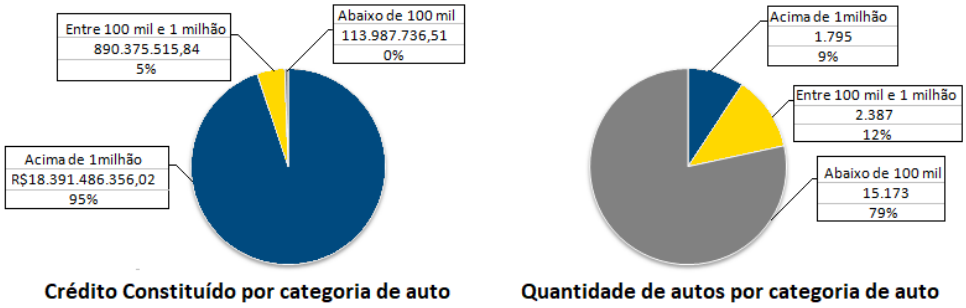
Inicialmente foram extraídos das bases de dados as informações relativas às 162.971 ações fiscais realizadas a partir de 2002 até 31/12/2019, a Tabela 5.4 apresenta o resultado por grupo tipo de ação fiscal definidos na metodologia.

Tabela 5.4: Resultado das ações fiscais por tipo de ação

Tipo de Ação	*Ações*	*Autos lavrados*	*%*	*Crédito Constituído (R$)*
Auditoria	3.770	2.262	60	10.651.846.890,79
AEC	15.809	4.975	31,46	6.495.970.062,21
Monitoramento	6.513	487	7,47	1.519.644.879,85
Diligência	133.361	10.521	7,88	441.744.086,95
Outras Ações	3.518	173	4,91	286.643.688,57
Total	*162.971*	*18.418*	*11,30*	*19.395.849.608,37*

A análise permite observar a participação da modalidade AEC, que engloba as ações decorrentes de algoritmos de mineração, no resultado total. É possível notar que mais indícios foram fiscalizados nesta modalidade, e que seu percentual de autuação no período correspondente a aproximadamente metade do obtido na modalidade auditoria.

A Figura 5.6 apresenta o perfil de autuações pelas categorias de valor e permite constatar que o valor dos créditos constituídos por auto de infração atende ao princípio de Pareto, sendo que 95% do valor do crédito constituído está relacionado aos autos de infração acima de R$ 1 milhão de reais, que correspondem a 9% das ocorrências.



Fonte: A autora (2020)

Figura 5.6: Perfil do crédito constituído por categoria de auto de infração

5.1.4 Análise dos dados selecionados para criação dos modelos

Com as limitações de escopo descritas na Seção 4.5, foram selecionados dados de 8.300 ações fiscais, com 12,64 bilhões de Reais em crédito constituído.

O conjunto de dados selecionados mantém o perfil de autuações, com 95,78% do crédito constituído associado a 15,14% das autuações, o percentual de empresas autuadas passou dos 11,30% iniciais para 45,85%, esse aumento decorre da retirada das ações fiscais do tipo diligência, cujo objetivo precípuo, conforme exposto na metodologia, não é a autuação.

A Tabela 5.5 apresenta o resumo dos dados considerando o valor da autuação, bem como as variáveis *dummy* por categoria de valor.

Tabela 5.5: Resumo das ações fiscais selecionadas após redução do escopo

Valor Autuado	*Ações*	*%*	*Valor Constituído*	*%*	*VN1*	*VN2*	*VN3*
Não autuado	4.578	55,15	0	0	0	0	0
até 100 mil	1.178	14,19	36.958.585,39	0,29	1	0	0
de 100 mil a 1 mi	1.287	15,5	496.605.526,54	3,92	0	1	0
acima de 1 mi	1.257	15,14	12.114.548.473,36	95,78	0	0	1
Total	*8.300*		*12.648.112.585,29*				

Dados cadastrais dos registros selecionados

Das 8.300 empresas fiscalizadas, 5.774 exercem como atividade principal o comércio, aliado ou não a uma atividade de serviço.

A Tabela 5.6 mostra o total de créditos constituídos por tipo de atividade de acordo com cinco das variáveis *dummy* criadas.

Tabela 5.6: Variáveis *dummy* para atividade econômica

Ações fiscais	*Autos*	*Crédito Constituído*	*VCme*	*VCmu*	*VInd*	*VTran*	*VSer*
3.317	1.862	5.288.592.719,15	1	0	0	0	0
2.457	1.052	3.290.757.582,02	1	0	0	0	1
148	66	1.199.212.484,84	0	1	0	0	1
16	5	810.056,82	0	1	0	0	0
217	99	1.160.770.147,33	0	0	1	0	1
368	109	340.213.802,95	0	0	1	0	0
152	67	529.910.007,83	0	0	0	1	1
44	14	6.817.368,4	0	0	0	1	0
1.509	423	799.084.587,39	0	0	0	0	1
49	15	26.215.889,97	0	0	0	0	0
8.300	*3.722*	12.648.112.585,29					

A Tabela 5.7 detalha o resultado das ações em empresas que exercem a atividade de comércio atacadista.

Tabela 5.7: Variáveis *dummy* para atividade de comércio atacadista

Ações Fiscais	*Autos*	*Crédito Constituído*	*VACme*	*VAtacado*
1.422	632	5.301.071.741,03	1	1
4.352	2.282	3.278.278.560,14	1	0
5.774	*2.914*	8.579.350.301,17		

A Tabela 5.8 apresenta as dummies criadas para identificar a localização das unidades da receita que atende aos contribuintes, e a quantidade de empresas por localidade.

Tabela 5.8: Variáveis *dummy* para localização

Qtd.	*VlBra*	*VlBrz*	*VlCei*	*VlGam*	*VlBan*	*VlOut*	*VlPla*	*VlSia*	*VlSob*	*VlTag*
1.547	0	0	0	0	0	0	0	0	0	1
232	0	0	0	0	0	0	0	0	1	0
1.497	0	0	0	0	0	0	0	1	0	0
149	0	0	0	0	0	0	1	0	0	0
381	0	0	0	0	0	1	0	0	0	0
295	0	0	0	0	1	0	0	0	0	0
606	0	0	0	1	0	0	0	0	0	0
501	0	0	1	0	0	0	0	0	0	0
65	0	1	0	0	0	0	0	0	0	0
3.027	1	0	0	0	0	0	0	0	0	0

Tipos de ações fiscais e indícios dos registros selecionados

A Tabela 5.9 mostra as variáveis *dummy* que identificam o tipo de ação fiscal e os resultados acumulados de cada tipo.

Tabela 5.9: Variáveis *dummy* para tipo de ação fiscal

Tipo de Ação	*Realizadas*	*Autuadas*	*Crédito Total*	*VAec*	*VAud*	*VMon*
AEC	4.057	2.864	6.335.670.423,03	1	0	0
Auditoria	841	620	6.049.841.529,42	0	1	0
Monitoramento	3.402	238	262.600.632,84	0	0	1

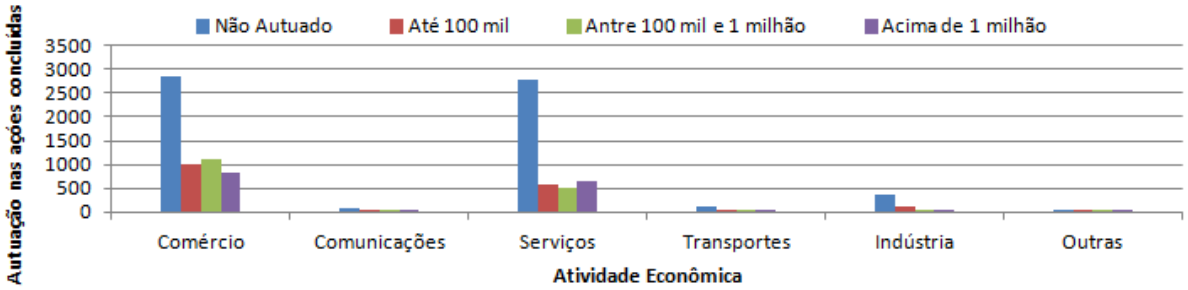
A Tabela 5.10 apresenta os resultados para as 6 dummies de indícios fiscalizados mais representativas em termos de valores autuados, a sétima linha apresenta o resultado global das outras 8 dummies (VAut, VCer, VCre, VCst, VImo, VLos e VWal).

Tabela 5.10: Variáveis *dummy* para tipo de indícios

Realizadas	*Autuadas*	*Crédito Total*	*VMis*	*VCar*	*VReg*	*VAnt*	*VDel*	*VSub*
553	481	3.170.265.085,48	1	0	0	0	0	0
52.107	1.720	1.533.434.502,74	0	1	0	0	0	0
48	41	230.306.685,91	0	0	1	0	0	0
207	182	200.196.818,69	0	0	0	1	0	0
49	39	147.871.677,54	0	0	0	0	1	0
61	56	131.045.055,23	0	0	0	0	0	1
1.032	345	922.550.597,44	0	0	0	0	0	0

Quantidade de fiscalizações e autuações por grupo de variáveis

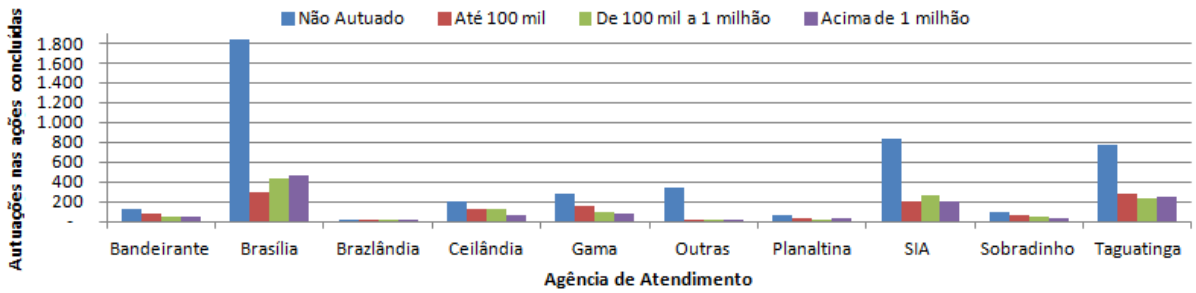
A Figura 5.7 permite notar que a maioria das ações fiscais ocorreu em empresas de comércio e serviços e que a proporção de autuações não é constante entre os setores.



Fonte: A autora (2020)

Figura 5.7: Perfil da quantidade de autos por atividade econômica

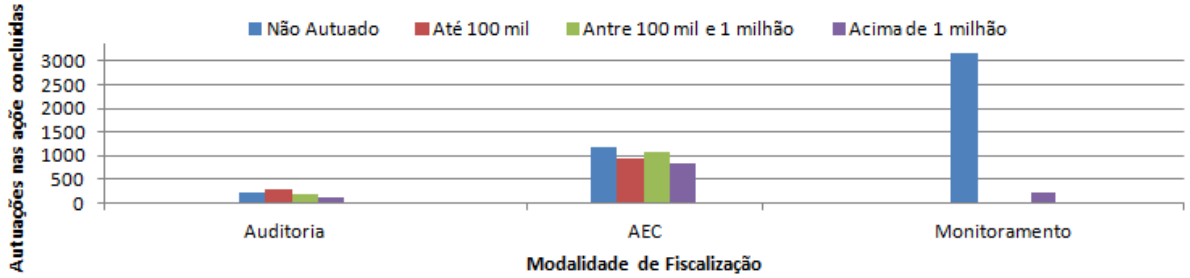
A Figura 5.8 apresenta a quantidade de autuações por localização. A proporção de empresas não autuadas varia com a localidade, mas o não é possível perceber um padrão.



Fonte: A autora (2020)

Figura 5.8: Perfil da quantidade de autos por endereço de atendimento da empresa

A Figura 5.9 apresenta a quantidade de autuações por tipo de fiscalização realizada. A baixa quantidade de autos lavrados na modalidade de monitoramento é reflexo dos objetivos da fiscalização, que visa ao adimplemento espontâneo das obrigações fiscais.



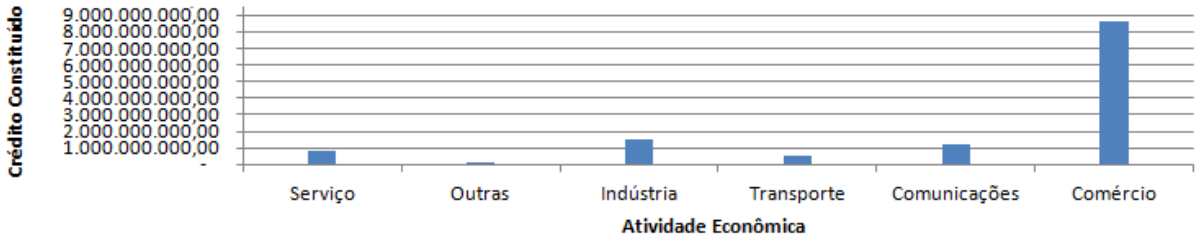
Fonte: A autora (2020)

Figura 5.9: Perfil da quantidade de autos por tipo de ação fiscal

A análise indica que os resultados das ações fiscais em termos de atividade econômica, localização da empresa e tipo de ação fiscal não apresentam uma uniformidade perceptível na análise visual, a qual não permite estabelecer um padrão de comportamento.

Crédito total constituído por grupo de variáveis

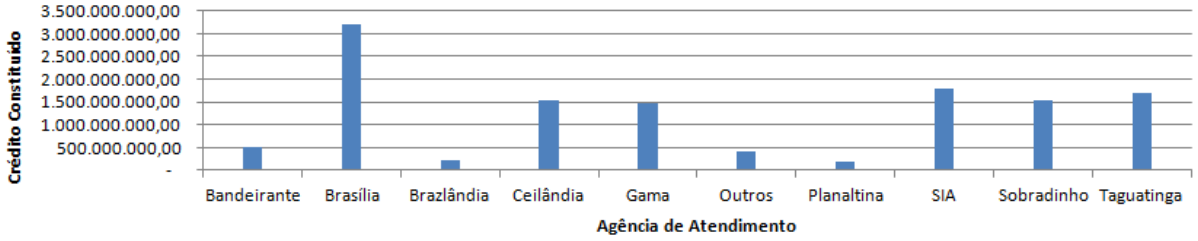
Na análise por atividade econômica, é possível constatar que a atividade comércio responde pela maior parcela o valor total do crédito constituído, conforme Figura 5.10



Fonte: A autora (2019)

Figura 5.10: Perfil dos valores constituídos por atividade econômica

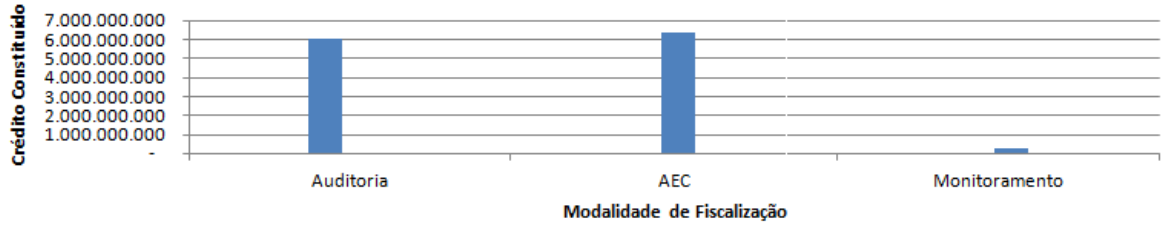
A análise por localização apresentada na Figura 5.11 permite notar que não existe uma relação linear entre o total constituído e a quantidade de empresas autuadas por localização que foi apresentada na Figura 5.8.



Fonte: A autora (2020)

Figura 5.11: Perfil dos valores constituídos por endereço de atendimento da empresa

A Figura 5.12 apresenta o crédito total constituído por tipo de fiscalização.



Fonte: A autora (2020)

Figura 5.12: Perfil dos valores constituídos por tipo de ação fiscal

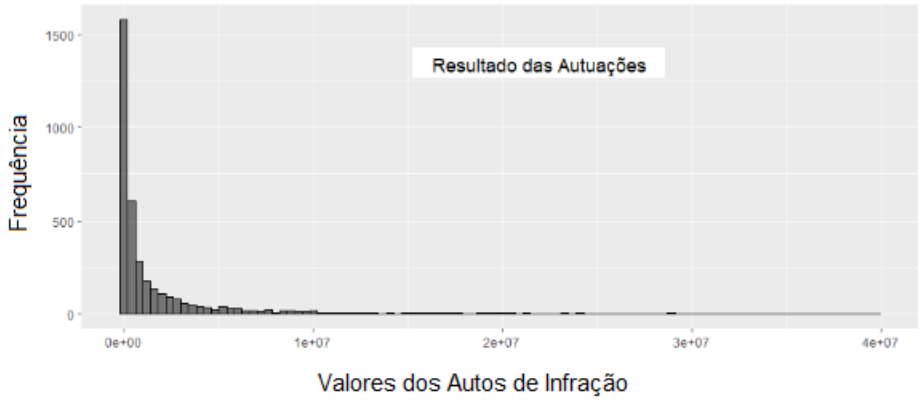
A exemplo observado para a quantidade de autos lavrados no tipo de ação fiscal Monitoramento, o valor constituído é inferior aos valores dos outros tipos de fiscalização, porque autuar não é o objetivo do monitoramento, sendo possível observar que o total do crédito constituído em Auditoria encontra-se em patamar similar aos valores resultantes de AECs.

As medidas resumo do crédito constituído no conjunto dos registros selecionados são apresentadas na Tabela 5.11 e demonstram que não existe simetria nas observações.

Tabela 5.11: Medidas de tendência central e de dispersão - variável crédito constituído

Média	*Moda*	*1º Quartil*	*Mediana*	*3º Quartil*	*Máximo*
1.523.868,99	0	0	0	224.347,80	582.702.580,63

A Figura 5.13 apresenta a distribuição de frequência por valor de auto lavrado, sem os *outliers* cujos valores superam 40 milhões de reais.



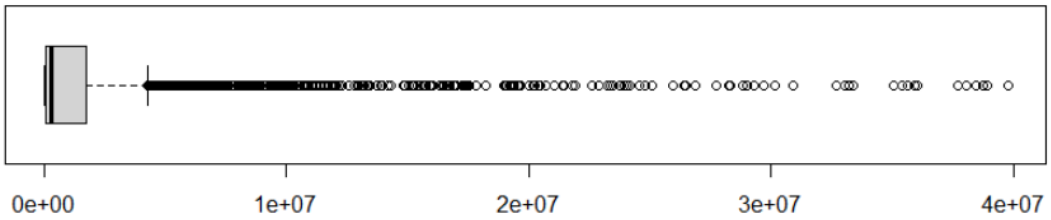
Fonte: A autora (2020) - Utilizando o Software R

Figura 5.13: Distribuição de frequência dos valores autuados sem os *outliers*

A curva permite observar a ocorrência das autuações, no que tange ao valor do crédito constituído, não pode ser modelada por uma função linear, nem pela curva normal.

A Figura 5.14 apresenta o Boxplot do conjunto de observações sem os *outliers* acima de 40 milhões de reais variável, permite observar a assimetria e a existência de diversos

valores extremos de forma mais clara do que a curva de distribuição de frequências da Figura 5.13.



Fonte: A autora (2020)

Figura 5.14: BoxPlot - crédito constituído sem os *outliers* acima de 40 milhões

As análises apresentadas para os valores dos autos em face da atividade econômica, localização e tipo de ação fiscal, a exemplo das análises referente às quantidades autuadas, não permitiram estabelecer um padrão entre as características listadas e o resultado das ações fiscais. Entretanto, tais análises não afastam a eventual existência de padrões ocultos, cuja detecção por meio de modelos preditivos de mineração de dados é objeto de estudo da presente pesquisa.

5.2 Identificação dos modelos preditivos aplicáveis

Uma vez que foi identificada a necessidade de utilizar o histórico do resultado das auditorias já realizadas como base para prever o grupo de contribuintes que terá maior chance de ser autuado, a revisão teórica permitiu verificar que a modalidade de aprendizado de máquina mais adequada ao problema corresponde ao aprendizado supervisionado, conforme descrito na Seção 3.2.2, página 19.

Quanto à escolha do modelo de mineração de dados, a necessidade identificada se amolda às características do modelo de regressão logística que, conforme pode ser visto na Seção 3.2.3, página 20, representa um processo de aprendizagem de relações entre entradas e saídas a partir de dados de exemplo a fim de efetuar previsões da categoria de saída para novas entradas. Desta forma o primeiro modelo preditivo escolhido para implementação foi o modelo de regressão logística binomial ou binária.

A revisão teórica também permitiu identificar o segundo modelo preditivo de mineração de dados aplicável ao caso, o modelo de redes neurais, em razão de sua capacidade de detectar relações subjacentes dentro de um conjunto de dados e de efetuar predições, conforme apontado na Seção 3.3.2, página 26.

A escolha de dois modelos foi influenciada pela indicação da necessidade do uso simultâneo de outros modelos quando se utiliza o modelo de redes neurais, para lidar com as incertezas decorrentes do uso de redes neurais, e pela identificação do uso do modelo

logístico associados a modelos de redes neurais em trabalhos correlacionados, conforme detalhado na Seção 3.3.2 do referencial Teórico, páginas 28 e 29.

5.3 Criação dos modelos preditivos

A presente seção detalha os modelos criados para procurar padrões ocultos na base de dados selecionada, de modo a verificar se é possível predizer se uma empresa será autuada, em face do histórico do resultado das ações fiscais, do tipo de ação fiscal, localização da empresa, atividade econômica e algoritmo utilizado para identificação dos indícios.

5.3.1 Modelos de regressão logística

Os modelos de regressões logísticas seguem a seguinte equação geral:

$$LogitP(Y) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + ...$$

Inicialmente, foi criado um modelo para as 8.300 observações, tendo como variável dependente Vautuado, que apresenta 1 para contribuinte "autuado" e 0 para "não autuado", e como variáveis explicativas todas as *dummies* relativas à modalidade de fiscalização, à localização, à atividade econômica e ao algoritmo de seleção dos indícios. Foram criadas variações do modelo inicial com a alteração da variável dependente para estudo das ações fiscais com autuação acima de 1 milhão de reais e para análise das ações fiscais da modalidade AEC, que correspondem a 4.057 observações, e com retirada de variáveis independentes estatisticamente pouco significativas.

Dentre as opções analisadas, merecem destaque os seguintes modelos:

- Modelo RL 1: Autuações de qualquer valor e base de dados com 8.300 registros.
- Modelo RL 2: Autuações acima de R$ 1milhão e base de dados com 8.300 registros.
- Modelo RL 3: Autuações de qualquer valor e base com 4.057 registros (só AEC).
- Modelo RL 4: Autuações acima de R$ 1milhão e base com 4.057 registros (só AEC).

As etapas de criação, treinamento, validação e predição e avaliação dos resultados foram realizadas com uso das funções descritas na metodologia. A técnica de validação cruzada foi implementada na proporção de 3/4 dos dados para treinamento e 1/4 para validação. Para cada modelo foi construída a matriz de confusão com obtenção das medidas de Acurácia, sensibilidade (*recall*), especificidade (*precision*) e (*F1-Score*).

Modelo RL 1: autuações de qualquer valor

A fórmula geral da regressão logística é:

$$LogitP(VAutuadoQQ) =$$
$$\beta_0 + \beta_1 VAcmu + \beta_2 VAserv + \beta_3 Vind + \beta_4 VAtacado + \beta_5 VAec + \beta_6 VAud + \beta_7 VBra +$$
$$\beta_8 VBrz + \beta_9 VCei + \beta_{10} VGam + \beta_{11} VBan + \beta_{12} VPla + \beta_{13} VSia + \beta_{14} VlSob + \beta_{15} VlTag$$

A Figura 5.15 apresenta o valor dos coeficientes e a significância de cada variável.

```
Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept)  4.06711    0.19657  20.691  < 2e-16 ***
Vcmu        -0.53093    0.22727  -2.336   0.0195 *
Vser         0.33753    0.06040   5.588 2.30e-08 ***
Vind         0.49802    0.12041   4.136 3.53e-05 ***
Vatacado     0.45848    0.07719   5.940 2.85e-09 ***
VAec        -3.45997    0.07868 -43.975  < 2e-16 ***
VAud        -3.69573    0.10702 -34.533  < 2e-16 ***
VlBra       -1.85824    0.19233  -9.662  < 2e-16 ***
VlBrz       -2.15215    0.36765  -5.854 4.81e-09 ***
VlCei       -2.05377    0.21381  -9.606  < 2e-16 ***
VlGam       -1.82734    0.20588  -8.876  < 2e-16 ***
VlBan       -2.25513    0.23836  -9.461  < 2e-16 ***
VlPla       -1.80412    0.26413  -6.830 8.46e-12 ***
VlSia       -1.57984    0.19420  -8.135 4.11e-16 ***
VlSob       -1.71603    0.23971  -7.159 8.14e-13 ***
VlTag       -1.82591    0.19378  -9.423  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Fonte: A autora (2020)

Figura 5.15: Coeficientes das variáveis explicativas Modelo RL 1 com o Software R

A matriz de confusão, Figura 5.15, permite observar que Acurácia do modelo é de 82,16%, a sensibilidade (recall) é de 77,31%, a especificidade (precision) é de 81,86% e a média harmônica entre sensibilidade e especificidade *(F1-SCore)* é de 81,91%.

```
pred          Autuado Não Autuado
  Autuado         719          159
  Não Autuado     211          985
```

```
 Accuracy
0.8216008
[1] 0.7731183
[1] 0.8189066
[1] 0.795354
```

Fonte: A autora (2020)

Figura 5.16: Matriz de confusão Modelo RL 1 com o Software R

A Tabela 5.12 apresenta um resumo da avaliação do resultado em termos de valores autuados, correspondentes às categorias previstas para autuação e não autuação na base de validação, sendo possível observar o modelo identificou corretamente 719 empresas autuadas e que estas ações respondem por 83,93% do crédito constituído (aproximadamente 2.1 bilhões de Reais em 2.6 bilhões de reais). Também foram identificadas de forma correta 985 empresas não autuadas.

Tabela 5.12: Modelo RL 1: crédito constituído x autuações previstas

VAutuadoQQ		Crédito	Qtd.	Total	%	%
Previsto	Realizado	constituído	Contrib.	Contrib.	Valor	Qtd
Autuado	Autuado	2.189.971.868,17	719	878	83,93	42,33
	Não Autuado	0	159			
Não Autuado	Autuado	419.106.935,39	211	1.196	16,06	57,66
	Não Autuado	0	985			
Total		2.609.078.803,56	2.074			

O modelo classificou 1.196 ações fiscais como não propensas a autuação, o que corresponde a 57,66% das ações realizadas. Assim, a aplicação do modelo implicaria em fiscalizar menos da metade das empresas (42,23%) e manter 83,93% do valor constituído.

Estes resultados indicam que o uso de mineração de dados poderia ter otimizado o aproveitamento dos recursos disponíveis na situação em análise, que poderiam ter sido utilizados para fiscalização de outros contribuintes, com a priorização de outros indícios disponíveis para averiguação.

Modelo RL 2: autuações acima de R$1milhão

O objetivo da análise está vinculado à avaliação preliminar dos dados, segundo a qual as autuações acima de R$1 milhão (aproximadamente 9% das autuações) correspondem a 95% do valor total, para verificar se é possível identificar padrões ocultos de relacionamentos entre as variáveis e o resultado específico de autuações acima de R$ 1 milhão.

A Figura 5.17 apresenta os coeficientes e a significância das 14 variável explicativas:

```
Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  5.61690    0.29562  19.000  < 2e-16 ***
VAec        -3.42544    0.30413 -11.263  < 2e-16 ***
VAud        -4.90775    0.29872 -16.429  < 2e-16 ***
Vcmu        -1.34370    0.24209  -5.550 2.85e-08 ***
Vtran       -0.83025    0.22872  -3.630 0.000283 ***
Vind        -0.61371    0.14869  -4.127 3.67e-05 ***
Vatacado    -0.61016    0.09883  -6.174 6.67e-10 ***
VCar        -0.70137    0.11562  -6.066 1.31e-09 ***
VDel        -2.03774    0.30620  -6.655 2.84e-11 ***
VMis        -2.80726    0.13317 -21.080  < 2e-16 ***
VReg        -2.29085    0.31686  -7.230 4.83e-13 ***
VSub        -1.11021    0.31580  -3.516 0.000439 ***
VlBra        0.36228    0.08995   4.027 5.64e-05 ***
VlSia        0.34189    0.10076   3.393 0.000691 ***
VlOut        1.00932    0.28050   3.598 0.000320 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Fonte: A autora (2020)

Figura 5.17: Coeficientes das variáveis explicativas Modelo RL 2 com o Software R

A fórmula geral é semelhante à fórmula do modelo RL1, substituindo a variável dependente por VautuadoN3 e as variáveis independentes por VAec, VAud, Vcme, Vcmu, Vtran, Vind, VAtacado, VlCei, VlGam, VlBan, VlSob e VlTag.

A matriz de confusão, Figura 5.18, permite calcular a Acurácia de 87,51%, a sensibilidade de 31,84%, a especificidade de 68,96%, e *F1-Score* de 43,57%.

```
pred          Autuado Não Autuado
  Autuado         100          45
  Não Autuado     214        1715
```

```
 Accuracy
0.8751205
[1] 0.3184713
[1] 0.6896552
[1] 0.4357298
```

Fonte: A autora (2020)

Figura 5.18: Matriz de confusão Modelo RL 2 com o Software R

A Tabela 5.13 apresenta a comparação entre as autuações efetivas e previstas:

Tabela 5.13: Modelo RL 2: crédito constituído (VN3) x autuações previstas

VAutuadoN3		Crédito constituído	Qtd. Contrib.	Total Contrib.	% Valor	% Qtd
Previsto	Realizado					
Autuado	Autuado	1.022.390.825,27	100	145	32,40	6,99
	Não Autuado	(*) 0	45			
Não Autuado	Autuado	2.132.983.197,4	214	1.929	67,60	93,00
	Não Autuado	(**) 0	1.715			
Total		3.155.374.022,67	2.074			

(*) 45 contribuintes com autuação total de R$9.967.450,67, em autos cujos valores individuais são inferiores a R$ 1 milhão.
(**) 1.715 contribuintes com autuação total de R$117.474.721,70, em autos cujos valores individuais são inferiores a R$ 1 milhão.

A sensibilidade indica que o modelo foi capaz de identificar 31,84% do total das autuações ocorridas (100 em 314), enquanto a Tabela 5.12 permitiu observar que estas ações fiscais respondem por 32,40% do total do crédito constituído. Deste modo, o erro na previsão resultaria em não autuação de 2,13 bilhões de Reais, caso o modelo tivesse sido utilizado como fator preponderante para escolha dos alvos da fiscalização.

Modelo RL 3: AECs e autos de qualquer valor

O modelo foi gerado com as 4.057 observações relacionadas AEC, com o objetivo de verificar se é possível identificar padrões ocultos de relacionamento considerando apenas os registros relacionados às auditorias especiais concentradas, sua fórmula geral apresenta a mesma configuração do Modelo RL1, sendo que as variáveis, juntamente com seus coeficientes e significância são apresentados na Figura 5.19.

```
Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  2.63633    0.40139   6.568 5.10e-11 ***
Vcmu         1.83985    0.49934   3.685 0.000229 ***
Vser         0.41178    0.08506   4.841 1.29e-06 ***
Vind         0.91785    0.17267   5.316 1.06e-07 ***
Vatacado     0.45100    0.12214   3.692 0.000222 ***
VAli        -2.62550    0.35174  -7.464 8.38e-14 ***
VAnt        -2.85578    0.23540 -12.132  < 2e-16 ***
VCar        -2.30815    0.10696 -21.579  < 2e-16 ***
VCer        -1.01221    0.26457  -3.826 0.000130 ***
VCre        -2.20133    0.29130  -7.557 4.13e-14 ***
VDel        -2.48632    0.37082  -6.705 2.02e-11 ***
VImo        -4.04280    0.90332  -4.475 7.62e-06 ***
VMis        -2.96260    0.15832 -18.713  < 2e-16 ***
VReg        -3.00012    0.42942  -6.986 2.82e-12 ***
VSub        -3.21042    0.47736  -6.725 1.75e-11 ***
VlBra       -1.97755    0.40560  -4.876 1.08e-06 ***
VlBrz       -2.11534    0.56905  -3.717 0.000201 ***
VlCei       -2.40348    0.42635  -5.637 1.73e-08 ***
VlGam       -2.22272    0.41832  -5.313 1.08e-07 ***
VlBan       -2.63902    0.45737  -5.770 7.93e-09 ***
VlPla       -2.36979    0.47481  -4.991 6.01e-07 ***
VlSia       -1.87993    0.40678  -4.621 3.81e-06 ***
VlSob       -2.53750    0.47042  -5.394 6.89e-08 ***
VlTag       -2.09299    0.40673  -5.146 2.66e-07 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Fonte: A autora (2020)

Figura 5.19: Coeficientes das variáveis explicativas Modelo RL 3 com o Software R

A matriz de confusão do modelo, Figura 5.20, permite calcular a acurácia do modelo, 81,36%, a sensibilidade, 91,62%, a especificidade, 83,56% e *F1-Score*, 87,40%:

```
pred          Autuado Não Autuado
  Autuado         656          129
  Não Autuado      60          169
```

```
Accuracy
0.8136095
[1] 0.9162011
[1] 0.8356688
[1] 0.8740839
```

Fonte: A autora (2020)

Figura 5.20: Matriz de confusão Modelo RL 3 com o Software R

A comparação das autuações previsões com as efetivas é apresentada na Tabela 5.14:

Tabela 5.14: Modelo RL 3: crédito constituído em AEC x autuações previstas

VAutuadoQQ		Crédito	Qtd.	Total	%	%
Previsto	Realizado	constituído	Contrib.	Contrib.	Valor	Qtd
Autuado	Autuado	1.135.119.762,29	656	785	90,10	77,42
	Não Autuado	0	129			
Não Autuado	Autuado	124.850.505,5	60	229	9,90	22,58
	Não Autuado	0	169			
Total		1.259.970.267,79	1.014			

A exemplo do modelo RL 1, o modelo RL3 possui todas as métricas de desempenho elevadas e a Tabela 5.14 mostra que as autuações previstas respondem por 90,10% do crédito constituído, indicando que seu uso teria otimizado os resultados da fiscalização.

Modelo RL 4: AECs e autos acima de R$ 1 milhão

O modelo apresenta a fórmula geral com a configuração do Modelo RL2, sendo que as variáveis, seus coeficientes e significância são apresentados na Figura 5.21.

```
Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept)   3.2113     0.2296  13.985  < 2e-16 ***
Vcme         -1.0309     0.2144  -4.809 1.52e-06 ***
Vtran        -1.6260     0.3679  -4.420 9.86e-06 ***
Vind         -1.1232     0.2660  -4.223 2.41e-05 ***
Vatacado     -0.3361     0.1208  -2.781 0.005418 **
VCar         -0.7399     0.1230  -6.013 1.82e-09 ***
VCre         -1.0577     0.2883  -3.668 0.000244 ***
VDel         -2.1842     0.3048  -7.166 7.74e-13 ***
VMis         -2.9517     0.1367 -21.600  < 2e-16 ***
VReg         -2.6651     0.3269  -8.153 3.54e-16 ***
VSub         -1.1085     0.3167  -3.500 0.000466 ***
VlBra         0.3425     0.1014   3.378 0.000730 ***
---
Signif. codes:
0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Fonte: A autora (2020)

Figura 5.21: Coeficientes das variáveis explicativas Modelo RL 4 com o Software R

A Figura 5.22 apresenta a matriz de confusão do modelo, sendo possível observar que Acurácia do modelo é de 82,13%, a sensibilidade (*recall*) é de 45,998%, a especificidade (*precision*) é de 67,28% e F1-Score é de 54,63%.

```
pred          Autuado Não Autuado
  Autuado         109          53
  Não Autuado     128         723
```

```
 Accuracy
0.8213228
[1] 0.4599156
[1] 0.6728395
[1] 0.5463659
```

Fonte: A autora (2020)

Figura 5.22: Matriz de confusão Modelo RL 4 com o Software R

A Tabela 5.15 apresenta a comparação entre as autuações efetivas e previstas:

Tabela 5.15: Modelo RL 4: crédito constituído em AEC (VN3) x autuações previstas

VAutuadoVN3		Crédito	Qtd.	Total	%	%
Previsto	Realizado	constituído	Contrib.	Contrib.	Valor	Qtd
Autuado	Autuado	668.422.469,22	109	162	48,85	15,99
	Não Autuado	(*) 0	53			
Não Autuado	Autuado	592.020.681,15	128	851	43,26	84,00
	Não Autuado	(**) 0	723			
Total		1.368.249.670,35	1.013			

(*) 53 contribuintes com autuação total de R$9.967.450,67, em autos cujos valores individuais são inferiores a R$ 1 milhão.
(**) 723 contribuintes com autuação total de R$96.269.201,73, em autos cujos valores individuais são inferiores a R$ 1 milhão.

A exemplo do que ocorreu com o modelo RL 2, o modelo apresenta a acurácia elevada, mas a sensibilidade e o F1-Score são baixos, o modelo não se mostrou eficaz para identificar os contribuintes autuados. Fato que reflete no resultado em termos de valores de autuação correspondente, que ficou abaixo de 50% do total.

Tais resultados indicam que este modelo, com as variáveis utilizadas, não é potencialmente útil na comparação de ações fiscais a realizar.

Análise dos resultados dos modelos de regressão logística

Nos modelos voltados para a predição do resultado autuação, a acurácia medida indica que o nível global de acertos é elevado em todos os modelos. Entretanto, a análise de sensibilidade demonstra que o número de acertos dos casos positivos é baixo quando a análise leva em conta apenas as autuações acima de 1 milhão de Reais, sendo inferior a 50%, independentemente de serem consideradas todas as ações fiscais ou apenas as ações do tipo AEC.

A Tabela 5.16 apresenta um resumo da avaliação dos modelos de regressão logística e comprova que a baixa identificação dos contribuintes autuados dos modelos RL2 e RL4 se reflete no valor do crédito que seria constituído, enquanto os modelos RL1 e RL3 apresentam valores elevados de acurácia, sensibilidade e manutenção do crédito constituído.

Tabela 5.16: Modelos de regressão logística

Modelo			*Acurácia*	*Sensibilidade*	*Especificidade*	*F1 Score*	*% Valor (Predição)*
Nome	*Ação*	*Autos*					
RL 1	Todas	Todos	81,43%	93,76%	72,72%	81,91%	83,93%
RL 2	Todas	>1 mi	87,51%	31,84%	68,96%	43,57%	32,40%
RL 3	AECs	Todos	81,36%	91,62%	83,56%	87,40%	90,10%
RL 4	AECs	>1 mi	82,13%	45,99%	67,28%	54,63%	48,85%

5.3.2 Modelos de redes neurais

Utilizando o mesmo processo de criação e as mesmas premissas dos modelos de regressão logística, de modo a permitir a comparação dos resultados, foram criados os seguintes os modelos preditivos de redes neurais:

- Modelo RN 1: Autuações de qualquer valor e base de dados com 8.300 registros.

- Modelo RN 2: Autuações acima de R$ 1milhão e base de dados com 8.300 registros.
- Modelo RN 3: Autuações de qualquer valor e base com 4.057 registros (só AEC).
- Modelo RN 4: Autuações acima de R$ 1milhão e base com 4.057 registros (só AEC).

Para facilitar a criação das redes neurais, as variáveis não usadas no modelo foram retiradas das bases de dados, assim os modelos que tratam de autuações de qualquer valor ficaram apenas com a variável numérica VAutuado e os modelos que tratam das autuações acima de R$ 1 milhão ficaram apenas com a variável numérica VN3, sendo que as demais variáveis relacionadas ao valor da autuação foram removidas dos modelos.

Para cada modelo foram usados ajustes diferentes de número de camadas ocultas, de épocas de treinamento, e de composição dos dados de teste e treinamento, cujos valores foram escolhidos de forma aleatória, respeitando sempre a proporção de 3/4 dos dados para treinamento e 1/4 para testes.

São apresentados os resultados para 1000 épocas de treinamento porque os modelos não mostraram variação significativa nas medidas de avaliação entre mil e dez mil épocas, foram detectadas variações entre os modelos de 1 e 2 camadas ocultas, onde os modelos com duas camadas apresentaram menos variação entre os testes.

As bases de dados foram preparadas com as variáveis utilizadas no modelo a variável dependente foi alterada em cada base de acordo com o modelo, mas as variáveis explicativas foram as mesmas, tendo sido usadas todas as dummies de atividade econômica e de localização, bem como as seguintes variáveis relacionadas aos indícios das AECs: *VAli* ; *VAnt* ; *VCar* ; *VCer* ; *VCre* ; *VCst* ; *VDel* ; *VImo* ; *VLos* ; *VMis* ; *VReg* ; *VSub* ; *VWal* e *VEcf*. As variáveis *VAfi* e *VAut* não foram utilizadas porque não possuem registros no período para os tipos de auditoria em análise.

Modelo RN 1: todas autuações e todas as ações

Neste modelo foram utilizados os 8.300 registros relacionados a auditorias e auditorias especiais concentradas, o modelo apresenta Vautuado como variável dependente, e visa verificar se é possível predizer o resultado autuação (independentemente do valor) com as variáveis disponíveis.

A Figura 5.23 mostra a matriz de confusão de um dos testes do modelo:

```
> matriz_confusao

previsoes   0   1
        0 951 114
        1 193 816
```

Cálculo das medidas de avaliação:

TN	FP	FN	TP	ACC	Recall	Precision	F1Score
951	193	114	816	85,20%	87,74%	80,87%	84,17%

Fonte: A autora (2020)

Figura 5.23: Matriz de confusão Modelo RN 1 com o Software R

A Tabela 5.17 apresenta as medidas de desempenho para alguns testes realizados, com ajustes na estrutura da rede e mudança (aleatória) dos registros que compõe as bases de treinamento e validação.

Tabela 5.17: Modelo RN 1: testes realizados

RN	Cam.Ocultas	TN	FP	FN	TP	ACC	Sensib.	Especif.	F1Score
1	2	957	187	117	813	85,34 %	87,41 %	81,3 %	84,24 %
1	2	933	211	101	829	84,95 %	89,13 %	79,71 %	84,16 %
1	1	952	192	106	824	85,63 %	88,60 %	81,10 %	84,68 %
1	1	957	187	128	802	84,81 %	86,23 %	81,09 %	83,58 %

A análise dos dados permite notar que tanto a acurácia do modelo quanto as demais medidas se alteram, mas mantém valores próximos, mesmo com alteração do número de camadas ocultas. Foram realizados testes com quantidade maior de épocas de treinamento, mas os resultados também não apresentaram mudança significativas.

A Tabela 5.18 apresenta a comparação entre as autuações efetivas e previstas, obtidas na rede neural de duas camadas cujo resultado na Tabela 5.17 obteve a menor acurácia, no caso 84,95%.

Tabela 5.18: Modelo RN 1: crédito constituído x autuações previstas

VAutuado		Crédito	Qtd.	Total	%	%
Previsto	Realizado	constituído	Contrib.	Contrib.	Valor	Qtd
Autuado(1)	Autuado(1)	2.683.902.509,29	829	1040	89,57	50,14
	Não Autuado(0)	0	211			
Não	Autuado(1)	312.322.526,79	101	1034	10,42	49,85
Autuado(0)	Não Autuado(0)	0	933			
Total		2.996.225.036,08	2074			

A exemplo dos modelos de regressão logística, estes dados foram obtidos a partir da comparação dos resultados previstos com as correspondentes autuações realizadas.

Se o modelo tivesse sido aplicado como critério para escolha das empresas a fiscalizar, a metade da amostra não teria sido fiscalizada, possibilitando a aplicação dos recursos em ações fiscais, sendo que o modelo previu corretamente a autuação de 84,95% dos casos (829 em 930), identificando empresas que responderam por 89,57% do crédito constituído.

Os resultados indicam que a aplicação deste modelo teria o potencial de otimizar a aplicação dos recursos disponíveis para fiscalização de empresas.

Modelo RN 2: autuações acima de R$ 1milhão

Neste modelo, foi utilizada a variável dependente VN3 e todas as 8.300 ações fiscais na base de dados. A Figura 5.24 traz a matriz de confusão e as medidas de avaliação correspondentes obtidas em um dos testes do modelo.

```
> matriz_confusao

previsoes    0    1
        0 1691  189
        1   70  125
```

Cálculo das medidas de avaliação:

TN	FP	FN	TP	ACC	Recall	Precision	F1Score
1091	70	189	125	82,44%	39,81%	64,10%	49,12%

Fonte: A autora (2020)

Figura 5.24: Matriz de confusão Modelo RN 2 com o Software R

Neste caso, a exemplo do que ocorreu com o modelo similar de regressão logística RL2, a medida *Recall* ficou abaixo de 50%, o que indica que o modelo não é bom para detecção de documentos relevantes (contribuintes autuados com autos acima de R$ 1 milhão), embora o desempenho geral do modelo esteja acima de 80%, nos testes realizados.

A Tabela 5.19 apresenta as medidas desempenho de testes realizados com ajustes de estrutura da rede e alteração aleatória das bases de treinamento e validação.

Tabela 5.19: Modelo RN 2: testes realizados

RN	Cam.Ocultas	TN	FP	FN	TP	ACC	Sensib.	Especif.	F1Score
2	2	1678	83	162	152	88,19 %	48,40 %	64,68 %	55,37 %
2	2	1675	86	174	140	87,46 %	44,58 %	61,94 %	51,85 %
2	1	1682	79	195	119	86,79 %	37,89 %	60,10 %	46,48 %
2	1	1699	62	192	122	87,75 %	38,85 %	66,30 %	48,99 %

É possível notar que, de modo análogo ao Modelo RN 1, alterações no número de camadas ocultas, nas épocas de treinamento e na seleção aleatória da base de dados de treinamento e validação não causaram diferenças significativas nas métricas do modelo.

A Tabela 5.20 apresenta a comparação entre as autuações efetivas e previstas, obtidas na rede neural de duas camadas cujo resultado na Tabela 5.19 obteve a menor acurácia.

Tabela 5.20: Modelo RN 2: crédito constituído (VN3) x autuações previstas

VN3		Crédito	Qtd.	Total	%	%
Previsto	Realizado	constituído	Contrib.	Contrib.	Valor	Qtd
Autuado(1)	Autuado(1)	1.404.118.707,31	140	226	47,89	10,89
	Não Autuado(0)	(*) 0	86			
Não Autuado(0)	Autuado(1)	1.527.726.055,73	174	1.849	52,1	89,1
	Não Autuado(0)	(**) 0	1.675			
Total		2.931.844.763,04	2.075			

(*) 86 contribuintes com autuação total de R$14.748.241,15, em autos cujos valores individuais são inferiores a R$ 1 milhão.
(**) 1.675 contribuintes com autuação total de R$124.575.152,88, em autos cujos valores individuais são inferiores a R$ 1 milhão.

O resultado indica que o modelo não teria otimizado a aplicação dos recursos disponíveis, porque o valor do crédito constituído associado aos casos em que houve a previsão correta de autuação corresponde a menos do que 50% do valor efetivamente autuado.

Modelo RN 3: AECs e autuações de qualquer valor

O modelo apresenta como Vautuado como variável dependente, e trabalha apenas com os 4.057 registros relacionados ao tipo de ação fiscal AEC. A Figura 5.25 traz a matriz de confusão e as medidas de avaliação do modelo obtida para um dos testes realizados.

```
> matriz_confusao

previsoes   0   1
        0 127  65
        1 171 651
```

Cálculo das Medidas de Avaliação:

TN	FP	FN	TP	ACC	Recall	Precision	F1Score
127	171	65	651	76,73%	90,92%	79,20%	84,66%

Fonte: A autora (2020)

Figura 5.25: Matriz de confusão Modelo RN 3 com o Software R

A Tabela 5.21 apresenta uma amostra dos resultados obtidos relativos em outros teste, com alterações no número de camadas ocultas e na separação da base de validação.

Tabela 5.21: Modelo RN 3: testes realizados

RN	Cam.Ocultas	TN	FP	FN	TP	ACC	Sensib.	Especif.	F1Score
3	2	150	148	89	627	76,62 %	87,56 %	80,90 %	84,10 %
3	2	139	159	50	666	79,38 %	93,01 %	80,72 %	86,43 %
3	1	268	30	513	203	46,44 %	28,35 %	87,12 %	42,78 %
3	1	133	165	46	670	79,19 %	93,57 %	80,23 %	86,39 %

A Tabela 5.22 apresenta a comparação entre as autuações efetivas e as previstas:

Tabela 5.22: Modelo RN 3: crédito constituído em AEC x autuações previstas

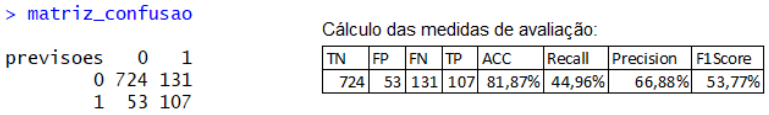
VAutuado		Crédito	Qtd.	Total	%	%
Previsto	Realizado	constituído	Contrib.	Contrib.	Valor	Qtd
Autuado(1)	Autuado(1)	1.516.644.057,63	627	775	86,52	76,42
	Não Autuado(0)	0	148			
Não Autuado(0)	Autuado(1)	236.219.363,58	89	239	13,47	23,57
	Não Autuado(0)	0	150			
Total		1.752.863.421,21	1.014			

A exemplo do resultado obtido no Modelo RN1, a acurácia e a sensibilidade do modelo são elevadas. O modelo indicou que 239 empresas não tinham propensão a autuação, sendo que tais ações correspondem a 23,57% das ações realizadas e a 13,47% do valor autuado.

Estes resultados indicam que o uso de mineração de dados poderia ter otimizado o aproveitamento dos recursos disponíveis com a priorização de outros indícios, embora o resultado não tenha alcançado as mesmas proporções do Modelo RN1.

Modelo RN 4: AECs e autuações acima de R$ 1milhão

Este modelo teve como VN3 como variável dependente, e foi rodado apenas com os registros relacionados ao tipo de ação fiscal do tipo AEC. A Figura 5.26 traz a matriz de confusão e as medidas de avaliação correspondentes a um dos testes realizados.



```
> matriz_confusao

previsoes   0   1
        0 724 131
        1  53 107
```

Cálculo das medidas de avaliação:

TN	FP	FN	TP	ACC	Recall	Precision	F1Score
724	53	131	107	81,87%	44,96%	66,88%	53,77%

Fonte: A autora (2020)

Figura 5.26: Matriz de confusão Modelo RN 4 com o Software R

A Tabela 5.23 apresenta as medidas de desempenho obtidas em outros testes, com ajustes na estrutura da rede e na seleção aleatória das bases treinamento e validação.

Tabela 5.23: Modelo RN 4: testes realizados

RN	Cam.Ocultas	TN	FP	FN	TP	ACC	Sensib.	Especif.	F1Score
4	2	719	58	141	97	80,39 %	40,75 %	62,58 %	49,36 %
4	2	713	64	139	99	80,00 %	41,59 %	60,73 %	49,37 %
4	1	723	54	136	102	81,28 %	42,85 %	65,38 %	51,77 %
4	1	730	47	147	91	80,88 %	38,23 %	65,94 %	48,40 %

A Tabela 5.24 apresenta a comparação entre as autuações efetivas e previstas, obtidas na rede neural de duas camadas cujo resultado na Tabela 5.23 obteve a menor acurácia.

Tabela 5.24: Modelo RN 4: crédito constituído (VN3) x autuações previstas

VN3		Crédito	Qtd.	Total	%	%
Previsto	Realizado	constituído	Contrib.	Contrib.	Valor	Qtd
Autuado(1)	Autuado(1)	662.326.142,33	100	171	54,24	16,84
	Não Autuado(0)	(*) 0	71			
Não Autuado(0)	Autuado(1)	558.647.142,04	138	844	45,75	83,15
	Não Autuado(0)	(**) 0	706			
Total		1.220.973.284,37	1.015			

(*) 71 contribuintes com autuação total de R$ 11.576.626,61 em autos cujos valores individuais são inferiores a R$ 1 milhão.
(**) 706 contribuintes com autuação total de R$ 94.609.874,65, em autos cujos valores individuais são inferiores a R$ 1 milhão.

A análise dos dados da Tabela 5.23 permite notar que a alteração no número de épocas de treinamento ou no número de camadas ocultas não melhora a capacidade do modelo de identificar os casos relevantes, que permanece abaixo de 50% em todos os casos, embora o desempenho geral do modelo tenha se mantido acima de 80% em todos os testes realizados.

O resultado indica que o modelo não teria otimizado a aplicação dos recursos disponíveis, uma vez que o crédito constituído dos casos em que houve a previsão correta de autuação não é muito superior a 50% do valor efetivamente autuado.

Análise dos resultados dos modelos de redes neurais

A Tabela 5.25 apresenta um resumo da avaliação dos modelos de redes neurais, considerando o menor valor de acurácia dos modelos com duas camadas ocultas e mil épocas de treinamento apresentados nas Tabelas 5.17, 5.19, 5.21 e 5.23:

Tabela 5.25: Modelos de redes neurais

Modelo			*Acurácia*	*Sensibilidade*	*Especificidade*	*F1 Score*	*% Valor (Predição)*
Nome	*Ação*	*Autos*					
RN 1	Todas	Todos	85,08 %	89,00 %	79,96 %	84,24 %	89,57 %
RN 2	Todas	>R$1mi	87,71 %	42,35 %	64,25 %	51,05 %	47,89 %
RN 3	AECs	Todos	78,20 %	95,39 %	78,41 %	86,07 %	86,52 %
RN 4	AECs	>R$1mi	80,00 %	41,59 %	60,73 %	49,37 %	54,24%

Nos modelos de redes neurais voltados para a predição do resultado autuação, a acurácia medida indica que o nível global de acertos é elevado, sendo que os resultados indicam que os modelos RN1 e RN3 apresentam alto grau de detecção dos contribuintes autuados e seu uso teria permitido diminuir a quantidade de contribuintes sem diminuir o crédito constituído em igual proporção, indicando que ambos poderiam otimizar a aplicação dos recursos disponíveis para fiscalização.

Entretanto, o Modelo RN1 mostrou maior potencial de otimização, visto que previu aproximadamente metade da amostra não seria autuada, e esta parte corresponde a apenas 10,42% do crédito constituído, o que implica em quase 90% de autuações seriam mantidas com a fiscalização de apenas 50% das empresas.

A análise de sensibilidade demonstra que o número de acertos dos casos positivos é baixo quando a análise leva em conta apenas as autuações acima de 1 milhão de Reais,

sendo que tanto no modelo RN2 quanto no RN4 o percentual de acerto de verdadeiros positivos ficou muito próximo a 50%.

5.3.3 Modelos logísticos x modelos de redes neurais

Conforme exposto no referencial teórico, na presente pesquisa foi adotado o uso em paralelo de redes neurais e de regressão logística como forma de lidar com as incertezas associadas à utilização de redes neurais apontadas por diversos estudos.

Modelos criados para autuações de qualquer valor

A Tabela 5.26 apresenta a comparação do resultado dos modelos de regressão logística e redes neurais relativos à predição do resultado autuação, independentemente do valor, sendo que a acurácia consta das Tabelas 5.16 e 5.25, e os demais dados das Tabelas 5.12, 5.18, 5.14 e 5.22.

Tabela 5.26: Comparação dos modelos preditivos para autuações de qualquer valor

Modelo			*Acurácia*	*Previsão*		*Crédito*	*Ações*
				Contr.	*Não*	*que seria*	*que seriam*
Ação	*Autos*	*Nome*	*(%)*	*Autuados*	*Autuados*	*Mantido*	*Evitadas (*)*
Todas	Todos	RL 1	81,43	878	1.196	83,93%	57,66%
		RN 1	85,08	1.040	1.034	89,57%	49,85%
AECs	Todos	RL 3	81,36	785	229	90,10%	22,58%
		RN 3	78,20	775	239	86,52%	23,57%

(*) os recursos liberados poderiam ser direcionados a outras ações fiscais.

A comparação dos modelos permite notar que o resultado das redes neurais é compatível com os resultados obtidos por meio das regressões logísticas e que os modelos que utilizam todas as ações fiscais (RN1 e RL1) mostraram maior potencial de otimização do emprego dos recursos disponíveis, visto que as ações que não seriam realizadas correspondem a aproximadamente metade das ações realizadas e o valor do crédito constituído se manteve acima de 80%, nos dois modelos.

Os modelos que utilizam apenas as AECs (RL3 e RN3), embora apresentem alta acurácia e preservam boa parte do crédito constituído apresentam uma quantidade menor de fiscalizações que poderiam ser evitadas, e consequentemente, um percentual menor de liberação de recursos para desenvolvimento de outras atividades.

Modelos criados para autuações acima de R$ 1 milhão

A Tabela 5.27 apresenta a comparação do resultado dos modelos de regressão logística e redes neurais relativos à predição do resultado autuação em valor superior a 1 milhão de Reais, sendo que a acurácia consta das Tabelas 5.16 e 5.25, e os demais dados das Tabelas 5.13, 5.20, 5.15 e 5.24,

Tabela 5.27: Comparação dos modelos preditivos para autuações acima de R$ 1 milhão

Modelo			*Acurácia*	*Previsão*		*Crédito*	*Ações*
				Contr.	*Não*	*que seria*	*que seriam*
Ação	*Autos*	*Nome*	*(%)*	*Autuados*	*Autuados*	*Mantido*	*Evitadas (*)*
Todas	>R$	RL 2	87,51	145	1.929	32,40%	93,00%
	1 milhão	RN 2	87,71	226	1.849	47,89%	89,10%
AECs	> R$	RL 4	82,13	162	851	48,85%	84,00%
	1 milhão	RN 4	80,00	171	844	52,24%	83,15%

(*) os recursos liberados poderiam ser direcionados a outras ações fiscais.

Embora apresentem valores elevados de acurácia e apresentem um percentual alto de ações fiscais que não seriam realizadas, os modelos não são bons porque sua eventual aplicação teria resultado em não constituição de grande parte do crédito que foi constituído com a realização de todas as ações fiscais em análise.

Desta forma, todos os modelos que utilizaram como variável dependente o resultado autuação acima de 1 milhão de Reais não se mostraram viáveis para a otimização da seleção de contribuintes, independentemente de terem considerado todas as ações fiscais, Modelo RN2 e RL2, ou apenas as ações fiscais do tipo Auditoria Especial Concentrada, RN4 e RL4, porque as predições deles decorrentes implicariam em auditar muitas empresas sem irregularidades e deixar de auditar muitas empresas com irregularidades, implicando em redução significativa do crédito constituído.

Capítulo 6

Conclusão e Trabalhos Futuros

O aumento do volume de informações digitais recebidas pelo Fisco, em razão da crescente informatização dos procedimentos relacionados ao cumprimento das obrigações acessórias, propicia um melhor conhecimento da realidade dos contribuintes e da complexidade de suas relações. Mas o tratamento destes dados requer estrutura, conhecimentos, técnicas e metodologia que se igualam em magnitude aos volumes recebidos, fato que impulsiona o aprimoramento constante dos processos de trabalho da Administração Tributária.

Otimizar o aproveitamento destes dados e sua conversão em informações úteis, que permitam nortear a fiscalização, representa um desafio contínuo, muito já foi feito neste sentido e ainda há muito por fazer, é neste contexto que se enquadra o presente estudo.

Os objetivos específicos adotados na pesquisa ajudaram a verificar que modelos preditivos de mineração de dados podem auxiliar na seleção dos indícios de irregularidades a fiscalizar e a entender como tais modelos podem ser utilizados para aprimorar o processo de gestão de riscos da Fiscalização Tributária do Distrito Federal.

A caracterização do processo de gestão de riscos aplicado pela Fiscalização Tributária do DF às informações apresentadas pelos contribuintes dos tributos indiretos, à luz das boas práticas de gestão de risco aplicáveis à Administração Pública Distrital, permitiu:

- Compreender o processo;
- Identificar em que etapa o uso de modelos preditivos de mineração de dados poderia ser investigado como meio de aprimorar o processo de gestão de riscos;
- Identificar a modalidade de aprendizado supervisionado de máquina como a adequada para tratamento da situação mapeada;
- Selecionar os modelos de mineração de dados baseados em redes neurais e regressão logística para tratamento das informações;

- Identificar as informações potencialmente úteis para criação, treinamento e validação dos modelos preditivos de mineração de dados, de regressão logística e de redes neurais, bem como os tratamentos necessários para preparação da base de dados selecionada.

Na etapa de caracterização do processo existente, as reuniões de trabalho com a equipe responsável pela seleção de contribuintes a fiscalizar permitiram identificar informações relevantes para o resultado das ações fiscais realizadas, e o questionário aplicado aos auditores que executam as fiscalizações demonstrou não ser possível estabelecer um padrão de comparação entre os diferentes tipos de ação fiscal baseado apenas na percepção pessoal, tendo como consequência a necessidade de tentar estabelecer uma base de comparação com o uso de informações relativas às ações realizadas no passado.

A análise das informações identificadas como potencialmente relevantes possibilitou definir o resultado das autuações passadas como um dos parâmetros para predição dos resultados de futuras ações; e estabelecer as estratégias de seleção de períodos e de dados a utilizar no treinamento dos modelos preditivos.

O estudo das técnicas de mineração de dados e de trabalhos realizados nas esferas financeira e tributária com tais recursos permitiu identificar os métodos e ferramentas necessárias para criação de modelos de mineração de dados baseados em regressão logística e em redes neurais aptos a tratar a base de dados selecionada à luz dos objetivos da pesquisa.

A criação e consequente avaliação dos modelos preditivos permitiu verificar que existem modelos preditivos de mineração de dados (RN1 e RL1), baseados em aprendizado de máquina supervisionado, capazes de auxiliar na priorização das auditorias a realizar e assim contribuir para o aprimoramento da gestão de riscos do Fisco Distrital.

A pesquisa mostrou que a aplicação dos recursos disponíveis para realização da fiscalização poderia ter sido otimizada com o uso dos modelos preditivos para a variável dependente resultado autuação, independentemente do valor autuado. Neste grupo, os dois modelos que utilizaram como base de treinamento e validação os dados de todas as ações fiscais realizadas a partir de 2009 apresentaram resultados mais promissores do que os modelos que utilizaram apenas as ações fiscais do tipo Auditoria Especial Concentrada, sendo que:

- O modelo preditivo de regressão logística (RL1) previu que 57,66% das fiscalizações não resultariam em autuação (1.196 ações fiscais), e quando comparado ao resultado realizado, as empresas que o modelo indicou para autuação responderam por 83,93% do crédito constituído da respectiva base de dados de validação (2.1 bilhões de Reais).

- O modelo preditivo de rede neural (RN1) indicou que 49,85% das empresas não seriam autuadas (1.034 ações fiscais), se o modelo tivesse sido aplicado o crédito constituído mantido, relativo às empresas que o modelo indicou para auditoria, corresponderia a 89,57% do total realizado da respectiva base de dados de validação (2.6 bilhões de reais).

Assim, os resultados do trabalho abrem a possibilidade de uso de tais modelos, na etapa de seleção de contribuintes a fiscalizar, como ferramentas para aprimorar a gestão de riscos ao abrir a possibilidade de incorporar modelos preditivos no processo de seleção de indícios de fraude fiscal a serem fiscalizados.

Um aspecto que chama a atenção quando se fala de melhorar a seleção de alvos é a tendência de ganho de produtividade, que implica em aplicar os recursos disponíveis de melhor forma e assim tende a contribuir na busca por eficiência. Outro aspecto relevante a considerar, o uso dos modelos possibilitaria escolher não auditar os contribuintes cujas previsões indicam que não haveria autuação (1.196 auditorias a menos no caso do modelo RL1 ou 1.034 auditorias a menos, no caso do RN1), o que permitiria concentrar esforços na fiscalização dos contribuintes inadimplentes, deste modo, este tipo de escolha tende a auxiliar a fiscalização a fazer apenas o que tem que ser feito o que contribuiria na busca pela eficácia da fiscalização. Estes dois aspectos indicam que a utilização de modelos preditivos encerra em si a possibilidade de melhorar a eficiência, a eficácia e a efetividade da fiscalização tributária.

Sugestões para trabalhos futuros:

- Incluir a execução dos modelos preditivos baseados de regressão logística e redes neurais, que considerem o histórico das fiscalizações realizadas, na etapa de seleção dos alvos a fiscalizar, e avaliar a evolução dos resultados na prática.
- Criação de novos modelos preditivos de mineração de dados que utilizem outras variáveis explicativas, como por exemplo o valor financeiro dos indícios.
- Investigação da possiblidade de uso de modelos preditivos para identificação de novos indícios de irregularidade ou de fraudes.

Referências

[1] CHOI, D. e K. LEE: *An artificial intelligence approach to financial fraud detection under iot environment: A survey and implementation.* Security and Communication Networks, 2018. 1, 30

[2] GONZÁLEZ, P. C. e VELÁSQUEZ, J.D.: *Characterization and detection of taxpayers with false invoices using data mining techniques.* Expert Systems with Applications, (40):1427–1436, Chile. 2012. Journal homepage: www.elsevier.com/locate/eswa. 1, 21, 29

[3] BABU, S.K. e S. VASAVI: *Predictive analytics as a service on tax evasion using gaussian regression process.* Helix, 7(5):1988–1993, Sep. 2017. 1, 29

[4] HAJEK, P. e HENRIQUES, R.: *Mining corporate annual reports for intelligent detection of financial statement fraud –a comparative study of machine learning methods.* Knowledge-Based Systems, (128):139–152, 2017. Journal homepage: www.elsevier.com/locate/knosy. 1, 29

[5] LI, Y.; JIANG, W.; YANG L. e WU T.: *On neural networks and learning systems for business computing.* Neurocomputing, (275):1150–1159, 2018. Journal homepage: www.elsevier.com/locate/neucom. 1, 26, 29

[6] ALINK, M. e KOMMER, V. van. Tradução Vinícius Pimentel de Freitas: *Manual de Administração Tributária.* IBFD - International Bureau of Fiscal Documentation, 2011. 2, 3

[7] MORALES, A. M. C.; PINEDA, C. M. R. e MONSALVE O. O. V.: *La primera reforma tributaria en la historia de la humanidad.* (Entramado, vol. 15, núm. 1), 2019. `http://www.redalyc.org/articulo.oa?id=265460762010`. 2

[8] MACHADO, C.H. e BALTHAZAR, U. C.: *A reforma tributária como instrumento de efetivação da justiça distributiva: uma abordagem histórica.* UFSC, 2017. `http://dx.doi.org/10.5007/2177-7055.2017v38n77p221`. 3

[9] TORRES, R. L.: *A Ideia de Liberdade no Estado Patrimonial e no Estado Fiscal.* Livraria e Editora Renovar LTDA, 1991. 3

[10] BRASIL: *Constituição Federal.* 1998. Art. 198, § 2º e Art. 212. 3

[11] SIQUEIRA, M. L. e RAMOS, F. S.: *A economia da sonegação.* Revista Econ. Contemporânea. RJ, 2005. http://www.scielo.br/pdf/rec/v9n3/v9n3a04.pdf. 3, 4

[12] COÊLHO, S. C. N.: *Curso de Direito Tributário Brasileiro*. Editora Forense, Rio de Janeiro, 2006. 3, 4

[13] ORAIR, R. e S. GOBETTI: *Reforma tributária no brasil: Princípios norteadores e propostas em debate*. 2018. https://doi.org/10.25091/s01013300201800020003. 4

[14] DISTRITO FEDERAL: Secretaria de Estado de Economia do Distrito Federal: *Balanço Geral 2019*, 2020. https://static.fazenda.df.gov.br//arquivos/12019BALGERALFINAL3003 2020FINAL.pdf. 5, 47, 89

[15] BRASIL: *Lei 5.472/1966 - Código Tributário Nacional*, 1966. 5, 6

[16] DISTRITO FEDERAL: *Lei 4.717/2011*, 2011. 5

[17] DISTRITO FEDERAL: *Decreto 35.565/2014: Regimento Interno da SEF/DF*, 2014. 6

[18] DISTRITO FEDERAL: Secretaria de Fazenda do Distrito Federal: *Carta de Serviços*, 2018. http://static.fazenda.df.gov.br/arquivos/pdf/ CartadeServicosFazendaDF2016.pdf. 6

[19] SILVA, A. A. e CERQUEIRA, A. F.: *Fraudes Contábeis Repercussões Tributárias Enfoque no ICMS*. Juruá Editora, 2018. 6

[20] DISTRITO FEDERAL: *Decreto 18.955/1997- Regulamento do ICMS*, 1997. Art. 205. 6

[21] DISTRITO FEDERAL: *Decreto 16.128/1994- Regulamento do ISS*, 1997. Art. 65. 6

[22] DISTRITO FEDERAL: *Decreto 26.529/2006 - Institui o LFE*, 2006. Art. 1º. 7

[23] DISTRITO FEDERAL: *Lei Complementar 772/2008*. 7, 39

[24] SECRETÁRIA DE ESTADO DE FAZENDA DO DISTRITO FEDERAL: *Portaria 234/2014*, 2014. 7

[25] BRASIL: Tribunal de Contas da União: *Referencial básico de Gestão de Riscos*, 2018. 11, 12, 13, 15, 16

[26] HILL, S. e DINSDALE, G. Traduzido por Vasconcelos L. M. B. L.: *Uma base para o desenvolvimento de estratégias de aprendizagem para a gestão de riscos no serviço público*. Cadernos ENAP, (23):80, 2003. https://repositorio.enap.gov.br/handle/1/692. 12, 15, 16

[27] OCDE, ORGANIZAÇÃO PARA A COOPERAÇÃO E DESENVOLVIMENTO ECONÔMICO: *Avaliação da ocde sobre o sistema de integridade da administração pública federal brasileira: gerenciando riscos por uma administração pública íntegra*. 2011. <http://www.cgu.gov.br/publicacoes/AvaliacaoIntegridadeBrasileiraOCDE/AvaliacaoIntegridadeBrasileiraOCDE.PDF. 12

[28] ARAÚJO JÚNIOR, J. B. e PINHO FILHO, L. C.: *Implantação da gestão de riscos no governo do distrito federal - gdf: Uma iniciativa de inovação da gestão pública.* Revista Processus de Estudos de Gestão, Jurídicos e Financeiros, 2019, ISSN 2237-2342 (IMPRESSO). 12, 15

[29] DISTRITO FEDERAL: *Decreto 37.302/2016*, 2016, Art. 2º. 13, 14

[30] BRASIL: Ministério Público da União e Controladoria Geral da União: *Instrução Normativa Conjunta MP/CGU No 01/2016*, 2016. 13

[31] ASSOCIAÇÃO BRASILEIRA DE NORMAS TÉCNICAS: *ISO31000: Gestão de riscos - Princípios e diretrizes*, 2009. 14, 16

[32] COMMITTEE OF SPONSORING ORGANIZATIONS OF THE TREADWAY COMMISSION – COSO: *COSO 1*, 2003. 15

[33] ENTERPRISE RISK MANARGEMENT COMMITTEE: *Overview of Enterprise Risk Management*, 2003. Pg.10. 15

[34] COMMITTEE OF SPONSORING ORGANIZATIONS OF THE TREADWAY COMMISSION – COSO – COSO: *COSO Gerenciamento de Riscos Corporativos - Estrutura Integrada Sumário - Executivo Estrutura*, 2007. `https://www.coso.org/Documents/COSO-ERM-Executive-Summary-Portuguese.pdf`. 15

[35] WITTEN, I. H; FRANK, E. e HALL M. A.: *Data Mining Practical Machine Learning Tools and Techniques.* Morgan Kaufmann Publishers - Elsevier, Burlington, MA , USA, 3th Ed, 2011. 16, 17, 18, 19, 21, 22, 23, 25, 27

[36] FAYYAD, U. e P PIATETSKY-SHAPIRO, G e SMYTH: *Knowledge discovery and data mining: Towards a unifying framework.* 1996. `https://www.aaai.org/Papers/KDD/1996/KDD96-014.pdf`. 17, 20, 21

[37] MARISCAL, G. E C. MARBÁN, Ó. E FERNÁNDEZ: *A survey of data mining and knowledge discovery process models and methodologies.* Knowledge Engineering Review, 25(2):137–166, 2010. `https://www.scopus.com/inward/record.uri?eid=2-s2.0-77958595644&doi=10.1017%2fS0269888910000032&partnerID=40&md5=5ca9b2e573e09b5d4679361ced62a98e`. 17

[38] IBM: *IBM SPSS Modeler CRISP-DM Guide.* `ftp://public.dhe.ibm.com/software/analytics/spss/documentation/modeler/18.0/en/ModelerCRISPDM.pdf`. 17

[39] MISHRA, P.: *R Data Mining Blueprints.* Packt Publishing Ltd., Birmingham, UK, 2016, ISBN 978-1-78398-968-38. 18, 19, 20, 24, 25, 26, 27

[40] BUSSAB, W de O. e MORETTIN, P.A.: *Estatística Básica.* Editora Saraiva, 2005, ISBN 8502034979. 18

[41] MONTOMERY, D. C. e RUNGER, G. C.: *Estatística Aplicada e Probabilidade para Engenheiros.* LTC - Livros Técnicos e Científicos Editora Ltda., 5ªª edição, 2012. 18, 23, 24, 25

[42] PRIETO N, A.; PRIETO, B.; ORTIGOSA E. M.; ROS E.; PELAYO F.; ORGEGA J. e ROJAS I.: *Neural networks: An overview of early research, current frameworks and newchallenges*. Knowledge-Based Systems, (128):139–152, 2017. Journal homepage: www.elsevier.com/locate/knosys. 19, 21, 30

[43] LECUN, Y.; BENGIO, Y. e HINTON G.: *Deep learning*. Nature, 521(7553):436–444, 2015. https://www.scopus.com/inward/record.uri?eid=2-s2.0-84930630277&doi=10.1038%2fnature14539&partnerID=40&md5=e324cb9ec992f892ebc74f3e06078083. 19, 21, 26, 27, 28

[44] ABIODUN, O.I.; JANTAN, A.; OMOLARA A.E.; DADA K.V.; MOHAMED N.A. e ARSHAD H.: *State-of-the-art in artificial neural network applications: A survey*. Heliyon, 4(11), 2018. https://www.scopus.com/inward/record.uri?eid=2-s2.0-85057050483&doi=10.1016%2fj.heliyon.2018.e00938&partnerID=40&md5=84cc31a1eae48ae403cfa37b56cf5962. 19, 27, 28, 30

[45] DONOHO, D.: *50 years of data science*. Journal of Computational and Graphical Statistics, (Dec.):745–766, 2017. 20, 41

[46] TKÁČ, M. e VERNER, R.: *Artificial neural networks in business: Two decades of research*. Applied Soft Computing Journal, 38:788–804, 2016. https://www.scopus.com/inward/record.uri?eid=2-s2.0-84946920504&doi=10.1016%2fj.asoc.2015.09.040&partnerID=40&md5=c91dd8b5047f31ebac2688a53dd56e9c. 21, 26, 28, 29

[47] GUO, Y.; LIU, Y.; OERLEMANS A.; LAO S.; WU S.; e LEW M.S.: *Deep learning for visual understanding: A review*. Neurocomputing, 187:27–48, 2016. https://www.scopus.com/inward/record.uri?eid=2-s2.0-84957837518&doi=10.1016%2fj.neucom.2015.09.116&partnerID=40&md5=c6b47408f1245c0319b820b3605e577a. 21, 26, 28, 29

[48] PRATI, R. C.; BATISTA, G. E. A. P. A. e MONARD M. C.: *Curvas roc para avaliação de classificadores*. IEEE LATIN AMERICA TRANSACTIONS. 22

[49] STULP, F. e SIGAUD, O.: *Many regression algorithms, one unified model: A review*. Neural Networks, 69:60–79, 2015. 24

[50] HUANG, W.; LAI, K.K.; NAKAMORI Y.; WANG S. e L. YU: *Neural networks in finance and economics forecasting*. International Journal of Information Technology and Decision Making, 6(1):113–140, 2007. https://www.scopus.com/inward/record.uri?eid=2-s2.0-33947658558&doi=10.1142%2fS021962200700237X&partnerID=40&md5=08db15343fa6ab2f64ce230b303a1878. 26, 28, 29

[51] FERNEDA, E: *Redes neurais e sua aplicação em sistemas de recuperação de informação*. Ci. Inf., Brasília, 35 jan./abr(1):25–30, 2006. 26, 27

[52] SILVA, L.S.: *Modelos preditivos para seleção de solicitações de compensação de crédito tributário*. Universidade de Brasília, 2016. 30

[53] SOUSA, R. M.: *Inteligência computacional aplicada ao controle externo da administração pública: aplicações da classificação de padrões utilizando redes neurais artificiais*. 30

[54] WU, R. S.; OU, C. S.; LIN H. Y.; CHANG S. I.; e YEN D. C.: *Using data mining technique to enhance tax evasion detection performance*. Expert Systems with Applications, (39):8769–8777, 2012. Journal homepage: www.elsevier.com/locate/eswa. 30

[55] BITTENCOURT NETO, S.: *Dissertação: Análise de "outliers" para o controle do risco de evasão tributária do icms*. http://repositorio.unb.br/handle/10482/33031, 2018. 30

[56] PRODANOV, C e FREITAS, E: *Metodologia do Trabalho Científico – Métodos e Técnicas da Pesquisa e do Trabalho Acadêmico*. Reading, Mass., 2013. Disponível em: https://aprender.ead.unb.br. 31, 32

[57] GIL, A. C: *Métodos e técnicas de pesquisa social*. Atlas, 2008. 31, 32

[58] GIL, A. C: *Como Elaborar Projetos de Pesquisa*. Atlas, 2002. 33

[59] DISTRITO FEDERAL: *Lei 1.254/1996- Lei do ICMS*, 1996. Art. 46, inciso XII. 39

[60] BRASIL: *Lei Complementar 123/2006 - Lei do Simples Nacional*, 2006. 39

[61] WICKHAM, H. e GROLEMUND, G.: *R for Data Science Import, Tidy, Transform, Visualize, and Model Data*. O'Reilly Media, Canadá, 2016. 41, 42

[62] KUHN, M.: *The caret Package*. 2019. 42

[63] LEDELL, E.; GILL, N.; AIELLO S.; FU A.; CANDEL A.; CLICK C.; KRALJEVIC T.; NYKODYM T.; ABOYOUN P.; KURKA M. E MALOHLAVA M.: *R Interface for the 'H2O' Scalable Machine Learning Platform*, volume https://cran.r-project.org/web/packages/h2o/h2o.pdf. 2020. 42

[64] DISTRITO FEDERAL: Secretaria de Fazenda do Distrito Federal: *Balanço Geral 2006 - Volume I*, 2007. http://static.fazenda.df.gov.br//arquivos/ZIP/GestaoContabil/balanco_geral____volume_i____2006.zip. 47, 89

[65] DISTRITO FEDERAL: Secretaria de Fazenda do Distrito Federal: *Balanço Geral 2011*, 2012. http://static.fazenda.df.gov.br//arquivos/ZIP/GestaoContabil/balanco_geral____2011.zip. 47, 89

[66] DISTRITO FEDERAL: Secretaria de Fazenda do Distrito Federal: *Balanço Geral 2016*, 2017. https://static.fazenda.df.gov.br//arquivos/balanco_geral_2016.pdf. 47, 89

[67] DISTRITO FEDERAL: Secretaria de Fazenda do Distrito Federal: *Balanço Geral 2018*, 2019. https://static.fazenda.df.gov.br//arquivos/BALANCO_GERAL_GDF_2018.pdf. 47

Anexo I

Resultados das Pesquisas Bibliométricas

I.1 Avaliação de risco e mineração de dados

A Figura I.1 apresenta o gráfico produzido pela plataforma *Web of Science* relativo à quantidade de documentos localizados por ano, e permite observar que o tema *risk assessment and data mining*) possui um aumento crescente do número de publicações anuais.

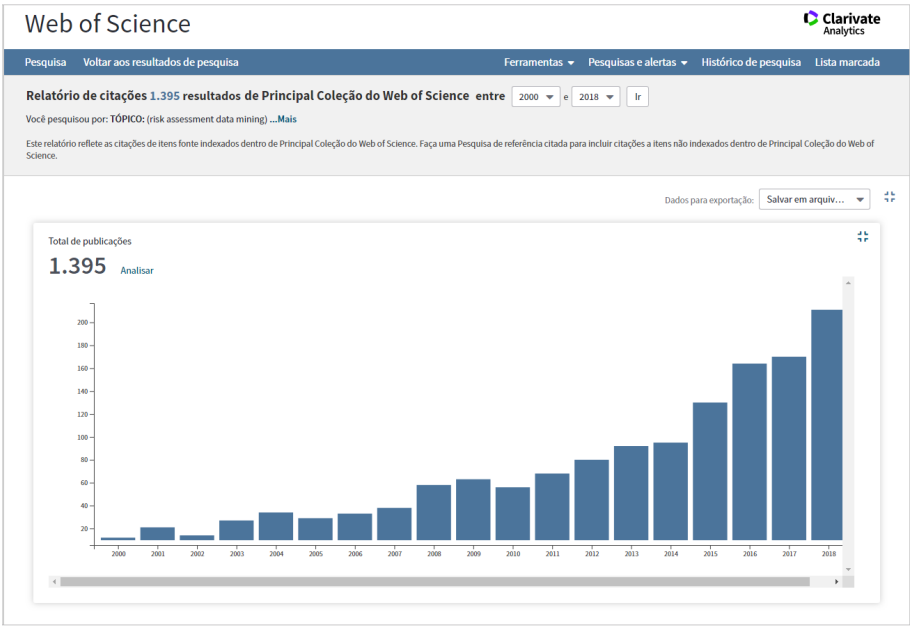


Figura I.1: Resultado da Pesquisa Web of Science - Documentos por Ano

A Figura I.2 apresenta o gráfico da plataforma *Web of Science* relativo à categorização dos trabalhos publicados, e permite observar que o tema mineração de dados e gestão de

riscos está associado a diversas categorias tais como computação, saúde, recursos hídricos.

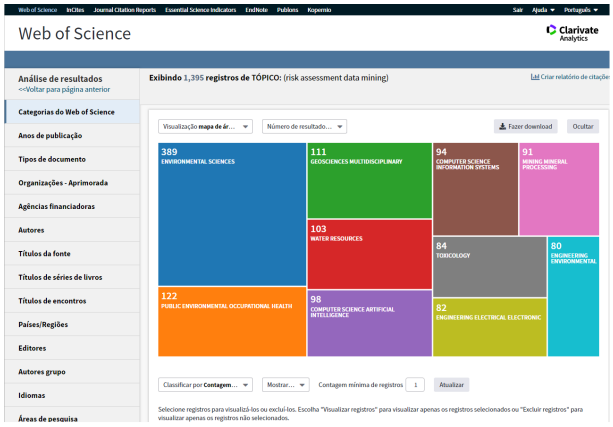


Figura I.2: Resultado da Pesquisa Web of Science - Área de Conhecimento

A Figura I.3 apresenta o gráfico produzido pela plataforma *Scopus* relativo à quantidade de documentos localizados por ano e, de modo análogo ao observado por meio da Figura I.1, permite observar que o tema possui um aumento crescente do número de publicações.

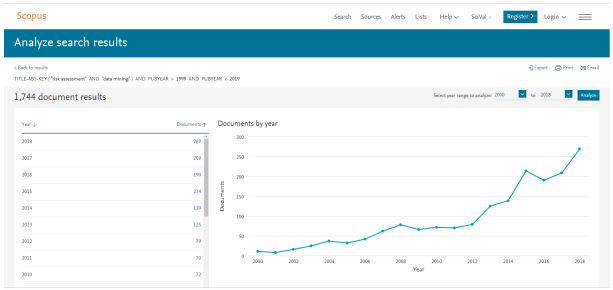


Figura I.3: Resultado da Pesquisa Scopus - Documentos por Ano

A Figura I.4 apresenta o gráfico da plataforma *Scopus* relativo à área de interesse dos trabalhos publicados e aponta que o tema está associado a diversas categorias.

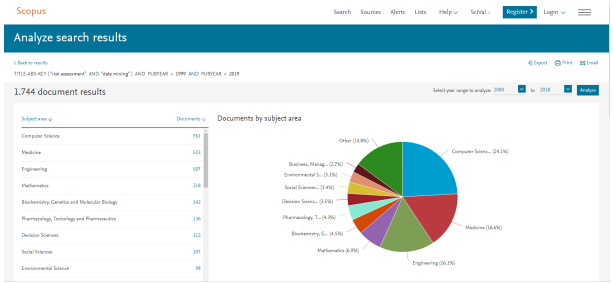


Figura I.4: Resultado da Pesquisa Scopus - Área de Conhecimento

A Figura I.5 apresenta a quantidade de documentos de interesse resultante após a filtragem, na plataforma *Scopus*.

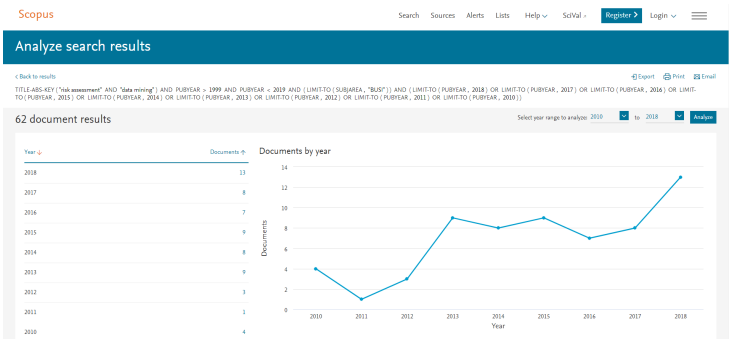


Figura I.5: Publicações filtradas Scopus - por Ano

A Figura I.6 apresenta o perfil por área de interesse das publicações selecionadas na plataforma *Scopus* após a filtragem.

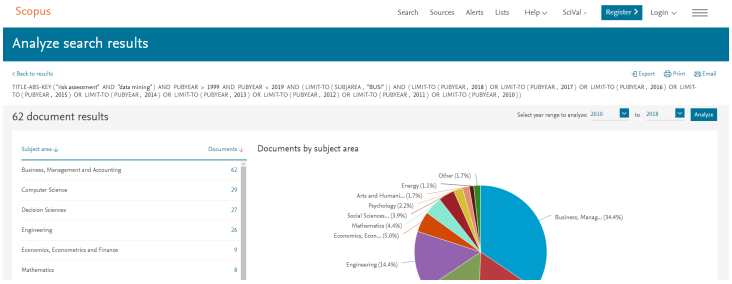


Figura I.6: Publicações filtrados Scopus - Área de Conhecimento - Filtrados

I.2 Redes Neurais

A Figura I.7 apresenta por ano os 617 documentos inicialmente localizados para o tema redes neurais na plataforma *Web of Science*.

A Figura I.8 detalha os 617 documentos por área de conhecimento.

A Figura I.9 apresenta por ano os 1.744 documentos inicialmente localizados para o tema redes neurais na plataforma *Scopus*.

A Figura I.10 detalha os 1.744 documentos da plataforma *Scopus* por área de conhecimento.

A Figura I.11 apresenta os documentos selecionados para leitura, na plataforma *Scopus*, relativos a redes neurais após a filtragem por área de conhecimento.

A Figura I.12 apresenta o perfil por área de interesse das publicações selecionadas na plataforma *Scopus*.

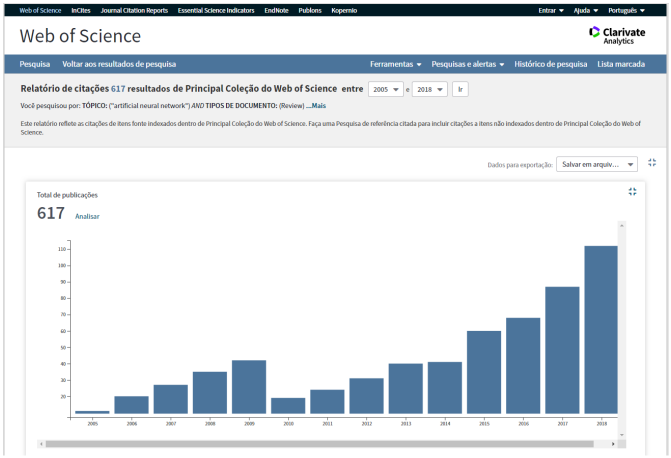


Figura I.7: Resultado da Pesquisa Web of Science - Documentos por Ano

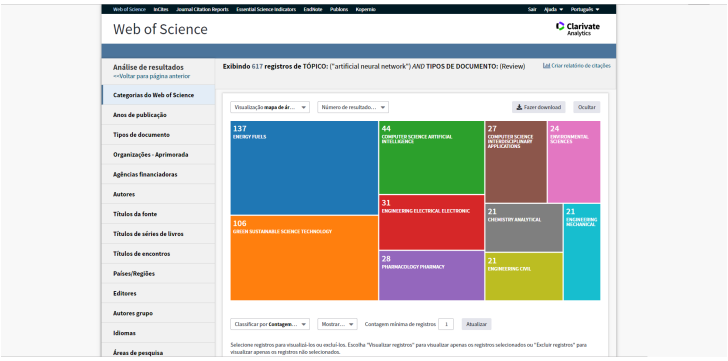


Figura I.8: Resultado da Pesquisa Web of Science - Área de Conhecimento

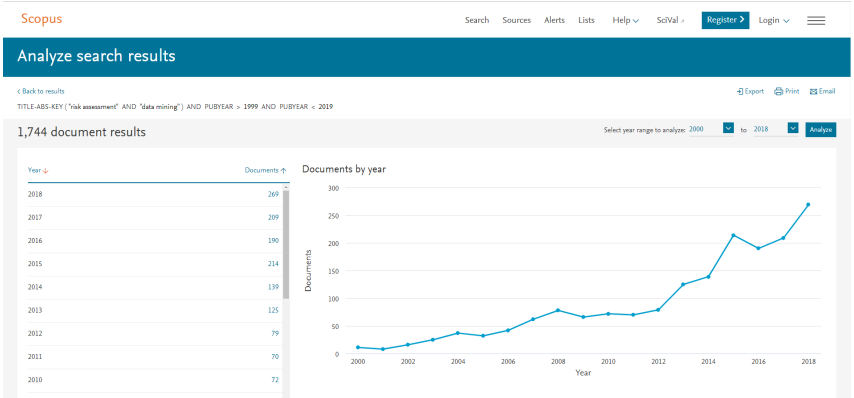


Figura I.9: Resultado da Pesquisa Scopus - Documentos por Ano

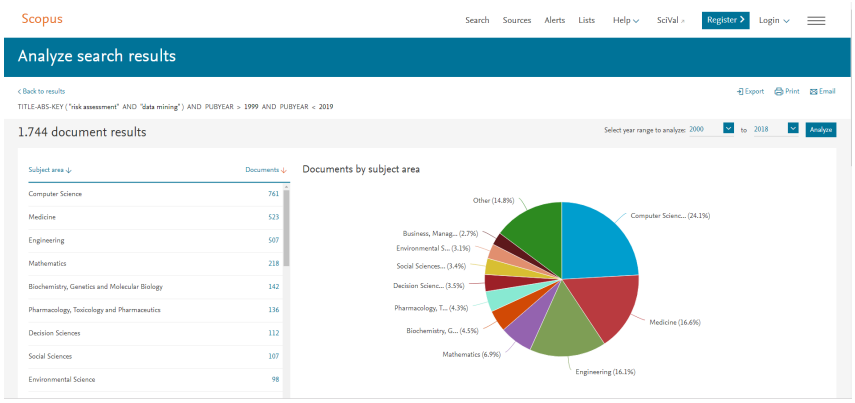


Figura I.10: Resultado da Pesquisa Scopus - Área de Conhecimento

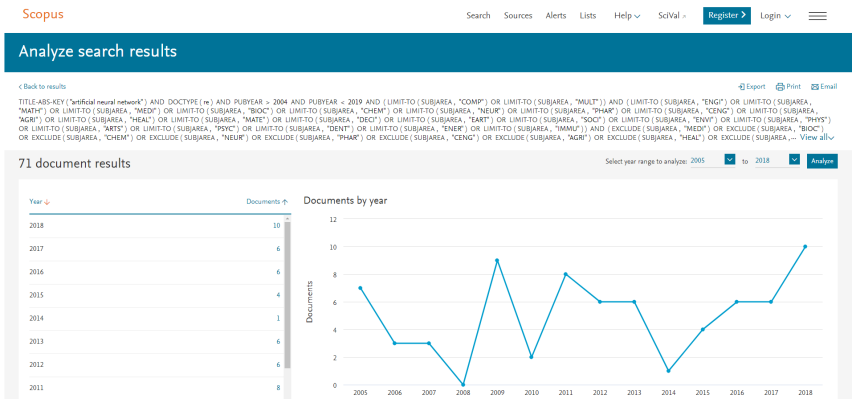


Figura I.11: Publicações filtradas Scopus - por Ano

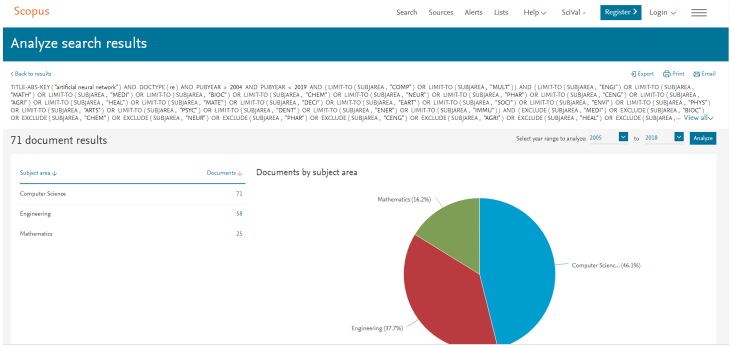


Figura I.12: Publicações filtrados Scopus - Área de Conhecimento - Filtrados

Anexo II

Série Histórica Arrecadação Tributária

Questionário aplicado aos auditores responsáveis pela execução das ações fiscais.

QUESTIONÁRIO PARA AVALIAÇÃO DOS TIPOS DE AÇÃO FISCAL

Finalidade: Verificar se é possível a partir da opinião dos auditores que executam as ações fiscais, estabelecer padrões que permitam a comparação das ações fiscais, tanto para fins de aprimoramento dos critérios de seleção das empresas para distribuição das auditorias quanto para avaliação de produtividade de que trata o Decreto do trabalho remoto.

Avaliação: A pontuação deve refletir a comparações entre as ações fiscais do ponto de vista de cada auditor que já executou a ação fiscal. Caso o auditor não tenha executado nenhuma auditoria do tipo em análise não deverá atribuir pontuação ao mesmo. Não é necessário que todos os níveis de pontuação sejam utilizados.

					CRITÉRIOS			
Avaliação: A pontuaçãoTipo de Ação	Anos	Ações	Autos	Principal	Total	Grau de Dificuldade	Potencial de Ampliação do Escopo	Tempo médio dispendido
AEC-CARTÃO	10	2263	1726	423.293.056,20	1.337.371.252,29			
AEC-MISSING	6	427	356	805.619.708,21	2.127.577.821,37			
AEC-CARTÃOS.NAC.	3	88	63	42.388.570,50	40.903.977,88			
AEC-CERBERUS	3	65	34	36.605.840,72	129.681.602,78			
AEC-ALÍ.FÓRA	6	62	42	8.988.211,87	26.010.896,00			
AEC-CARTÓRIOS	1	59	35	1.683.164,46	7.179.508,81			
AEC-DELETA	5	57	39	50.477.146,04	147.871.677,54			
AEC-SUBPRIME	3	57	45	6.089.555,18	15.902.830,38			
AEC-MANU	2	47	27	5.379.441,08	18.946.851,27			
AEC-CRÉDITOPODRE	5	32	22	21.626.547,71	76.548.926,45			
AEC-CRÉD.IMPRÓPRIOS	3	25	24	16.151.006,57	43.313.201,11			
AEC-REG.ESPECIAIS	3	20	17	16.788.624,63	54.859.118,58			
AEC-EX-REA	1	19	20	35.349.095,93	106.711.908,30			
AEC-CUSTOMERS	1	15	6	441.113,82	1.088.256,28			
AEC-WALKINGDEAD	2	13	10	8.086.135,87	21.130.187,76			
AEC-IMOBILIZADO	1	13	13	19.537.386,02	60.094.718,05			
AEC-TRANSP.SLDCREDOR	2	12	4	36.896,13	132.320,30			
AEC-PRÓ-DF	4	11	8	21.316.712,95	68.735.659,03			
PERÍCIAS								
AUDITORIA	20	4004	2919	3.316.199.789,68	10.405.745.087,82			

Escala a ser usada em todos os critérios

Valor	Grau
1	Muito baixo
3	Baixo
5	Médio
7	Alto
9	Muito alto
2,4,6,8	Valores intermediários também podem ser usados

Auditor:

Figura II.1: Questionário

Anexo III

Questionário:

As Figuras III.1, III.2, III.3 e III.4 apresentam tabelas elaboradas com dados obtidos nos relatórios Balanço Geral dos exercícios de 2006 [64], 2011 [65], 2016 [66] e 2019 [14] publicados na página da Secretaria de Estado de Economia do DF (SEEC). Trata-se de dados relativos à arrecadação de receitas tributárias em valores nominais e a primeira linha apresenta a soma das receitas tributárias identificadas na primeira coluna de cada linha.

A Figura III.1 apresenta os valores das receitas tributárias arrecadadas nos exercícios de 2002 a 2006, conforme Relatório Balanço Geral 2006, páginas 15 a 18.

	2002	2003	2004	2005	2006
RECEITAS TRIBUTÁRIAS	2.939.353.841,44	3.499.530.743,84	4.180.317.729,85	4.790.379.225,42	5.552.654.882,62
IPTU	167.942.031,11	182.929.730,23	208.141.798,08	235.883.233,92	257.601.482,26
IR	368.762.206,21	429.744.726,79	532.911.395,48	631.819.070,92	841.159.469,48
IPVA	157.379.065,64	172.134.596,27	215.592.064,58	266.011.562,27	318.722.226,17
ITBIM	6.565.755,81	9.595.126,85	10.423.449,94	12.183.696,77	14.939.361,74
ITBIV	49.520.367,61	52.892.499,61	59.389.510,09	70.970.363,75	91.559.253,84
ICMS	1.793.745.713,65	2.192.768.089,56	2.600.321.927,24	2.906.600.539,58	3.290.372.834,75
ISS	332.912.571,65	381.837.934,25	453.796.391,02	555.279.474,69	607.776.273,30
ICM/ISS/ SIMPLES - LEI FEDERAL	22.514.877,36	27.227.846,46	31.040.907,96	38.632.389,94	41.542.325,56
TAXAS	40.011.252,40	50.400.193,82	68.700.285,46	72.998.893,58	88.981.655,52

http://static.fazenda.df.gov.br//arquivos/ZIP/GestaoContabil/balanco_geral___volume_i___2006.zip

Figura III.1: Arrecadação Tributária 2002 a 2006

A Figura III.2 apresenta os valores das receitas tributárias arrecadadas nos exercícios de 2007 a 2011, constantes no Relatório Balanço Geral 2011, páginas 29 a 32.

A Figura III.3 apresenta os valores das receitas tributárias arrecadadas nos exercícios de 2012 a 2016, conforme constam no Relatório Balanço Geral 2016, páginas 30 a 33.

A Figura III.4 apresenta os valores das receitas tributárias arrecadadas nos exercícios de 2017 a 2019, conforme Relatório Balanço Geral 2019, páginas 25 a 27.

	2007	2008	2009	2010	2011
RECEITAS TRIBUTÁRIAS	6.074.155,40	7.102.139,31	7.392.530,03	8.352.921,76	9.366.541,16
IPTU	276.625,59	340.217,38	364.849,23	400.008,66	446.247,25
IR	1.036.243,44	1.293.924,25	1.287.418,47	1.503.877,44	1.742.844,02
IPVA	373.357,24	448.113,30	535.887,62	537.171,20	622.809,85
ITBIM	20.758,81	25.517,61	25.597,05	33.193,73	38.648,83
ITBIV	121.292,92	148.536,38	172.358,16	209.861,72	208.675,14
ICMS	3.433.791,26	3.941.222,99	3.983.560,66	4.493.608,95	5.008.748,86
ISS	642.762,61	675.049,07	759.201,39	856.498,34	941.303,82
ICM/ISS/SIMPLES - LEI FEDERAL	67.114,01	114.655,04	149.804,62	193.124,53	221.398,85
TAXAS	102.209,52	114.903,29	113.852,84	125.577,19	135.864,54

http://static.fazenda.df.gov.br//arquivos/ZIP/GestaoContabil/balanco_geral___2011.zip

Figura III.2: Arrecadação Tributária 2007 a 2011

	2012	2013	2014	2015	2016
RECEITAS TRIBUTÁRIAS	10.287.231,70	11.443.797,28	12.665.997,73	13.155.461,65	14.355.152,35
IPTU	474.722,43	525.284,09	550.371,77	596.069,68	704.910,33
IR	1.957.895,67	2.165.085,20	2.612.009,02	2.862.950,59	2.858.090,76
IPVA	554.372,40	598.893,68	696.590,25	782.035,14	918.686,27
ITBIM	53.009,42	153.145,32	89.086,12	133.417,38	109.201,18
ITBIV	276.616,05	329.701,42	318.060,67	306.014,17	322.855,12
ICMS	5.494.095,34	5.987.377,33	6.540.460,06	6.481.462,21	7.375.552,48
ISS	1.083.337,50	1.238.746,01	1.375.364,13	1.459.916,28	1.503.032,65
ICM/ISS/SIMPLES - LEI FEDERAL	243.914,89	275.985,74	312.771,01	339.868,38	347.899,25
TAXAS	149.268,00	169.578,49	171.284,70	193.727,82	214.922,91

https://static.fazenda.df.gov.br//arquivos/balanco_geral_2016.pdf

Figura III.3: Arrecadação Tributária 2012 a 2016

	2017	2018	2019
RECEITAS TRIBUTÁRIAS	14.779.734,08	15.811.232,57	16.581.617,79
IPTU	722.355,83	794.122,16	1.040.544,21
IR	2.790.541,72	3.168.567,44	3.080.033,67
IPVA	993.058,25	1.057.738,94	1.314.322,99
ITBIM	138.874,25	113.105,17	146.414,04
ITBIV	368.596,70	411.462,60	415.020,61
ICMS	7.557.718,63	7.988.958,10	8.173.794,51
ISS	1.623.386,69	1.651.240,49	2.013.620,28
ICM/ISS/SIMPLES - LEI FEDERAL	369.302,51	401.705,76	0,00
TAXAS	215.899,51	224.331,92	378.612,65
Outros Impostos	-	0,00	19.254,83

https://static.fazenda.df.gov.br//arquivos/12019BALGERALFINAL3003202OFINAL.pdf

Figura III.4: Arrecadação Tributária 2017 a 2019

A Figura III.5 apresenta a transcrição das receitas relativas ao ICMS e ao ISS nos exercícios de 2002 a 2019, e sua totalização anual, de modo que é possível perceber a participação do ICMS e do ISS no valor total arrecadado.

	2002	2003	2004	2005	2006
ICMS	1.793.745.713,65	2.192.768.089,56	2.600.321.927,24	2.906.600.539,58	3.290.372.834,75
ISS	332.912.571,65	381.837.934,25	453.796.391,02	555.279.474,69	607.776.273,30
ICM/ISS/ SIMPLES - LEI FEDERAL	22.514.877,36	27.227.846,46	31.040.907,96	38.632.389,94	41.542.325,56
Total ICMS e ISS	**2.149.173.162,66**	**2.601.833.870,27**	**3.085.159.226,22**	**3.500.512.404,21**	**3.939.691.433,61**

	2007	2008	2009	2010	2011
ICMS	3.433.791,26	3.941.222,99	3.983.560,66	4.493.608,95	5.008.748,86
ISS	642.762,61	675.049,07	759.201,39	856.498,34	941.303,82
ICM/ISS/SIMPLES - LEI FEDERAL	67.114,01	114.655,04	149.804,62	193.124,53	221.398,85
Total ICMS e ISS	**4.143.667,88**	**4.730.927,10**	**4.892.566,67**	**5.543.231,82**	**6.171.451,53**

	2012	2013	2014	2015	2016
ICMS	5.494.095,34	5.987.377,33	6.540.460,06	6.481.462,21	7.375.552,48
ISS	1.083.337,50	1.238.746,01	1.375.364,13	1.459.916,28	1.503.032,65
ICM/ISS/SIMPLES - LEI FEDERAL	243.914,89	275.985,74	312.771,01	339.868,38	347.899,25
Total ICMS e ISS	**6.821.347,73**	**7.502.109,08**	**8.228.595,20**	**8.281.246,87**	**9.226.484,38**

	2017	2018	2019
ICMS	7.557.718,63	7.988.958,10	8.173.794,51
ISS	1.623.386,69	1.651.240,49	2.013.620,28
ICM/ISS/SIMPLES - LEI FEDERAL	369.302,51	401.705,76	-
Total ICMS e ISS	**9.550.407,83**	**10.041.904,35**	**10.189.433,79**

Figura III.5: Arrecadação de ICMS e ISS: 2002 a 2019