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

Redes neurais artificiais aplicadas à identificação de riscos de inadimplência fiscal de ICMS e ISS no Distrito Federal

Vinícius Di Oliveira

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

Orientador
Prof. Dr. Ricardo Matos Chaim

Brasília
2019

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

OOL48r

Oliveira, Vinícius Di
Redes neurais artificiais aplicadas à identificação de riscos de inadimplência fiscal de ICMS e ISS no Distrito Federal / Vinícius Di Oliveira; orientador Ricardo Matos Chaim. -- Brasília, 2019.
68 p.

Dissertação (Mestrado - Mestrado Profissional em Computação Aplicada) -- Universidade de Brasília, 2019.

1. redes neurais artificiais. 2. auditoria fiscal. 3. identificação de riscos. I. Chaim, Ricardo Matos, orient. II. Título.

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

Redes neurais artificiais aplicadas à identificação de riscos de inadimplência fiscal de ICMS e ISS no Distrito Federal

Vinícius Di Oliveira

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

Prof. Dr. Ricardo Matos Chaim (Orientador)
FGA/UnB

Prof. Dr. Rosalvo Ermes Streit
UCB

Prof. Dr. João Carlos Félix Souza
EPR/UnB

Prof.a Dr.a Ana Carla Bittencourt Reis
EPR/UnB

Prof.a Dr.a Aletéia Patrícia Favacho de Araújo
Coordenadora do Programa de Pós-graduação em Computação Aplicada

Brasília, 08 de julho de 2019

Dedicatória

Aos meus pais, Nivaldo e Aleth, pelo amor, dedicação e exemplo de vida.

À mulher da minha vida, Claudia, e aos meus filhos, Jordana, Pedro e João Paulo por me ensinar dia a dia a maior de todas as virtudes: *amar*.

Agradecimentos

Em primeiro lugar à Deus,

À minha família pelo apoio integral, durante o período de ausência de sua presença, nos momentos em que me dedicava a este estudo,

Aos colegas alunos do PPCA-UnB, pelos momentos de motivação, incentivo e de descontração,

Ao excelente corpo docente e técnico da UnB, em especial ao meu orientador Professor Dr. Ricardo Matos Chaim.

Resumo

O presente trabalho busca estudar a aplicação de redes neurais artificiais na identificação de riscos de inadimplência fiscal no Distrito Federal. Foi empregada na pesquisa a base de dados do cadastro fiscal do DF, o qual agrega mais de 300 mil empresas, a modelagem estatística foi feita com dois modelos de predição: regressão LOGIT e redes neurais do tipo perceptron multicamadas. Essa pesquisa procura, como objetivo geral, verificar como o uso de redes neurais artificiais pode auxiliar na identificação de riscos de inadimplência fiscal de ICMS e ISS. O estudo bibliográfico realizado mostrou que as técnicas de modelagem utilizadas na avaliação de risco de crédito, na sua predição de inadimplência, tem semelhanças com a predição de inadimplência fiscal, sugerindo sua aplicação neste problema. A meta-análise bibliométrica foi usada para o embasamento teórico. A metodologia utilizada foi de natureza aplicada, com uma abordagem predominantemente quantitativa, com seu objetivo principal explicativo e uma estratégia do tipo *ex-post facto*. Por fim, a pesquisa constatou que é possível identificar riscos de inadimplência fiscal através da modelagem preditiva utilizando redes neurais artificiais. O resultado alcançado foi um modelo que obteve, na tarefa de predição, uma taxa de erro menor que 11%. A maior dificuldade da rede neural desenvolvida foi a interpretação da influência das variáveis modeladas no resultado da predição, o que foi resolvido pelo modelo de regressão LOGIT.

Palavras-chave: redes neurais artificiais, auditoria fiscal, identificação de riscos

Abstract

This paper aims to study the application of artificial neural networks to identify risks of tax default in the Federal District of Brazil. The survey used the database of the DF tax register, which aggregates more than 300 thousand companies, the statistical modeling was done with two prediction models: LOGIT regression and multilayer perceptron neural networks. This research seeks, as a general objective, to verify how the use of artificial neural networks can help in identifying the risks of ICMS and ISS tax default. The bibliographic study showed that the modeling techniques used on credit risk assessment, on its default prediction, have similarities with the prediction of tax default, suggesting its application on this problem. Bibliometric meta-analysis was used for the theoretical basis. The methodology used was on applied nature, with a predominantly quantitative approach, with its explanatory main objective and an *ex-post facto* strategy type. Finally, the research found that it is possible to identify risks of tax default through predictive modeling using artificial neural networks. The results achieved were a model that obtained, in the prediction task, an error rate of less than 11%. The greatest difficulty of the developed neural network was the interpretation of the modeled variables influence on the prediction result, which was solved by the LOGIT regression model.

Keywords: artificial neural networks, tax audit, risk identification

Sumário

Lista de Figuras

Lista de Tabelas

Lista de Abreviaturas e Siglas

COSO Committee of Sponsoring Organizations of the Treadway Commission.

ERM Enterprise Risk Management.

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.

LFE Livro Fiscal Eletrônico.

SEFP Secretaria de Estado de Fazenda, Planejamento e Gestão do DF.

Capítulo 1

Introdução

1.1 Definição do Problema

Diante do objetivo da fiscalização tributária que é detectar a falta de pagamentos devidos por sonegação ou por erros, assim planejando, executando e gerenciando medidas de prevenção e repressão aos ilícitos identificados, *o principal risco da fiscalização tributária é a falha na percepção do não-pagamento do tributo devido.* A grande quantidade de informações à disposição e a complexidade da legislação tributária torna indispensável o uso de tecnologia de ponta e técnicas científicas para gerenciar o risco de sonegação fiscal, consequentemente surge a necessidade de adoção de padrões de gestão e ferramentas validados cientificamente.

Algumas tendências já podem ser observadas no gerenciamento de riscos corporativos, especialmente: Ferramentas avançadas de análise e visualização de dados e desenvolvimento de inteligência artificial e automação. Na medida em que mais informações são disponibilizadas e a velocidade de sua análise aumenta, o gerenciamento de riscos corporativos precisará se adaptar. Os dados vêm de fora da organização bem como de dentro dela e serão estruturados de diversas formas. Assim, ferramentas avançadas de análise e visualização de dados, em evolução contínua, serão muito úteis para entender o risco e seu impacto, tanto positivo quanto negativo. É fundamental que as práticas de gerenciamento de riscos corporativos observem o impacto dessas e de futuras tecnologias e tirem proveito de seu potencial. Relações, tendências e padrões não reconhecíveis anteriormente podem agora ser observados, proporcionando uma rica fonte de informações críticas para o gerenciamento do risco, a inteligência artificial e a automação viabilizam estes avanços, [3].

No contexto atual da fiscalização tributária, onde se busca detectar as falhas no pagamento dos tributos devidos pelos contribuintes, o foco no que tange à repressão das irregularidades está na observação do passado fiscal das empresas e, consequentemente,

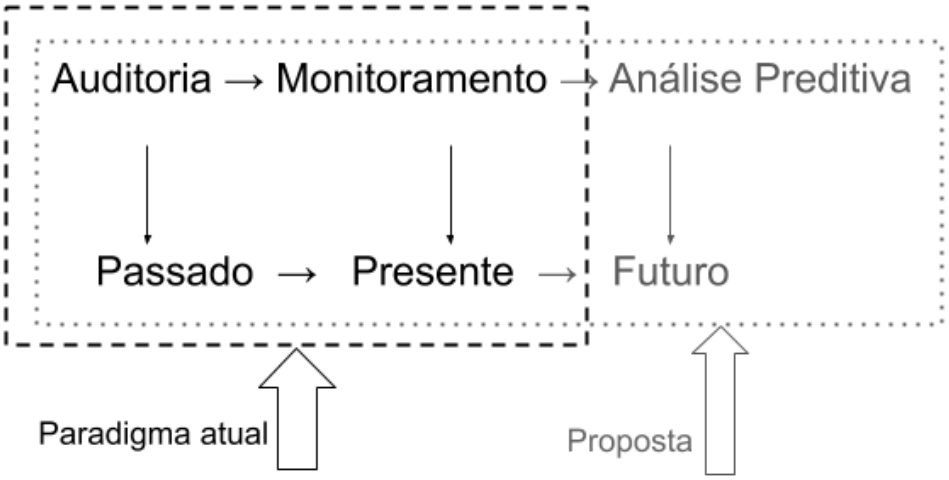

Figura 1.1: Paradigmas da Fiscalização Tributária (Fonte: Autor).

do cumprimento das obrigações tributárias pertinentes. Esta ação é feita na auditoria fiscal. Por outro lado, o monitoramento fiscal observa o cumprimento das obrigações tributárias no presente, mês corrente e meses anteriores recentes, principalmente nas grandes empresas e contribuintes de alto impacto na arrecadação. Diante dos avanços tecnológicos e da grande disponibilidade de dados, seria viável predizer o comportamento fiscal de contribuintes? Seria possível começar a olhar para o futuro na abordagem da gestão de riscos? Este estudo se propõe a iniciar esta discussão apresentando uma novo paradigma para a Fiscalização Tributária. Esta perspectiva é mostrada na Figura 1.1.

Assim, apresenta-se a definição do problema abordado neste estudo: *É possível predizer o comportamento de contribuintes, quanto à inadimplência fiscal, através da aplicação de redes neurais artificiais?*

1.2 Justificativa

Com o incremento do uso dos documentos fiscais eletrônicos bem como da escrituração fiscal eletrônica, o volume de informações disponíveis às administrações tributárias vem crescendo exponencialmente sem uma contrapartida equivalente de crescimento em tecnologia e métodos para monitoramento e controle dessas transações.

Esta lacuna tem destacado, cada vez mais, a necessidade de aprimoramento das técnicas de fiscalização tributária e do desenvolvimento de novas ferramentas aplicáveis ao vultoso volume de dados gerados, de tal forma que o objetivo do FISCO seja alcançado, qual seja a arrecadação e cobrança dos tributos de forma justa e efetiva promovendo a justiça fiscal.

O uso de algoritmos de redes neurais artificiais na avaliação de risco de inadimplência fiscal *buscará aumentar a percepção da população quanto à isenção do FISCO na seleção dos contribuintes a serem fiscalizados* e, posteriormente, no desenvolvimento das ações fiscais. Além de oferecer aos órgãos de Controle Interno e aos Tribunais de Contas um padrão consolidado de referência, proporcionando, assim, maior compreensão e controle dos atos administrativos afetos para que se possa ter maior segurança nos processos de controle externo e interno tanto para os executores quanto para os controladores.

A proposta deste trabalho é pioneira no âmbito das administrações tributárias brasileiras. Não foi identificada iniciativa semelhante nas outras unidades da federação. No segundo semestre de 2018 foi feito contato com outras administrações tributárias estaduais, através do Grupo de Trabalho de Modernização da Fiscalização do ENCAT (Encontro Nacional de Coordenadores e Administradores Tributários) e do Grupo de Trabalho de Meios Eletrônicos de Pagamento do CONFAZ (Conselho Nacional de Política Fazendária), sendo constatado a relevância desta abordagem, pois nenhuma das unidades federadas indicou a implantação de sistemática de classificação de contribuintes quanto à inadimplência fiscal com o uso de técnicas estatísticas. A unidade mais avançada, até então, é a SEFAZ do Estado de Pernambuco que criou um setor para estudos relacionados à Ciência de Dados e Inteligência Artificial mas, até o momento, não teria nada semelhante implantado ou em desenvolvimento.

No que tange à produção acadêmico-científica, também não foram encontrados trabalhos correlatos nas bases de dados pesquisadas.

1.3 Hipótese

Diante da lacuna observada, quanto a estudos que tratem da modelagem da inadimplência fiscal por modelos preditivos, se propôs utilizar modelos de análise de risco de crédito como referência para a tarefa de predição da inadimplência fiscal. Nas seções "3.1 Compreensão do negócio" e "3.2 Compreensão dos dados" serão tratadas as semelhanças da análise de risco de crédito com a análise de risco de inadimplência fiscal.

Em se tratando dos modelos de análise de risco de crédito, são utilizados vários algoritmos para a modelagem preditiva, como será mostrado na Revisão Teórica deste estudo. Foi escolhida a modelagem por redes neurais artificiais, segundo a pesquisa bibliográfica realizada e apresentada no Capítulo 2, por estes algoritmos obterem a melhor taxa de acertos na tarefa de predição. Dessa forma se formou a hipótese de estudo deste trabalho.

A hipótese a ser testada é **"um modelo de redes neurais artificiais pode predizer, com significância estatística, quais contribuintes ficarão inadimplentes quanto às suas obrigações tributárias."**

1.4 Objetivos

Como objetivo deste estudo, pretende-se investigar como o uso de modelos preditivos baseados em redes neurais artificiais podem auxiliar na identificação de riscos da fiscalização tributária. Consequentemente, no mesmo contexto, pretende-se investigar como o uso da mineração de dados pode auxiliar o gerenciamento de projetos de fiscalização tributária.

Objetivo Geral Verificar como o uso de redes neurais artificiais pode auxiliar na identificação de riscos de inadimplência fiscal de ICMS e ISS.

Objetivos Específicos

1. Propor um processo de tratamentos de dados fiscais de forma a discretizar variáveis qualitativas e quantitativas.
2. Identificar fatores de riscos associados às características observadas nas informações fiscais da Administração Tributária através da interpretação associada dos modelos de regressão LOGIT e de redes neurais artificiais.
3. Predizer o comportamento fiscal de empresas do DF (inadimplência ou não) considerando os fatores de risco identificados e o tratamento dos dados fiscais através do uso de modelos preditivos baseados em redes neurais artificiais.

1.5 Estrutura do Estudo

A revisão teórica deste trabalho abordará o conceito do Gerenciamento de Risco Corporativo, ou *Enterprise Risk Management (ERM)*, apresentando em seguida modelos e ferramentas de avaliação do risco utilizadas no ERM. O painel de boas práticas do *Committee of Sponsoring Organizations of the Treadway Commission (COSO)* será abordado trazendo seus conceitos e evolução ao longo do tempo, para depois mostrar o elementos básicos utilizados hoje na gestão de riscos na auditoria fiscal. Na sequência é apresentada a meta-análise bibliográfica e conceito de Análise de Risco de Crédito, a qual demonstrará a relevância dos modelos de predição estatística nesta tarefa, para então seguir para a apresentação do conceito de mineração de dados e do método a ser utilizado neste estudo, por fim serão apresentados os dois tipos de modelos preditivos estudados neste trabalho, regressão logística e redes neurais artificiais.

O capítulo Metodologia apresentará a classificação metodológica do presente estudo, a Compreensão do Negócio, onde será apresentada a conexão do problema de estudo com a solução proposta, a Compreensão dos Dados, a apresentar as características das informações utilizadas e definição das variáveis, a Preparação e Tratamento dos Dados, a propor

um método de discretização das variáveis qualitativas e quantitativas, a Modelagem, que mostrará a construção dos modelos regressão LOGIT e rede neural.

Na Avaliação dos Resultados será feita a identificação e avaliação dos fatores de risco, a discussão da solução proposta, bem como a validação do modelo RNA. Por fim na seção Conclusões e Estudos Futuros o trabalho será concluído apresentando o resultado do cumprimento dos objetivos propostos para este estudo.

Capítulo 2

Revisão Teórica

Nesta seção serão estudadas as dimensões nas quais se inserem os objetivos deste estudo. Será feita a apresentação do conceito de Gestão de Risco Corporativo ou *Enterprise Risk Management (ERM)*, no contexto estudado, bem como uma abordagem da sua estrutura dentro do modelo do *Committee of Sponsoring Organizations of the Treadway Commission (COSO)*. Também será mostrado como a fiscalização tributária do Distrito Federal se encaixa na sistemática de gestão de riscos destacando sua metodologia e diretrizes. Em seguida será apresentado um panorama geral da meta-análise bibliográfica da produção acadêmica relacionada a análise de risco de crédito de 2015 a 2018. Por fim, será exibida uma visão geral da teoria que envolve as técnicas estatísticas de modelos preditivos baseados em redes neurais artificiais.

2.1 Conceito de ERM no contexto estudado

Uma das primordiais tentativas de conceituar *Enterprise Risk Management (ERM)* foi a do *Price Waters Coopers (PwC)* no seu *framework* em 2003 *"Enterprise Risk Management: A framework for success"* onde trata o ERM como uma abordagem completa e sistemática que busca ajudar as organizações, independentemente da seu tamanho ou missão, a identificar eventos, mensurar, priorizar e responder aos seus desafios de risco dos projetos e iniciativas que assumem. O ERM permite às organizações determinar o nível de risco que podem ou querem aceitar na sua procura de gerar valor para os investidores [4]. O "ERM oferece uma estrutura para gerir eficazmente a incerteza, respondendo aos riscos e explorando as oportunidades que surjam".

Em setembro de 2004 o *Committee of Sponsoring Organizations of the Treadway Commission (COSO)*, em parceria com a *PwC*, divulga aquele que é o documento base e de referência sobre o ERM, o *Enterprise Risk Management – Integrated Framework*, que o conceitua como: "Um processo, efetuado pela Comissão Executiva, a gestão e outros co-

laboradores, aplicado na definição de estratégias e em toda a entidade, desenhado para identificar eventos potenciais que possam afetar a entidade, e gerir o risco dentro dos limites definidos, de modo a fornecer uma segurança razoável no que respeita ao alcançar dos objetivos da entidade" [5].

É um conceito intencionalmente abrangente para encapsular definições relevantes sobre como as empresas e outras organizações gerem o risco, entregando uma plataforma para o seu uso através das mais diversas entidades [4]. Conforme Beasley et al., [6] muitas organizações estão usando processos de ERM para aumentar a eficácia dos seus processos de gestão de riscos, com o objetivo final de gerar valor para os *Stakeholders*.

Diversas organizações têm implementado programas de gestão de risco do negócio de forma progressiva, empresas de consultoria estabelecem equipes especializadas em ERM, e os estabelecimentos de ensino têm criado disciplinas relacionadas com o ERM. Ao contrário da gestão de risco tradicional que analisa categorias de risco separadamente, o ERM permite às entidades fazer a gestão de riscos de forma integrada e sistêmica [7]. Ou seja, o *Enterprise Risk Management (ERM)* propõe que as empresas abordem todos os seus riscos de forma abrangente e coerente, em vez de gerenciá-los individualmente. Outra visão semelhante foi trazida por Wu et al. [8], "O ERM é uma abordagem integrada para gerenciar os riscos enfrentados por uma organização, buscando as formas mais eficazes de lidar com os riscos".

Profissionais da gestão de risco precisam entender como indivíduos e grupos diferentes dentro da organização definem riscos, como atuam os possíveis desvios na avaliação de riscos e os desafios na implementação de iniciativas de gerenciamento de riscos. Estudos ainda precisam demonstrar benefícios consistentes do ERM. A história recente levanta dúvidas sobre a eficácia do gerenciamento de risco, conforme praticado anteriormente. Afinal na crise econômica causada pela "crise do *subprime*" em 2008, os profissionais mais avançados no gerenciamento de risco (por exemplo, os bancos de Wall Street) sofreram mais, causando grandes danos à economia internacional [9].

2.2 COSO

O *Committee of Sponsoring Organizations of the Treadway Commission (COSO)* é um *framework* desenvolvido para auxiliar o gerenciamento na implementação do ERM. A estrutura básica de idéias do COSO é implementar boas práticas de governança, definir seus objetivos, identificar os riscos importantes e, ao mesmo tempo, controlar e monitorar ativamente esses riscos [3].

Foi instituído no Distrito Federal, em abril de 2016, o modelo de Gestão de Riscos e Controle Interno - Estrutura Integrada - 2013 do COSO como instrumento de boa prática técnica e gerencial [10].

Em agosto de 2018 a Secretaria de Estado de Fazenda, Planejamento e Gestão do DF (SEFP) instituiu o seu Comitê Superior de Gestão de Riscos, em consonância com a diretiva distrital, tendo o COSO e a Norma ABNT NBR ISO 31000 como referências de trabalho [11]. Assim, a gestão de risco na administração tributária do DF deve ser construída de acordo com estas diretrizes.

O COSO publicou ao longo dos últimos anos seus documentos tidos como referências em termos de boas práticas de gestão os quais destacamos a seguir.

2.2.1 COSO-IC

Publicado em 1994, o modelo orienta as organizações quanto a princípios e melhores práticas de controle interno – processo projetado e implementado pelos gestores para mitigar riscos e alcançar objetivos, onde risco é entendido como a possibilidade de ocorrência de evento capaz de afetar o alcance dos objetivos.

Atualmente, este *framework* é o mais usado pelas companhias com ações em bolsa nos Estados Unidos, devido à necessidade de atendimento de requisitos legais [12], podendo ser aplicado na avaliação dos controles internos relacionados com as operações e conformidade legal/regulatória de diversas organizações. Em 2013, foi publicada uma atualização do COSO-IC, denominada Controle Interno - Estrutura Integrada - COSO.

No que tange aos projetos de fiscalização tributária, as boas práticas do COSO seguidas na Programação Fiscal da Subsecretaria de Receita da SEFP dizem respeito à definição e alinhamento dos objetivos operacionais (criação de projeto de fiscalização), de comunicação (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 ficam restritas - em face da obrigação de manter o sigilo fiscal das empresas.

2.2.2 COSO-ERM

Publicado em 2004, o COSO-ERM incorpora o COSO-IC, sem substituí-lo, abrangendo a realização normal das atividades - operacionais, administrativas e de suporte – e o planejamento voltado à definição da estratégia da organização, proporcionando um enfoque mais robusto sobre o tema da gestão de riscos.

Para orientar o estabelecimento do processo de gestão de riscos, esse *framework* incorporou novos componentes e elementos, como a definição do apetite de risco e os níveis de tolerância a riscos da organização.

A Figura 2.1 apresenta a relação entre controle interno, gestão de riscos corporativos e a governança definida na versão 2013 do COSO-IC, bem como a diferença da representação das estruturas do COSO-IC e do COSO-ERM.

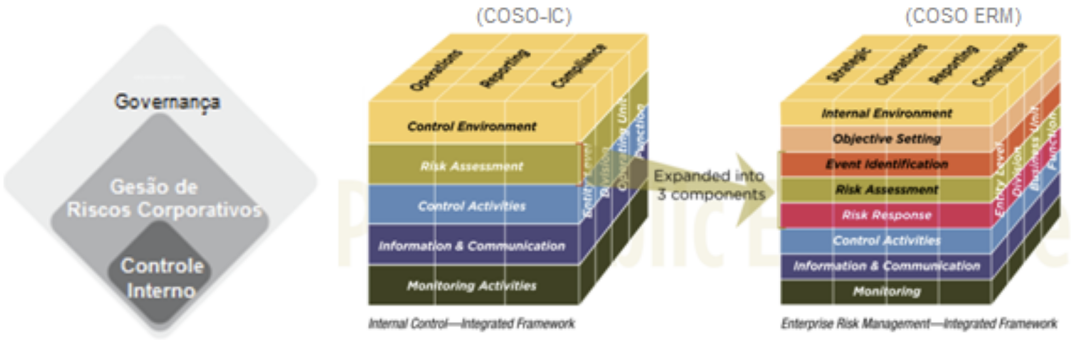

Figura 2.1: Relação entre a representação do COSO-IC e do COSO-ERM (Fonte: Autor).

O COSO-ERM - *Integrating with Strategy and Performance* [3], publicado em 2017, representa a atualização deste *framework* e traz em destaque a importância de considerar os riscos tanto no processo de estabelecimento da estratégia quanto na condução no desempenho da execução da organização.

Uma novidade da publicação é a estrutura de gerenciamento de riscos representada por duas figuras e não em forma de cubo. A primeira dessas figuras demonstra a importância de se ter objetivos estratégicos alinhados à missão, visão e valores da organização como forma de fortalecer o desempenho da organização. Como pode ser visto na Figura 2.2.

Figura 2.2: Relação entre missão, visão e valores da organização, atividades, estratégias, objetivos e performance (Fonte:[1]).

A Figura 2.3, a seguir, demonstra que o sucesso para um desempenho operacional gerador de riqueza acontece através da integração e equilíbrio de todos os departamentos e funções.

Figura 2.3: Importância da integração e equilíbrio de todos os componentes para o desempenho operacional (Fonte:[1]).

Semelhante à estrutura do COSO-IC, esta nova visão estabelece vinte princípios organizados entre os cinco componentes inter-relacionados (Figura 2.4). Estes princípios cobrem todo o entendimento entre as atividades de governança e monitoramento, auxiliando as corporações se prepararem para estarem aderentes a esta melhor prática.

Figura 2.4: Correlação entre Componentes e Princípios do COSO (Fonte:[1]).

A aplicação de modelos preditivos baseados em redes neurais artificiais na gestão de risco de projetos de auditoria fiscal, objeto do presente estudo, estaria inserida no componente "Performance" dentro das diretrizes do COSO-ERM, mais especificamente no item "10- Identifica o risco" (*10. Identifies Risk*), caso fosse implementado pela administração tributária do DF. O uso das redes neurais, nos moldes dos algoritmos de análise de risco de crédito, poderiam indicar contribuintes com alto risco de inadimplência fiscal para o

FISCO, assim como indicam mutuários com alto risco de inadimplência financeira para os operadores de crédito, ou seja, o modelo estaria a identificar o risco apontando quais contribuintes são inadimplentes em potencial.

Na fiscalização tributária, as boas práticas do COSO-ERM seguidas na Programação Fiscal dizem respeito à observação dos princípios relacionados aos componentes *Strategy & Objective-Setting* (Estratégia e definição de objetivos); *Performance* (Desenho e performance); e *Review & Revision* (Análise e revisão). Sendo que os princípios relacionados a *Governance & Culture* (Governança e cultura) estão mais alinhados com as atribuições da Coordenação de Fiscalização Tributária e *Information, Communication & Reporting* (Objetivos de Informação, comunicação e divulgação) possue ações limitadas aos relatórios de acompanhamento, faltando um política mais abrangente de divulgação de resultados, o que prejudica que os bons resultados aumentem a sensação de risco por parte dos contribuintes que descumprem suas obrigações.

2.3 Modelos e ferramentas de avaliação de risco

Nesta seção será apresentada uma revisão sobre as ferramentas de modelagem utilizadas no processo de tomada de decisão no contexto do *Enterprise Risk Management* feita por Wu et al. [8] no artigo *"Decision making in enterprise risk management: A review and introduction to special issue"*.

Muitos tipos de modelos diferentes foram propostos para apoiar o gerenciamento de riscos dentro das organizações, [13] apud [8]. A abordagem tradicional é desenvolver um modelo analítico com a intenção de identificar uma política ótima, [14] apud [8]. Devido à incerteza envolvida, a análise estatística e a simulação são muito apropriadas para analisar o risco.

A **análise bayesiana** já foi proposta para modelar a integração de informação e conhecimento dentro de redes complexas, [15] apud [8]. O foco no gerenciamento de riscos tem sido dado há muito tempo ao campo das finanças, com muitas ferramentas científicas de gestão para ajudar investidores e seguradoras. Modelos de Simulações diversas já foram propostos em vários estudos, para incluir modelagem de eventos discretos estimando a sobrevivência no gerenciamento de risco financeiro, [16] apud [8].

A **simulação de Monte Carlo** é usada para avaliar os riscos associados à seleção de fornecedores, seguindo modelos semelhantes de várias fontes, [17] apud [8].

Os **modelos da Dinâmica de Sistemas** têm sido amplamente utilizados, especialmente no que diz respeito ao "efeito chicote", na cadeia de abastecimento, [18] apud [8], e para modelar questões de risco relacionadas ao ambiente, à rede logística e ao planejamento de estoques, [19] apud [8].

Como observado acima, vemos um espectro de modelagem aplicado nas ciências empresariais, variando de modelos analíticos destinados a identificar decisões ótimas sob suposições geralmente bastante rígidas, modelos matemáticos de integração até modelos simulando incertezas de forma numérica [8].

As últimas décadas também viram um tremendo crescimento na **mineração de dados** como uma ferramenta para lidar com a enorme quantidade de dados gerados pela tecnologia em constante mudança. Todos esses tipos de modelagem foram aplicados à modelagem de risco.

Ainda segundo Wu et al. [8] a mineração de dados emergiu como um campo muito ativo aplicado a praticamente todos os campos da ciência. Há uma grande variedade de ferramentas de mineração de dados disponíveis e várias técnicas diferentes em cada área de pesquisa em evolução. *Uma aplicação clássica de mineração de dados é a classificação para aprovação de crédito, da qual surgiu a hipótese de pesquisa deste trabalho.* Outros exemplos são a aplicação atual neste campo usando algoritmos de máquinas de vetores de suporte (*Support Vector Machines*) *fuzzy*, [20] apud [8], ou ainda, árvores de decisão aplicadas, redes neurais e regressão logística aplicada à análise de risco de contas a receber, [21] apud [8].

A **mineração de texto** lida com palavras em vez de números. Pesquisadores aplicaram a mineração de texto à indústria de serviçoes financeiros, [22] apud [8]. Outro campo emergente é a mineração de processos aplicada à verificação de conformidade, [23] apud [8], e ao gerenciamento de riscos corporativos em geral, [24] apud [8].

Com o crescimento cada vez maior do volume e da variedade dos dados disponíveis, a mineração de dados tem ganho relevância como ferramenta na gestão de risco, em especial pode-se destacar o seu uso na análise de risco de crédito, *credit scoring*, amplamente usadas por estabelecimentos financeiros e de venda a varejo. A partir dessa aplicação surge a hipótese de uso de modelos de *credit scoring* para avaliação de risco de inadimplência tributária.

2.4 Gestão do Risco em Projetos de Auditoria Tributária

Por norma [10, 11], o planejamento, a execução das atividades relacionadas à elaboração e do refinamento dos projetos de fiscalização apoia-se em ferramentas de uso consagrado, apresentadas na ABNT NBR ISO 31010 (XV) [25], dentre as quais podemos destacar as seguintes.

2.4.1 Estrutura de tratamento dos riscos

A fiscalização tributária do DF utiliza a estrutura de processos da NBR ISO 31000, Representada pela lista 7Rs e 4Ts do gerenciamento de risco - *hazard* [26]:

1. Reconhecimento o risco
2. Ranquear os riscos
3. Responder a riscos significativos
    - Tolerar
    - Tratar
    - Transferência
    - Terminar
4. Recursos Controlados
5. Reações Planejadas
6. Relatórios de monitoramento de desempenho
7. Revisão

2.4.2 Ferramentas utilizadas para identificar e avaliar riscos

A seguir são apresentadas as ferramentas de identificação e avaliação de riscos apontadas na NBR ISO 31000 [25] implementadas na gestão de projetos de auditoria fiscal da administração tributária do Distrito Federal. Para cada item é feita uma breve descrição da ferramenta e da forma como é utilizada na fiscalização tributária.

Brainsstorming Estimular e incentivar o livre fluxo de conversação entre um grupo de pessoas conhecedoras para identificar os modos de falha potenciais e os perigos e riscos associados. Na Programação Fiscal, as denúncias fornecem subsídios necessários às discussões e análises da realidade à luz de novos relatos e indícios.

Análise de Cenários Desenvolver modelo descritivo do futuro e os conjuntos de cenários (por exemplo: melhor caso, pior caso e caso esperado) podem ser usados para mapear os potenciais riscos futuros. Como exemplo de uso desta técnica temos a análise para decidir se um grupo de contribuintes com indícios de irregularidade receberá um comunicado ou uma notificação. A notificação não cumprida implica em atuação, portanto é preciso avaliar se temos recursos para autuar todos os contribuintes caso o indício se comprove para todos.

Análise Custo Benefício Avaliar os riscos ponderando os custos totais contra os benefícios totais esperados a fim de escolher a melhor ou a mais rentável opção. Esta ferramenta é utilizada porque os recursos para auditoria são limitados pelo número de auditores, desta forma é preciso comparar os indícios de irregularidade de diversos projetos e contribuintes, antes de distribuir as ordens de serviço para os auditores.

Matriz de Probabilidade e Consequência Combinar classificações qualitativas ou semi-quantitativas de consequências e probabilidades a fim de produzir uma classificação de risco Na fiscalização distrital, a matriz é utilizada para selecionar o modelo de fiscalização adotado por atividade econômica. A Figura 2.5 apresenta um exemplo: neste caso, os riscos extremos são monitorados pela Gerência de Monitoramento (GEMAE), os riscos altos são tratados com fiscalizações pontuais pela GEPRO e pelas gerências de Auditoria (GEAUT) e de fiscalização de mercadorias em trânsito (GEFMT); os riscos moderados e baixo são tratados pela Gerência da Malha Fiscal e pela GEFMT.

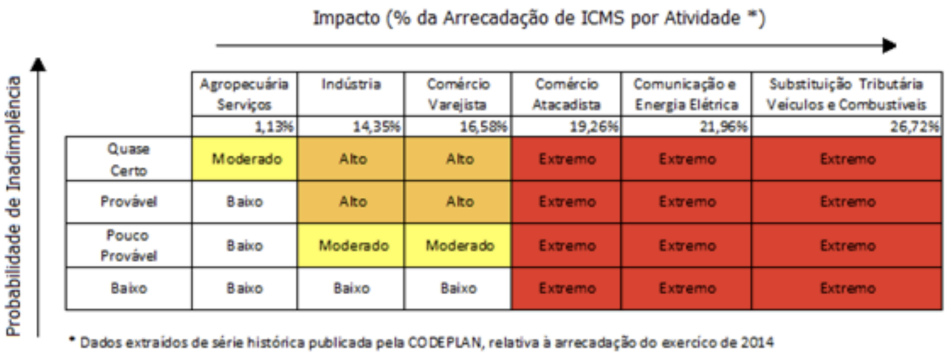

Figura 2.5: Matriz Probabilidade de Inadimplência x Impacto da arrecadação (Fonte: Autor).

2.5 Análise de Risco de Crédito - Meta-análise bibliográfica e conceito

"A pontuação de crédito, ou classificação de crédito, é o conjunto de modelos de decisão e suas técnicas subjacentes que ajudam os credores a julgar se uma solicitação de crédito deve ser aprovada ou rejeitada" Thomas et al. [27].

Para a meta-análise bibliométrica deste trabalho foi usada a base de dados *Web Of Science*. Segundo Mariano [28], atualmente os bancos de dados bibliográficos mais im-

portantes são *ISI Web of Science* - WoS[1], Scopus[2], Google Scholar[3] e o MEDLINE da NLM[4]. A escolha do *Web Of Science* se deu pela maior extensão temporal abrangida, pela maior disponibilidade de material nas áreas das ciências e pela maior disponibilidade de monografias e anais de conferências.

Foi feita a pesquisa de ocorrências na base WoS com as palavras-chave "*credit scoring*" no período dos últimos 5 anos sendo selecionados mais de 1.100 trabalhos acadêmicos, dos quais quase 900 são artigos. A relação de tipos de documentos produzidos está mostrada na Figura 2.6.

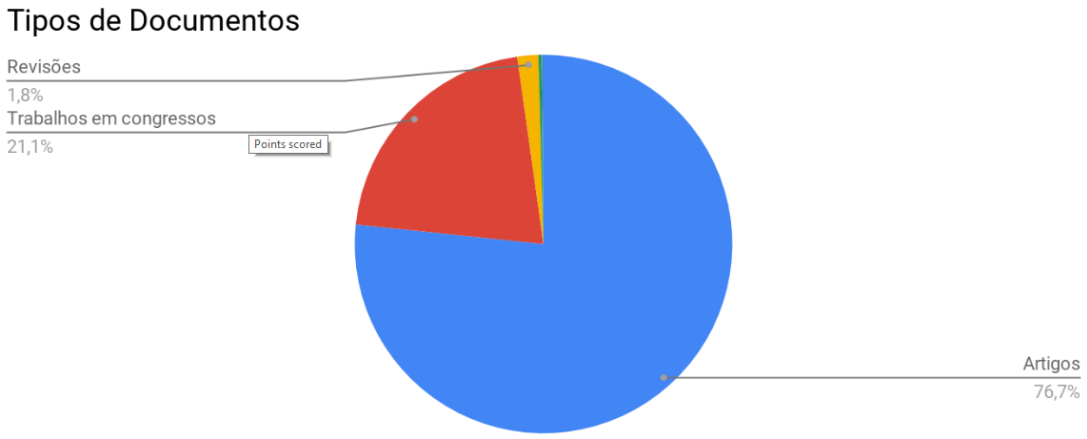

Figura 2.6: Relação dos Tipos de documentos encontrados na base *Web of Science* (Fonte: Autor).

A Figura 2.7 apresenta a distribuição dos trabalhos publicados pelas categorias do *Web of Science*. Pode ser verificado o destaque das categorias "Economia" (*Economics*) e "Finanças de Negócios" (*Business Finance*) que se relacionam ao negócio finalístico do *credit scoring*.

A evolução das citações em contraste com a crescimento das publicações pode ser visualizado na Figura 2.8, a qual demonstra o aumento exponencial das citações dos artigos relacionados ao *credit scoring*.

A Figura 2.9 apresenta o grafo das relações de co-citação das publicações mineradas na base *WoS*. Por meio do software VOSviwer foram elaborados grafos de relações, facilitando a visualização da análise de co-citação, palavras-chave e fonte de publicações, com base nos registros encontrados em *Web of Science* do tema *credit scoring*. Na análise de co-citação, Figura 2.9, é possível compreender quais autores costumam ser citados simultaneamente, indicando similaridade entre as linhas de pesquisa dos mesmos [28, 29, 30], e na análise

[1]Mais informações *ISI Web of Science*: `http://www.webofknowledge.com`
[2]Mais informações *Scopus*: `http://www.scopus.com`
[3]Mais informações *Google Scholar*: `http://scholar.google.com`
[4]Mais informações *MEDLINE*: `http://www.ncbi.nlm.nih.gov/pubmed`

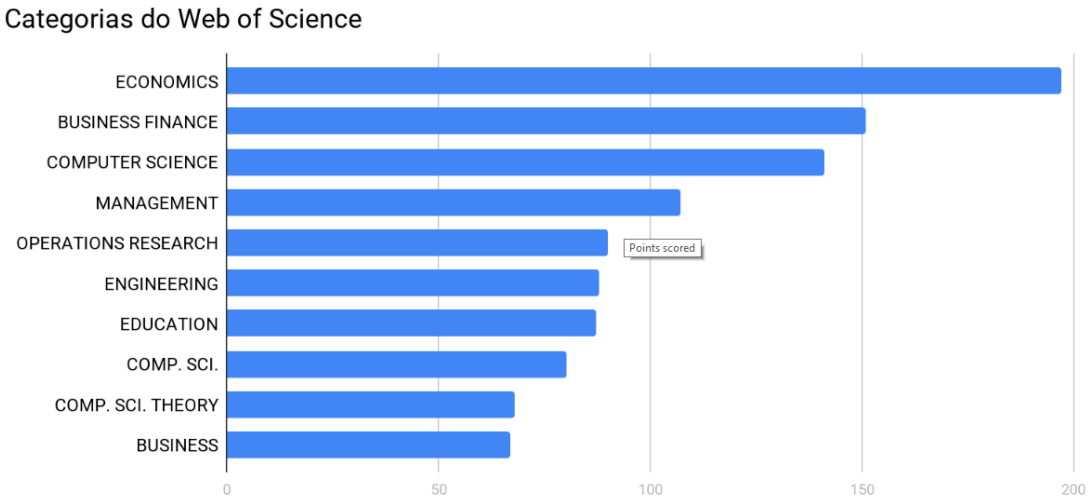

Figura 2.7: Distribuição dos trabalhos na Base WoS publicados por categoria (Fonte: Autor).

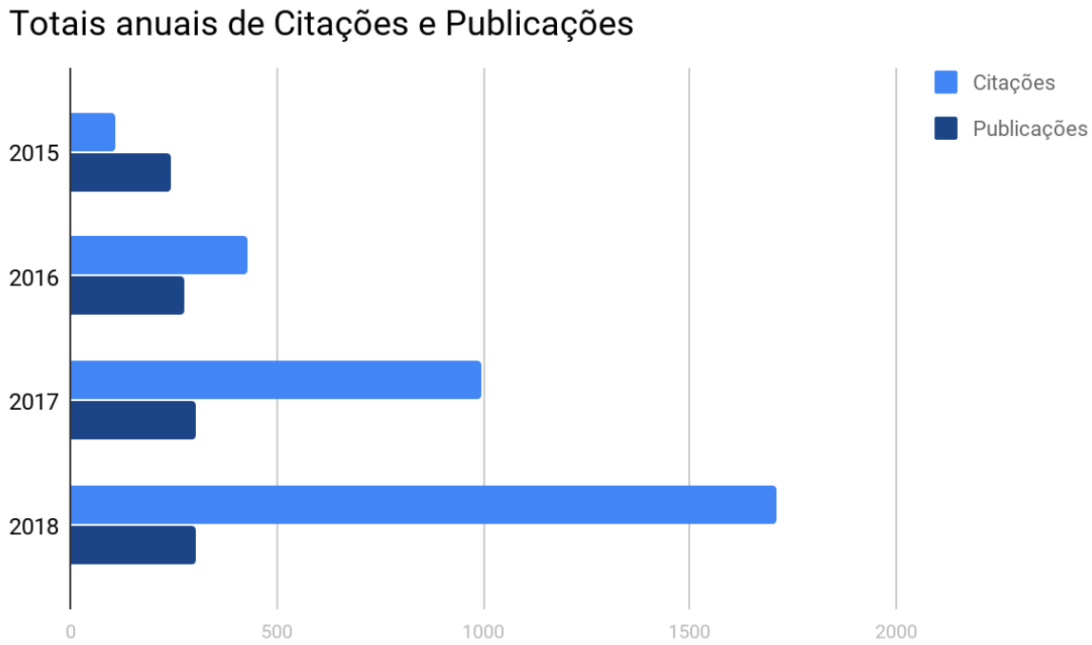

Figura 2.8: Evolução dos totais anuais de citações e publicações dos trabalhos publicados (Fonte: Autor).

das palavras-chave, Figura 2.10, pode-se observar quando dois ou mais trabalhos fazem referência a palavras-chaves em comum, indicando que há chances de que os trabalhos tenham assuntos em comum [28, 31], as relações de publicação, Figura 2.11, mostram as fontes com mais de 5 publicações na base *WoS* e os *links* mostram as relações entre elas [32, 33].

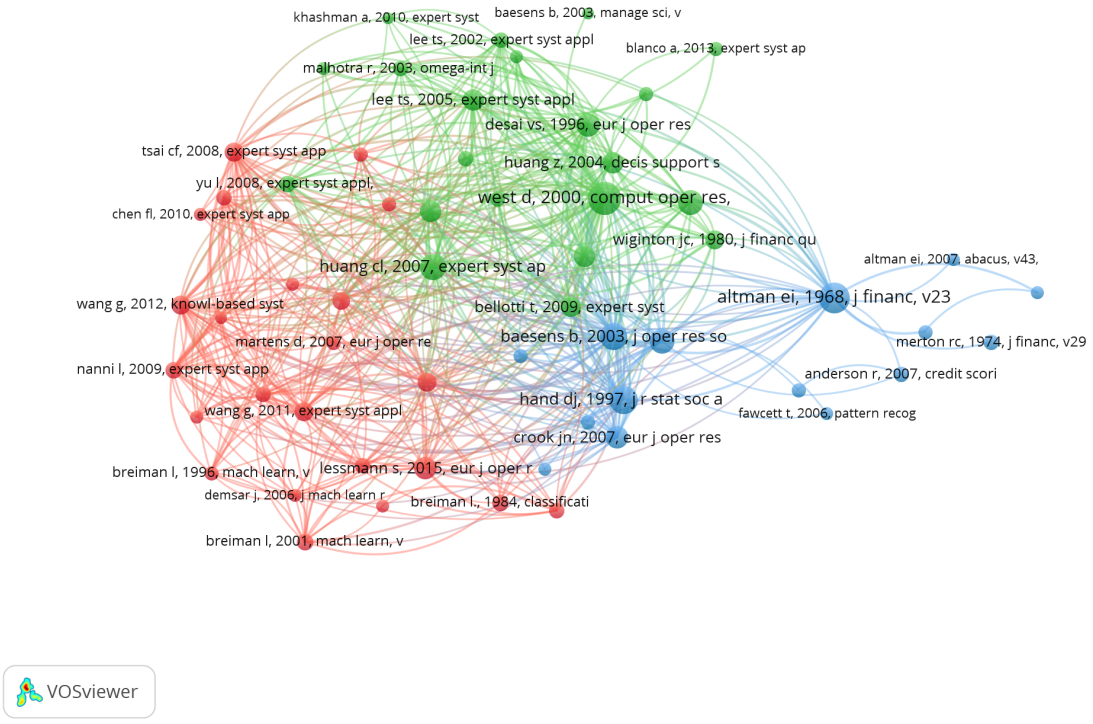

Figura 2.9: Grafo de co-citações dos trabalhos publicados na base WoS (Fonte: Autor).

Pode ser observado na Figura 2.10, grafo das relações entre palavras-chave das publicações destacadas da base WoS com mais de 30 citações, a formação de três *clusters*: (1) Financeiro, lado direito do grafo em verde; (2) Gestão de Risco, lado inferior em azul; e (3) Modelagem preditiva, lado esquerdo e superior em vermelho. Neste último *cluster* é possível perceber o destaque de dois algoritmos de predição e classificação mais utilizados no *credit scoring*: *neural-networks* e *support vector machines*.

As redes neurais são o tipo algoritmo mais citado em publicações (135) e, como pode ser visto na Figura 2.10, tem relação mais próxima com o *credit scoring*. O que corrobora, dentre outros aspectos tratados no item 2.7.2, a escolha deste algoritmo como objeto deste estudo.

No que tange às fontes de publicação, podem ser destacadas áreas de interesse do *credit scoring*: Economia e finanças, Gestão de Risco e Pesquisa Operacional e computação aplicada, conforme se observa nos clusters vermelho, verde e azul, respectivamente, apresentados da Figura 2.11.

O conjunto de trabalhos identificados foi ordenado em ordem decrescente pelo número

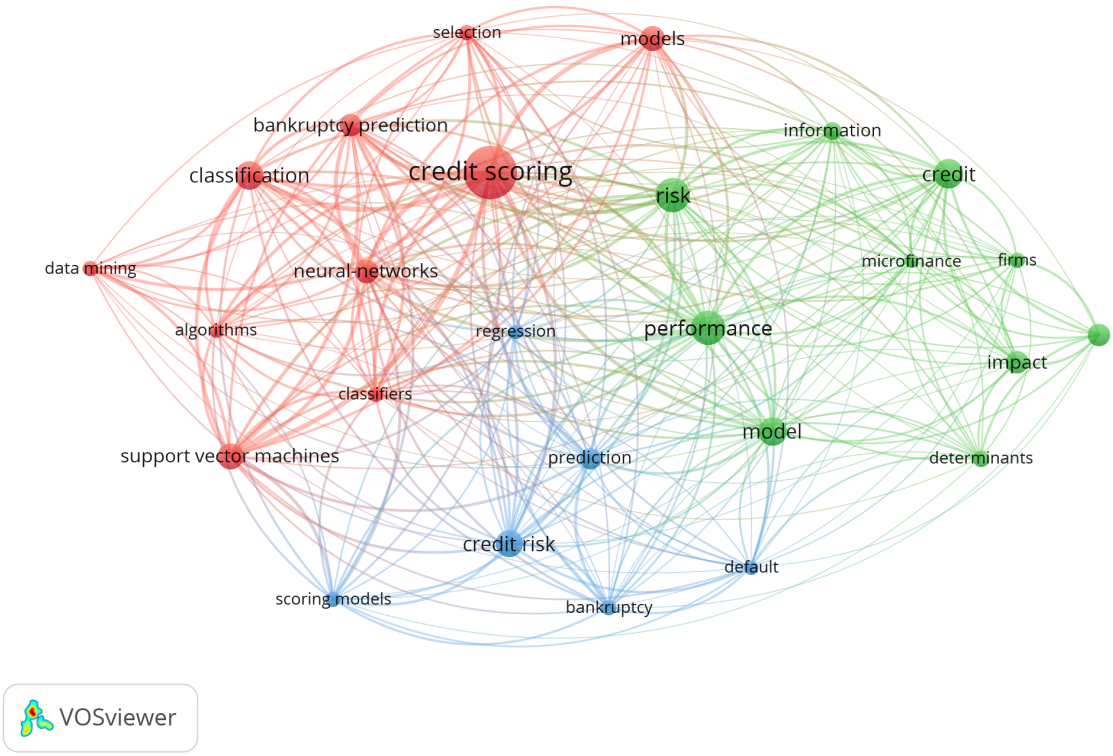

Figura 2.10: Grafo das relaçoes de palavras-chave dos trabalhos publicados na base WoS (Fonte: Autor).

de citações para destacar os artigos mais citados. A seguir são mostrados os 10 artigos com mais citações.

1. Benchmarking state-of-the-art classification algorithms for credit scoring: An update of research

   - Autores: Lessmann, Stefan; Baesens, Bart; Seow, Hsin-Vonn; Thomas, Lyn.
   - Publicação: EUROPEAN JOURNAL OF OPERATIONAL RESEARCH Volume: 247 Edição: 1 Páginas: 124-136 Publicado: NOV 16 2015
   - Citações: 83

2. The Scored Society: Due process for automated predictions

   - Autores: Citron, Danielle Keats; Pasquale, Frank
   - Publicação: WASHINGTON LAW REVIEW Volume: 89 Edição: 1 Páginas: 1-33 Publicado: MAR 2014

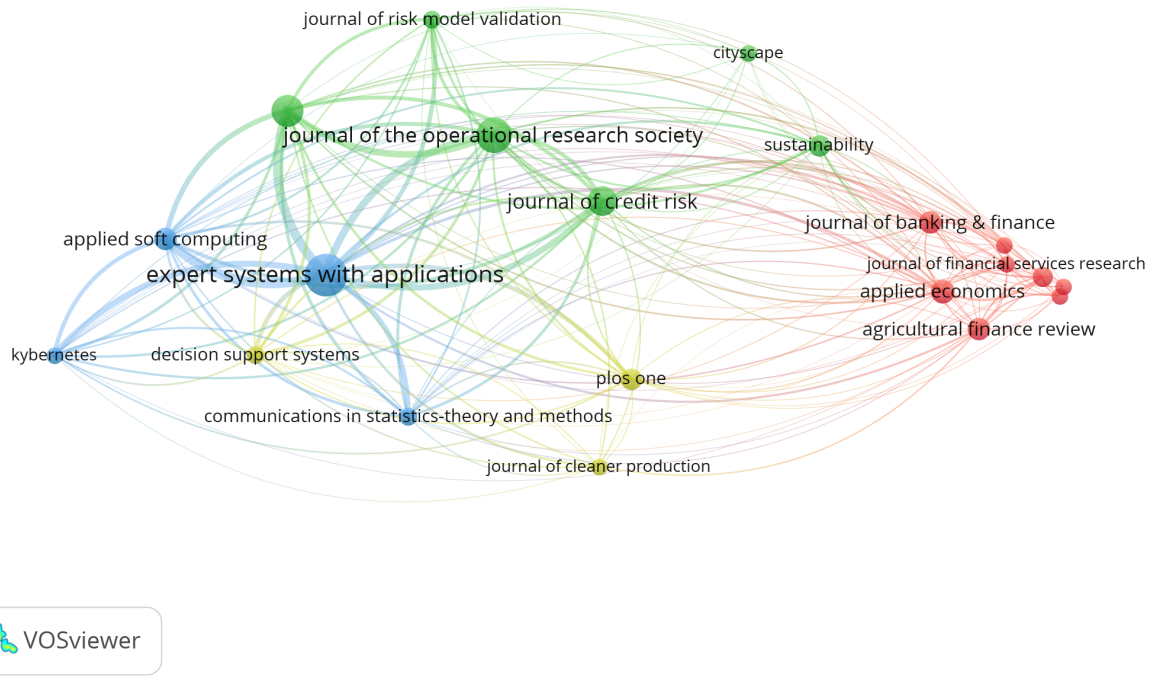

Figura 2.11: Grafo das fontes de publicação e suas relações na base WoS (Fonte: Autor).

- Citações: 73

3. How the machine 'thinks': Understanding opacity in machine learning algorithms
   - Autores: Burrell, Jenna
   - Publicação: BIG DATA SOCIETY Volume: 3 Edição: 1 Páginas: 1-12 Publicado: JAN 6 2016
   - Citações: 66
4. Credit scoring using the clustered support vector machine
   - Autores: Harris, Terry
   - Publicação: EXPERT SYSTEMS WITH APPLICATIONS Volume: 42 Edição: 2 Páginas: 741-750 Publicado: FEB 1 2015
   - Citações: 52

5. APATE: A novel approach for automated credit card transaction fraud detection using network-based extensions

    - Autores: Van Vlasselaer, Veronique; Bravo, Cristian; Caelen, Olivier; et al.
    - Publicação: DECISION SUPPORT SYSTEMS Volume: 75 Páginas: 38-48 Publicado: JUL 2015
    - Citações: 44

6. Demand-side vs. supply-side technology policies: Hidden treatment and new empirical evidence on the policy mix

    - Autores: Guerzoni, Marco; Raiteri, Emilio
    - Publicação: RESEARCH POLICY Volume: 44 Edição: 3 Páginas: 726-747 Publicado: APR 2015
    - Citações: 43

7. Regulating Consumer Financial Products: Evidence from Credit Cards

    - Autores: Agarwal, Sumit; Chomsisengphet, Souphala; Mahoney, Neale; et al.
    - Publicação: QUARTERLY JOURNAL OF ECONOMICS Volume: 130 Edição: 1 Páginas: 111-164 Publicado: FEB 2015
    - Citações: 43

8. Evaluating credit risk and loan performance in online Peer-to-Peer (P2P) lending

    - Autores: Emekter, Riza; Tu, Yanbin; Jirasakuldech, Benjamas; et al.
    - Publicação: APPLIED ECONOMICS Volume: 47 Edição: 1 Páginas: 54-70 Publicado: JAN 2 2015
    - Citações: 43

9. Business intelligence in banking: A literature analysis from 2002 to 2013 using text mining and latent Dirichlet allocation

    - Autores: Moro, Sergio; Cortez, Paulo; Rita, Paulo
    - Publicação: EXPERT SYSTEMS WITH APPLICATIONS Volume: 42 Edição: 3 Páginas: 1314-1324 Publicado: FEB 15 2015
    - Citações: 42

10. Combining cluster analysis with classifier ensembles to predict financial distress

    - Autores: Tsai, Chih-Fong

- Publicação: INFORMATION FUSION Volume: 16 Edição especial: SI Páginas: 46-58 Publicado: MAR 2014
- Citações: 42

Destaca-se o artigo "*Benchmarking state-of-the-art classification algorithms for credit scoring: An update of research*", devido à sua relevância e afinidade com o tema do presente trabalho, este artigo terá seus conceitos e contribuições abordadas neste trabalho.

Segundo Lessmann et al [34], o *credit scoring* (classificação de risco de crédito) está preocupado com o desenvolvimento de modelos empíricos para apoiar a tomada de decisão no negócio de crédito de varejo. A classificação de risco de crédito é um modelo base para estimar a probabilidade de um mutuário demonstrar algum comportamento não desejado no futuro. Na aplicação da pontuação, por exemplo, credores aplicam modelos preditivos, chamados de quadros de resultado (*scorecards*), para estimar a probabilidade de uma candidato ficar inadimplente. Os modelos de risco corporativo utilizam dados de balanços, índices financeiros ou indicadores macroeconômicos, enquanto os modelos de varejo usam dados de formulários de solicitação, dados demográficos de clientes e dados transacionais do histórico do cliente.

O problema associado à classificação de risco de crédito é a categorização de potenciais tomadores de empréstimos em bons ou maus pagadores. Os modelos são desenvolvidos para ajudar os bancos a decidir sobre a concessão ou não de um empréstimo para um novo tomador usando os dados de suas características. Apesar da evolução da tecnologia, a regressão linear ainda é o modelo de referência padrão da indústria usado para construir modelos de classificação de risco de crédito [35], os estudos pesquisados demonstraram que técnicas de inteligência artificial (IA), como redes neurais (Neural networks - NN), máquinas de vetores de suporte (support vector machine - SVM), árvores de decisão (decisions trees - DT), florestas aleatórias (random forests - RF) e Bayes (naïve Bayes - NB), podem ser substitutos para abordagens estatísticas na construção de modelos de pontuação de crédito [36]. Ou seja, nos últimos anos, a inteligência artificial mostrou suas vantagens na pontuação de crédito em comparação com modelos lineares de probabilidade, análise discriminante e outras técnicas estatísticas [37].

2.6 Mineração de Dados

O conceito de Mineração de Dados não é claro na literatura, pode ser uma fase do processo de Descoberta de Conhecimento em Bases de Dados (*Knowledge Discovery in Databases* - KDD) ou pode ser um conjunto de meios de detecção de padrões nos dados [38]. A atenção sobre o tema cresceu com o surgimento e disseminação de grande bancos de

dados em diversas áreas, onde a obtenção de conhecimento útil sobre estes dados depende de técnicas adequadas para tal tarefa. Os maiores esforços acadêmicos para criar um padrão comum na mineração de dados são no sentido do desenvolvimento de um idioma único, assim como o SQL está para os bancos de dados relacionais. Já na indústria, os esforços são no sentido de padronizar processos e metodologias de mineração de dados [39].

Nesta seção, conforme pesquisa bibliográfica, foram escolhidos para análise três processos de mineração de dados, KDD, SEMMA e CRISP-DM , porque são considerados os mais populares e realmente são utilizados na prática [39].

2.6.1 KDD

O processo *Knowledge-Discovery in Databases* - KDD é o processo de usar métodos de mineração de dados para extrair conhecimento de forma específica usando um banco de dados em conjunto com processos de manipulação destes dados. O KDD tem cinco etapas [39, 40]:

1. **Seleção** - determinação dos dados alvo da pesquisa;
2. **Pré-processamento** - limpeza e pré-processamento de dados;
3. **Transformação** - transformação dos dados reduzindo suas dimensões ou transformando-os;
4. **Mineração de dados** - busca por padrões de interesse em uma forma de visualização;
5. **Interpretação / Avaliação** - interpretação e avaliação dos padrões extraídos.

O processo KDD é interativo e repetitivo, envolvendo inúmeros passos com muitas decisões, sendo precedido pela compreensão do aplicativo, do conhecimento prévio relevante e dos objetivos do usuário final. Deve haver a consolidação do conhecimento, incorporando-o no sistema [39, 41].

2.6.2 SEMMA

O acrônimo SEMMA significa Sortear/Amostrar (*Sample*), Explorar (*Explore*), Modificar (*Modify*), Modelar (*Model*), Avaliar (*Assess*) e representa um processo de mineração de dados. É um ciclo de 5 etapas [39, 42]:

1. **Amostra** - amostragem dos dados de forma a conter a informação significativa facilmente manipulável;

2. **Explorar** - procurar por tendências e anomalias imprevistas;
3. **Modificar** - criar, selecionar e transformar variáveis;
4. **Modelo** - modelar os dados com processo automatizado nativo do programa;
5. **Avaliação** - avaliar os dados, observando a utilidade e a confiabilidade dos resultados e estimando o desempenho.

O processo SEMMA oferece um método fácil de entender, permitindo um desenvolvimento e manutenção organizados e adequados de projetos de mineração de dados numa estrutura clara para sua concepção, criação e evolução apresentando soluções para problemas [39, 42].

2.6.3 CRISP-DM

Conforme Azevedo [39], O *CRoss Industry Standard Process for Data Mining* - CRISP-DM consiste em um ciclo de seis etapas:

1. **Compreensão do negócio** - compreensão dos objetivos e requisitos do projeto na perspectiva do negócio, definição do problema e de um plano preliminar para atingir os objetivos;
2. **Compreensão de dados** - coleta de dados e seu entendimento, identificar problemas, buscar *insights* sobre os dados e/ou subconjuntos relevantes para revelar informações ocultas;
3. **Preparação de dados** - atividades para construir o conjunto de dados final;
4. **Modelagem** - várias técnicas de modelagem são selecionadas e aplicadas e seus parâmetros são calibrados;
5. **Avaliação** - o modelo obtido é avaliado detalhadamente e revisado para que atinja os objetivos comerciais;
6. **Implantação** - geralmente não é o fim do projeto. Mesmo que o objetivo do modelo seja aumentar o conhecimento dos dados, o conhecimento adquirido precisará ser organizado e apresentado de forma útil para o usuário/cliente.

O CRISP-DM é um padrão amplamente documentado e completo de forma a facilitar o seu entendimento e revisão [43].

Tabela 2.1: Quadro resumo das correspondências entre KDD, SEMMA e CRISP-DM.

KDD	**SEMMA**	**CRISP-DM**
Pré KDD	–	Compreensão do negócio
Seleção	Amostra	Compreensão dos dados
Pré-processamento	Explorar	–
Transformação	Modificar	Preparação dos dados
mineração de dados	Modelo	Modelagem
Interpretação / Avaliação	Avaliação	Avaliação
Pós KDD	–	Implantação

2.6.4 Análise comparativa dos processos de mineração de dados

A Tabela 2.1 apresenta a comparação entre as etapas de cada processo de mineração de dados analisado, com correspondência de suas etapas ordenadas.

Ao se comparar KDD e SEMMA, percebe-se que são equivalentes: a etapa "Amostra" do SEMMA pode ser identificada com a etapa "Seleção" do KDD; "Explorar", SEMMA, com "Pré processamento", KDD; "Modificar" com "Transformação", SEMMA e KDD respectivamente; O "Modelo" do SEMMA com "mineração de dados" do KDD; "Avaliar" com "Interpretação / Avaliação" SEMMA e KDD respectivamente. Examinando-as detalhadamente, percebe-se que as cinco etapas do SEMMA são como uma implementação das cinco etapas do KDD.

Por outro lado, comparar os estágios do KDD com os estágios do CRISP-DM não é tão direto quanto na situação de comparação com o SEMMA. No entanto, observa-se que a metodologia CRISP-DM incorpora as etapas que devem preceder e seguir o processo KDD, ou seja: a fase "Compreensão do negócio" pode ser identificada com o "pré KDD"; A fase de "Implantação" com a consolidação do conhecimento no sistema. Quanto aos estágios restantes: A fase de "Compreensão de Dados" do CRISP-DM pode ser identificada como a combinação de "Seleção" e "Pré-processamento" do KDD; A "Preparação de dados" do CRISP-DM pode ser identificada com "Transformação" do KDD; A fase de "Modelagem", CRISP-DM, com "mineração de dados", KDD; A fase de "Avaliação" do CRISP-DM com "Interpretação / Avaliação" do KDD.

Concluiu-se que SEMMA e CRISP-DM podem ser vistos como uma implementação do processo KDD. Aparentemente o CRISP-DM é mais completo do que o SEMMA. No entanto, podemos integrar a compreensão da aplicação, do conhecimento prévio relevante e das necessidades do usuário final, na etapa de amostragem da SEMMA, porque os dados não podem ser amostrados se não houver um conhecimento real dos dados. Com relação à Implantação, incorporação do conhecimento no sistema, está presente no KDD, porque é o real motivo para fazê-lo. Assim os padrões foram alcançados, em relação ao processo geral: SEMMA e CRISP-DM, assim como o KDD, orientam as pessoas a saber como a

mineração de dados pode ser aplicada na prática em sistemas reais.

No contexto da literatura estudada a estrutura de projeto *CRoss Industry Standard Process for Data Mining* - CRISP-DM se mostrou a mais robusta em termos de uso e facilidade de entendimento. O KDD se mostrou uma boa metodologia, mas menos completa que o CRISP-DM, já o SEMMA, apesar de sua clareza e objetividade está restrito ao programa SAS [39, 42]. A grande permeabilidade do CRISP-DM no mercado e no meio acadêmico torna a sua utilização mais experimentada e documentada [39, 43]. O padrão KDD mostra ter como corpo a metodologia CRISP-DM acrescido de um detalhamento maior nas fases de trabalho, com destaque no entendimento e definição do escopo da mineração de dados e na implantação da aplicação final [39, 40, 41].

Assim, de acordo com o propósito deste trabalho, conforme as constatações supracitadas, selecionamos o processo de mineração de dados CRISP-DM como mais adequado para este estudo.

2.7 Modelos Preditivos

A Inferência Estatística é um instrumento bastante utilizado para, com base em padrões e informações de uma base de dados, alargar os resultados aferidos para a população de origem dos dados [44]. Predizer o futuro fundamentado nos eventos do passado é a principal propriedade chamativa para o uso de modelos preditivos na solução dos mais diversos problemas. Para gerar modelos preditivos é possível usar tanto algoritmos de classificação quanto de regressão.

As técnicas estatísticas mais utilizadas para a análise de risco de inadimplência (*credit scoring*), segundo Abdou [45], são: Regressão Linear, Análise Discriminante, Análise de Probit, Regressão Logística, Árvores de Decisão, Sistemas Expert, Redes Neurais Artificiais e Algoritmos Genéticos. Após estudo breve sobre as características de cada um e do problema proposto, bem como pela dimensão do assunto, no presente trabalho serão abordados as técnicas de modelagem Regressão Logística e Redes Neurais Artificiais. As seções seguintes trazem os fundamentos de cada algoritmo.

2.7.1 Regressão Logística

As regressões logísticas são usadas para descobrir relações entre características observadas (variáveis independentes) e um resultado analisado (variável dependente). Muitos problemas podem ser modelados por regressões logísticas, a ser utilizadas para criar modelos preditivos a partir de bases de dados com conhecimento prévio dos resultados.

Segundo Hosmer Jr. et al. [46], a Regressão Logística é o método estatístico mais utilizado quando a variável independente é discreta, pois, é um tipo de regressão esta-

tística onde a variável dependente é dicotômica (aceita apenas valores “0” ou “1”). A função Logística tem uma forma de “S” conseguindo apresentar uma boa sensibilidade na predição. O uso de tal ferramenta facilita a modelagem de problemas relativos à criação de modelos preditivos de classificação, diferente de outros tipos de regressão que resultam em valores quantitativos. Mesmo sendo uma regressão em termos estatísticos, na mineração de dados este algoritmo é muitas vezes indicado como sendo de classificação, quando se busca delinear os relacionamentos entre variáveis independentes, discreta ou contínua, com uma variável dependente discreta. Na Figura 2.12 é mostrada a diferença gráfica da regressão logística com a regressão linear simples.

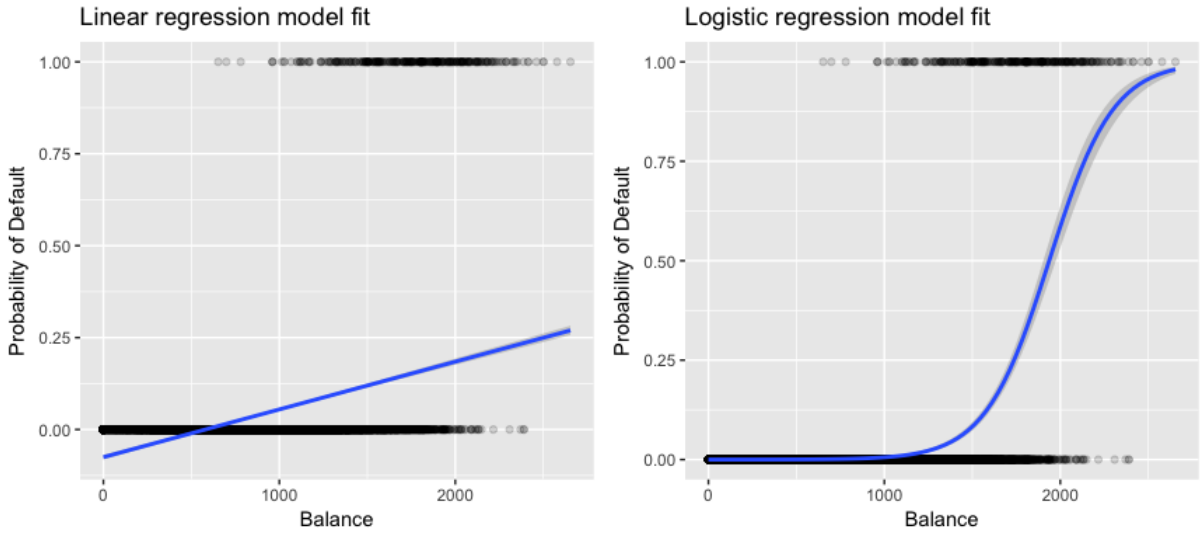

Figura 2.12: Regressão Linear e Regressão Logística (Fonte: `bit.ly/2YloKpk`).

A função logística ou curva logística tem uma forma em “S” ou curva sigmóide, com a equação:

$$f(x) = \frac{L}{1 + e^{-k(x-x_0)}} \tag{2.1}$$

Onde “e” é a base do logaritmo natural (ou número de Euler), x_0 é o valor x do ponto médio do sigmóide, L, o valor máximo da curva e “k”, a inclinação da curva.

Com alguma manipulação matemática, o inverso da função da curva sigmóide se assemelha com a expressão da probabilidade de um evento ocorrer, sobre a probabilidade de não ocorrer: $P(X)/(1 - P(X))$, cuja aplicação da função logarítmica gera a equação da expressão da função logística, ou *logit*.

$$logitP(X) = log(\frac{P(X)}{1 - P(X)}) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + ... + \beta_n X_n \tag{2.2}$$

Assim, a Regressão Logística, no que tange a criação de modelos preditivos, pode ser usada para identificar e analisar as influências das variáveis explicativas sobre a variável dependente. A partir dos parâmetros de avaliação conhecidos pode-se aplicar a fórmula para predizer o valor de $P(X)$, ou Y, a partir dos valores de $X_1, X_2...X_n$.

Segundo Chan et al [47], os parâmetros do modelo de regressão LOGIT podem ser interpretados como a relação da probabilidade de ocorrência e da não-ocorrência daquela característica, usando a função *odds ratio - OR*, ou ainda chamada de "razão de chances". Este conceito é fundamental para determinar o modelo LOGIT que consiste no logaritmo do odds (chance), conforme mostrado na Fórmula 2.2 acima.

2.7.2 Redes Neurais Artificiais

Redes neurais artificiais são técnicas matemáticas inspiradas nas operações do cérebro humano atuando na resolução de problemas. Redes neurais também podem ser caracterizadas por uma aplicação de resolução de problemas de inteligência artificial que aprende através de um processo de treinamento de tentativa e erro. Portanto, a construção de redes neurais requer um processo de treinamento, e as variáveis lineares ou não lineares no procedimento de treinamento ajudam a distinguir variáveis para um melhor resultado de tomada de decisão. Na área de *credit scoring*, redes neurais podem ser distinguidas de outras técnicas estatísticas. Utilizando redes neurais, se os resultados forem inaceitáveis, os valores estimados dos parâmetros serão alterados pelas redes até se tornarem aceitáveis ou até atingirem o valor ideal de cada parâmetro [45].

Recentemente, as redes neurais surgiram como uma tecnologia prática, com aplicações bem-sucedidas em muitos campos das instituições financeiras em geral e dos bancos em particular. Segundo Abdou [45], aplicações como fraude de cartão de crédito, previsão de insolvência de um banco, previsão de falência, aplicação de hipoteca, precificação de opções e outros usos, foram sugeridos como redes neurais podem ser usadas com sucesso na área financeira. Eles abordam muitos problemas, como o reconhecimento de padrões, e fazem uso da arquitetura de redes de *feed-forward*, tais como as redes de alimentação multicamada e redes neurais probabilísticas, representando a maioria dessas aplicações.

Geralmente, **os modelos com redes neurais têm a maior taxa média de classificação correta** quando comparados com outras técnicas tradicionais, [35, 36, 37], embora existam trabalhos recomendando a regressão logística como mais acurada na predição [48]. Assim, este estudo optou pela rede neural perceptron multicamada como algoritmo de predição, mas utilizará a regressão LOGIT como auxiliar na interpretação da significância das variáveis (conforme especificado nos objetivos específicos deste trabalho).

Segundo Pandey et al. [2], uma rede neural perceptron multicamada - MLP (*Multilayer Perceptron*) é uma rede neural artificial usada para classificação, reconhecimento

de padrões e previsão. O MLP consiste de uma camada de entrada, camadas ocultas e camada de saída, o número de mudanças na camada oculta depende da complexidade dos dados. A camada de entrada recebe os dados de entrada que depois de multiplicados pelo peso são encaminhados para camada oculta. Na camada oculta é usada a função de ativação não linear que converte a forma não linear de um problema em sua forma linear de modo que seja facilmente separável, diferentes funções de ativação são adquiridas para diferentes redes para melhor desempenho. A função de ativação mais comum para o MLP são a tangente sigmóide e a hiperbólica. Todos os nós na camada oculta usam a mesma ativação. O MLP usa uma técnica de aprendizado supervisionado em que a saída desejada é conhecida pela rede. Uma arquitetura típica de MLP é mostrada na Figura 2.13. A saída da rede MLP, "y" (dados de saída), é calculada como

$$y = f_3(\sum_{j=1}^{N} w_{j3} h_j - T) \tag{2.3}$$

Onde, N é o número de neurônios na camada oculta, w_{j3} é o peso entre a camada oculta e o neurônio de saída, h_j é a saída do neurônio j, T é o limiar do neurônio de saída e f_3 é a função de ativação sigmóide do neurônio de saída.

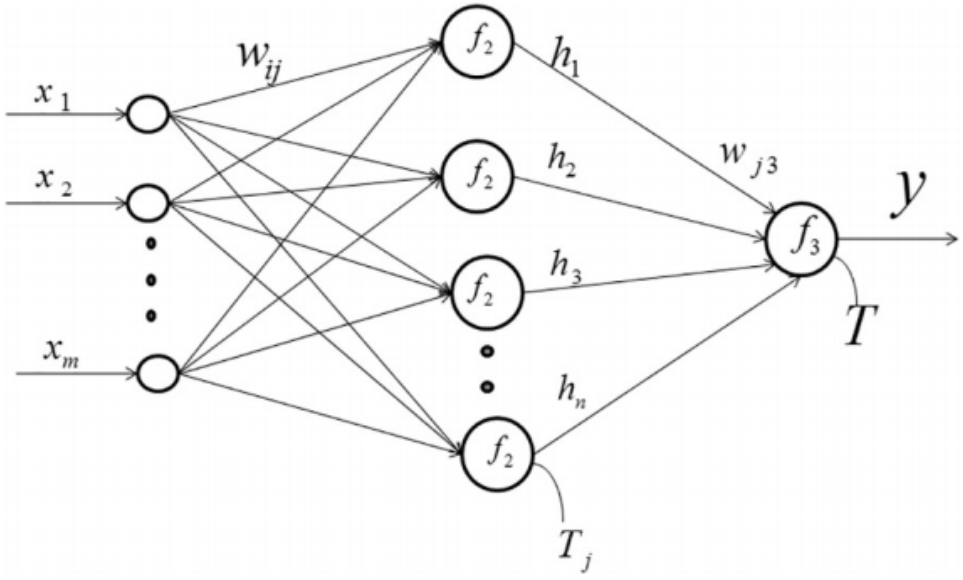

Figura 2.13: A arquitetura genérica da rede neural (Fonte: [2]).

Para as análises desenvolvidas neste estudo foram utilizados os algoritmos disponíveis na plataforma H2O aplicadas na linguagem R através do programa R Studio. Esta escolha se deve a gratuidade da aplicação, sua facilidade e disponibilidade de uso.

Capítulo 3

Metodologia

Para a classificação metodológica do presente trabalho serão apontados os preceitos propostos por Gil [49] e Prodanov [50] no contexto de pesquisas científicas e tecnológicas, especialmente em relação a estruturação da pesquisa, onde será usado um modelo padrão para trabalhos de mineração de dados aplicável a diferentes áreas de conhecimento, conforme foi apresentado seu conceito neste estudo na Revisão Teórica, Mineração de dados, CRISP-DM.

Segundo Gil [49], método é uma forma de pensar para se chegar à natureza de um determinado problema, quer seja para estudá-lo ou explicá-lo. Compreender o método usado é muito importante para utilizar padrões de geração de conhecimento acadêmico e tecnológico nas pesquisas científicas.

Qualquer pesquisa verdadeiramente científica ou tecnológica pode ser classificada em diversas formas em diferentes dimensões. As classificações podem ser feitas observando a natureza da pesquisa, a abordagem, os objetivos do estudo, a estratégia utilizada e as técnicas de coleta de informações que subsidiarão o estudo [49, 50]. Diante dessa perspectiva o presente estudo se enquadraria da seguinte forma:

- **Quanto a Natureza**: Observando o caráter claramente tecnológico, objetivando a melhoria de um processo de trabalho existente, a pesquisa é classificada como de natureza aplicada. Gil [49] afirma que pesquisas "puras" e "aplicadas" não são excludentes, ou seja, não podem ser tratadas como se fossem inteiramente diferentes. Apesar dessa conclusão, essa classificação ainda é muito usada na literatura.
- **Quanto a Forma de Abordagem**: Este estudo terá uma abordagem predominantemente quantitativa. Pois, o uso de técnicas de mineração de dados, baseados em métodos estatísticos, torna a pesquisa substanciada em interpretações numéricas dos fenômenos descritos e verificados.

- **Quanto aos Objetivos**: Apesar do presente trabalho demonstrar certo conhecimento descritivo, especialmente nas partes que tratam do conceito e contextualização do COSO-ERM e da Análise de Risco de Crédito, este estudo traz em seu Objetivo principal um caráter explicativo, afinal se enquadra em trabalhos cujas variáveis influenciam em determinado resultado.
- **Quanto à Estratégia**: Este trabalho se enquadra como uma pesquisa *ex-post facto*, pois esta classificação se refere às pesquisas que se caracterizam por avaliar a influência de um conjunto de variáveis independentes com outra dependente a partir de fatos já ocorridos.

Em suma, tem-se que, o presente estudo apresenta sua natureza **aplicada**, uma abordagem **predominantemente quantitativa**, com seu objetivo principal **explicativo** e uma estratégia do tipo *ex-post facto*.

Universo de pesquisa - A coleta de dados realizada teve um caráter censitário, ou seja, não foram feitas amostragens, os dados foram colhidos na sua totalidade. A base de dados deste estudo englobou:

1. Todo o cadastro fiscal do Distrito Federal, 305.685 empresas inscritas;
2. Toda a base de dados do Livro Fiscal Eletrônico declaradas entre os anos de 2012 a 2017;
3. Todo o banco de dados de registros financeiros da SEFP, o qual se refere a todos os recolhimentos dos contribuintes do DF no período de 2012 a 2017;
4. Todo o registro de inscrição em Dívida Ativa do DF, 79.548 Certidões de Dívida Ativa relativas ao período de 2012 a 2017; e
5. Toda a base de autos de infração lavrados pela fiscalização tributária entre os anos de 2012 a outubro de 2018, sendo observadas as infrações cometidas até dezembro de 2017.

Limitações do escopo da pesquisa - Este trabalho se limitou aos dados disponíveis nos bancos de dados da Secretaria de Estado de Fazenda, Planejamento e Gestão do DF (SEFP), não foram utilizados dados de outras fontes, por exemplo: dados de emprego, dados previdenciários ou tributários referentes à competência da União (tributos federais).

Não foram avaliados modelos para análise de risco de crédito que utilizem mais de um algoritmo operando em conjunto, ou seja, atuando na modelagem, em paralelo ou em série, no tarefa de predição. Neste estudo os modelos de redes neurais e a regressão LOGIT são abservados simultaneamente apenas na tarefa de identificação dos fatores de risco.

Também não foi analisada a possibilidade de interação entre a proposta deste estudo com os vários sistemas em uso da fiscalização tributária do DF ou em qual sistema seria viável a implantação dos modelos estudados, o estudo limitou-se aos modelos em si.

Os dados referentes aos documentos fiscais eletrônicos e escrituração fiscal eletrônica bem como todos os demais bancos de dados oficiais da Subsecretaria de Receita - SUREC foram disponibilizados para este estudo devido à sua intenção primordial que é melhorar a gestão de projetos de fiscalização. Foram utilizadas diversas ferramentas eletrônicas para consulta e manipulação dos dados, tais como Oracle (BD), QlikView (BI), R Studio e outros softwares próprios da SEFP. Logo, o acesso direto aos dados foi autorizado e facilitado.

Conforme exposto, o modelo será construído dentro da técnica de mineração de dados CRISP-DM, já detalhado na seção "Mineração de dados", a qual se apresenta a seguir.

3.1 Compreensão do negócio

Foi tratada na seção Revisão Teórica a "Gestão do Risco em Projetos de Auditoria Tributária", a contextualização da auditoria fiscal na administração tributária do Distrito Federal. Na seção "Modelos preditivos", foram apresentadas as características dos algoritmos de predição estudados utilizados na análise de risco de crédito. Por sua vez, na seção "Modelos e ferramentas de avaliação de risco", foi exposta a ideia deste estudo sendo o uso dos modelos de predição de risco de crédito na predição de inadimplência tributária.

A prospecção de indícios se apresenta como uma forma de buscar pistas de algum tipo de irregularidade que gere pagamento a menor do imposto devido, esta é distinta da identificação de irregularidades, que aponta exatamente a falha, fraude ou omissão por parte do contribuinte que gerou a evasão de tributos. Como exemplo de situações de identificação de irregularidades tem-se a não entrega de declarações, o não pagamento deliberado de imposto já lançado, a não escrituração de documentos fiscais ou ainda a mera não emissão de notas fiscais.

A prospecção de indícios também busca a descoberta de irregularidades inusitadas ou singulares que ainda não tenham sido previstas e documentadas. A forma mais utilizada na administração tributária do DF é a tentativa de inferir o faturamento de determinados contribuintes, em seguida estimar o montante de recolhimentos esperados para aqueles indivíduos e compara-se com o total efetivamente recolhido, onde houver divergência significativa são apartados para inspeção mais acurada, para então se identificar uma possível irregularidade que justifique alguma ação fiscal.

Mas e nos casos de omissões de receita que impossibilitem uma boa inferência de faturamento? Ou ainda, e nos casos de manobras contábeis inovadoras com o intuito

de sonegação? Os métodos tradicionais não seriam efetivos. Portanto o uso de modelos preditivos de risco de inadimplência surgem como uma alternativa, ou ainda complemento, aos métodos utilizados atualmente de forma a possibilitar a identificação de casos de evasão fiscal não reconhecidos nas rotinas adotadas no presente.

3.2 Compreensão dos dados

Para alimentar os modelos, estão dispostas as informações dos contribuintes objeto de estudo que demonstram suas características próprias individuais bem como suas características econômico-fiscais. Numa comparação com os modelos de risco de crédito utilizados por prestadores de serviço financeiro, as características próprias individuais de um ente solicitante de empréstimo seriam as ligadas à sua pessoa, como idade, sexo ou profissão e suas características econômico-fiscais seriam sua renda, seus gastos ou seu histórico de comportamento como mutuário [34]. Paralelamente para um ente contribuinte fiscal as características próprias individuais seriam as ligadas à sua personalidade jurídica, como Atividade Econômica, tipo de empresa ou regime fiscal e as características econômico-fiscais seriam seu faturamento, seus recolhimentos ou seu histórico fiscal.

As características próprias individuais das entidades "contribuintes" neste estudo foram extraídas do Cadastro Fiscal do Distrito Federal - CFDF. São qualidades intrínsecas à personalidade jurídica da empresa e de sua constituição legal, registradas no momento no qual a empresa é criada, atualizados a cada alteração cadastral ao longo da existência da empresa. Os atributos exclusivos, como o número de identificação cadastral, o número do CNPJ, endereço e os nomes dos sócios foram excluídos para manter o sigilo fiscal, o que não trará prejuízo ao presente estudo.

Foram utilizados nos modelos as características genéricas que se repetem ao longo do cadastro em diversas empresa onde as mesmas podem ser agrupadas em conjuntos diferentes conforme o atributo sob análise. Assim, foram usadas a seguintes características: tipo do contribuinte, tempo de atividade, atividade econômica do ICMS, atividade econômica do ISS, forma de cálculo do ICMS e/ou ISS, quantidade de sócios da empresa e, por fim se algum dos sócios é pessoa jurídica ou não.

Variáveis de estudo:

- TC. Tipo do contribuinte - é a especificação da forma societária da empresa bem como de sua natureza jurídica. Ex.: Sociedade Empresária Limitada, Sociedade Anônima, Sociedade de Economia Mista, Cooperativa, Associação Privada, etc.
- TA. Tempo de atividade - é o tempo de funcionamento da empresa desde sua inscrição no cadastro fiscal do DF.

- AEICMS. Atividade econômica do ICMS - é a indicação da atividade econômica realizada pela empresa relacionada ao ICMS de acordo com a Classificação Nacional de Atividades Econômicas [1] adotada pelo Sistema Estatístico Nacional do Brasil e pelos órgãos públicos e demais instituições no Brasil. É indicado por o código alfanumérico de dez caracteres. A seguir, alguns exemplos do código CNAE e seu respectivo significado: "D351230000 - Transmissão de energia elétrica"; "D351310000 - Comércio atacadista de energia elétrica"; "G465240000 - Comércio atacadista de componentes eletrônicos e equipamentos de telefonia e comunicação". Neste estudo foi considerada a família de cada tipo de atividade representada pela primeira letra do código CNAE conforme mostrado a seguir.
  - A. agricultura, pecuária, produção florestal, pesca e aqüicultura
  - B. indústrias extrativas
  - C. indústrias de transformação
  - D. eletricidade e gás
  - E. água, esgoto, atividades de gestão de resíduos e descontaminação
  - F. construção
  - G. comércio; reparação de veículos automotores e motocicletas
  - H. transporte, armazenagem e correio
  - I. alojamento e alimentação
  - J. informação e comunicação
  - K. atividades financeiras, de seguros e serviços relacionados
  - L. atividades imobiliárias
  - M. atividades profissionais, científicas e técnicas
  - N. atividades administrativas e serviços complementares
  - O. administração pública, defesa e seguridade social
  - P. educação
  - Q. saúde humana e serviços sociais
  - R. artes, cultura, esporte e recreação
  - S. outras atividades de serviços
  - T. serviços domésticos

[1] Fonte para consulta pública no sítio da RFB: `http://receita.economia.gov.br/orientacao/tributaria/cadastros/cadastro-nacional-de-pessoas-juridicas-cnpj/classificacao-nacional-de-atividades-economicas-2013-cnae`

  - U. organismos internacionais e outras instituições extraterritoriais
  - Z. outras atividades

- AISS. Atividade econômica do ISS - Mesmo que a Atividade econômica do ICMS, só que aplicada ao ISS. A seguir, alguns exemplos do código CNAE e seu respectivo significado: "J611080100 - Serviços de telefonia fixa comutada - STFC"; "J612050100 - Telefonia móvel celular"; "J612059900 - Serviços de telecomunicações sem fio não especificados anteriormente". Da mesma forma tratada para o ICMS, a atividade do ISS será representada pela família a qual pertence conforme a primeira letra do código CNAE.
- CALC. Forma de cálculo do ICMS e/ou ISS - é a descrição da maneira pela qual o imposto será aplicado sobre a atividade econômica da empresa. Ex.: Regime normal, Simples Nacional, Microempreendedor Individual - SIMEI, Produtor Rural, Sociedade uniprofissional, Autônomo, etc.
- SOCI. Quantidade de sócios da empresa - é o número de sócios registrados no contrato social da empresa e indicado no CFDF.
- SOPJ. Possui sócio pessoa jurídica - é a indicação de que pelo menos um dos sócios da empresa é um pessoa jurídica.

As características sócio-econômicas das entidades "contribuintes" neste estudo foram extraídas de suas declarações fiscais, de seus registros financeiros de recolhimentos e de seu histórico fiscal. Todas estas informações foram disponibilizadas para o presente trabalho nos bancos de dados da Secretaria de Estado de Fazenda, Planejamento e Gestão do DF (SEFP). As declarações fiscais apontam qualidades relacionadas ao faturamento da empresa, consequentemente ligadas às suas receitas, para tal, foram observados os Livros Fiscais eletrônicos - LFE entregues. Os seus registros financeiros de recolhimentos foram observados como parte dos custos da empresa. O comportamento fiscal foi verificado através do histórico de ações fiscais sofridas pelo contribuinte que resultaram em cobrança de imposto (crédito tributário constituído).

No modelo foram utilizados as somas dos valores totais anuais verificados individualmente pelas empresas no que tange aos valore declarados e aos valores de recolhimento, já para o histórico fiscal foi observado se a empresa foi autuada ou não. Neste caso, "autuada" significa que a empresa sofreu alguma ação fiscal no passado e a mesma resultou em algum tipo de cobrança de imposto devido e não recolhido. Assim, foram usadas a seguintes características: Valor total de saídas sobre o valor total de entradas declaradas no Livro Fiscal Eletrônico, valor total de imposto recolhido e por fim se foi autuado ou não.

- RLFE. Valor total de saídas sobre o valor total de entradas declaradas no Livro Fiscal Eletrônico (LFE) - é a soma dos valores das saídas (vendas e/ou prestações de serviço) dividido pela soma das entradas (compras e/ou serviços tomados) declaradas no Livro Fiscal eletrônico entregue no sistema da Secretaria de Fazenda do DF.
- REC. Valor total de imposto recolhido - é a soma dos valores recolhidos ao tesouro do DF referentes aos impostos ICMS e ISS.
- AUTO. Foi autuado ou não - é a indicação se a empresa já sofreu alguma ação fiscal cujo resultado tenha comprovado alguma irregularidade fiscal e consequentemente cobrado impostos e multas devidas.

Os modelos se propõem a predizer a futura situação de inadimplência ou não de determinados contribuintes. Para representar tal predição, será adotada como variável explicada, o registro de inscrição em Dívida Ativa do Distrito Federal. Este cadastro mostra a relação das empresas inadimplentes em relação às suas obrigações perante a Administração Tributária do DF bem como as características do débito inscrito, ex.: valor, ano de reverência, natureza do débito, etc. É importante destacar que débitos inscritos em Dívida Ativa já foram cobrados administrativamente, com o devido processo administrativo fiscal, e são encaminhados para execução judicial.

- DA (Variável Y). Inclusão em Dívida Ativa. Temos que a variável a ser predita nos modelos, ou variável explicada, será a ocorrência da empresa no cadastro da Dívida Ativa do DF.

Quanto ao aspecto temporal os dados trazem as informações das empresas ativas no DF verificadas no mês de outubro de 2018, assim como a relação de empresas inscritas em Dívida Ativa até esta data, mas foram considerados apenas os débitos que se referem a 2017. No que tange aos dados econômico-financeiros foram observados os registros dos anos de 2012 a 2017.

Assim tendo exposto o conjunto de considerações para promover o "entendimento do negócio" de acordo com a metodologia CRISP-DM, bem como definidas as informações pertinentes a serem usadas nos modelos, a seguir será discutido detalhadamente o tratamento dos dados utilizados nos modelos.

3.3 Preparação e tratamento dos dados

Foram extraídos os dados de 305.685 empresas inscritas no **Cadastro Fiscal** do DF, dentre essas 79.548 tem débito inscrito em Dívida Ativa por alguma cobrança referente

aos anos de 2012 a 2017. É possível que uma mesma empresa tenha débitos distintos inscritos em DA referentes a mais de um ano. A seguir tem-se o relato de como os dados foram **discretizados** para a composição da base dos modelos, conforme descrito por Silva [51].

3.3.1 Variáveis Explicativas Discretas

Todas as variáveis discretas foram tratadas da mesma forma, a intenção deste tratamento é apartar, para cada variável, as instâncias mais significativas das demais. Para tal, foi aplicada a regra 80/20, de forma que as instâncias responsáveis por 80% das ocorrências foram destacadas, sendo as demais agrupadas numa única identificação. O processo seguiu os seguintes passos:

1. Identificação das instâncias observadas e respectivas frequências (número de ocorrências);
2. Ordenação das instâncias pela frequência em ordem decrescente;
3. A cada instância será atribuída sua frequência relativa e acumulada;
4. Atribuir uma identificação para cada instância, sequencialmente na ordem do item anterior, até que a frequência acumulada seja 80% do total;
5. Agrupar as demais instâncias em um mesmo grupo identificado, por exemplo: Outros.

TC. Tipo do contribuinte

As ocorrências mais encontradas foram de "Empresário (individual)" e "Sociedade empresária limitada", quando somadas representam mais de 80% dos registros.

Para a inclusão no modelo, considerando a regra de Pareto (Ver Figura 3.1), foram consideradas as ocorrências com maior relevância, cuja soma seja mais de 80% do total, e as demais foram agrupadas como "Outros" da seguinte forma:

Tipo de contribuinte "Empresário (individual)" = TC1
Tipo de contribuinte "Sociedade Empresária Limitada" = TC2
Demais ocorrências de Tipo de contribuinte "Outros" = TC3

TA. Tempo de atividade

Seguindo o mesmo raciocínio aplicado ao campo Tipo de Contribuintes, indicando as ocorrências cuja as somas sejam superiores a 80% e agrupando o restante em outros

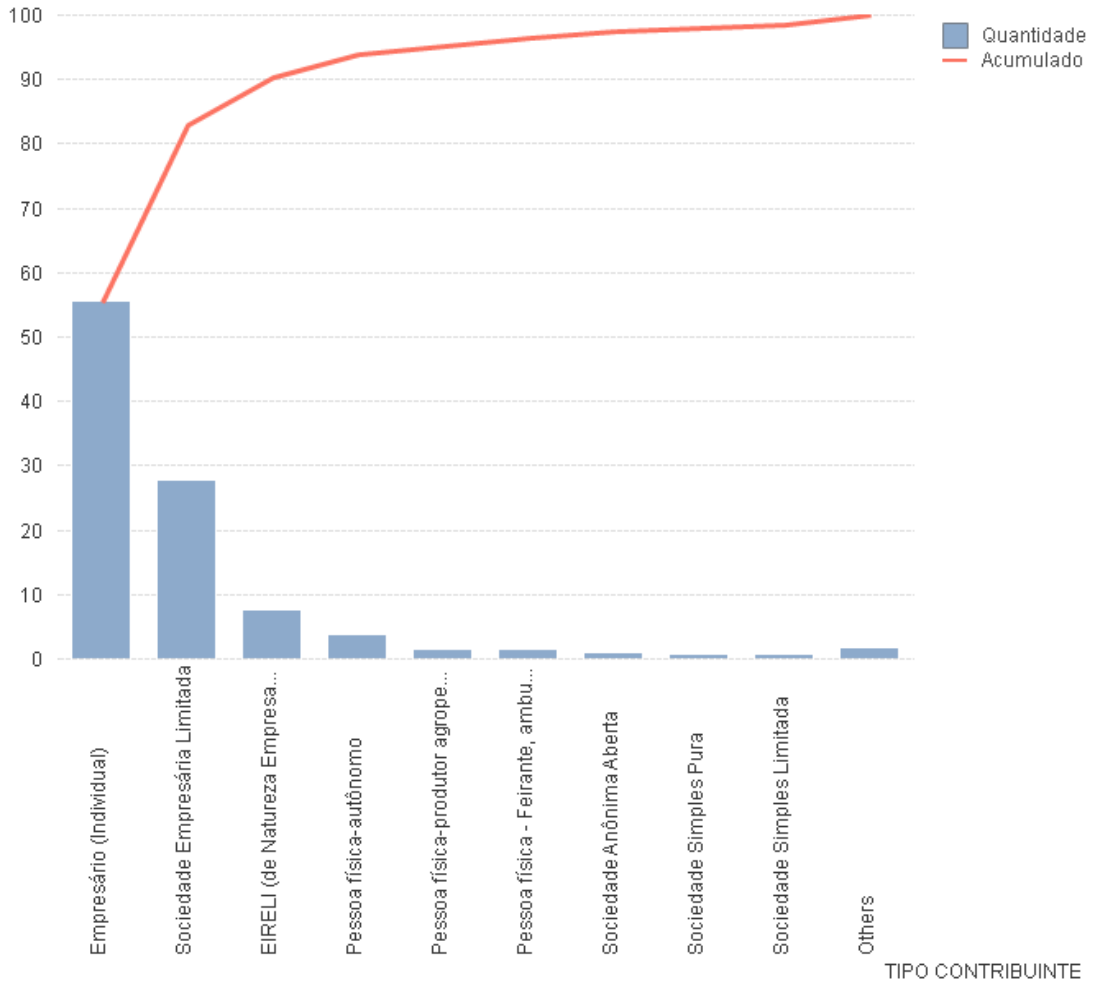

Figura 3.1: Gráfico da regra de Pareto para o campo “Tipo de Contribuintes” (Fonte: Autor).

seriam geradas 11 variáveis (sendo “0” a “9”, mais “10 e demais”), conforme mostrado na Figura 3.2.

Assim, para evitar uma quantidade excessiva de variáveis, foi proposta a seguinte configuração, também mostrada graficamente na Figura 3.3:

Tempo de atividade de 0 a 4 anos = TA1

Tempo de atividade de 5 a 9 anos = TA2

Tempo de atividade 10 anos ou mais = TA3

AEICMS. Atividade econômica do ICMS

Para a inclusão no modelo, considerando a regra de Pareto (Ver Figura 3.4), foram consideradas as ocorrências com maior relevância, cuja soma seja mais de 80% do total, e as demais foram agrupadas como “Outros” da seguinte forma:

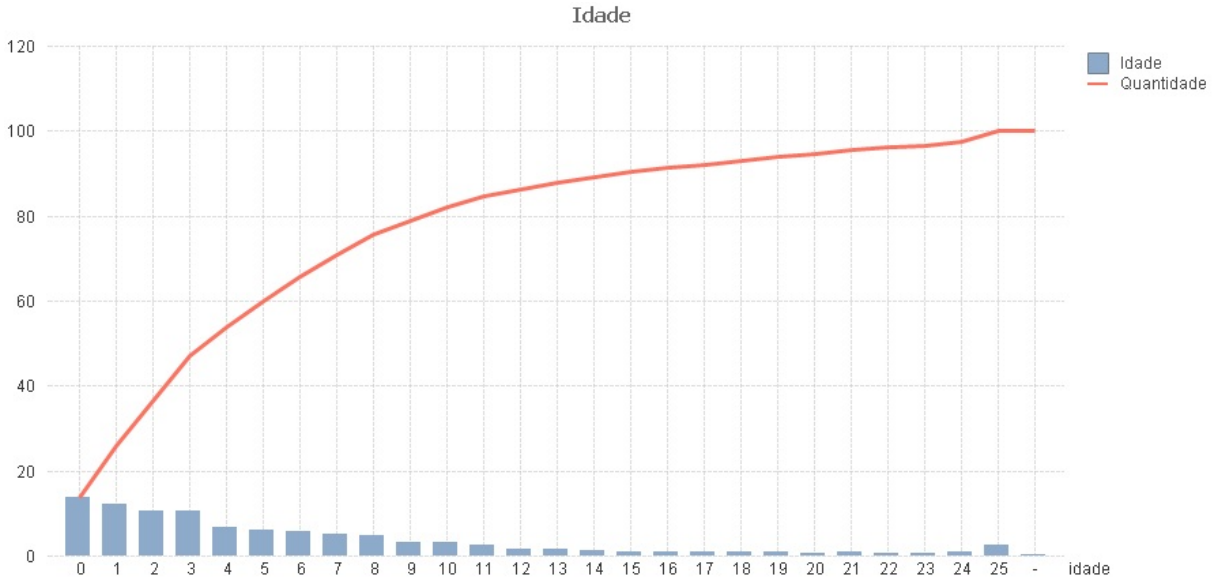

Figura 3.2: Gráfico da Regra de Pareto para o tempo de funcionamento de empresas (Fonte: Autor).

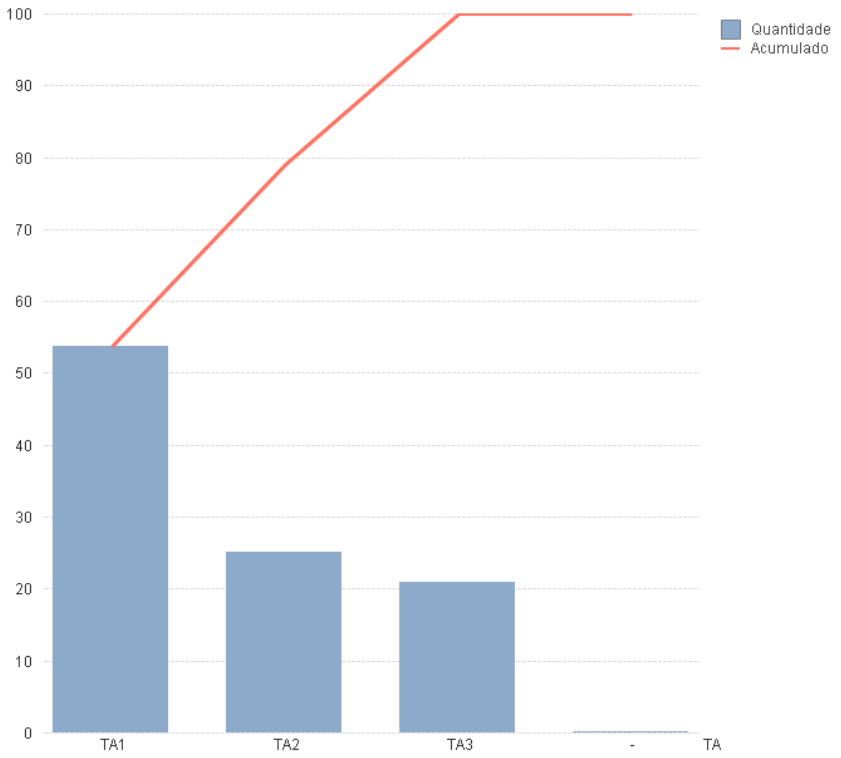

Figura 3.3: Gráfico da Regra de Pareto para a variável Tempo de atividade (Fonte: Autor).

Atividade econômica do ICMS de Comércio (G): AICMS1

Atividade econômica do ICMS de Alojamento e Alimentação (I): AICMS2

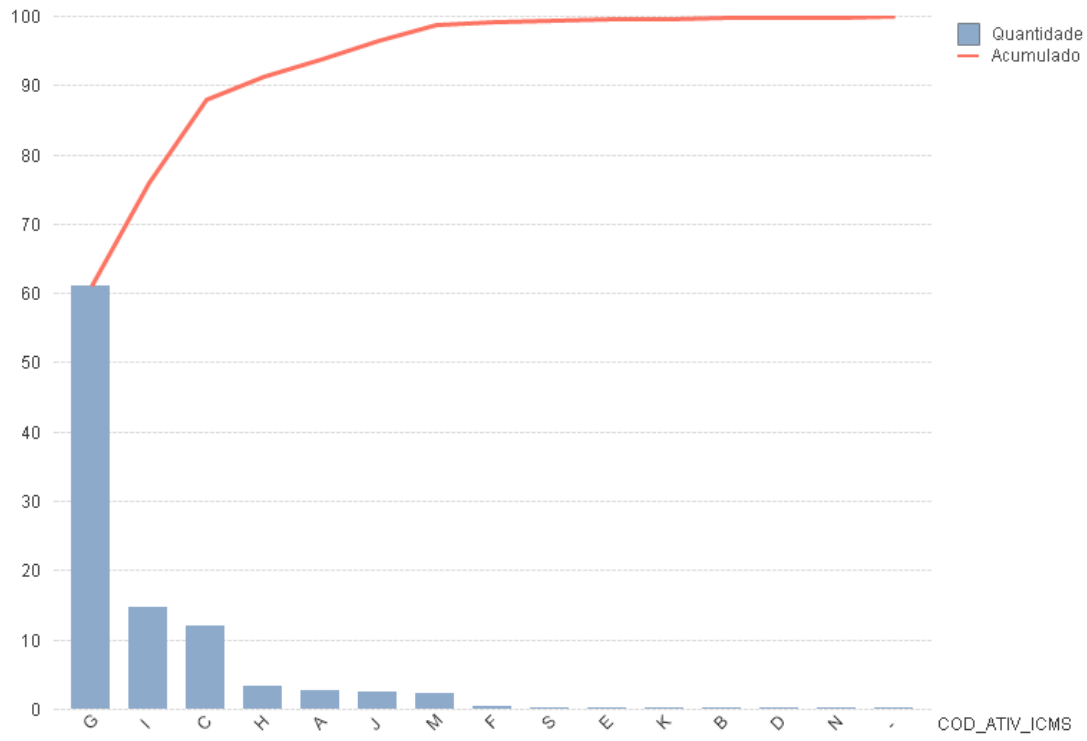

Figura 3.4: Gráfico da Regra de Pareto para a Atividade Econômica do ICMS (Fonte: Autor).

Atividade econômica do ICMS de Indústria (C): AICMS3
Atividade econômica do ICMS demais atividades "Outros": AICMS4

AISS. Atividade econômica do ISS

Da mesma forma que o item anterior, para a inclusão no modelo, considerando a regra de Pareto (Ver Figura 3.5), foram consideradas as ocorrências com maior relevância, cuja soma seja mais de 80% do total, e as demais foram agrupadas como "Outros" da seguinte forma:

Atividade econômica do ISS de Atividades de Serviços (S): AISS1
Atividade econômica do ISS de Construção (F): AISS2
Atividade econômica do ISS de Atividades Administrativas (N): AISS3
Atividade econômica do ISS de Atividades Profissionais (M): AISS4
Atividade econômica do ISS de Comércio (G): AISS5
Atividade econômica do ISS de Indústria (C): AISS6
Atividade econômica do ISS de Transporte (H): AISS7
Atividade econômica do ISS de Autônomos (Z): AISS8
Atividade econômica do ISS de Educação (P): AISS9

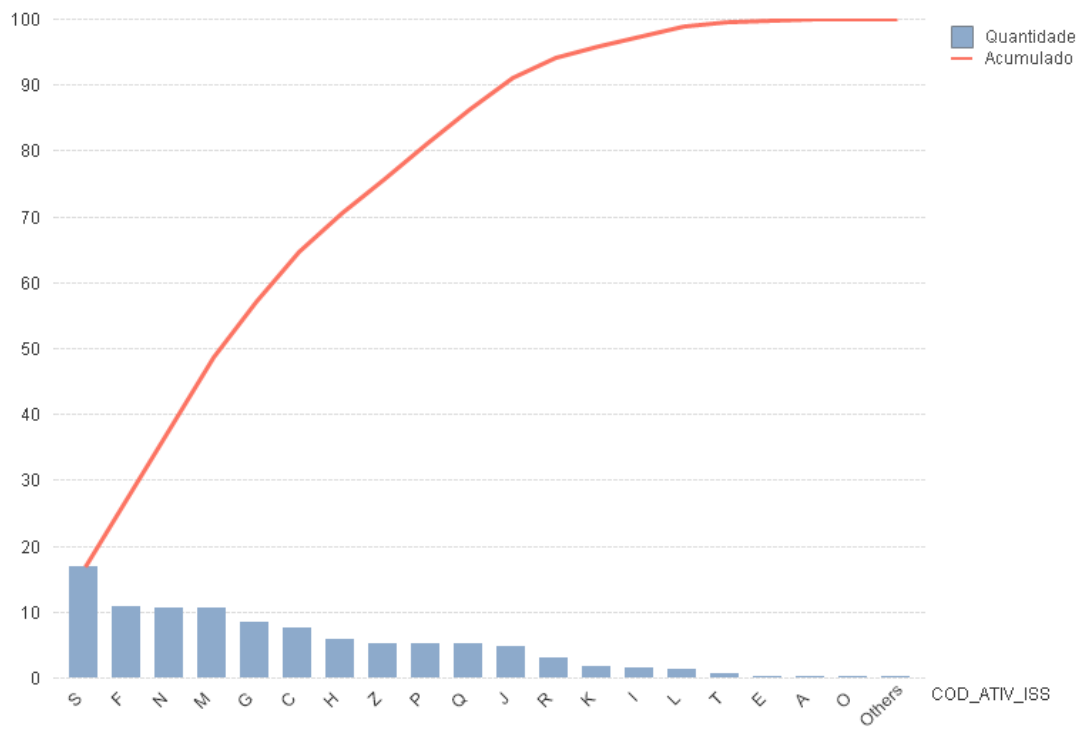

Figura 3.5: Gráfico da Regra de Pareto para a Atividade Econômica do ISS (Fonte: Autor).

Tabela 3.1: Distribuição de ocorrências da Forma de Cálculo do imposto.

ICMS	**contribuintes**	**ISS**	**contribuintes**
SIMEI	89.468	SIMEI	92.494
Simples Nacional	50.737	Simples Nacional	51.219
Normal	41.261	Normal	45.500
Outras	13.851	Outras	12.731
Total	*195.317*	*Total*	*201.944*

Atividade econômica do ISS demais atividades: AISS10

CALC. Forma de cálculo do ICMS e/ou ISS

Uma empresa pode ser contribuinte do ICMS ou do ISS, ou ainda de ambos simultaneamente, conforme as atividades econômicas desenvolvidas. Sempre é mantida a mesma forma de cálculo para os dois impostos nos casos em que a empresa é contribuinte de ambos. Ou seja, se uma empresa está no regime Simples Nacional para o ICMS, necessariamente estará para o ISS também. As formas de cálculo dos impostos mais recorrentes são os regimes Normal, Simples Nacional e Microempreendedor Individual - SIMEI. A distribuição das ocorrências desse campo é mostrada na Tabela 3.1.

Foi proposta a seguinte configuração:

Forma de cálculo do regime SIMEI: CALC1
Forma de cálculo do regime Simples Nacional: CALC2
Forma de cálculo do regime Normal: CALC3
Forma de cálculo de outros regimes: CALC4

SOCI. Quantidade de sócios da empresa

Foi verificados dentre o total de 305.685 registros, 200.367 (65%) tem apenas um sócio no contrato social da empresa. 67.990 (22%) empresas tem 2 sócios e o restante 3 ou mais sócios (13%). Foi proposta a seguinte configuração:

Apenas um sócio: SOCI1
Dois sócios: SOCI2
Três ou mais sócios: SOCI3

SOPJ. Possui sócio pessoa jurídica

Dentre o total de 305.685 registros, 281.372 (92%) constam apenas um sócio no contrato social, 5.185 (1,7%) tem pessoas jurídicas na sua composição societária, já 19.128 não constam registros, nestes casos foi considerado para todos os efeitos deste estudo como não havendo pessoa jurídica como sócio (ver Figura 3.6).

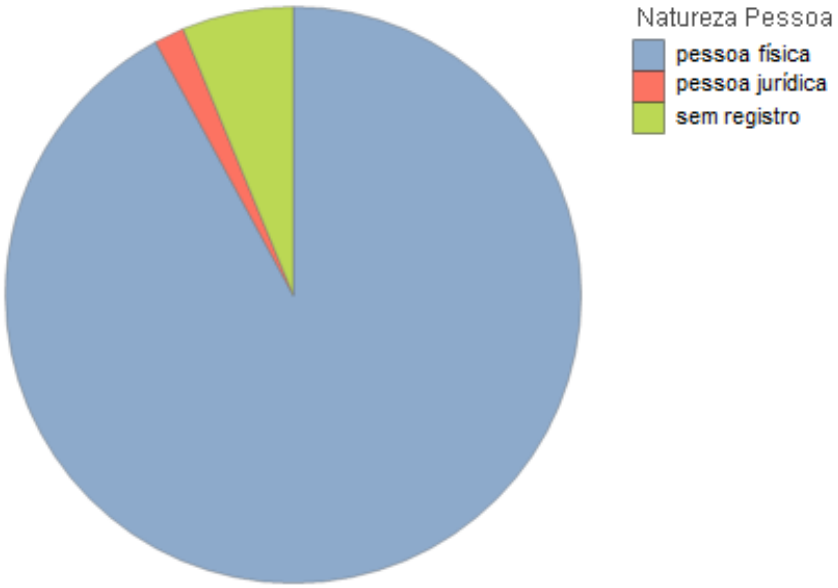

Figura 3.6: Gráfico de Pizza da Natureza Jurídica dos sócios da empresas (Fonte: Autor).

Dessa forma foi proposta a seguinte configuração:
Sem pessoa jurídica: SOPJ1

Tabela 3.2: Estatísticas da série de valores da fração RLFE.

Mínimo	1º Quartil	Mediana	Média	3º Quartil	Máximo
0,0	0,1	0,4	225,4	0,8	2.802.499,9

Com pessoa jurídica: SOPJ2

AUTO - Se a empresa já foi autuado ou não

Foi observado um total de 1.597 autuadas de alguma forma no período observado. Assim foi proposta a seguinte configuração para esta variável:

Para empresas que não foram autuadas no período: AUTON

Para empresas que foram autuadas no período: AUTOS

3.3.2 Variáveis Explicativas Contínuas

As variáveis contínuas apresentadas a seguir, RLFE e REC, são oriundas das bases de dados da SEFP, as quais são estruturadas, tratadas e auditadas regularmente, logo assume-se que são dados consistentes e seguros. Considerando que os eventos de evasão tributária são de natureza dissociativa e de ocorrência rara em relação ao comportamento normal da população, optou-se por manter as ocorrências de menor frequência, *outliers*, porquanto estas são essenciais para descrição do comportamento dos contribuintes [52].

RLFE - Valor total de saídas sobre o valor total de entradas declaradas no Livro Fiscal Eletrônico

Na base de dados extraída, 305.685 registros, foram verificados valores nulos ou ausentes em 261.382 registros, dessa forma as estatísticas dessa distribuição consideraram apenas os valores observados.

Nota-se pelas estatísticas da série que a distribuição está concentrada em valores menores do que 1 (3º quartil = 0,8), sendo o restante distribuído esparsamente até o valor máximo de “2.802.499,9”, onde a média é “225,4” (Tabela 3.2). Pode-se dizer que a série é uma distribuição geométrica semelhante à Distribuição Gama, os valores são concentrados numa região da distribuição de frequência com grande dispersão nos valores mais altos [53].

Assim foi considerado, para o tratamento dessa série, o logaritmo decimal dos valores observados de forma a aproximar a distribuição de frequência da série a forma da distribuição normal [54]. Este recurso matemático é comum em estudos econométricos, pois altera a escala da série de valores tratando as ocorrências dispersas e facilitando a visuali-

Tabela 3.3: Estatísticas da série de valores do *Log10* da fração RLFE.

Mínimo	**1º Quartil**	**Mediana**	**Média**	**3º Quartil**	**Máximo**
-6,27	-0,59	-0,27	-0,35	-0,05	6,45

zação da distribuição de frequência dos dados [55, 56], como pode ser visto na Tabela 3.3 e na Figura 17 a seguir.

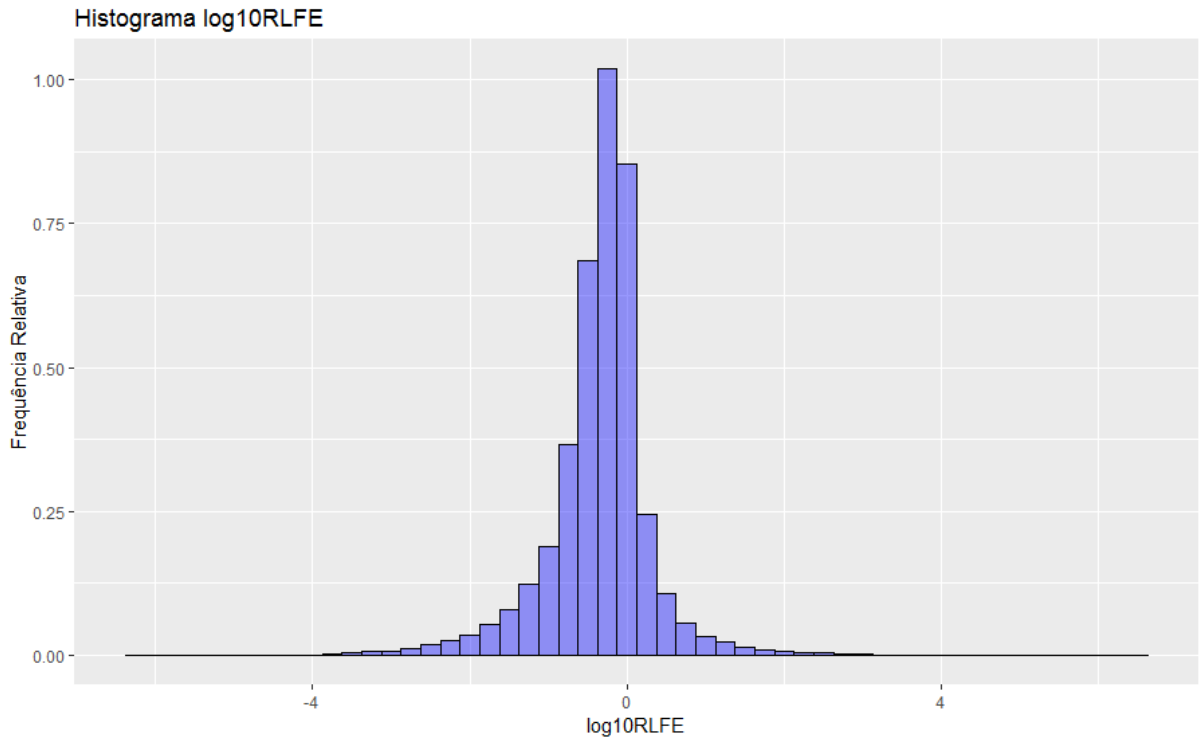

Figura 3.7: Histograma da série *Log10* da fração RLFE (Fonte: Autor).

Dessa forma foi proposta a seguinte configuração para a variável RLFE:
Onde havia valor nulo ou ausente: **RLFE0**
Onde $Log10(RLFE) < (-0,59)$; 1º Quartil: **RLFE1**
Onde $(-0,59) \leq Log10(RLFE) < (-0,27)$; 2º Quartil: **RLFE2**
Onde $(-0,27) \leq Log10(RLFE) < (-0,05)$; 3º Quartil: **RLFE3**
Onde $Log10(RLFE) \geq (-0,05)$; 4º Quartil: **RLFE4**

REC - Valor total de imposto recolhido

Assim como a série de valores da fração RLFE, os valores dos recolhimentos totais, REC, se assemelham a uma Distribuição Gama, valores concentrados numa região da distribuição de frequência com grande dispersão nos valores mais altos, conforme se observa na Tabela 3.4.

Tabela 3.4: Estatísticas da série de valores recolhidos REC.

Mínimo	1º Quartil	Mediana	Média	3º Quartil	Máximo
0,0	0,0	35,01	156.324,57	1.2610,09	5.645.030.723

Tabela 3.5: Estatísticas da série de valores do *Log10* de REC.

Mínimo	1º Quartil	Mediana	Média	3º Quartil	Máximo
-2	0,0	1,54	1,77	3,1	9,75

Dessa forma foi adotada a mesma estratégia anterior, a conversão dos valores para o logaritmo decimal. Na Tabela 3.5 se apresentam as novas estatísticas da série e na Figura 3.8 observa-se o histograma de frequência da série *Log10* de REC.

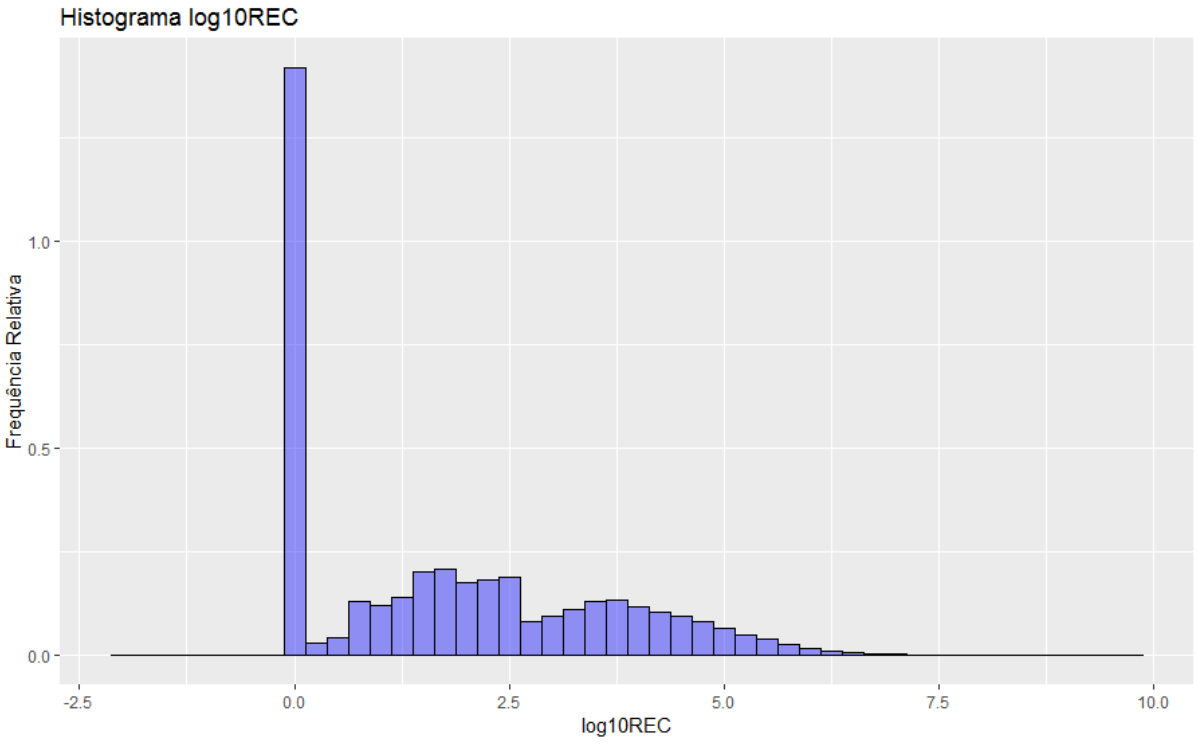

Figura 3.8: Histograma da série *Log10* de REC (Fonte: Autor).

Em atenção ao histograma da série *Log10* de REC, destaca-se a alta incidência de valores iguais a zero, o que na série original retratam as empresas que não efetuaram recolhimento algum no período observado. Também percebem-se alguns valores entre "0" e "-2", que representam as somas de recolhimentos entre um centavo e um real. Estas situações representam coincidentemente o 1º Quartil da distribuição, sendo o restante dos valores distribuídos de forma semelhante à normal nos demais quartis.

Foi proposta a seguinte configuração para a variável REC:

Onde $Log10(REC) \leq 0$; 1º Quartil: **REC1**

Tabela 3.6: Visualização dos dados após o tratamento de discretização.

Registro	**1**	**2**	**3**	**...**	**305.686**
Tipo	TC2	TC1	TC2	...	TC1
TA	TA3	TA1	TA3	...	TA1
AICMS	AICMS1	AICMS4	AICMS4	...	AICMS2
AISS	AISS1	AISS9	AISS2	...	AISS10
CALC	CALC3	CALC1	CALC3	...	CALC1
SOCI	SOCI2	SOCI1	SOCI2	...	SOCI1
SOPJ	SOPJ1	SOPJ1	SOPJ1	...	SOPJ1
RLFE	RLFE0	RLFE0	RLFE0	...	RLFE0
REC	REC1	REC1	REC4	...	REC1
AUTO	AUTON	AUTON	AUTON	...	AUTON
DA	1	0	1	...	0

Onde $0 < Log10(REC) \leq (1,54)$; 2º Quartil: **REC2**
Onde $(1,54) < Log10(REC) \leq (3,1)$; 3º Quartil: **REC3**
Onde $Log10(REC) > (3,1)$; 4º Quartil: **REC4**

DA - Variável explicada, se foi inscrito em Dívida Ativa no período

No total de 305.686 registros observados, ocorreram 79.548 empresas inscritas em dívida ativa em algum momento no período observado, Assim para a variável explicada foi adotada a configuração padrão para algoritmos de classificação, ou seja, o valor "1" para as ocorrências positivas e "0" para as negativas.

- Foi inscrito em Dívida Ativa: "1"
- Não foi inscrito em Dívida Ativa: "0"

Dado o tratamento supra à base de dados extraída, pode ser observado na Tabela 3.6 a visualização dos dados tratados.

3.4 Modelagem

Na construção da base de dados para a modelagem, para ambos os algoritmos, foi feita a fatorização da base de dados, o que transformou as variáveis discretizadas em variáveis numéricas do tipo binária ("0" ou "1"), exceto a variável explicada, "DA", a qual já é binária, conforme mostrado na Tabela 3.7.

Para a criação dos subconjuntos da base de dados que servirão para treinamento e teste dos modelos, é recomendada tradicionalmente, a divisão das instâncias de forma aleatória sendo 70% dos exemplares no conjunto de treinamento e os 30% restantes no

Tabela 3.7: Visualização da base de dados modelada - Variáveis binárias transformadas.

Variável original	**Variável Binária**	**reg. 1**	**reg. 2**	**reg. 3**	**...**	**reg. 305.686**
TC	TC1	1	0	0	...	0
TC	TC2	0	0	1	...	1
TC	TC3	0	1	0	...	0
TA	TA1	0	1	0	...	0
TA	TA2	1	0	0	...	0
TA	TA3	0	0	1	...	1
AICMS	AICMS1	0	1	0	...	0
AICMS	AICMS2	0	0	0	...	1
AICMS	AICMS3	1	0	0	...	0
AICMS	AICMS4	0	0	0	...	0
AISS	AISS1	0	1	0	...	0
AISS	AISS2	0	0	0	...	0
AISS	AISS3	0	0	0	...	0
AISS	AISS4	0	0	0	...	0
AISS	AISS5	0	0	0	...	0
AISS	AISS6	1	0	0	...	0
AISS	AISS7	0	0	0	...	1
AISS	AISS8	0	0	0	...	0
AISS	AISS9	0	0	0	...	0
AISS	AISS10	0	0	0	...	0
CALC	CALC1	1	0	0	...	0
CALC	CALC2	0	1	0	...	1
CALC	CALC3	0	0	0	...	0
CALC	CALC4	0	0	1	...	0
SOCI	SOCI1	1	0	0	...	1
SOCI	SOCI2	0	0	1	...	0
SOCI	SOCI3	0	1	0	...	0
SOPJ	SOPJ1	1	1	0	...	1
SOPJ	SOPJ2	0	0	1	...	0
RLFE	RLFE0	0	1	0	...	0
RLFE	RLFE1	1	0	0	...	0
RLFE	RLFE2	0	0	1	...	0
RLFE	RLFE3	0	0	0	...	1
RLFE	RLFE4	0	0	0	...	0
REC	REC1	1	0	0	...	0
REC	REC2	0	1	0	...	0
REC	REC3	0	0	0	...	1
REC	REC4	0	0	1	...	0
AUTO	AUTON	1	1	0	...	1
AUTO	AUTOS	0	0	1	...	0
DA	DA	0	1	0	...	1

conjunto de teste, alternativamente, as porcentagens 60% e 40% podem ser usadas [51]. Considerando a tamanho elevado da base de dados, mais de 300 mil exemplares, optou-se pelas porcentagens de 50% e 50% para treinamento e teste de tal forma que as estatísticas de validação aplicadas ao subconjunto de teste serão mais significativas. Dessa forma a base de dados foi dividida de forma aleatória em duas partes: (1) Treinamento - 50% dos registros e (2) Validação - 50% dos registros [54].

Em se tratando da distribuição do conjunto de dados em relação à classificação, empresas regulares e empresas inadimplentes, nota-se uma diferença no tamanho dessas classes. As empresas regulares (Y = "0") são 74% do total e as inadimplentes (Y = "1") são 26%. Este é um caso no qual o conjunto de dados possui uma distribuição de classes desbalanceada, logo foi verificada essa proporção nos subconjuntos de treinamento e teste. Este controle mostra que a distribuição de classes associadas aos exemplares distribuídos nos subconjuntos segue as mesmas proporções existentes na base total [51]. A princípio, a distribuição efetivamente aleatória dos elementos nos subconjuntos manterá essa proporção.

Os algoritmos utilizados na plataforma R Studio para a modelagem dos dados foram: (1) Para a regressão LOGIT, "H2O Generalized Linear Models - GLM", apesar do nome remeter a modelo linear, traz a opção de configuração para a LOGIT; e (2) para a Rede Neural Artificial, "H2O Deep Learning - DL", configurado para uma rede *multilayer perceptron - MLP*. Os códigos produzidos (*script*) na linguagem R utilizados para a construção dos modelos são mostrados no Anexo I. A escolha destes algoritmos se deu pela conveniência e oportunidade dos mesmos, são gratuitos, de fácil acesso, intuitivos e de boa aceitação nos meios profissionais e acadêmicos. Os modelos LOGIT e MLP foram desenvolvidos conforme mostrado nos itens a seguir.

3.4.1 Regressão LOGIT

O modelo LOGIT desenvolvido é da família binomial (*Logit Elastic Net*, regularização $\alpha = 0,5, \lambda = 0,000241$) com 41 (quarenta e um) coeficientes preditores, correspondentes a cada uma das variáveis binárias, dentre estes 32 foram ativos, ou seja, 9 não tiveram significância no modelo. A Figura 3.9 mostra os 20 coeficientes de maior significância. Essa informação será útil na identificação dos fatores de risco a serem descritos conforme um dos objetivos específicos deste trabalho.

O modelo de regressão obteve um $R^2 = 0,5921$ e a área sob a curva ROC foi $AUC = 0,9485$, conforme pode ser observado na Figura 3.10. A Matriz de confusão do modelo é apresentada na Tabela 3.8. Esta matriz apresenta os valores reais e os preditos, bem como a taxa de erro para cada ocorrência ("0" e e"1") e a taxa de erro total [54], assim

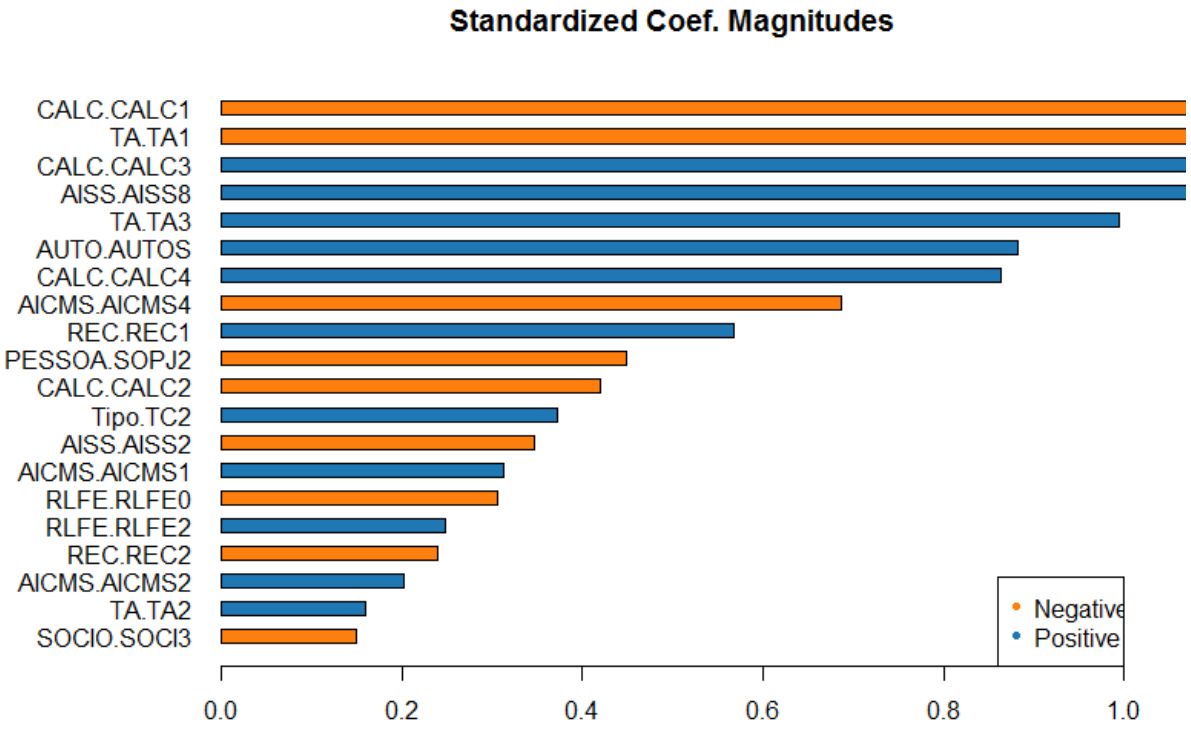

Figura 3.9: Magnitude dos 20 maiores coeficientes do modelo LOGIT (Fonte: Autor).

Tabela 3.8: Matriz de confusão do modelo LOGIT.

↓ **Real x Predito** →	**0**	**1**	**Erro**	**Taxa**
0	103.432	9.312	0,082594	= 9.312/112.744
1	7.139	32.848	0,178533	= 7.139/39.387
Totais	110.571	42.160	0,107712	= 16.451/152731

este recurso se apresenta como uma ótima forma de avaliar e comparar a assertividade de modelos preditivos [51].

3.4.2 Rede Neural Artificial

Na configuração da rede neural artificial - RNA foram testadas diversas formas de desenho da rede, variando de uma a três camadas internas bem como o número de neurônios por camada (10, 20, 30, 60 e 100 neurônios; 10, 20, 100 e 1000 ciclos). Por fim chegou-se ao arranjo ótimo de duas camadas internas com 30 neurônios cada com 20 ciclos, ou épocas, de treinamento, conforme mostrado na Figura 3.11, onde as configurações das camadas de entrada, classificação e saída estão indicadas na Tabela 3.9.

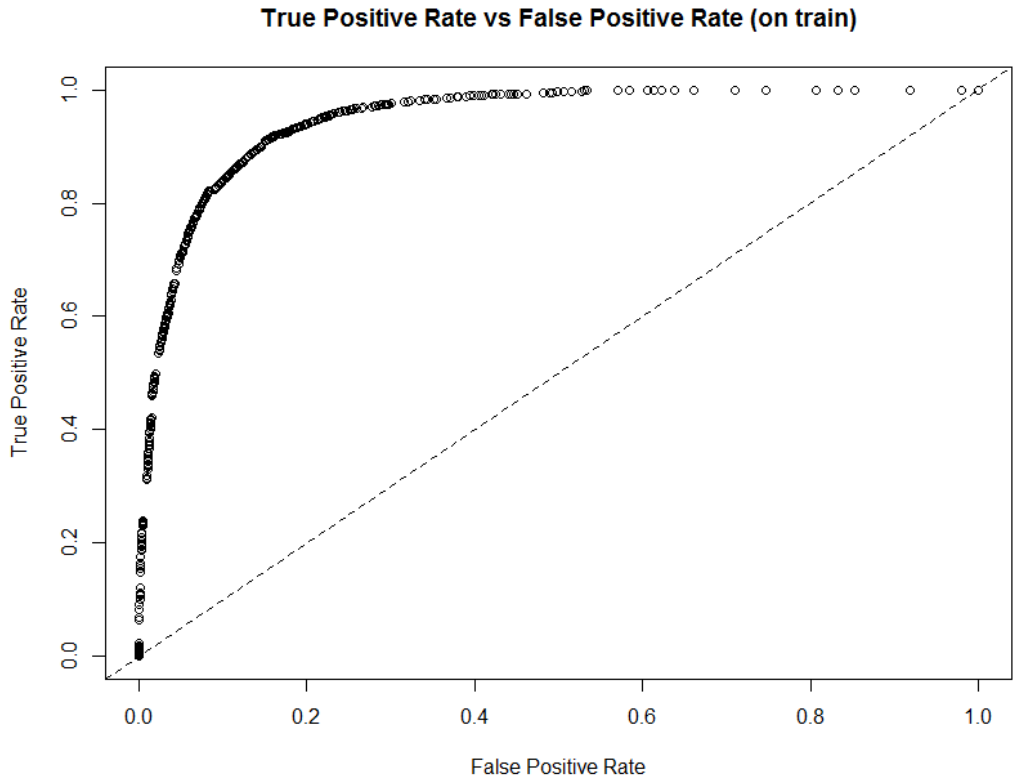

Figura 3.10: Curva ROC do modelo LOGIT (Fonte: Autor).

O modelo de redes neurais obteve um $R^2 = 0,6231$ e a área sob a curva ROC foi $AUC = 0,9568$, conforme pode ser observado na Figura 3.12. A Matriz de confusão do modelo é apresentada na Tabela 3.10.

As 20 variáveis de maior significância do modelo MLP estão indicadas na Figura 3.13. Em associação com a magnitude dos coeficientes do modelo LOGIT, esta informação será utilizada para descrição dos fatores de risco identificados pela modelagem, conforme os objetivos específicos deste trabalho.

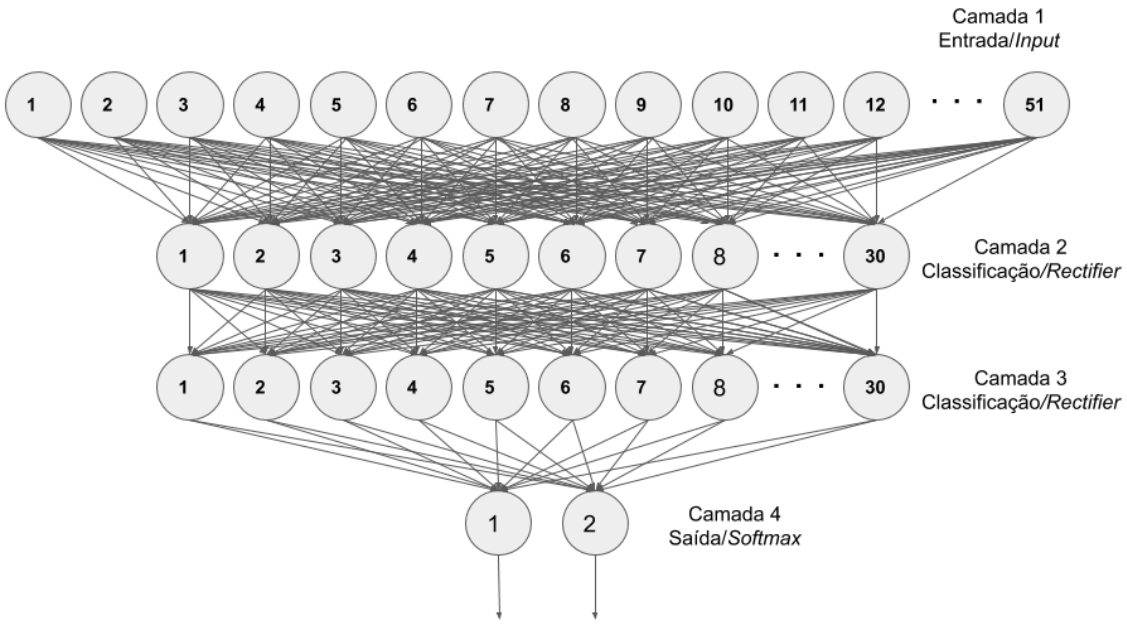

Figura 3.11: Representação gráfica da RNA ótima adotada no modelo (Fonte: Autor).

Tabela 3.9: Descrição das camadas de neurônios da RNA do modelo MLP.

Camada	**1**	**2**	**3**	**4**
Qtd. Neurônios	51	30	30	2
Tipo	Input	Rectifier	Rectifier	Softmax
Exclusões	0,00%	0,00%	0,00%	NA
Taxa média	NA	0,210636	0,040880	0,000998
Taxa RMS	NA	0,406258	0,095798	0,001849
Momentum	NA	0,00	0,00	0,00
Peso médio	NA	-0,011367	0,199399	-0,715722
Peso RMS	NA	1,286634	1,430612	1,180140
Viés médio	NA	1,743134	-0,095577	-0,030709
Viés RMS	NA	4,412186	0,564915	0,744752

Tabela 3.10: Matriz de confusão da RNA adotada.

↓ Real x Predito →	**0**	**1**	**Erro**	**Taxa**
0	101.499	11.245	0,099739	= 11.245/112.744
1	5.438	34.549	0,135994	= 5.438/39.987
Totais	106.937	45.794	0,109231	= 16.683/152731

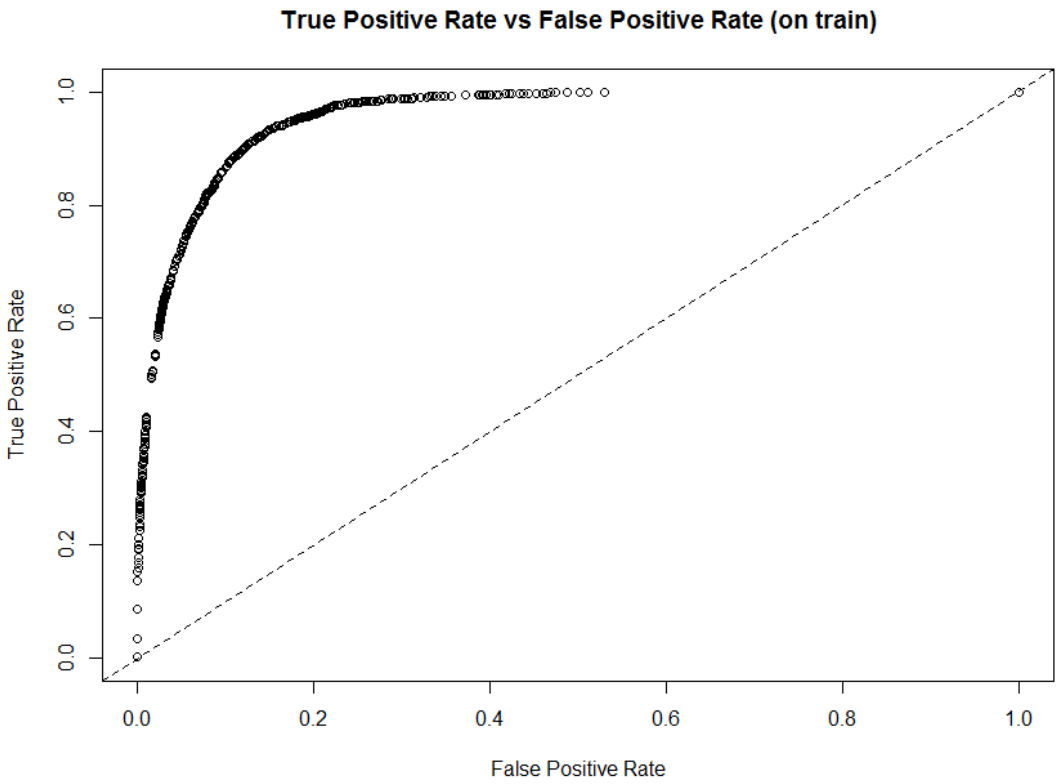

Figura 3.12: Curva ROC do modelo MLP (Fonte: Autor).

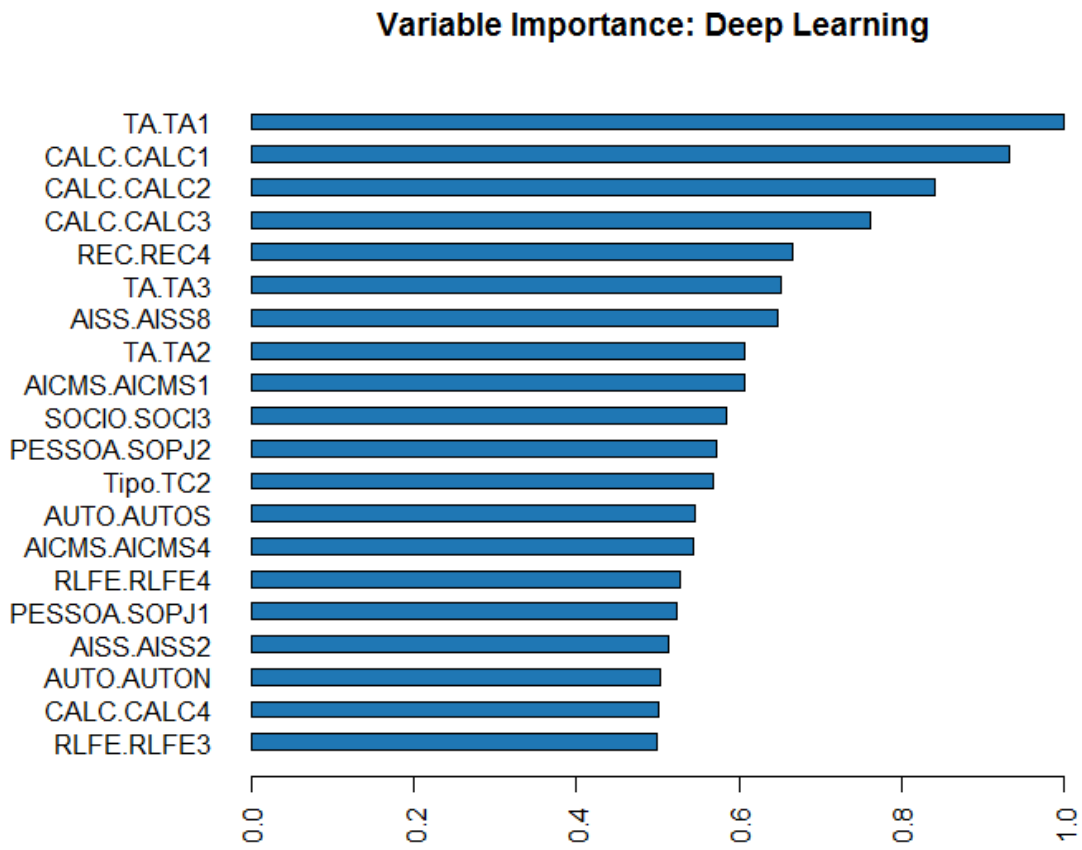

Figura 3.13: Importância das variáveis do modelo MLP (20 maiores) (Fonte: Autor).

Capítulo 4

Avaliação dos Resultados

Nas seções a seguir serão apresentadas a avaliação e discussão dos resultados, a comparação dos dois modelos (LOGIT e MLP), a identificação e avaliação dos fatores de risco (conforme observado na importância das variáveis dos modelos), e em seguida, as conclusões alcançadas neste estudo bem como sugestões para estudos futuros.

4.1 Proposta de tratamento de dados fiscais

Destacou-se no estudo a simplicidade da metodologia de discretização das variáveis com base na regra de Pareto (80/20). Apesar de trabalhoso este método proporciona uma uniformidade no tratamento dos dados de todas as variáveis. O quadro resumo das variáveis é apresentado na Tabela 4.1.

A discretização das variáveis qualitativas, referentes às características cadastrais e econômicas das empresas, seguiu uma sequência clara e definida de passos no seu trata-

Tabela 4.1: Resumo descritivo das variáveis modeladas.

Variável	**Descrição**	**Natureza**	**Especificidades**
Tipo	Tipo do contribuinte	Categórica	Multiv., 3 instâncias
TA	Tempo de atividade	Discreta	Multiv., 3 instâncias
AICMS	Atividade econômica do ICMS	Categórica	Multiv., 4 instâncias
AISS	Atividade econômica do ISS	Categórica	Multiv., 10 instâncias
CALC	Forma de cálculo do ICMS e/ou ISS	Categórica	Multiv., 4 instâncias
SOCI	Quantidade de sócios da empresa	Discreta	Multiv., 3 instâncias
SOPJ	Possui sócio pessoa jurídica	Discreta	Multiv., 2 instâncias
RLFE	Relação Livro Fiscal eletrônico	Contínua	Número Real
REC	Valor total de imposto recolhido	Contínua	Número Real
AUTO	Foi autuado ou não	Discreta	Multiv., 2 instâncias
DA	Inclusão em Dívida Ativa	Discreta	Binária, 0 ou 1

mento de forma a priorizar as características mais recorrentes e agrupar as menos recorrentes.

Por outro lado a discretização das variáveis quantitativas, referentes aos valores declarados (RLFE) e valores recolhidos (REC) pelas empresas, foi feita com a divisão dos dados em quartis conforme sua distribuição de frequência. O artifício matemático da utilização do logaritmo decimal dos valores mostrou-se eficaz como pode ser observado na relevância dessas variáveis em ambos os modelos.

Existe ume grande diversidade de informações disponíveis a administração tributária, logo a escolha de quais dados serão modelados e a definição de como esta atividade será feita são tarefas difíceis e onerosas. A proposta apresentada neste trabalho tem o intuito de trazer uma direção para o desafio de modelagem das bases de informações tributárias.

Este conjunto de procedimentos poderá ser adotado como padrão de tratamento dos dados fiscais, associado ao CRISP-DM, o que sugere a consolidação desta prática como padrão de abordagem na configuração de bases de dados para algoritmos de aprendizado de máquina a serem utilizados na gestão tributária.

4.2 Identificação e Avaliação dos fatores de risco

A modelagem dos dados através dos modelos preditivos utilizados neste estudo destacaram algumas variáveis em relação a outras. Este destaque foi observado pela importância das variáveis verificadas nos modelos (Figura 3.9 e Figura 3.13). A diferença da mensuração da magnitude das variáveis entre os modelos é que na regressão (LOGIT) é possível observar o sinal, positivo ou negativo, do coeficiente ligado à cada variável, já na rede neural (MLP), como sua característica própria, se verifica apenas quais são as variáveis mais significativas para o modelo, o que impede observar se essa influência é positiva ou negativa. Assim a interpretação da significância das variáveis nos dois modelos de forma conjunta pode fornecer a identificação de quais fatores representam risco de inadimplência fiscal [35].

Tendo em vista que o objetivo da predição dos modelos é a indicação dos contribuintes com maior chance de inadimplência, sendo a variável explicada Y (inscrição em Dívida Ativa) determinada em “1” para os inadimplentes, no modelo de regressão (LOGIT) as variáveis explicativas com coeficientes positivos contribuem para que o valor de Y seja mais próximo de “1”, já as variáveis explicativas com coeficientes negativos contribuem para que o valor de Y seja mais próximo de “0”. Assim as variáveis com maior valor positivo são as que oferecem o maior risco de inadimplência.

Conforme observado em conjunto nas Figura 3.9 e Figura 3.13, as variáveis de maior magnitude no modelo MLP e de sinal positivo no modelo LOGIT são: CALC3, TA3,

AISS8, TA2, AICMS1, TC2, e AUTOS. Portanto essas variáveis ficam apontadas como os fatores de risco de inadimplência observados neste estudo. A seguir apresenta-se a descrição destes fatores.

CALC3 - *Forma de cálculo do imposto "Normal".* Nesta categoria se enquadram as empresas com regime de apuração Normal dos impostos ICMS e/ou ISS, em sua maioria, empresa de faturamento anual superior a R$3.600.000,00 (três milhões e seiscentos mil reais).

AISS8 - *Atividade econômica do ISS de Autônomos.* Aqui se enquadram os profissionais autônomos prestadores de serviços sujeitos à cobrança do ISS. Em sua maioria são pessoas físicas autônomas atuando em atividades de pequeno porte econômico.

TA2 - *Tempo de atividade de 5 a 9 anos* e **TA3** - *Tempo de atividade 10 anos ou mais.* Neste conjunto estão as empresas com 5 anos ou mais de funcionamento. Vale destacar essa variável, pois indica que quanto mais tempo de funcionamento mais propensas à inadimplência fiscal estão as empresas. De certa forma, se trata de uma inferência lógica, porque quanto mais tempo de atividade, mais tempo a empresa estaria exposta ao risco de não pagar impostos por erro ou dificuldades financeiras.

AUTOS - *Empresas que foram autuadas no período.* No universo estudado foram observadas mais de 1.500 empresas que em algum momento dentro do período analisado foram autuadas pelo FISCO, sendo que dentre estas, 1.400, foram inscritas em dívida ativa em algum momento no período observado. Este é um fator muito interessante para o FISCO pois mostra uma forte tendência das empresas já multadas se tornarem inadimplentes também. Esta relação demonstra a necessidade de se desenvolver programas de monitoramento das empresas autuadas para evitar que as mesmas se tornem inadimplentes.

AICMS1 - *Atividade econômica do ICMS de Comércio.* Nesta categoria se enquadram todas as empresas com a atividade econômica de comércio, a mais recorrente no universo perquisado e mais sujeita às oscilações econômicas. Como é a instância mais frequente nesta variável, não foi constado indicativo que demonstre tendência de inadimplência fiscal relativa a outras características.

TC2 - *Tipo de contribuinte "Sociedade Empresária Limitada".* São as empresas cujo corpo societário é definido e limitado no contrato social de constituição da sociedade. Tradicionalmente possuem a terminação "LTDA" em suas razões sociais, a qual indica a palavra "limitada".

Tabela 4.2: Comparação entre os modelos LOGIT e MLP.

Modelo	R^2	**AUC**	**Taxa de Erro**
LOGIT	0,5902	0,9485	0,107477
MLP	0,6203	0,9568	0,109231

4.3 Discussão da solução

A Meta-análise Bibliográfica sobre a Análise de Risco de Crédito, apresentada na revisão teórica deste trabalho, mostrou os algoritmos mais usados na modelagem de risco de inadimplência utilizados. São eles: regressão linear, regressão LOGIT, *support vector machines*, árvores de decisão e redes neurais artificiais.

Foi escolhida a rede neural artificial *multi-layer perceptron*, conforme a literatura pesquisada, pela sua alta taxa de acerto nas predições em relação aos outros algoritmos. Conforme demonstrado ao longo do estudo, também foi construído um modelo de regressão LOGIT para auxiliar na interpretação da influência das variáveis estudadas na tarefa de predição.

O modelo de regressão logística (LOGIT) se mostrou uma boa opção para identificar as variáveis mais importantes na predição, a sua análise em consonância com a importância das variáveis na rede neural (MLP) apresenta uma perspectiva interessante para identificação dos fatores de risco inerentes ao problema da inadimplência fiscal. O "poder explicativo" da regressão LOGIT é destacado pelos coeficientes de cada variável, $\beta_0, \beta_1, \beta_2, ...\beta_n$, equação (2.12), onde, conforme mostrado na Revisão Teórica, podem ser interpretados como a relação da probabilidade de ocorrência e/ou da não-ocorrência daquela característica modelada.

Por outro lado, a rede neural (MLP), teve uma melhor aderência ao modelo em relação à regressão, como pode ser observado nos valores do Coeficiente de Determinação R^2 dos modelos (ver Tabela 4.2) e melhor distribuição na taxa de acerto da predição, conforme as matrizes de confusão dos modelos (ver Tabela 3.8 e Tabela 3.10).

Os modelos estudados tiveram uma performance bastante similar, conforme mostra a Tabela 4.2. Ambos obtiveram uma bom desempenho na tarefa de predição com destaque para superioridade da rede neural (MLP).

Diante da ausência de publicações tratando da predição de inadimplência fiscal, surgiu a necessidade de buscar uma referência teórica para tal tarefa. Este estudo propôs a predição de inadimplência financeira, vista na Análise de Risco de Crédito, como referência de modelagem. O que, depois de obsevados os resultados, se comprovou uma hipótese viável.

Tabela 4.3: Indícios de irregularidade fiscal nas empresas indicadas pela RNA.

Descrição do indício de irregularidade fiscal encontrado	**Nº de Empresas**
I - Aproveitamento indevido de crédito	377
II - Não escrituração de débitos	146
III - Omissão de receitas	677
IV - Pelo menos dois dos itens acima	58
V - Os três indícios I, II e III	1

4.4 Validação do modelo RNA

Para a validação do modelo de Redes Neurais Artificiais (MLP), utilizando o algoritmo H2O *Deep Learning*, a base de contribuintes ativos do Distrito Federal (mais de 270 mil empresas) foi classificada pelo modelo desenvolvido. O resultado desta modelagem gerou um conjunto de 32.105 empresas classificadas como "possíveis inadimplentes" ($Y = 1$). Em seguida foram isoladas as empresas do Regime Normal de Apuração, pois é o mais relevante em termos de arrecadação, restando 7.573 empresas para verificação.

Esta lista de contribuintes foi submetida aos três mais usados algoritmos de detecção primária de indícios de irregularidade fiscal utilizados na Secretaria de Fazenda do DF e o resultado foi a indicação positiva em 1.004 empresa, 13,25% das empresas apontadas, conforme mostrado na Tabela 4.3.

Em consulta ao sistema de Dívida Ativa da SEF foi constatado uma média anual de 8.763 empresas, entre 2012 e 2017, consideradas inadimplentes por imposto declarado e não pago, ou seja no período estudado (5 anos) foram 43.815 empresas, aproximadamente 16% do total de 270 mil. Esta situação é diversa das mostradas acima, pois é a mais recorrente nos casos de inadimplência e não implica em irregularidade fiscal ou sonegação, apenas o não pagamento do imposto.

No grupo de 7.573 empresas selecionadas para verificação 2.471 apresentaram divergências de declarações e pagamentos, 33% do total, ou seja, um forte indício de não recolhimento de imposto declarado. Esta porcentagem é o dobro da porcentagem observada na população, 16%, o que demonstra a importância da classificação do modelo estudado. Assim, diante dos resultados observados foi considerada convalidada a relevância da classificação feita pela RNA validando o modelo.

Capítulo 5

Conclusões e Estudos Futuros

O presente estudo teve a perspectiva de confirmar a possibilidade da utilização de técnicas estatísticas de inferência na prospecção de indícios de inadimplência fiscal de empresas. A tarefa de seleção de empresas para auditoria fiscal, com o auxílio da predição das redes neurais, se tornará progressivamente, mais eficiente, imparcial e com menos custos para o Estado.

Este estudo propõe o uso contínuo desta técnica, e sua consequente validação, no intuito de aprimorar o trabalho de investigação de indícios de inadimplência tributária, ou seja, a associação dos métodos tradicionais de prospecção de indícios de sonegação com métodos de predição. Isto criará novas possibilidades de sondagem e descobertas de irregularidades. Todo este processo poderá, num futuro próximo, fazer parte de uma sistema de inteligência artificial de detecção de fraudes fiscais.

Em consonância com o Objetivo Específico 1, foi proposto um processo de tratamentos de dados fiscais que discretizou as variáveis qualitativas e quantitativas. Este método de discretização, baseado na regra 80/20 (Pareto), se mostrou efetivo para tratar os dados cadastrais dos contribuintes do ICMS e ISS no DF de forma simples e de fácil implementação. A proposta da regra 80/20 destacou as características mais significativas individualmente e agrupou as menos relevantes. Este método tornou o tratamento dos dados mais racional e intuitivo, logo, espera-se que seja adotado continuamente pela administração tributária em projetos ao porvir.

O Cadastro Fiscal apresenta uma grande variedade de informações sobre as empresas, todas as características formais estão lá representadas, mas esta diversidade de dados torna difícil a tarefa de modelar estas características. O método proposto neste estudo visa simplificar esta tarefa sem perder a assertividade da modelagem estatística. O resultados obtidos mostraram a eficácia desta proposta.

O comportamento das variáveis explicativas foi analisado de forma conjunta entre os modelos de redes neurais e da regressão LOGIT. Esta associação proporcionou verificar

quais as variáveis mais impactantes nos resultados e como se deu a influência de cada uma delas. Desta forma, as variáveis que indicam o maior risco de inadimplência foram apontadas como principais fatores de risco. Assim, conforme o Objetivo Específico 2, os fatores de risco associados às características observadas nas informações fiscais foram identificados e descritos no Capítulo 4 Avaliação dos Resultados. Destacam-se as variáveis:

(1) a "Forma de cálculo do imposto *Normal*", variável CALC3, pois confirmou o entendimento do senso comum da necessidade da simplificação dos regimes de tributação. Apesar do nome "Normal", esta forma de apuração é trabalhosa e complicada, tanto para a apuração do próprio contribuinte em suas declarações como para a auditoria nos processos de fiscalização; e

(2) "Empresas que foram autuadas no período", variável AUTOS, porque na população estudada foram observadas mais de 1.500 empresas autuadas pelo FISCO, sendo que dentre estas, 1.400, foram inscritas em dívida ativa, o que destacaria o interesse do FISCO neste grupo. Isto mostra uma tendência de empresas já multadas se tornarem inadimplentes também. Esta constatação demonstra a necessidade de programas de monitoramento de empresas autuadas para evitar sua eventual inadimplência fiscal.

Os resultados obtidos no modelo de redes neurais artificiais demonstraram a viabilidade de predizer o comportamento fiscal de empresas do DF, quanto a inadimplência ou não, através do uso de modelos preditivos baseados neste algoritmo, de acordo com o Objetivo Específico 3. Observou-se que a taxa de erro na predição ficou menor que 11% e o desempenho geral da rede neural foi superior a da regressão (Tabela 4.2), sendo esta a referência para análise de risco de crédito. Finalmente a efetividade da predição foi validada submetendo os contribuintes classificados aos algoritmos de prospecção de indícios de irregularidades utilizados na fiscalização tributária, conforme mostrado na Avaliação dos Resultados, em especial na Tabela 4.3. Portanto, diante dessas considerações, o Objetivo Geral deste estudo foi atingido, qual seja "Verificar como o uso de redes neurais artificiais pode auxiliar na identificação de riscos de inadimplência fiscal de ICMS e ISS".

Atualmente a fiscalização tributária não dispõe de modelos de predição de inadimplência fiscal. A seleção de empresas é feita pela detecção de indícios de irregularidades, processo este baseado em algoritmos de otimização e de cruzamentos das informações disponíveis nas bases de dados da Administração Tributária. Os comportamentos de evasão fiscal antes não detectados são difíceis de serem identificados, pois essa sistemática é programada empiricamente de acordo com o conhecimento existente.

A aplicação dos modelos de predição baseados em redes neurais na seleção de empresas para auditoria pode mudar este cenário. Espera-se que a associação dos métodos tradicionais utilizados hoje com os métodos de inferência, em especial o proposto neste estudo, possa aprimorar a identificação dos casos de imposto devido e não recolhido. Um futuro

sistema de Inteligência Artificial de detecção de fraude fiscais, necessariamente, executará tarefas de predição de inadimplência fiscal, logo o presente trabalho tem sua contribuição propondo uma metodologia de mineração dos dados e um método de modelagem preditiva, através do uso de redes neurais artificiais associadas à regressão LOGIT.

Para futuros estudos propõem-se avaliar formas de implantação dessa metodologia nas atividades precípuas da administração tributária: Uma proposta seria submeter o modelo a uma plataforma de monitoramento de empresas classificadas como possíveis inadimplentes; e outra proposta seria uma plataforma de classificação de empresas recém criadas para indicação de possíveis diligências fiscais.

Outra abordagem seria a inclusão de novas variáveis, como a emissão de notas fiscais, o faturamento com vendas em cartão de crédito e a variação do número de empregados. Também pode ser testada a variação do modelo de redes neurais artificiais para grupos de RNAs associadas a outros algoritmos de aprendizado de máquina, operando em conjunto, simultaneamente ou em série.

Referências

[1] COMMITTEE OF SPONSORING ORGANIZATIONS OF THE TREADWAY COMMISSION, (COSO): *COSO-ERM - Executive Summary*, 2017. x, 9, 10

[2] Pandey, T., Jagadev A. Dehuri S. e S. Cho: *A novel committee machine and reviews of neural network and statistical models for currency exchange rate prediction: An experimental analysis.* Journal of King Saud University – Computer and Information Sciences, 2018. x, 27, 28

[3] COMMITTEE OF SPONSORING ORGANIZATIONS OF THE TREADWAY COMMISSION, (COSO): *COSO-ERM - Integrating with Strategy and Performance*, 2017. 1, 7, 9

[4] Arena, M., Arnaboldi M. Azzone G.: *The organizational dynamics of enterprise risk management.* Accounting, Organizations and Society, 35(7):659–675, 2010. 6, 7

[5] COMMITTEE OF SPONSORING ORGANIZATIONS OF THE TREADWAY COMMISSION, (COSO): *COSO-ERM - Enterprise Risk Management - integrated framework*, 2004. 7

[6] Beasley, M. S., Clune R. e Hermanson D. R.: *Enterprise risk management: An empirical analysis of factors associated with the extent of implementation.* Journal of Accounting and Public Policy, nd(11/12):521–531, 2005. 7

[7] Moeller, R. R.: *COSO Enterprise Risk Management: Understanding the new integrated ERM framework.* John Wiley Sons, Inc., New Jersey, USA, 2007. 7

[8] Wu, D., Olson D. L. Dolgui A.: *Decision making in enterprise risk management: A review and introduction to special issue.* Omega, 57(Part A):1–4, 2015. 7, 11, 12

[9] Bromiley, P., McShane M. Nair A. Rustambekov E.: *Enterprise risk management: Review, critique, and research directions.* Long Range Planning, 48(4):265–276, 2015. 7

[10] DISTRITO FEDERAL, DF: *Decreto n. 37.302, de 29 de abr. de 2016. Estabelece os modelos de boas práticas gerenciais em Gestão de Riscos e Controle Interno a serem adotados no âmbito da Administração Pública do Distrito Federal, Brasília, DF*, 2016. 8, 12

[11] SECRETARIA DE ESTADO DE FAZENDA DO DISTRITO FEDERAL, SEFP: *Portaria Conjunta n. 11, de 28 de ago. de 2018. Institui o Comitê Superior de Gestão*

de Riscos, no âmbito da Secretaria de Estado de Fazenda do Distrito Federal e dá outras providências, Brasília, DF, 2018. 8, 12

[12] COMMITTEE OF SPONSORING ORGANIZATIONS OF THE TREADWAY COMMISSION, (COSO): *COSO-IC - Control - integrated framework*, 1994. 8

[13] Olson, W., Dash. D. et al: *Enterprise risk management models.* Springer, 2010. 11

[14] Wang, S. e G.H. Huang: *An integrated approach for water resources decision making under interactive and compound uncertainties.* Omega, 44:32 – 40, 2014. 11

[15] Choi, Y., Ye X. Zhao L. et al.: *Optimizing enterprise risk management: a literature review and critical analysis of the work of wu and olson.* Annals of Operations Research, 237(1/2):281–300, 2015. 11

[16] Sun, Yunpeng, Daniel W. Apley e Jeremy Staum: *Efficient nested simulation for estimating the variance of a conditional expectation.* Operations Research, 59(4):998–1007, 2011. 11

[17] Wu, Desheng e David L. Olson: *Supply chain risk, simulation, and vendor selection.* International Journal of Production Economics, 114(2):646 – 655, 2008. Special Section on Logistics Management in Fashion Retail Supply Chains. 11

[18] Towill, D. e S. Disney: *Managing bullwhip-induced risks in supply chains.* International Journal of Risk Assessment and Management (IJRAM), 10(3), 2008. 11

[19] Peng, Min, Yi Peng e Hong Chen: *Post-seismic supply chain risk management: A system dynamics disruption analysis approach for inventory and logistics planning.* Computers Operations Research, 42:14 – 24, 2014. 11

[20] Chaudhuri, Arindam: *Modified fuzzy support vector machine for credit approval classification.* AI Communications, 27:189–211, 2014. 12

[21] Wu, D. D., D. L. Olson e C. Luo: *A decision support approach for accounts receivable risk management.* IEEE Transactions on Systems, Man, and Cybernetics: Systems, 44(12):1624–1632, 2014. 12

[22] Groth, Sven S. e Jan Muntermann: *An intraday market risk management approach based on textual analysis.* Decision Support Systems, 50(4):680 – 691, 2011. Enterprise Risk and Security Management: Data, Text and Web Mining. 12

[23] Caron, F., Vanthienen J. e B. Baesens: *Comprehensive rule-based compliance checking and risk management with process mining.* Decision Support Systems, 54(3):1357 – 1369, 2013. 12

[24] Caron, F., Vanthienen J. e B. Baesens: *A comprehensive investigation of the applicability of process mining techniques for enterprise risk management.* Computers in Industry, 64(4):464 – 475, 2013. 12

[25] ASSOCIAÇÃO BRASILEIRA DE NORMAS TÉCNICAS, (ISO 31000:2018): *ISO 31000 - Gestão de Riscos - Diretrizes*, 2018. 12, 13

[26] Insurance & Risk Managers (AIRMIC), The Public Risk Management Association (Alarm), The Institute of Risk Management (IRM) The Association of: *A structured approach to Enterprise Risk Management (ERM) and the requirements of ISO 31000*, 2010. 13

[27] Thomas, L., Crook J. e D. Edelman: *Credit scoring and its applications*, volume 2. Siam, 2017. 14

[28] Mariano, A. e M. Rocha: *Revisão da literatura: Apresentação de uma abordagem integradora.* Em *In AEDM International Conference–Economy, Business and Uncertainty: Ideas for a European and Mediterranean industrial policy*, Reggio Calabria, Italia, 2017. 14, 15, 16

[29] Mariano, A. e L. F. Diaz: *A importância da aceitação e uso da tecnologia em aplicativos de mobilidade urbana: contribuições da literatura científica.* Em *VII Congresso de Engenharia de Produção*, Ponta Grossa, Brasil, 2017. 15

[30] Mariano, A. e A. F. Gomes: *Endividamento com cartão de crédito: um estudo exploratório por meio da teoria do enfoque meta analítico consolidado.* Em *VII Congresso de Engenharia de Produção*, Ponta Grossa, Brasil, 2017. 15

[31] Calazans, A. T. S., E. T. S. Masson e A. M. Mariano: *Uma revisão sistemática da bibliografia sobre inovação bancária utilizando o enfoque meta-analítico.* Revista Espacios: Revista arbitrada de gestion tecnologica, 36(15):8–31, 2015. 16

[32] Garrido, J. A. M. e A. R. Rodríguez: *La investigación sobre las relaciones interorganizativas: un estudio bibliométrico.* Investigaciones Europeas de Dirección y Economía de la Empresa, 10(1):149–163, 2004. 16

[33] Correa, P. R. e R. G. Cruz: *Meta-análisis sobre la implantación de sistemas de planificación de recursos empresariales (erp).* Journal of information Systems and Technology Management, 2(3):245–273, 2005. 16

[34] Lessmann, S., Baesens B. Seow H. et al.: *Benchmarking state-of-the-art classification algorithms for credit scoring: An update of research.* European Journal of Operational Research, 247(1):124–136, 2015. 21, 32

[35] Gouvêa M., Gonçalves E., Mantovani D.: *Análise de risco de crédito com aplicação de regressão logística e redes neurais.* Revista Contabilidade Vista Revista, 24(4):96–123, 2013. 21, 27, 54

[36] Maher A., Maysam F.: *Classifiers consensus system approach for credit scoring.* Knowledge-Based Systems, 104(1):89–105, 2016. 21, 27

[37] Zhao Z., Xu S., Kang B. Kabir M. J. Liu Y. Wasinger R.: *Investigation and improvement of multi-layer perceptron neural networks for credit scoring.* Expert Systems with Applications, 1(42):3508–3516, 2015. 21, 27

[38] Fayyad, U., Piatetsky G. Smyth P.: *From data mining to knowledge discovery in databases*. AI Magazine, 17(3):18, 1996. 21

[39] Azevedo, A., Santos M.: *Kdd, semma and crisp-dm: A parallel overview*. IADS European Conference Data Mining, nd(11), 2008. 22, 23, 25

[40] Ristoski, Petar e Heiko Paulheim: *Semantic web in data mining and knowledge discovery: A comprehensive survey*. Journal of Web Semantics, 36:1–22, 2016. 22, 25

[41] Gardiner, Eleanor J. e Valerie J. Gillet: *Perspectives on knowledge discovery algorithms recently introduced in chemoinformatics: Rough set theory, association rule mining, emerging patterns, and formal concept analysis*. Journal of Chemical Information and Modeling, 55(9):1781–1803, 2015. 22, 25

[42] Corrales, D. C., Agapito L. Corrales J. C.: *A conceptual framework for data quality in knowledge discovery tasks (fdq-kdt): A proposal*. Journal of Computers, 10:396–405, 2015. 22, 23, 25

[43] *A data mining based approach to a firm's marketing channel*. Procedia Economics and Finance, 27:77 – 84, 2015. 22nd International Economic Conference of Sibiu 2015, IECS 2015. 23, 25

[44] Caffo, B.: *Regression Models for Data Science in R*, volume 1. Leanpub, 2015. 25

[45] Abdou, H., Pointon J.: *Credit scoring, statistical techniques and evaluation criteria: a review of the literature*. Intelligent Systems in Accounting, Finance Management, 18(2/3):59–88, 2011. 25, 27

[46] Hosmer Jr., D., Lemeshow S.: *Applied logistic regression*, volume 1. John Wiley Sons, 2004. 25

[47] Chan, B., L.P.L. Fávero, F.L. Da Silva e P. Belfiore: *Análise de dados: modelagem multivariada para tomada de decisões*. Elsevier, 2009. 27

[48] West, D.: *Neural network credit scoring models*. Computers Operations Research, 27(11):1131 – 1152, 2000. 27

[49] Gil, A. C.: *Como elaborar projetos de pesquisa*. Ed. Atlas, São Paulo, 4ª edição, 2002. 29

[50] Prodanov, C., Freitas E.: *Metodologia do trabalho científico: métodos e técnicas da pesquisa e do trabalho acadêmico*. Ed. Feevale, Novo Hamburgo, 2ª edição, 2013. 29

[51] Silva, L. A., Peres S. M. Boscarioli C.: *Introdução a Mineração de Dados: com aplicações em R*. Ed. Elsevier, Rio de Janeiro, 1ª edição, 2016. 36, 47, 48

[52] Bittencourt Neto, S. A. P.: *Análise de "outliers" para o controle do risco de evasão tributária do icms*. Mestrado profissional em computação aplicada, Universidade de Brasília, Brasília, 2018. 42

[53] Montgomery, D., Runger G.: *Estatística aplicada e probabilidade para engenheiros.* Ed. LTC, Rio de Janeiro, 5ª edição, 2014. 42

[54] Oliveira, P., Guerra S.: *Ciência de Dados com R – Introdução.* Ed. IBPAD, Brasília, 1ª edição, 2018. 42, 47

[55] Sartoris, A.: *Estatística e introdução à econometria.* Ed. Saraiva, São Paulo, 2ª edição, 2013. 43

[56] Wooldridge, J. M.: *Introdução à econometria: uma abordagem moderna.* Ed. Cengage Learning, São Paulo, 4ª edição, 2014. 43

Anexo I

Códigos da linguagem R para construção dos modelos

```
## Início
library(dplyr)
library(ggplot2)
library(readxl)
library(h2o)

BaseModelo <- read_excel("BaseTratada.xlsx",
            col_types = c("skip", "text", "text","text", "text", "text",
            "text", "text","text", "text", "text", "numeric"))

View(BaseModelo )
summary(BaseModelo)
str(BaseModelo)

BaseModelo$Tipo <- as.factor(BaseModelo$Tipo)
BaseModelo$TA <- as.factor(BaseModelo$TA)
BaseModelo$AICMS <- as.factor(BaseModelo$AICMS)
BaseModelo$AISS <- as.factor(BaseModelo$AISS)
BaseModelo$CALC <- as.factor(BaseModelo$CALC)
BaseModelo$SOCIO <- as.factor(BaseModelo$SOCIO)
BaseModelo$PESSOA <- as.factor(BaseModelo$PESSOA)
BaseModelo$RLFE <- as.factor(BaseModelo$RLFE)
BaseModelo$REC <- as.factor(BaseModelo$REC)
BaseModelo$AUTO <- as.factor(BaseModelo$AUTO)
BaseModelo$DA <- as.factor(BaseModelo$DA)
```

```
localH2O = h2o.init(nthreads= -1)

BaseCodificada.h <- as.h2o(BaseModelo)
options(OutDec= ".")
data.split <- h2o.splitFrame(data = BaseCodificada.h, ratios = c(0.5), seed = 4321)
data.train <- data.split[[1]]
data.valid <- data.split[[2]]
myY <- "DA"
myX <- setdiff(names(data.train), c(myY, "ID"))

#DeepLearning
dl.model <- h2o.deeplearning(myX, myY,
                             training_frame = data.train ,
                             hidden = c(30, 30), epochs = 20,
                             validation_frame = data.valid,
                             model_id = "dl_DA")
h2o.confusionMatrix(dl.model@model$validation_metrics)
dl.model@model$validation_metrics@metrics$AUC
dl.model@model$validation_metrics@metrics$r2

# MODELO GlM
glm.model <- h2o.glm(myX, myY,
                     training_frame = data.train ,
                     validation_frame = data.valid,
                     family = "binomial",
                     model_id = "glm_DA")
h2o.confusionMatrix(glm.model@model$validation_metrics)
glm.model@model$validation_metrics@metrics$AUC
glm.model@model$validation_metrics@metrics$r2

# Para plotar o gráfico de importância das variáveis
h2o.varimp_plot(dl.model, 20)
h2o.varimp_plot(glm.model, 20)

# Dados dos modelos
dl.model@model$model_summary

# Plotando a curva ROC
# Modelo DL Rede Neural
plot(h2o.performance(dl.model))
```

```
# Modelo GLM Regressão Linear
plot(h2o.performance(glm.model))

## Desligar o H2O
h2o.shutdown(prompt=FALSE)

## Fim
```